การทำนายธุรกรรมที่หลอกลวง โดยใช้เทคนิคการเรียนรู้แบบกลุ่ม ร่วมกับการสกัดคุณลักษณะเด่น

วราภรณ์ พิมา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2564

Fraudulent Transactions Predictions using Ensemble Method with Features Extractions Techniques

Warabhorn Pima

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

2021

ใบรับรองงานวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ การทำนายธุรกรรมที่หลอกลวง โดยใช้เทคนิคการเรียนรู้แบบกลุ่ม ร่วมกับการสกัดคุณลักษณะเด่น

เสนอโดย วราภรณ์ พิมา

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(ดร.สรรพฤทธิ์ มฤคทัต)

.....กรรมการและอาจารย์ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น)

.....กรรมการ

(ดร.ธนภัทร ฆังคะจิตร)

.....กรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ดร.ชัยพร เขมะภาตะพันธ์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 27 เดือน พ.ค. พ.ศ. 2564

หัวข้อวิทยานิพนธ์	การทำนายธุรกรรมที่หลอกลวงโดยใช้เทคนิคการเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะเด่น
ชื่อผู้เขียน	วราภรณ์ พิมา
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2563

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการทำนายธุรกรรมที่หลอกลวง โดยใช้เทคนิคการเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะเด่น โดยทำการศึกษาข้อมูลการซื้อขายบิทคอยน์จากมหาวิทยาลัยสแตนฟอร์ด ซึ่งเผยแพร่ให้บุคคลทั่วไปสามารถนำข้อมูลไปใช้ประโยชน์ได้ จำนวน 59,788 ระเบียน ทำการแบ่งวิธีการทำนายธุรกรรมที่หลอกลวงทั้งหมด 3 วิธีได้แก่ 1. การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และค่าความเป็นศูนย์กลาง 2. การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และ 3. การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วยค่าความเป็นศูนย์กลาง ในแต่ละวิธีทำการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) และการแบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดลด้วยวิธี 10-fold cross-validation และสร้างโมเดลด้วยการเรียนรู้แบบกลุ่ม จากนั้นทำการวัดประสิทธิภาพของโมเดลด้วยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) ความแม่นยำ (Precision) ความระลึก (Recall) ผลการวิเคราะห์ในแต่ละวิธี พบว่า วิธีที่ 3 มีประสิทธิภาพในการทำนายถูกต้องสูงสุด

Thesis Title	Fraudulent Transactions Predictions using Ensemble Method with Feature Extractions Techniques
Author	Warabhorn Pima
Thesis Advisor	Asst. Prof. Dr. Duangjai Jitkongchuen
Department	Big Data Engineering
Academic Year	2020

ABSTRACT

This research was to purpose the fraudulent transactions prediction using ensemble technique with feature extraction techniques. 59,788 sets of transactional data were public data from Stanford University were used as datasets in the experiment. We conducted 3 methods of an experiment: 1. Ensemble technique with Node2Vec and Centrality feature extraction techniques, 2. Ensemble technique with Node2Vec feature extraction technique and 3. Ensemble technique with Centrality feature extraction technique. The imbalance class was handled by Synthetic Minority Oversampling Technique (SMOTE) and was validated using 10-fold cross-validation.

The performance of each model was evaluated with the accuracy, precision, recall, and F1 score. From the experimental results, From the experimental results, it was found that the proposed model is effective in competing with the models with the best performance.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความช่วยเหลืออย่างดียิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษา แนะนำ แก้ไขข้อบกพร่องต่างๆ ทั้งให้กำลังใจ และติดตามสอบถามเกี่ยวกับการเรียนด้วยความห่วงใยแก่ผู้เขียนเป็นอย่างดีมาโดยตลอด ผู้เรียนขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.วรพล พงษ์เพ็ชร อาจารย์ประจำหลักสูตรการวิเคราะห์ธุรกิจและวิทยาการข้อมูล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์(นิด้า) ที่ได้ให้ข้อเสนอแนะแนวทางการทำวิทยานิพนธ์ รวมถึงคำแนะนำที่เป็นประโยชน์ในการทำวิจัยอย่างมีค่ายิ่ง

ขอขอบพระคุณ คณาจารย์ประจำสาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ได้กรุณาประสิทธิ์ประสาทวิชาความรู้แก่ผู้เขียน

ขอขอบพระคุณ บิดามารดา และน้องชายของผู้เขียน ที่ให้การสนับสนุนและให้กำลังใจในด้านการเรียน และการทำวิทยานิพนธ์ด้วยดีมาโดยตลอด

ขอขอบพระคุณ เพื่อนๆ ทุกคนที่ให้กำลังใจ ช่วยเหลือ และเป็นที่ปรึกษาในด้านการเรียนด้วยดีมาโดยตลอด และขอบพระคุณผู้เกี่ยวข้องทุกท่านที่ไม่ได้ระบุนามในที่นี้ ที่กรุณาช่วยเหลือแก่ผู้เขียนมาโดยตลอด

วราภรณ์ พิมา

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญภาพ

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ด้วยความก้าวหน้าทางเทคโนโลยีอินเทอร์เน็ตในปัจจุบัน ส่งผลให้วิถีชีวิตและพฤติกรรมของผู้บริโภคเปลี่ยนแปลงไป ผู้ประกอบการธุรกิจต่าง ๆ ต้องปรับตัวให้เข้ากับยุคดิจิทัล และสนองตอบต่อลูกค้าให้ได้มากที่สุด จากการซื้อขายที่ผู้ขาย (Sellers) และผู้ซื้อ (Buyers) ต้องเผชิญหน้ากัน กลายเป็นร้านค้าออนไลน์ ที่ซึ่งแสดงรายละเอียดสินค้าผ่านทางเว็บไซต์ หรือเรียกว่า พาณิชย์อิเล็กทรอนิกส์ (Electronic Commerce) หรืออีคอมเมิร์ซ (E-Commerce) พาณิชย์อิเล็กทรอนิกส์จึงเป็นช่องทางการตลาดขนาดใหญ่ของโลกไร้พรมแดนที่สามารถเข้าถึงกลุ่มผู้บริโภคเป้าหมายได้โดยตรงอย่างรวดเร็ว ไร้ขีดจำกัดของเรื่องเวลาและสถานที่ ทำให้พาณิชย์อิเล็กทรอนิกส์กลายเป็นโอกาสทางธุรกิจที่สำคัญ

เมื่อการพาณิชย์อิเล็กทรอนิกส์ได้รับความนิยมมากขึ้น จึงมีการนำเอาแนวคิดดังกล่าวมาสร้างตลาดกลางในรูปแบบออนไลน์ เสมือนเป็นห้างสรรพสินค้า ทำให้เกิดเป็น E-Marketplace หรือตลาดกลางพาณิชย์อิเล็กทรอนิกส์ ซึ่งจะทำหน้าที่เป็นตัวกลางให้ผู้ซื้อ และผู้ขาย ได้พบกัน เพื่อทราบโอกาสทางธุรกิจของแต่ละฝ่าย ผู้ดำเนินการเว็บไซต์ตลาดกลางพาณิชย์อิเล็กทรอนิกส์จะควบคุมการเปลี่ยนข้อมูล สินค้า บริการ และการจ่ายเงิน ตลอดจนเป็นผู้กำหนดกติกา กฎระเบียบในการเข้าใช้บริการตลาดกลางพาณิชย์อิเล็กทรอนิกส์ เพื่อแลกเปลี่ยนสินค้าและบริการด้วยระบบที่มีความน่าเชื่อถือ ปลอดภัย เพื่ออำนวยความสะดวกและสร้างความไว้วางใจให้ผู้ใช้บริการต่อการทำธุรกรรม (สรีพร โพธิ์งาม, 2560)

ผลการสำรวจมูลค่าพาณิชย์อิเล็กทรอนิกส์ในประเทศไทยในปี พ.ศ. 2561 มีมูลค่าทั้งสิ้น 3,150,232.96 ล้านบาท ซึ่งมีอัตราการเติบโตเพิ่มขึ้นจากปี พ.ศ. 2560 คิดเป็นร้อยละ 14.04 (สำนักงาน

พัฒนาธุรกรรมทางอิเล็กทรอนิกส์, 2562) ข้อมูลดังกล่าวแสดงให้เห็นว่าธุรกิจพาณิชย์อิเล็กทรอนิกส์มีทิศทางเติบโตเพิ่มขึ้นทุกปี และเป็นกลไกสำคัญในการขับเคลื่อนเศรษฐกิจ

ปัจจัยหนึ่งที่ทำให้ธุรกิจพาณิชย์อิเล็กทรอนิกส์เติบโต คือ ความสามารถเข้าถึงอินเทอร์เน็ตได้ทุกที่ทุกเวลา รวมถึงอุปกรณ์ที่ทันสมัยในราคาที่ตอบโจทย์ ความจำเป็น และความสะดวกสบายในการสื่อสารข้อมูล จึงทำให้มีผู้ใช้อินเทอร์เน็ตเพิ่มขึ้น จากผลสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทย ปี 2561 พบว่า คนไทยใช้อินเทอร์เน็ตนานขึ้นเฉลี่ยวันละ 10 ชั่วโมง 5 นาที เพิ่มขึ้นจากปีที่แล้ว 3 ชั่วโมง 30 นาที กิจกรรมการใช้งานผ่านอินเทอร์เน็ตเพื่อซื้อสินค้าและบริการผ่านพาณิชย์อิเล็กทรอนิกส์ติดอันดับ 1 ใน 5 ของกิจกรรมการใช้งานผ่านอินเทอร์เน็ตยอดนิยมต่อเนื่องเป็นปีที่ 2 ในร้อยละที่เพิ่มสูงขึ้น จากปี 2560 ที่มีร้อยละ 50.8 เพิ่มเป็นร้อยละ 51.3 ในปี 2560 (สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์, 2561) มูลเหตุแห่งการตัดสินใจเพื่อซื้อสินค้าและบริการผ่านพาณิชย์อิเล็กทรอนิกส์อันดับหนึ่ง มาจากการโฆษณาผ่านสื่อออนไลน์ต่างๆ คิดเป็นร้อยละ 55.9 รองลงมาคือ การรีวิวและคอมเมนต์ของผู้เคยใช้สินค้า คิดเป็นร้อยละ 54.9 ส่วนลดและของแถม คิดเป็นร้อยละ 47.5 และอันดับของเว็บไซต์จากการค้นหาทาง Search Engine คิดเป็นร้อยละ 41.9 ตามลำดับ (กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม, 2560)

ถึงแม้ผลสำรวจกิจกรรมการใช้งานผ่านอินเทอร์เน็ตที่ผ่านมา จะสะท้อนให้เห็นถึงความนิยมและความสำคัญของธุรกิจพาณิชย์อิเล็กทรอนิกส์ แต่ในทางกลับกันการพาณิชย์อิเล็กทรอนิกส์ก็มีช่องโหว่ให้มิจฉาชีพแฝงตัวเข้ามาได้ เช่น การสร้างร้านค้าปลอม การหลอกให้โอนเงินแต่ไม่ส่งมอบสินค้า การลักลอบนำข้อมูลส่วนตัวของลูกค้าไปใช้งาน หรือนำไปขายให้กับองค์กรต่าง ๆ ตลอดจนการโจรกรรมต่าง ๆ ที่อาจนำมาซึ่งความเสียหายต่อชีวิตและทรัพย์สิน (สำนักพัฒนาธุรกรรมทางอิเล็กทรอนิกส์, 2557)

การป้องกันและรับมือกับการทุจริตบนโลกออนไลน์จึงต้องอาศัยความรวดเร็ว ทันเหตุการณ์ ด้วยเทคโนโลยีที่ทันสมัย เพราะเมื่อผู้ร้ายถูกจับได้ ก็จะมีการเปลี่ยนรูปแบบการโกงรูปแบบใหม่ซึ่งยากแก่การตรวจสอบมากขึ้น เช่น คนร้ายสั่งซื้อสินค้ากับเจ้าของบัญชีตัวจริง เพื่อจะนำบัญชีดังกล่าวไปหลอกเยื่อ เมื่อคนร้ายหลอกขายสินค้ากับเหยื่อ โดยอ้างบัญชีจริงของเจ้าของสินค้า จากนั้นเหยื่อโอนเงินเข้าบัญชีของเจ้าของสินค้าตัวจริง เนื่องจากเข้าใจว่าเป็นบัญชีของคนร้าย เจ้าของตัวจริงจัดส่งสินค้าให้กับคนร้าย แต่คนร้ายกลับไม่ส่งสินค้าให้กับเหยื่อ ทำให้เหยื่อได้รับความเสียหาย จึงแจ้งความดำเนินคดีกับเจ้าของบัญชี ในขณะที่คนร้ายเมื่อได้รับสินค้าแล้ว ก็ทำการย้ายที่อยู่ หรือใช้ที่อยู่ของ

คนรู้จักในการจัดส่งสินค้า ซึ่งเป็นการหลอกลวงอีกต่อหนึ่ง (สำนักงานตำรวจแห่งชาติ, 2562) รายละเอียดดังภาพที่ 1.1

ภาพที่ 1.1 แสดงการหลอกลวงซื้อขายสินค้าออนไลน์รูปแบบใหม่

ด้วยรูปแบบการทุจริตข้างต้น จะดูเหมือนไม่มีความผิดปกติแต่อย่างใด จนกระทั่งเหยื่อไม่ได้รับสินค้า และมีการแจ้งความดำเนินคดี ทำให้เกิดความเสียหายต่อเหยื่อและเจ้าของบัญชีตัวจริงไปมากแล้ว แต่ถ้านำข้อมูลรายธุรกรรมเข้าสู่กระบวนการพิจารณาร่วมกันทั้งหมดจึงจะเห็นความผิดปกติขึ้น

ดังนั้น งานวิจัยนี้จะนำเสนอวิธีการทำนายธุรกรรมที่หลอกลวง โดยใช้เทคนิคการเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะเด่น มาช่วยทำนายธุรกรรมที่มีแนวโน้มจะหลอกลวง เพื่อหาทางป้องกันธุรกรรมที่จะหลอกลวงในอนาคต

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาลักษณะธุรกิจพาณิชย์อิเล็กทรอนิกส์
2. เพื่อศึกษาลักษณะของการหลอกลวงการซื้อขายออนไลน์
3. เพื่อวิเคราะห์หาวิธีการทำนายธุรกรรมที่หลอกลวง (Fraudulent transactions)

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1. เข้าใจลักษณะธุรกิจพาณิชย์อิเล็กทรอนิกส์
2. เข้าใจลักษณะของการหลอกลวงการซื้อขายออนไลน์
3. สามารถวิเคราะห์หาวิธีการทำนายธุรกรรมที่หลอกลวง และนำวิธีการที่ได้ไปใช้เพื่อช่วยให้การสร้างโมเดลประเภทต่าง ๆ ให้มีประสิทธิภาพมากยิ่งขึ้น

1.4 ขอบเขตของงานวิจัย

1. เป็นการศึกษาลักษณะของการหลอกลวงการซื้อขายออนไลน์
2. เป็นการศึกษาเพื่อเสนอวิธีสกัดคุณลักษณะเด่น (Feature Extraction) จากข้อมูลรายธุรกรรม
3. เป็นการศึกษาเพื่อเพิ่มประสิทธิภาพความถูกต้องแม่นยำของโมเดลการเรียนรู้ของที่ใช้ในการจำแนกข้อมูล ในกรณีข้อมูลที่ศึกษาเป็นข้อมูลรายธุรกรรม

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 พาณิชย์อิเล็กทรอนิกส์และตลาดกลางพาณิชย์อิเล็กทรอนิกส์

จุดกำเนิดของการพาณิชย์อิเล็กทรอนิกส์ (Electronic Commerce) หรือ อีคอมเมิร์ซ (E-Commerce) เติบโตมาในช่วงปี ค.ศ. 1970 ด้วยเทคโนโลยีที่เรียกว่า “การโอนย้ายทุนทางอิเล็กทรอนิกส์ (Electronic Funds Transfer : EFT)” ที่ใช้ในธุรกิจธนาคารเพื่อการโอนข้อมูลเกี่ยวกับบัญชีของลูกค้าผ่านเครือข่ายภายในของธนาคาร จากนั้นข้อมูลอื่นที่นอกเหนือจากนั้น หรือข้อมูลอื่นที่นอกเหนือจากข้อมูลทางการเงินก็สามารถโอนย้ายระหว่างบริษัทหนึ่งไปยังอีกบริษัทหนึ่งได้ด้วยเทคโนโลยีที่เรียกว่า “การแลกเปลี่ยนข้อมูลทางอิเล็กทรอนิกส์ (Electronic Data Interchange : EDI) โดยข้อมูลส่วนใหญ่จะเป็นใบกำกับภาษี คำสั่งซื้อสินค้า และเอกสารเกี่ยวกับการขนส่ง เป็นต้น เทคโนโลยีข้อมูลทางอิเล็กทรอนิกส์นี้ ถูกใช้ในกระบวนการสั่งซื้อสินค้า และสร้างสายสัมพันธ์กับผู้ส่งมอบสินค้าและบริการ (Suppliers) ในอุตสาหกรรมการผลิต การบริการ ธุรกิจค้าปลีกและอื่นๆ ให้มีประสิทธิภาพมากยิ่งขึ้น หลังจากนั้นการพาณิชย์อิเล็กทรอนิกส์ถูกพัฒนาอย่างเรื่อยมาจนเป็นที่นิยมอย่างเช่นทุกวันนี้

องค์กรความร่วมมือและพัฒนาทางเศรษฐกิจ (Organization for Economic Co-Operation and Development : OECD) ได้นิยาม การพาณิชย์อิเล็กทรอนิกส์ (Electronic Commerce) หรือ อีคอมเมิร์ซ (E-Commerce) ว่าหมายถึง ธุรกิจที่มีการขายสินค้า หรือบริการให้ลูกค้าผ่านอินเทอร์เน็ต หรือหมายถึงมีการให้ลูกค้าส่งคำสั่งซื้อ สั่งจองสินค้า หรือบริการผ่านทางอินเทอร์เน็ต (ผ่านทางหน้าเว็บไซต์ หรือทางอีเมล) ส่วนการชำระเงินและการจัดส่ง จัดทำผ่านช่องทางใดก็ได้ ซึ่งจะนับรวมคำสั่งซื้อที่ได้รับจาก Internet application เช่น เว็บไซต์ หรือโปรแกรมอื่นๆ ที่ทำงานผ่านทางอินเทอร์เน็ต เช่น EDI ที่ผ่านทางอินเทอร์เน็ต, Minitel ที่ผ่านทางอินเทอร์เน็ต หรือผ่านทางเว็บไซต์อื่นๆ โดยไม่คำนึงถึงวิธีการที่เข้าถึงเว็บไซต์เหล่านั้น (เช่น เข้าเว็บไซต์ผ่านมือถือ หรือโทรทัศน์ เป็นต้น) แต่จะไม่

รวมคำสั่งซื้อที่ได้รับทางโทรศัพท์ โทรสาร หรือการโต้ตอบผ่านทางอีเมล์ (กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม, 2560) จุดเด่นของพาณิชย์อิเล็กทรอนิกส์คือ ต้นทุนการลงทุนต่ำ ลดค่าใช้จ่ายในการดำเนินงาน เพิ่มประสิทธิภาพในการดำเนินธุรกิจ และช่วยลดองค์ประกอบธุรกิจที่มองเห็นจับต้องได้ เช่น การใช้อาคารสำหรับจัดแสดงสินค้า คลังสินค้า พนักงานขาย เป็นต้น จึงเป็นการลดการใช้ทรัพยากรในองค์กร ลดข้อจำกัดของระยะทางและเวลาลงได้

ตลาดกลางพาณิชย์อิเล็กทรอนิกส์ (Electronics Marketplace : E-Marketplace) ก็เป็นหนึ่งในประเภทธุรกิจของพาณิชย์อิเล็กทรอนิกส์ ที่เป็นตลาดกลางรวบรวมสินค้า ร้านค้า หรือบริษัทจำนวนมาก เพื่อเป็นสื่อกลางในการซื้อขายสินค้าระหว่างกัน โดยรูปแบบของตลาดกลางพาณิชย์อิเล็กทรอนิกส์จะเป็นการบริการในรูปแบบของเว็บไซต์ ที่เปิดให้บริการโดยสามารถนำข้อมูลธุรกิจและข้อมูลสินค้าไปใส่ไว้ในตลาดกลางพาณิชย์อิเล็กทรอนิกส์เหล่านั้นได้ในรูปแบบของการสร้างเว็บไซต์ แคตตาล็อกสินค้า และส่วนใหญ่ในเว็บไซต์เหล่านี้จะเป็นแหล่งที่มีคนเข้ามาหาข้อมูลสินค้าอยู่เป็นประจำมากมายในแต่ละวันเหมือนกับตลาดนัด แต่เป็นตลาดนัดออนไลน์ขนาดใหญ่ รูปแบบธุรกรรมทางพาณิชย์อิเล็กทรอนิกส์ที่มีบทบาทในการให้บริการใน 2 รูปแบบ (Efraim Turban et al., 2018) คือ

1. เป็นตัวกลางให้ผู้ซื้อ (Buyers) และผู้ขาย (Seller) ได้พบปะ เพื่อทราบถึงโอกาสในทางธุรกิจของแต่ละฝ่าย

2. เป็นตัวกลางในการแลกเปลี่ยนข้อมูล สินค้า บริการ และการจ่ายเงินตามธุรกรรมที่ได้เกิดขึ้นจากการซื้อขายสินค้าและบริการ

3. เป็นผู้กำหนดกติกา กฎระเบียบในการเข้าใช้บริการตลาดกลางพาณิชย์อิเล็กทรอนิกส์เพื่อแลกเปลี่ยนสินค้า บริการ และธุรกรรม เพื่ออำนวยความสะดวก และความเชื่อมั่นให้ผู้ใช้บริการมีความมั่นใจ และสะดวกในการใช้บริการมากขึ้น

ซึ่งตลาดกลางพาณิชย์อิเล็กทรอนิกส์ เป็นการดำเนินงานที่เปลี่ยนแปลงรูปแบบของการทำธุรกรรมเชิงพาณิชย์ การค้าขายแลกเปลี่ยน และห่วงโซ่อุปทาน (Supply chain) ของตลาดรูปแบบดั้งเดิมดังต่อไปนี้

1. เนื่องจากการประยุกต์ระบบเทคโนโลยีสารสนเทศของตลาดกลางพาณิชย์อิเล็กทรอนิกส์ทำให้การเข้าถึงข้อมูลเพื่อประโยชน์ในการซื้อขาย แลกเปลี่ยน และทำธุรกรรมได้อย่างรวดเร็ว และมีประสิทธิภาพ

2. ตลาดกลางพาณิชย์อิเล็กทรอนิกส์ช่วยลดเวลาในการค้นหาข้อมูลต่ำลง

3. ลดความเหลื่อมล้ำและความไม่เท่าเทียมกันของข้อมูลระหว่างผู้ซื้อและผู้ขาย

4. ลดระยะเวลาของการติดต่อซื้อขายแลกเปลี่ยนสินค้า และการตัดสินใจทำธุรกรรม

5. ตลาดกลางพาณิชย์อิเล็กทรอนิกส์ช่วยเพิ่มความสะดวกในการติดต่อระหว่างผู้ซื้อและผู้ขาย โดยที่ไม่ต้องพบปะในสถานที่เดียวกัน

2.2 การสกัดคุณลักษณะ (Feature Extraction)

การสกัดคุณลักษณะ (Feature Extraction) เป็นวิธีการค้นหาคุณลักษณะใหม่ที่มีจำนวนมิติข้อมูลลดลง และยังสามารถกำจัดข้อมูลที่ไม่เกี่ยวข้องหรือไม่สัมพันธ์กับผลลัพธ์ ทำให้สามารถเพิ่มประสิทธิภาพของโมเดลได้อีกด้วย การสกัดคุณลักษณะสามารถแบ่งได้ 2 ประเภทได้แก่

1. การสกัดคุณลักษณะแบบไม่มีผู้สอน (Unsupervised Feature Extraction) เป็นการสกัดคุณลักษณะโดยที่ไม่รู้ผลลัพธ์มาก่อน แบ่งออกเป็น 2 ประเภท (Laurens van der Maaten et al., 2009) ได้แก่

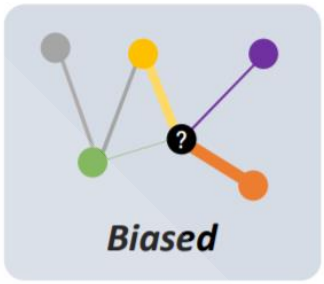
1.1 วิธีคอนเวกซ์ (Convex) วิธีนี้จะมีฟังก์ชันจุดประสงค์ที่มีค่าเหมาะสมโดยรวม (Global Optima) กับค่าเหมาะสมเฉพาะที่เป็นค่าเดียวกัน เช่น PCA (Principle Components Analysis) Kernel-PCA หรือ MDS (Multi-Dimensional Scaling) เป็นต้น ข้อดีของวิธีนี้คือ สามารถนำไปใช้ได้ง่ายและรับประกันผลลัพธ์ตามข้อสันนิษฐาน

1.2 วิธีไม่คอนเว็กซ์ วิธีนี้จะไม่รับประกันว่าผลลัพธ์จะได้ตามสมมติฐานหรือไม่ เนื่องจากปัญหาไม่คอนเว็กซ์จะหาค่าที่เหมาะสมที่สุดได้ยาก เช่น Multilayer Autoencoders หรือ LLC (Locally Linear Coordination) เป็นต้น

2. การสกัดคุณลักษณะแบบมีผู้สอน (Supervise Feature Extraction) โดยทั่วไปวิธีการนี้จะทำงานได้ดีกว่าวิธีการสกัดคุณลักษณะแบบไม่มีผู้สอน เนื่องจากนำข้อมูลประเภทที่ทราบอยู่ก่อนมาใช้ในขั้นตอนการเทรน ทำให้โมเดลที่ได้มีความแม่นยำมากขึ้น

2.3 Node2Vec

การสกัดคุณลักษณะเมื่อข้อมูลอยู่ในรูปแบบระเบียนข้อมูลวิธีหนึ่งคือ Node2Vec เทคนิคนี้จะทำการสร้าง Low-dimensional ในแต่ละโหนด โดยจำลองการสุ่มเดินแบบลำเอียง (Random biased walks) ดังภาพที่ 2.1



ภาพที่ 2.1 แสดงการสุ่มเดินแบบลำเอียง

จากนั้น คำนวณค่าความน่าจะเป็นของเส้นเชื่อมในแต่ละโหนด

$$P(e)=weight(e)\cdot\alpha_{pq}$$

โดยที่

$$\alpha_{pq}(t,x)=\begin{cases}\frac{1}{p} & \text{if } d_{tx}=0\\ 1 & \text{if } d_{tx}=1\\ \frac{1}{q} & \text{if } d_{tx}=2\end{cases}$$

เมื่อ d_{tx} คือ เส้นทางเดินที่สั้นที่สุดระหว่างโหนด t และ x

p และ q คือ ค่าคงที่ hyper-parameters

ข้อดีของ Node2Vec คือ สามารถคำนวณค่าความน่าจะเป็นของเส้นเชื่อมในแต่ละโหนด เพื่อใช้เป็นแนวเดินในกราฟ และจัดกลุ่มโหนดที่มีคุณลักษณะคลายคลึงกัน (Homophily) เข้าไว้ด้วยกันได้ดีมากยิ่งขึ้น

2.4 ทฤษฎีกราฟ (Graph Theory)

การวิเคราะห์เครือข่ายทางสังคม (Social Network Analysis : SNA) เริ่มจากการนำทฤษฎีความรู้ด้านต่างๆ เช่น สังคม คณิตศาสตร์ สถิติ และการคำนวณมาใช้ร่วมกัน โดยทฤษฎีกราฟ (Graph Theory) ถูกนำมาใช้ในการวิเคราะห์เครือข่ายทางสังคม เพื่อระบุส่วนประกอบของเครือข่าย บ่งบอกลักษณะความสัมพันธ์ หรือเปรียบเทียบความแตกต่างในเครือข่าย โดยสามารถใช้ทฤษฎีกราฟอธิบายลักษณะในแต่ละโหนดได้ สามารถขียนนิยามของกราฟได้ดังนี้ (วราภรณ์ พิมา, 2561)

$$G = (V, E)$$

โดยที่ V คือ โหนด หรือสมาชิกในเครือข่าย

E คือ เส้นเชื่อม

การวัดความสัมพันธ์ภายในเครือข่าย เป็นการค้นหาโหนดที่มีตำแหน่งเป็นศูนย์กลาง (Central) เป็นตำแหน่งที่อำนาจเหนือกว่าโหนดอื่น และมีศักยภาพในการควบคุมเครือข่าย สามารถจำแนกรูปแบบของความเป็นศูนย์กลางได้ดังนี้ (ธนพล พุกเส็ง, 2563)

1. Degree Centrality เป็นการวัดจำนวนการเชื่อมโยงที่เข้าสู่ (In-Degree) และออกจาก (Out-Degree) โหนดในเครือข่าย คำนวณได้ดังสมการ

$$d(i) = \sum_j m_{ij}$$

โดยที่ d(i) คือ ค่าความเป็นศูนย์กลาง โดยวัดจากดีกรีของการเชื่อมโยงที่โหนด i ใดๆ

m_{ij} จะมีค่าเท่ากับ 1 ถ้าหากมีการเชื่อมโยงระหว่างจุดยอด และจะมีค่าเท่ากับ 0 ถ้าหากไม่มีการเชื่อมโยงระหว่างกัน

2. Closeness Centrality เป็นการวัดระยะที่ใกล้ที่สุดจากโหนดหนึ่งไปยังโหนดอื่นๆ ในเครือข่าย การมี Closeness centrality สูง หมายถึง โหนดที่มีความใกล้ชิดกับโหนดอื่นๆ ในเครือข่ายสูง มีความสามารถในการติดต่อสื่อสาร หรือเข้าถึงโหนดอื่นในเครือข่ายได้อย่างรวดเร็ว คำนวณได้ดังสมการ

$$c(i) = \sum_j n_{ij}$$

โดยที่ c(i) คือ ค่าความเป็นศูนย์กลาง โดยวัดจากความใกล้ชิดของการเชื่อมโยงที่โหนด i ใดๆ

n_{ij} คือ จำนวนเส้นเชื่อมโยงในเส้นทางที่สั้นที่สุดจากโหนดหนึ่งไปยังอีกโหนดหนึ่ง

3. Betweenness Centrality เป็นการวัดจุดคั่นกลาง หรือตำแหน่งที่เป็นสะพาน (Bridges) เชื่อมโหนดต่างๆ เข้าหากัน การมี Betweenness centrality สูง หมายถึง โหนดที่เป็นทางผ่านของคู่อื่นๆ ในเครือข่ายสูง คำนวณได้ดังสมการ

$$b(i) = \sum_{j,k} \frac{g_{jik}}{g_{jk}}$$

โดยที่ b(i) คือ ค่าความเป็นศูนย์กลาง โดยวัดจากการคั่นกลางของการเชื่อมโยงที่โหนด i ใดๆ

g_{jk} คือ จำนวนเส้นทางที่สั้นที่สุดจากโหนด j ไปยังโหนด k (j, k≠i)

g_{jik} คือ จำนวนเส้นทางที่สั้นที่สุดจากโหนด j ไปยังโหนด k ที่ต้องผ่านโหนด i

4. Eigenvector Centrality เป็นการวัดจุดศูนย์กลางโดยวัดจากเวกเตอร์ลักษณะเฉพาะ เป็นการคำนวณค่าความเป็นจุดศูนย์กลางของเครือข่ายจากการวัดค่าอิทธิพลของโหนดในเครือข่าย โดยหากโหนดนั้นเชื่อมโยงกับโหนดอื่นที่มีค่าอิทธิพลสูงอยู่แล้ว ก็จะมีค่าเวกเตอร์ลักษณะเฉพาะสูงกว่าโหนดที่มีค่าอิทธิพลต่ำ คำนวณได้ดังสมการ

$$ev(i) = \frac{1}{\lambda}\left(\sum_{t \in V(t)} t\right)$$

โดยที่ ev(i) คือ ค่าความเป็นศูนย์กลาง โดยวัดจากเวกเตอร์ลักษณะเฉพาะของการเชื่อมโยงที่โหนด i ใดๆ โดยที่ t เป็นสมาชิกของ V(t)

V(i) คือ ชุดของโหนดที่เชื่อมโยงไปยังโหนด i

λ คือ ค่าคงที่

2.5 การจำแนกประเภทข้อมูล (Classification)

เป็นการสร้างโมเดลสำหรับการจำแนกประเภทข้อมูลจากแอททริบิวต์และลาเบลคำตอบ เพื่อจัดข้อมูลให้อยู่ในกลุ่มที่กำหนด

1. ต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคที่นำการตัดสินใจที่มีลักษณะคล้ายต้นไม้ มีการแตกแขนงกิ่งไปตามเงื่อนไข ถ้า (เงื่อนไข) แล้ว (ผลลัพธ์) หรือ if-then rule สำหรับโครงสร้างต้นไม้ประกอบไปด้วย

1. โหนด (Node) แสดงถึงคุณลักษณะที่นำมาใช้ในการแบ่งกลุ่มของข้อมูล โดยมีโหนดราก (Root Node) อยู่บนสุด และเป็นโหนดที่มีอิทธิพลต่อการจำแนกกลุ่มมากที่สุดของโครงสร้าง

2. กิ่ง (Branch) เป็นตัวเชื่อมระหว่างโหนด ที่ใช้เป็นเงื่อนไข หรือทางเลือกของการกระทำ

3. ใบ (Leaf Node) เป็นโหนดแสดงผลลัพธ์ของเงื่อนไข หรือการกระทำตามเงื่อนไขที่เกิดขึ้น

ต้นไม้ตัดสินใจที่นิยมใช้งานอย่างแพร่หลายโดยเฉพาะ C4.5 และ C5.0 ซึ่งความแตกต่างระหว่างสองแบบนี้คือ ความเร็วในการทำงานและการใช้ทรัพยากรของคอมพิวเตอร์ โดยที่ C5.0 มีความสามารถที่ดีกว่า C4.5 รวมถึงความแม่นยำในการจำแนก เพราะ C5.0 มีการนำค่าความผิดพลาดในการจำแนก (Variable Misclassification Costs) มาใช้แบบแยกส่วนไปตามแต่ละตัวอย่างข้อมูล ในขณะที่ C4.5 มองว่าค่าความผิดพลาดของแต่ละตัวอย่างนั้นมีลักษณะเหมือนกัน

ข้อดีของต้นไม้ตัดสินใจที่ชัดเจนที่สุดคือเป็นโมเดลที่มีความซับซ้อนต่ำ จึงทำให้ผลลัพธ์ที่ได้จากการทำงานสามารถแปรผลได้ง่าย สามารถทำงานกับข้อมูลที่มีตัวแปรมีความสัมพันธ์กันแบบไม่เป็นเส้นตรงได้ดี นอกจากนี้ยังสามารถฝึกได้อย่างรวดเร็ว อย่างไรก็ดี เนื่องจากโมเดลนี้ใช้หลักการแบ่งแยกข้อมูลโดยอ้างอิงตามค่าของข้อมูลเป็นหลัก ทำให้บางครั้งประสิทธิภาพความแม่นยำของโมเดลไม่สามารถเพิ่มขึ้นตามขนาดของข้อมูลได้ ถ้าหากตัวแปรที่ใช้มีค่าความสัมพันธ์ระหว่างกันต่ำเป็นจำนวนมากเกินไปหรือตัวแปรที่ใช้เป็นค่าต่อเนื่องก็อาจทำให้ประสิทธิภาพของโมเดลนี้ลดลงได้

2. การวิเคราะห์ถดถอยโลจิสติก (Logistics Regression) เป็นเทคนิคการวิเคราะห์การถดถอยรูปแบบหนึ่ง ที่ลาเบลคำตอบเป็นข้อมูลเชิงคุณภาพ ส่วนแอทริบิวต์เป็นได้ทั้งข้อมูลเชิงปริมาณและเชิงคุณภาพ โดยวัตถุประสงค์หลักของการวิเคราะห์การถดถอยโลจีสติก คือ การประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (David W Hosmer, 1989)

การวิเคราะห์การถดถอยโลจีสติกแบ่งออกเป็น 2 ประเภท คือ

1. Binary Logistics Regression คือ เทคนิคที่ลาเบลคำตอบเป็นข้อมูลเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า เช่น ป่วย หรือ ไม่ป่วย

2. Multinomial Logistics Regression คือ เทคนิคที่ลาเบลคำตอบเป็นข้อมูลเชิงคุณภาพที่มีค่ามากกว่า 2 ค่า เช่น มะเร็งขั้นต้น ขั้นกลาง ขั้นสุดท้าย

ข้อตกลงเบื้องต้นของการวิเคราะห์ถดถอยแบบโลจิสติก คือ

1. แอททริบิวต์ หรือลาเบลคำตอบต้องเป็นข้อมูลระดับช่วง (Interval Scale) เป็นอย่างน้อย
2. ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์
3. แอททริบิวต์ไม่มีความสัมพันธ์กัน หรือไม่เกิดปัญหา Multicollinearity

สมการการวิเคราะห์การถดถอยโลจีสติก จะเป็นสมการแสดงความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (Probability of Event) เขียนได้ดังนี้

$$P(Event)= \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

เมื่อ β_0 คือ ค่าคงที่

β_i คือ ค่าสัมประสิทธิ์ของแต่ละแอททริบิวต์

X_i คือ แอททริบิวต์

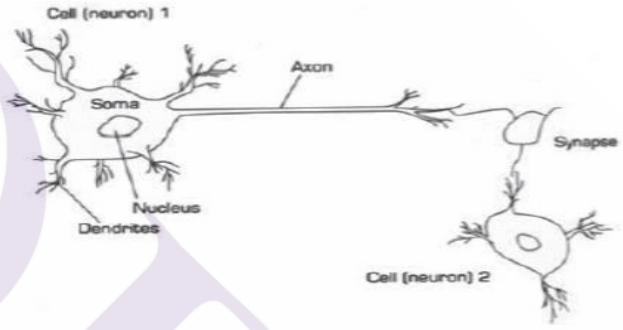
e คือ Exponential function

จากสมการข้างต้น แสดงความสัมพันธ์ระหว่างแอททริบิวต์และลาเบลคำตอบไม่ได้อยู่ในรูปเชิงเส้นตรง จึงต้องปรับให้อยู่ในรูปเชิงเส้นโดยใช้การแปลงลอการิทึม (Logarithm transformation) และสามาถแสดงให้อยู่ในรูปสมการถดถอยเชิงเส้นได้ดังนี้

$$\log(odds) = \beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n$$

การวิเคราะห์การถดถอยโลจีสติกนี้ สามารถทำงานได้รวดเร็วและมีประสิทธิภาพความแม่นยำค่อนข้างสูง สามารถนำสมการผลลัพธ์ที่ได้ไปใช้ในการเลือกแอททริบิวต์ที่สำคัญได้ และยังทนทานต่อข้อมูลที่แปลกแยก (Outlier) ในระดับหนึ่ง แต่ข้อเสียคือไม่สามารถทำงานได้ดี หากแอททริบิวต์มีความสัมพันธ์เชิงเส้นตรงระหว่างกันเองสูงและจำนวนมากเกินไป (Multicollinearity)

3. โครงข่ายประสาทเทียม (Neural Network) เป็นเทคนิคที่จำลองการทำงานของเซลล์สมองของมนุษย์ ที่แต่ละเซลล์ประสาทจะประกอบไปด้วยเดนไดรท์ (Dendrite) ทำหน้าที่นำกระแสประสาทเข้าสู่เซลล์ประสาท หรือเป็นอินพุต (Input) ของเซลล์ นิวรอน (Neurons) ทำหน้าที่ในการประมวลผลต่างๆ ไซแนป (Synapses) ทำหน้าที่ในการเชื่อมต่อโครงข่ายประสาทเข้าด้วยกัน ด้วยการส่งสัญญาณระหว่างเซลล์ประสาท และแอกซอน (Axon) ทำหน้าที่ส่งกระแสประสาทไปจากตัวเซลล์ หรือเป็นเอาท์พุต (Output) ไปยังเซลล์ตัวอื่น โครงสร้างของเซลล์ประสาทดังแสดงในภาพที่ 2.2



ภาพที่ 2.2 แสดงโครงสร้างของเซลล์ประสาท

หลักการทำงานของโครงข่ายประสาทเทียม จะทำการรวบรวมข้อมูล (Knowledge) โดยผ่านกระบวนการเรียนรู้ (Learning Process) และจัดเก็บความรู้เหล่านั้นในโครงข่ายประสาทเทียมในรูปของค่าน้ำหนัก (Weight) ซึ่งสามารถปรับเปลี่ยนค่าได้ เมื่อมีการเรียนรู้ใหม่ๆ เกิดขึ้น องค์ประกอบที่สำคัญของโครงข่ายประสาทเทียม มีดังนี้

1. ข้อมูลอินพุต (Input) จะต้องเป็นข้อมูลเชิงปริมาณ หากเป็นข้อมูลเชิงกลุ่ม ต้องแปลงให้เป็นข้อมูลเชิงปริมาณก่อน

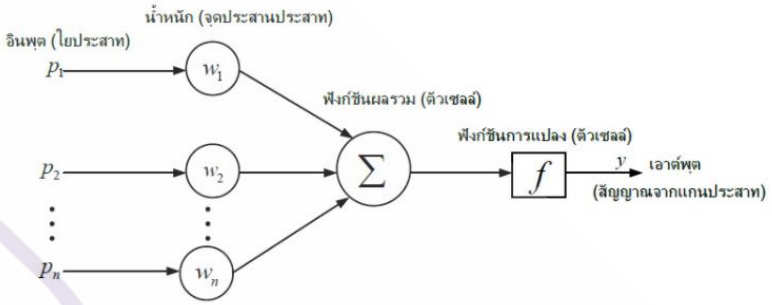
2. ค่าน้ำหนัก (Weight) คือ สิ่งที่ได้จากการเรียนรู้ของโครงข่ายประสาทเทียม หรือค่าความรู้ (Knowledge) ค่านี้จะถูกเก็บเป็นทักษะเพื่อใช้ในการจดจำข้อมูลอื่นๆ ที่อยู่ในรูปแบบเดียวกัน

3. ข้อมูลเอาท์พุต (Output) คือ ผลลัพธ์ที่เกิดขึ้น

4. ฟังก์ชันผลรวม (Summary function) เป็นผลรวมของข้อมูลอินพุตและค่าน้ำหนัก

5. ฟังก์ชันกระตุ้น (Activation function) หรือฟังก์ชันการแปลง (Transfer function) ทำหน้าที่รวมค่าเชิงตัวเลขจากเอาท์พุตของนิวรอน แล้วทำการตัดสินใจว่าจะส่งสัญญาณเอาท์พุตออกไป

ในรูปแบบใด ฟังก์ชันกระตุ้นมีหลากหลายรูปแบบ เช่น ฟังก์ชันเชิงเส้น (Linear function) ฟังก์ชันซิกมอยด์ (Sigmoid function) หรือฟังก์ชันไฮเปอร์โบลิกแทนเจนต์ (Hyperbolic tangent function)



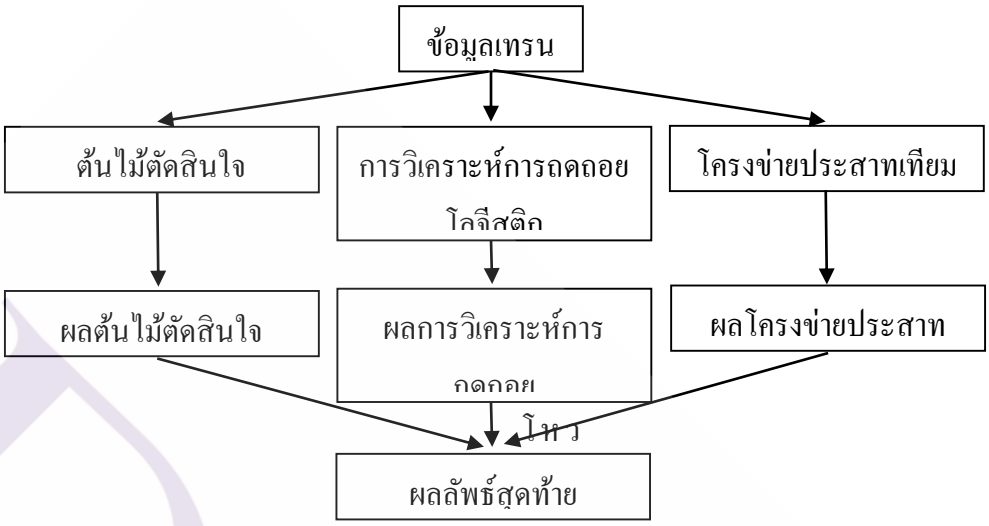
ภาพที่ 2.3 หลักการทำงานของโครงข่ายประสาทเทียม

ข้อดีของเทคนิคนี้คือ มีความยืดหยุ่นในการค้นหาผลลัพธ์ ไม่ว่าข้อมูลอินพุตจะอยู่ในรูปเชิงเส้นตรงหรือไม่ รวมไปถึงความแปรปรวนของข้อมูลคงที่หรือไม่ก็ตาม แต่เนื่องจากการเรียนรู้ได้อย่างละเอียดของเทคนิคนี้ จะทำให้มีโอกาสเกิดภาวะการเรียนรู้มากเกินไปได้ง่าย มากไปกว่านั้นเทคนิคนี้มีความซับซ้อนสูง จึงส่งผลให้การแปรผลทำได้ยากตามไปด้วย ทำให้การออกแบบโครงสร้างและการกำหนดพารามิเตอร์ต่าง ๆ มีความสำคัญ เพราะจะส่งผลต่อประสิทธิภาพความถูกต้องแม่นยำโดยตรง

2.6 การเรียนรู้แบบกลุ่ม (Ensemble)

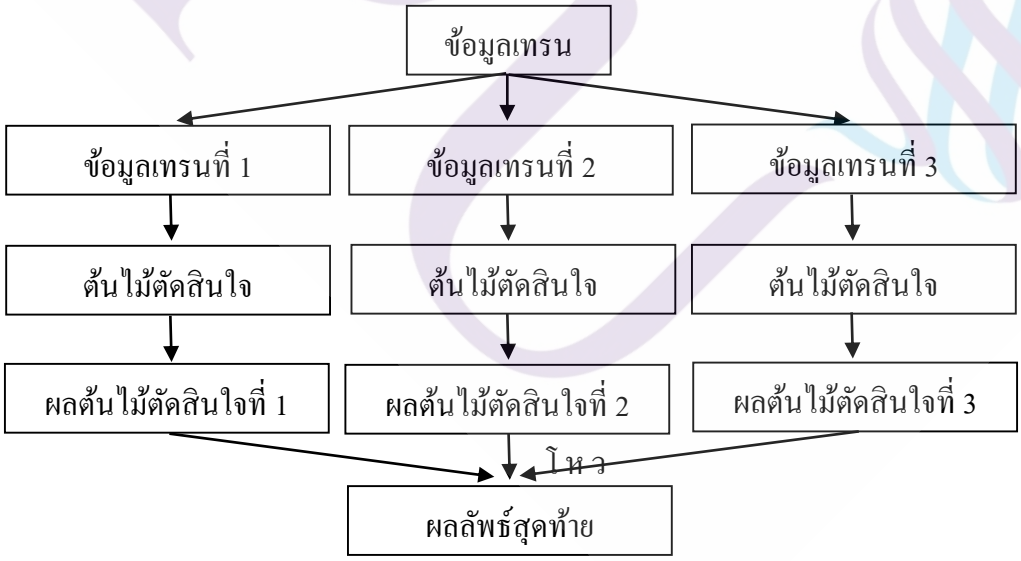
การเรียนรู้แบบกลุ่ม เป็นการรวมเทคนิคการจำแนกประเภทข้อมูลหลาย ๆ เทคนิคเข้าด้วยกัน โดยใช้ข้อมูลชุดเดียวกันในการทำงาน แต่ต่างกันที่กระบวนการทำงานในแต่ละเทคนิคออกไป จากนั้นทำการตัดสินผลที่ได้จากแต่ละเทคนิค เพื่อให้ได้ผลลัพธ์สุดท้ายเพียงผลลัพธ์เดียว การเรียนรู้แบบกลุ่มแบ่งออกเป็น 3 ประเภท ดังนี้

1. เทคนิค Vote เป็นเทคนิคที่ใช้การจำแนกประเภทข้อมูลหลาย ๆ เทคนิคเข้าด้วยกัน และตัดสินผลที่ได้ในแต่ละเทคนิคด้วยวิธีโหวต รายละเอียดดังแสดงในภาพที่ 2.4



ภาพที่ 2.4 ขั้นตอนการทำงานของเทคนิค Vote

2. เทคนิค Bootstrap Aggregating (Bagging) เป็นเทคนิคที่ใช้การจำแนกประเภทข้อมูลเพียงเทคนิคเดียว แต่ใช้ข้อมูลในการเทรนแตกต่างกัน รายละเอียดดังแสดงในภาพที่ 2.5



ภาพที่ 2.5 ขั้นตอนการทำงานของเทคนิค Bagging

3. เทคนิค Random Forest หลักการทำงานคล้ายเทคนิค Bagging แต่แตกต่างกันตรงที่เทคนิค Random Forest จะมีการสุ่มแอททริบิวต์ด้วย

2.7 การคัดเลือกคุณลักษณะ (Feature Selection)

การคัดเลือกคุณลักษณะเป็นขั้นตอนสำคัญ (Jiawei Han et al., 2006) ในการเตรียมข้อมูลก่อนการสร้างโมเดล นอกจากจะช่วยลดระยะเวลาในการสร้างโมเดลแล้ว ยังช่วยตัดคุณลักษณะที่ไม่จำเป็นต่อการสร้างโมเดลอีกด้วย การคัดเลือกคุณลักษณะแบ่งออกเป็น 2 วิธี (อัจจิมา มณฑาพันธุ์, 2562) ได้แก่

1. Filter Approach เป็นการคัดเลือกคุณลักษณะโดยการคำนวณค่าน้ำหนัก หรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ โดยคุณลักษณะที่เกี่ยวข้องน้อยจะถูกคัดออกไม่นำมาสร้างโมเดล ตัวอย่างของ Filter Approach เช่น เทคนิค Correlation Based Feature Selection (CFS) เทคนิค Information Gain (IG) เทคนิค Gain Ratio (GR) เทคนิค Chi-Square เป็นต้น

ข้อดี คือ เป็นเทคนิคที่ใช้ได้รวดเร็ว เหมาะกับคุณลักษณะที่มีจำนวนมาก เพราะไม่ต้องใช้อัลกอริทึมในการสร้างโมเดลมาเกี่ยวข้อง และทำการคำนวณหาความเกี่ยวข้องของคุณลักษณะเพียงครั้งเดียว จึงไม่ต้องใช้ทรัพยากรในการคำนวณ

ข้อเสีย คือ คุณลักษณะแต่ละตัวจะถูกวัดคะแนนความเกี่ยวข้องแยกกันกับการสร้างโมเดล ดังนั้นคุณลักษณะอาจจะไม่เหมาะที่จะนำไปสร้างโมเดลก็ได้ (ลวณา มสารกรัณฑ์, 2561)

2. Wrapper Approach เป็นการคัดเลือกคุณลักษณะโดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่ม ตัวอย่างของ Wrapper Approach ได้แก่

2.1 Forward Selection เป็นการคัดเลือกคุณลักษณะเข้าทีละตัว เพื่อดูว่าโมเดลมีประสิทธิภาพดีขึ้นหรือไม่ หากมีประสิทธิภาพดีขึ้นก็จะเก็บคุณลักษณะนั้น ทำซ้ำจนไม่สามารถเพิ่มคุณลักษณะที่ทำให้โมเดลมีประสิทธิภาพดีขึ้นได้

2.2 Recursive Feature Elimination (RFE) เป็นการนำคุณลักษณะเข้าทุกตัว แล้วคัดคุณลักษณะที่สำคัญน้อยที่สุดออก แล้วใช้คุณลักษณะที่เหลือมาสร้างโมเดลใหม่ ทำซ้ำจนไม่สามารถเพิ่มคุณลักษณะที่ทำให้โมเดลมีประสิทธิภาพดีขึ้นได้

ข้อดี คือ เป็นการหาคุณลักษณะที่มีความเกี่ยวข้องกับการสร้างโมเดลโดยตรง

2.8 การวัดประสิทธิภาพของโมเดล

การวัดประสิทธิภาพของโมเดล เป็นการวัดความแม่นยำของโมเดลก่อนการนำไปใช้งานจริง โดยทั่วไปตัววัดที่นิยมใช้มีดังนี้

1. ตารางแจกแจงผลลัพธ์ (Confusion Matrix) คือ การประเมินผลการทำนายของโมเดลเทียบกับผลที่เกิดขึ้นจริง อธิบายได้ดังนี้

1.1 True Positive (TP) คือ ค่าที่ทำนายว่าจริง และผลที่เกิดขึ้นนั้นจริง

1.2 True Negative (TN) คือ ค่าที่ทำนายว่าไม่จริง และผลที่เกิดขึ้นนั้นไม่จริง

1.3 False Positive (FP) คือ ค่าที่ทำนายว่าจริง แต่ผลที่เกิดขึ้นนั้นไม่จริง

1.4 False Negative (FN) คือ ค่าที่ทำนายว่าไม่จริง แต่ผลที่เกิดขึ้นนั้นจริง

ตารางที่ 2.1 ลักษณะตารางแจกแจงผลลัพธ์

		ผลที่เกิดขึ้นจริง	
		คำตอบเป็นบวก	**คำตอบเป็นลบ**
ผลการทำนาย	**คำตอบเป็นบวก**	ถูกต้องในเชิงบวก (True Positive)	ผิดพลาดในเชิงบวก (False Positive)
	คำตอบเป็นลบ	ผิดพลาดในเชิงลบ (False Negative)	ถูกต้องในเชิงลบ (True Negative)

2. ค่าความถูกต้อง (Accuracy) คือ ค่าที่ทำนายถูกโดยรวมที่มีต่อข้อมูลทั้งหมด คำนวณได้ดังสมการ

$$\text{accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

3. ค่าความผิดพลาด (Error) คือ ค่าที่ทำนายผิดโดยรวมที่มีต่อข้อมูลทั้งหมด คำนวณได้ดังสมการ

$$\text{Error} = 1 - \text{accuracy}$$

4. ค่าความแม่นยำ (Precision) คือ ค่าที่ทำนายว่าจริง และผลที่เกิดขึ้นนั้นจริง เทียบกับผลของการทำนายเป็นบวกทั้งหมด คำนวณได้ดังสมการ

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

5. ค่าความระลึก (Recall) คือ ค่าที่ทำนายว่าจริง และผลที่เกิดขึ้นนั้นจริง เทียบกับผลที่เกิดขึ้นจริงเป็นบวกทั้งหมด คำนวณได้ดังสมการ

$$\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

6. ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) คือ ค่าเฉลี่ยของค่ากลางของผลจากการหารจำนวนข้อมูลทั้งหมด คำนวณได้ดังสมการ

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

7. ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) คือ ค่าเฉลี่ยของค่ากลางของผลจากการหารจำนวนข้อมูลทั้งหมด คำนวณได้ดังสมการ

$$\text{GM} = \sqrt{\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \times \frac{\text{True Positive}}{\text{True Positive} + \text{Positive}}}$$

8. การตรวจสอบความถูกต้องของโมเดลโดยการแบ่งข้อมูลฝึกแบบหลายชุด (Cross Validation) คือ วิธีการที่ใช้ทดสอบความถูกต้องของโมเดลโดยการแบ่งข้อมูลออกหลายส่วนเพื่อสอนโมเดล แล้วจึงใช้ข้อมูลบางส่วนจากข้อมูลเหล่านั้นมาใช้ทดสอบบประสิทธิภาพความถูกต้องของโมเดล จำนวนการแบ่งข้อมูลสามารถแทนได้ด้วยค่า k เช่น หากต้องการแบ่งข้อมูลเพื่อทดสอบ 5 ส่วน ดังนั้น k จะมีค่าเท่ากับ 5 โดยโมเดลจะถูกฝึกด้วยข้อมูล 4 ส่วน แล้วทดสอบด้วยข้อมูล 1 ส่วนที่เหลือ ขั้นตอนนี้จะทำงานวนไปจนกระทั่งสามารถใช้ข้อมูลทุกส่วนมาฝึกและทดสอบโมเดล จากนั้นจึงนำผลการทดสอบทั้งหมด 5 ครั้งมาหาค่าเฉลี่ยเพื่อแสดงค่าความถูกต้องเฉลี่ยของโมเดล

2.9 วรรณกรรมที่เกี่ยวข้อง

ปัจจุบันข้อมูลที่มีการเชื่อมโยงกัน (Connected data) หรือข้อมูลที่สามารถระบุต้นทางและปลายทางได้ เช่น ข้อมูล Social network การเงิน การขนส่ง หรือการโทรคมนาคม เป็นต้น มีมากขึ้น การนำทฤษฎีกราฟมาประยุกต์ใช้ จึงมีส่วนสำคัญในการช่วยปรับปรุงอัลกอลิทึมต่าง ๆ ให้ทำนายผลแม่นยำยิ่งขึ้น ยกตัวอย่างงานวิจัย เช่น

เชษฐพงศ์ ปัญญาชนกุล และ อานนท์ ศักดิ์วรกุล (2559) ได้ทำการพยากรณ์การสูญเสียลูกค้าด้วยเทคนิคการวิเคราะห์เครือข่ายทางสังคมในธุรกิจโทรคมนาคม โดยนำทฤษฎีกราฟมาใช้เพื่อหาค่าระดับความเป็นศูนย์กลาง (Centrality) จากลักษณะการโทร (Voice Call) ระหว่างลูกค้า เพื่อระบุผู้ทรงอิทธิพล (Influencer) จากนั้นทำการลดขนาดของข้อมูล โดยการใช้เทคนิค K-Mean เลือกกลุ่มที่สร้างรายได้เฉลี่ยจากการโทรมากที่สุด 2 อันดับแรก ใช้ค่าความเป็นศูนย์กลาง ได้แก่ Degree, Closeness, Betweenness และ Eigenvector เป็นตัวแปรในการสร้างโมเดล และใช้เทคนิคการวิเคราะห์องค์ประกอบหลัก ในการลดจำนวนตัวแปร และใช้เทคนิค Logistics Regression ในการสร้างโมเดล และแบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดลด้วยวิธี Split test ในอัตราส่วน 70:30 ผลการศึกษาพบว่า การพยากรณ์มีความถูกต้อง (Accuracy) ร้อยละ 72.70 ความไว (Sensitivity) ร้อยละ 99.44 ความจำเพาะ (Specificity) ร้อยละ 39.23 และค่าพื้นที่ใต้กราฟ (Area under the receiver operating curve :AUC) ร้อยละ 85.81

Aditya Grover และ Jure Leskovec (2016) ได้เสนอเทคนิค node2vec ในการสกัดคุณลักษณะเด่นจากโหนดในเครือข่าย เพื่อทำนายเส้นเชื่อมเครือข่าย (Link prediction) โดยใช้ข้อมูล 3 ชุดข้อมูล ได้แก่ Facebook, PPI และ arXiv และทำการการสกัดคุณลักษณะเด่นด้วยกัน 4 วิธี ได้แก่ Spectral Clustering, Deep Walk, LINE และ node2vec และใช้ Bootstrapping ในการทำนายเส้นเชื่อม ผลพบว่า node2vec ให้ค่าพื้นที่ใต้กราฟสูงที่สุดทั้ง 3 ชุดข้อมูล (Facebook เท่ากับร้อยละ 96.80 PPI เท่ากับร้อยละ 77.19 และ arXiv เท่ากับร้อยละ 93.66) รองลงมาคือ Deepwalk (Facebook เท่ากับร้อยละ 96.80, PPI เท่ากับร้อยละ 74.41 และ arXiv เท่ากับร้อยละ 93.40) LINE (Facebook เท่ากับร้อยละ 94.90, PPI เท่ากับร้อยละ 72.49 และ arXiv เท่ากับร้อยละ 89.02) และ Spectral Clustering (Facebook เท่ากับร้อยละ 61.92, PPI เท่ากับร้อยละ 49.20 และ arXiv เท่ากับร้อยละ 57.40) ตามลำดับ

Jayesh Soni และ Himanshu Upadhyay (2019) ได้เสนอเทคนิค Deepwalk on weigh graph (DW-WG) ในการสกัดคุณลักษณะเด่นจากโหนดในเครือข่าย ซึ่งพัฒนาต่อยอดมาจากเทคนิค Conventional Deepwalk (C-DW) เพื่อสามารถสกัดคุณลักษณะของกราฟแบบมีทิศทาง เช่น ทิศทาง

เดียว หลายทิศทาง หรือกราฟที่มีน้ำหนักได้ จากนั้นทำการเปรียบเทียบค่าความถูกต้องระหว่างเทคนิคเก่าคือ C-DW และเทคนิคใหม่ คือ DW-WG กับข้อมูล Google Play Store และ Bitcoin OTC+Alpha และใช้เทคนิค SVM, Random Forest และ K-NN ในการสร้างโมเดล และแบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดลด้วยวิธี K-Fold Cross Validation ผลการศึกษาพบว่า เทคนิค DW-WG มีค่าความถูกต้องสูงกว่า C-DW ทั้ง 2 ชุดข้อมูล (Google Play store: SVM ร้อยละ 87.0 รองลงมาคือ Random Forest ร้อยละ 85.0 และ KNN ร้อยละ 81.0 ตามลำดับ ส่วนข้อมูล Bitcoin OTC+Alpha: SVM ร้อยละ 89.0 รองลงมาคือ Random Forest ร้อยละ 84.0 และ KNN ร้อยละ 84.0)

Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos และ V.S. Subrahmanian (2018) ได้เสนอเทคนิค REV2 เพื่อระบุผู้ใช้งานออนไลน์ที่หลอกลวงจากการวัดในด้าน Fairness of user, Reliability of rating และ Goodness of product โดยใช้ข้อมูล 5 ชุดข้อมูล ได้แก่ Bitcoin OTC, Bitcoin Alpha, Amazon, Flipkart และ Epinions ผลพบว่า REV2 สามารถระบุผู้ใช้งานหลอกลวงได้ดี โดยเฉพาะอย่างยิ่งในชุดข้อมูล Flipkart ที่สามารถระบุผู้ใช้งานหลอกลวงถูก 127 รายจาก 150 ราย คิดเป็นร้อยละ 84.6

Jahir Gutierrez ได้ทำนายการทุจริตในข้อมูลแบบ peer-to-peer transaction โดยใช้ node2vec และค่า in-degree และ out-degree ในการสกัดคุณลักษณะเด่นจากโหนดในเครือข่าย และใช้ Deep Learning เพื่อใช้ในการทำนายธุรกรรมที่หลอกลวง โดยใช้ข้อมูล Bitcoin OTC+Alpha และมีการนำคะแนนความพึงพอใจมาช่วยในการแบ่งคลาสเป็น Honest และ Fraudulent จากนั้น แบ่ง Layer ทั้งหมด 5 Layers โดยใช้ ReLU เป็น Activation function ใน 4 Layers แรก และ ใช้ Sigmoid ใน Layer สุดท้าย ผลการศึกษาพบว่า การพยากรณ์มีความถูกต้อง (Accuracy) ร้อยละ 90.0 ความแม่นยำ (Precision) ร้อยละ 93.0 ความระลึก (Recall) ร้อยละ 90.0 และค่า ROC Curve ร้อยละ 93.0

บทที่ 3
ระเบียบวิธีวิจัย

ทำนายธุรกรรมที่หลอกลวง (Fraudulent transactions) โดยการนำเทคนิค Node2Vec และค่าความเป็นศูนย์กลาง มาประยุกต์ใช้ในการสกัดคุณลักษณะ (Feature Extraction) ของข้อมูลรายธุรกรรม จากนั้นใช้เทคนิคการเรียนรู้แบบกลุ่ม ในการทำนายธุรกรรมที่คาดว่าจะหลอกลวง เพื่อเพิ่มประสิทธิภาพความแม่นยำของโมเดล

3.1 แนวทางการศึกษา

3.1.1 ศึกษาทฤษฎีการสกัดคุณลักษณะ (Feature Extraction) ของข้อมูลรายธุรกรรม เนื่องจากการสกัดคุณลักษณะมีหลากหลายรูปแบบ ผู้วิจัยจึงต้องค้นคว้าหาข้อมูลเพื่อที่จะทำความเข้าใจการทำงานของแนวทางเหล่านั้น แล้วนำมาประยุกต์ใช้กับข้อมูลที่ใช้ในงานวิจัยนี้

3.1.2 ศึกษาทฤษฎีการจำแนกประเภทข้อมูล เพื่อเลือกโมเดลให้ความสอดคล้องกับลาเบลคำตอบ

3.1.3 ศึกษาลักษณะของข้อมูลที่นำมาใช้ในการสกัดคุณลักษณะเด่นและการทำนายการทุจริต

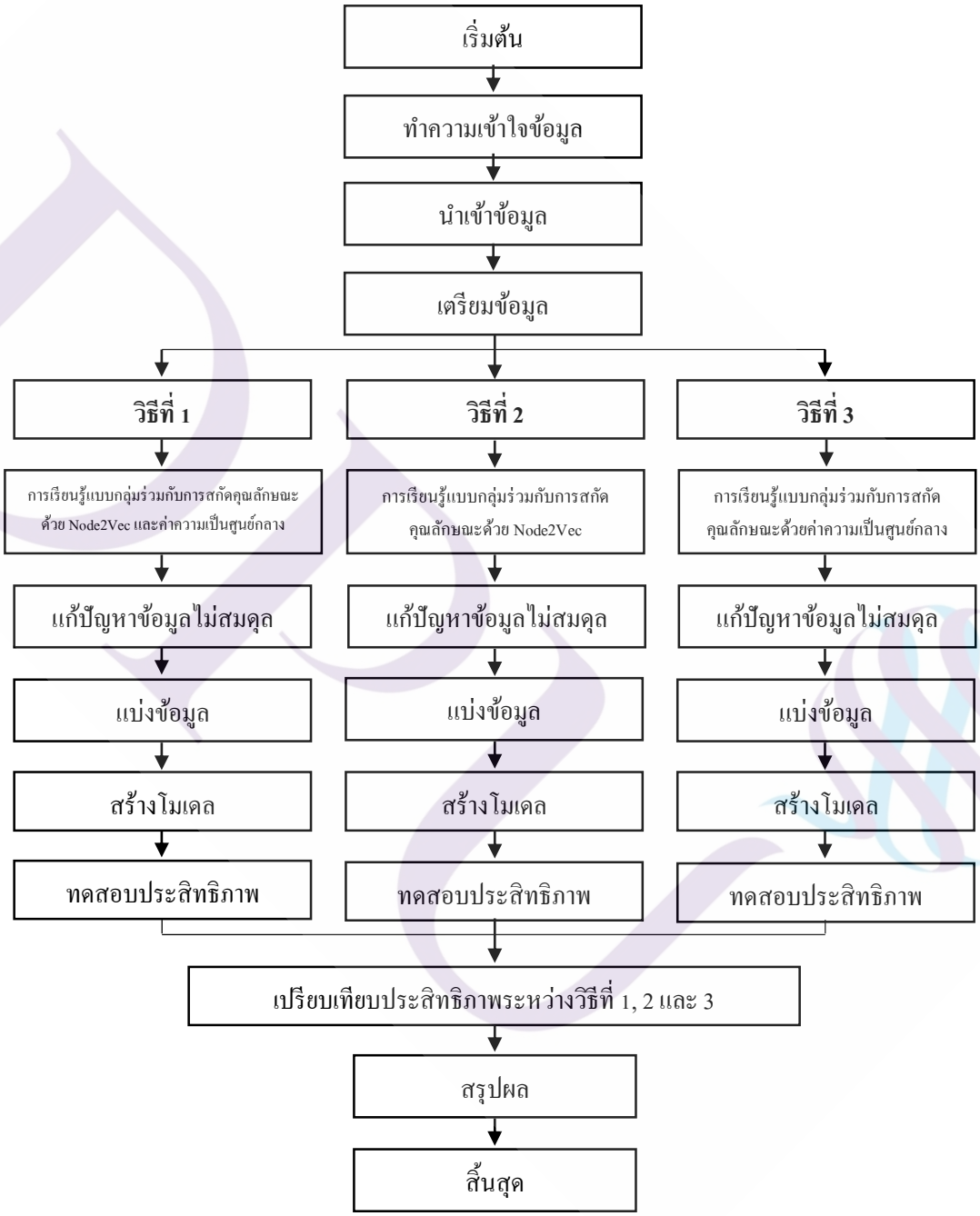
3.1.4 นำข้อมูลธุรกรรมไปสกัดคุณลักษณะเด่น และทำนายธุรกรรมที่หลอกลวง

3.1.5 ดำเนินการประมวลผลข้อมูลและบันทึกผล

3.1.6 สรุปผล

3.2 ขั้นตอนการทำงานโดยสังเขป

มีขั้นตอนดังภาพ



ภาพที่ 3.1 ขั้นตอนการทำงานโดยสังเขป

3.3 ขั้นตอนการทำงานโดยละเอียด

ผู้วิจัยดำเนินงานวิจัยดังนี้

3.3.1 ทำความเข้าใจข้อมูล (Data Understanding) เริ่มจากการทำความเข้าใจลักษณะของข้อมูลรายธุรกรรม หลังจากนั้นค้นหาแหล่งข้อมูลที่มีข้อมูลรายธุรกรรมเผยแพร่ให้บุคคลทั่วไปสามารถนำข้อมูลไปใช้ประโยชน์ได้ สำหรับข้อมูลที่นำมาใช้ในการวิเคราะห์ครั้งนี้ได้จาก Stanford Large Network Dataset Collection มหาวิทยาลัยสแตนฟอร์ด ชื่อชุดข้อมูล Bitcoin Alpha trust weighted signed network ซึ่งเป็นข้อมูลการซื้อขาย Bitcoin ของตลาดกลางพาณิชย์อิเล็กทรอนิกส์ชื่อ Bitcoin Alpha

3.3.2 นำเข้าข้อมูล (Input data) เป็นการนำข้อมูลเข้าสู่คอมพิวเตอร์ เพื่อสร้างชุดข้อมูล (Dataset)

3.3.3 เตรียมข้อมูล (Data preprocessing) เป็นขั้นตอนที่ใช้เวลาค่อนข้างนานเนื่องจากผลของโมเดลจะแม่นยำมากเพียงใด ขึ้นอยู่กับคุณภาพของข้อมูลด้วยเช่นกัน สำหรับข้อมูลการซื้อขาย Bitcoin นี้ มีจำนวนแอททริบิวต์ทั้งหมด 4 แอทริบิวต์ ได้แก่

- Source คือ โหนดต้นทางและเป็นผู้ให้คะแนนความพึงพอใจในการซื้อขาย
- Target คือ โหนดปลายทางและเป็นผู้ถูกให้คะแนนความพึงพอใจในการซื้อขาย
- Rating คือ คะแนนความพึงพอใจ มีค่าตั้งแต่ -10 ถึง 10
- TimeStamp คือ วันและเวลาของการให้คะแนนความพึงพอใจ หน่วยเป็น Epoch

ตัวอย่างข้อมูลการซื้อขาย Bitcoin ดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างข้อมูลการซื้อขาย Bitcoin รายธุรกรรม

SOURCE	TARGET	RATING	TIMESTAMP
1	2	-8	1296629344
1	3	6	1308241841
1	4	10	1343107173
4	5	-4	1289710643
4	5	8	1308242031

ขั้นตอนการเตรียมข้อมูลประกอบด้วยขั้นตอนย่อย ๆ ได้ดังนี้

ภาพที่ 3.2 ขั้นตอนการเตรียมข้อมูล

3.3.3.1 การคัดเลือกข้อมูล (Data selection) เป็นการคัดเลือกข้อมูลที่จะนำมาวิเคราะห์ ในงานวิจัยนี้ใช้ข้อมูลทั้งหมด เพื่อนำไปสร้างแอททริบิวต์ใหม่ และนำไปใช้ในการสร้างโมเดลให้ครบทุกแง่มุม

3.3.3.2 การกลั่นกรองข้อมูล (Data cleaning) เป็นขั้นตอนการทำข้อมูลให้มีความถูกต้อง เช่น การลบข้อมูลที่มีความซ้ำซ้อนกัน ซ่อมแซมข้อมูลที่ขาดหายไป รวมถึงการแก้ไขข้อมูลที่มีความผิดพลาด เป็นต้น สำหรับการกลั่นกรองข้อมูลของงานวิจัยนี้ พบว่า ข้อมูลไม่มีความซ้ำซ้อน ไม่มีข้อมูลสูญหาย จึงไม่ทำการแก้ไขข้อมูล

3.3.3.3 การแปลงรูปข้อมูล (Data transformation) เป็นการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมไปใช้ในการวิเคราะห์ข้อมูล ในที่นี้จะทำการแปลงข้อมูลดังนี้

3.3.3.3.1 คะแนนความพึงพอใจ เพื่อใช้เป็นลาเบลคำตอบ ซึ่งคะแนนความพึงพอใจจัดเก็บในรูปแบบของข้อมูลเชิงปริมาณ (Numerical data) ที่มีค่าตั้งแต่ -10 ถึง 10 จัดให้เป็นข้อมูลเชิงคุณภาพ (Category data) โดยค่าตั้งแต่ 0 ถึง -10 จะจัดเป็น 1 กลุ่ม ซึ่ง Label ว่า Fraud ส่วนค่าตั้งแต่ 1 ถึง 10 จะจัดเป็นอีก 1 กลุ่ม ซึ่ง Label ว่า Non Fraud

3.3.3.3.2 วันและเวลาของการให้คะแนนความพึงพอใจ เพื่อใช้เป็นแอททริบิวต์ในการสร้างโมเดล โดยทำการแปลงหน่วยจาก Epoch ให้อยู่ในรูปวันและเวลาตามมาตรฐานสากล จากนั้นสร้างแอททริบิวต์วัน เดือน ปี และเวลาที่ให้คะแนน ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 แอททริบิวต์ที่ได้จากการแปลงข้อมูลวันและเวลาของการให้คะแนนความพึงพอใจแล้ว

TIMESTAMP	DAY_RATE	MONTH_RATE	YEAR_RATE	TIME_RATE
1296629344	Monday	1	2010	5
1343107173	Thursday	6	2013	15
1308242031	Sunday	12	2016	23

หลังจากนั้นผู้วิจัยใช้เทคนิค One Hot Key เพื่อแปลงค่าในแอททริบิวต์ DAY_RATE และ YEAR_RATE ให้เป็นแอททริบิวต์ใหม่ที่มีค่าเป็น 0 หรือ 1 เช่น DAY_RATE สร้างแอททริบิวต์ใหม่ดังแสดงในตารางที่ 3.3

ตารางที่ 3.3 ตัวอย่างแอททริบิวต์ DAY_RATE

DAY_RATE	DAY_RATE_MONDAY	DAY_RATE_TUESDAY	…	DAY_RATE_SUNDAY
Monday	1	0		0
Thursday	0	0		0
Sunday	0	0		1

แอททริบิวต์ YEAR_RATE สร้างแอททริบิวต์ใหม่ดังแสดงในตารางที่ 3.4

ตารางที่ 3.4 ตัวอย่างแอททริบิวต์ YEAR_RATE

YEAR_RATE	YEAR_RATE_2010	YEAR_RATE_2011	…	YEAR_RATE_2016
2010	1	0		0
2013	0	0		0
2016	0	0		1

แอททริบิวต์ TIME_RATE ได้ทำการจัดกลุ่มใหม่ {Morning = 5am-11am, Afternoon = 12am-4pm, Evening = 5pm-4am} โดยอ้างอิงตามข้อมูลการซื้อขาย Bitcoin ของตลาด US ดังแสดงในตารางที่ 3.5

ตารางที่ 3.5 ตัวอย่างแอททริบิวต์ TIME_RATE

TIME_RATE	TIME_OF_DAY_MORNING	TIME_OF_DAY_AFTERNOON	TIME_OF_DAY_EVENING
5	1	0	0
21	0	0	0
23	0	0	1

3.3.4 หลังจากขั้นตอนการเตรียมข้อมูลเสร็จเรียบร้อยแล้ว ผู้วิจัยจะทำการเปรียบเทียบผลการทำนายธุรกรรมที่หลอกลวงด้วยการสกัดคุณลักษณะ 3 วิธี คือ

วิธีที่ 1 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และค่าความเป็นศูนย์กลาง

วิธีที่ 2 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec

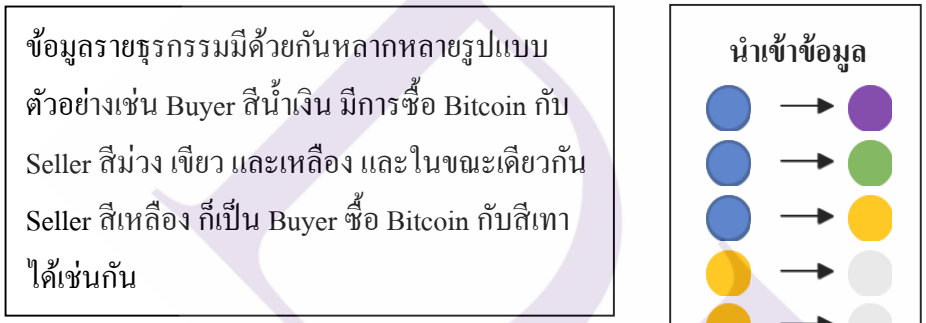
วิธีที่ 3 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วยค่าความเป็นศูนย์กลาง

3.3.5 สกัดคุณลักษณะ (Feature Extraction) การสกัดคุณลักษณะโดยทั่วไปเหมาะกับข้อมูลที่จัดเก็บเป็นระเบียนข้อมูล (Record) ที่นำเอาหลายๆ แอททริบิวต์ (Attribute) มารวมกัน เพื่อเกิดเป็นข้อมูลเรื่องใดเรื่องหนึ่ง เช่น ข้อมูลลูกค้า ข้อมูลนักศึกษา ข้อมูลพนักงาน เป็นต้น แต่สำหรับข้อมูลรายธุรกรรมนั้น มีการเก็บข้อมูลในรูปแบบความสัมพันธ์ระหว่างต้นทางไปยังปลายทางภายในระเบียนข้อมูลเดียวกัน ดังแสดงในตารางที่ 3.1 จึงจำเป็นต้องใช้วิธีการอื่นในการช่วยสกัดคุณลักษณะ

ในงานวิจัยนี้ใช้วิธี Node2Vec เนื่องจากเป็นเทคนิคที่ทำการสร้าง Low-dimensional ในแต่ละโหนด โดยจำลองการสุ่มเดินแบบลำเอียง (Random biased walks) และคำนวณค่าความน่าจะเป็นของเส้นเชื่อมในแต่ละโหนด เพื่อใช้เป็นแนวเดินในกราฟ และจัดกลุ่มโหนดที่มีคุณลักษณะคลายคลึงกัน (Homophily) เข้าไว้ด้วยกันได้ดีมากยิ่งขึ้น โดยกำหนดให้จำนวนการเดินเท่ากับ 25 และจำนวน iterations เท่ากับ 15

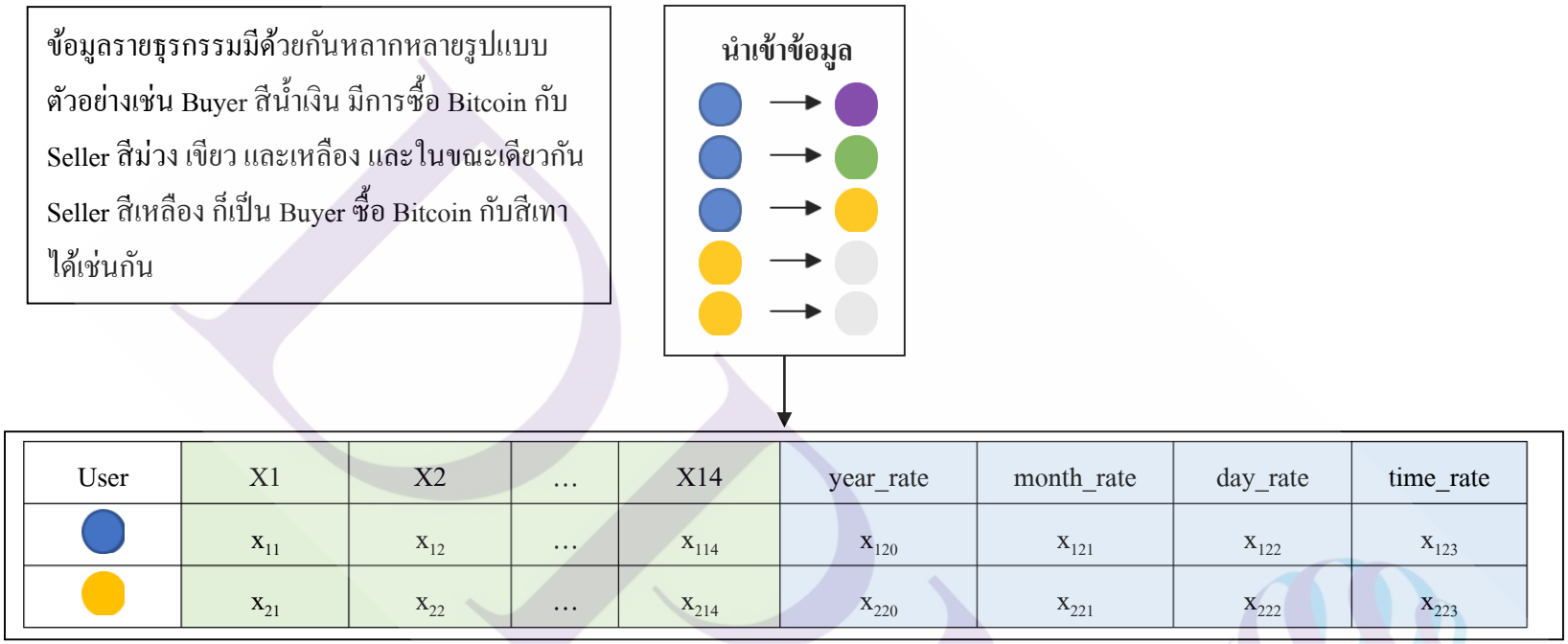
นอกจากนี้ จะทำ Feature Engineering โดยใช้ค่าความเป็นศูนย์กลาง (Degree Centrality) จาก Social Network Analysis เนื่องจากเป็นค่าที่แสดงถึงการเข้าถึง การเชื่อมต่อความสัมพันธ์กับโหนดอื่นๆ ค่าความเป็นศูนย์กลางที่จะนำมาใช้ คือ in-degree, out-degree, Closeness, Betweenness และ Eigenvector เป็นต้น และแอททริบิวต์ TIMESTAMP เช่น YEAR_RATE_2010, MONTH_RATE, DAY_RATE_MONDAY, TIME_OF_DAY_MORNING เป็นต้น

จากนั้นทำการเตรียมแอททริบิวต์ทั้งหมดที่ได้ข้างต้น ตามวิธีการทั้ง 3 วิธี ดังแสดงในภาพที่ 3.3-3.5

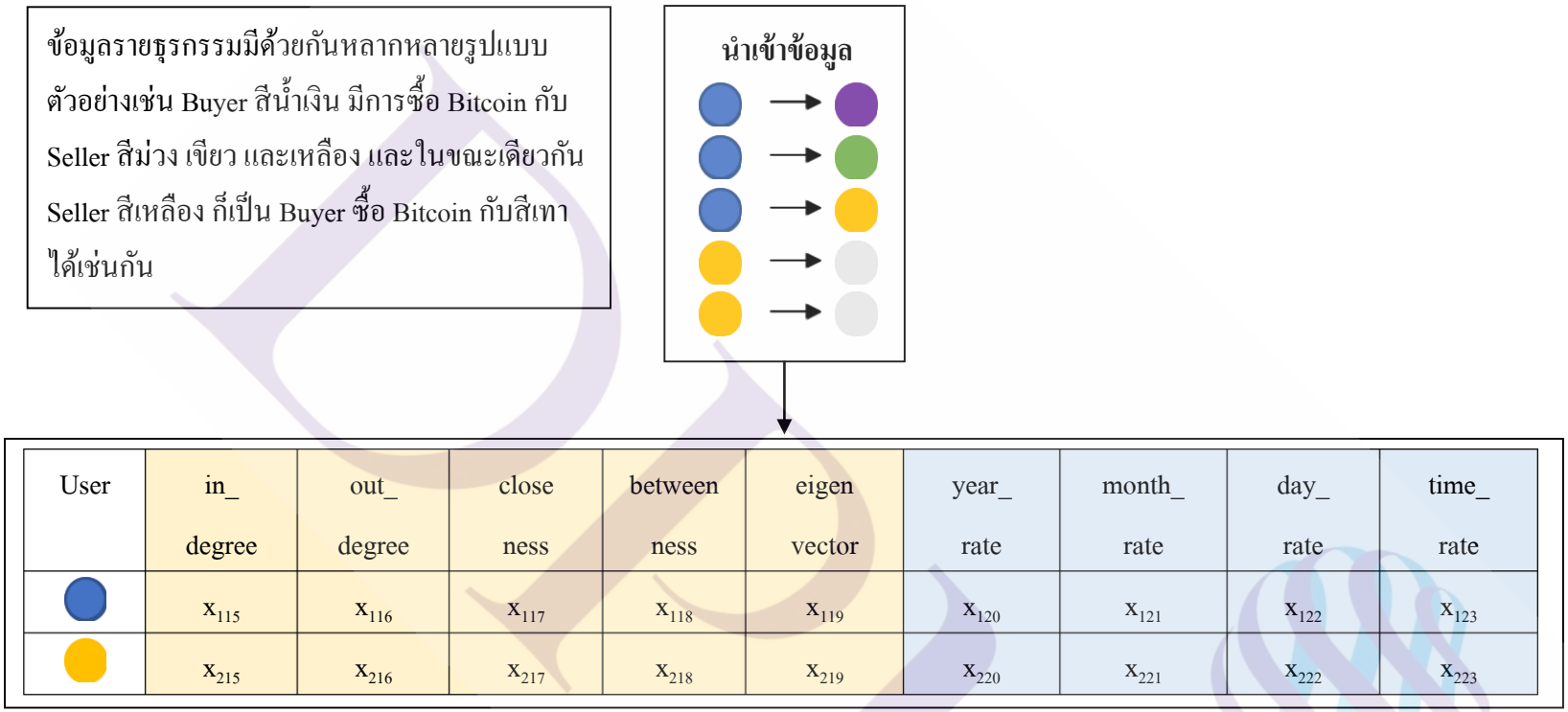


User	X1	X2	…	X14	in_ degree	out_ degree	close ness	between ness	eigen vector	year_ rate	month_ rate	day_ rate	time_ rate
	x_{11}	x_{12}	…	x_{114}	x_{115}	x_{116}	x_{117}	x_{118}	x_{119}	x_{120}	x_{121}	x_{122}	x_{123}
	x_{21}	x_{22}	…	x_{214}	x_{215}	x_{216}	x_{217}	x_{218}	x_{219}	x_{220}	x_{221}	x_{222}	x_{223}

ภาพที่ 3.3 การเตรียมแอททริบิวต์ของวิธีที่ 1 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และค่าความเป็นศูนย์กลาง



ภาพที่ 3.4 การเตรียมแอททริบิวต์ของวิธีที่ 2 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec



ภาพที่ 3.5 การเตรียมแอททริบิวต์ของวิธีที่ 3 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วยค่าความเป็นศูนย์กลาง

หลังจากได้แอททริบิวต์ในแต่ละวิธีแล้ว นำแอททริบิวต์ดังกล่าวมาจัดรูปแบบข้อมูล ในรูปข้อมูลต้นทาง (Source) และข้อมูลปลายทาง (Target) เรียงตามระเบียนข้อมูลนั้นๆ แต่เนื่องจากมีแอททริบิวต์ที่สร้างจาก TIME_RATE ซ้ำกัน (Redundant) ผู้วิจัยจึงตัดแอททริบิวต์ที่ซ้ำกันออก เพื่อให้เหลือเพียง 1 ชุด จากนั้นนำแอททริบิวต์ที่เหลือไปใช้วิเคราะห์ข้อมูลในลำดับถัดไป ดังแสดงในตารางที่ 3.6 – 3.8

ตารางที่ 3.6 วิธีที่ 1 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และค่าความเป็นศูนย์กลาง

Transaction ID	Source								Target								year_ rate	month_ rate	day_ rate	time_ rate
	X1	…	X14	in_ degree	out_ degree	close ness	between ness	eigen vector	X1	…	X14	in_ degree	out_ degree	close ness	between ness	eigen vector				
1	x_{11}	…	x_{114}	x_{115}	x_{116}	x_{117}	x_{118}	x_{119}	x_{120}	…	x_{134}	x_{135}	x_{136}	x_{137}	x_{138}	x_{139}	x_{140}	x_{141}	x_{142}	x_{143}
2	x_{21}	…	x_{214}	x_{215}	x_{216}	x_{217}	x_{218}	x_{219}	x_{220}	…	x_{234}	x_{235}	x_{236}	x_{237}	x_{238}	x_{239}	x_{240}	x_{241}	x_{242}	x_{243}

ตารางที่ 3.7 วิธีที่ 2 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec

TransactionID	Source			Target			year_rate	month_rate	day_rate	time_rate
	X1	…	X14	X1	…	X14				
1	x_{11}	…	x_{114}	x_{120}	…	x_{134}	x_{140}	x_{141}	x_{142}	x_{143}
2	x_{21}	…	x_{214}	x_{220}	…	x_{234}	x_{240}	x_{241}	x_{242}	x_{243}

ตารางที่ 3.8 วิธีที่ 3 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วยค่าความเป็นศูนย์กลาง

TransactionID	Source					Target					year_ rate	month_ rate	day_ rate	time_ rate
	in_ degree	out_ degree	close ness	between ness	eigen vector	in_ degree	out_ degree	close ness	between ness	eigen vector				
1	x_{115}	x_{116}	x_{117}	x_{118}	x_{119}	x_{135}	x_{136}	x_{137}	x_{138}	x_{139}	x_{140}	x_{141}	x_{142}	x_{143}
2	x_{215}	x_{216}	x_{217}	x_{218}	x_{219}	x_{235}	x_{236}	x_{237}	x_{238}	x_{239}	x_{240}	x_{241}	x_{242}	x_{243}

3.3.6 การคัดเลือกคุณลักษณะ (Feature Selection) เป็นการลดขนาดหรือมิติของข้อมูล โดยยังคงลักษณะสำคัญของข้อมูลไว้ วิธีการคัดเลือกคุณลักษณะที่ใช้คือ Recursive Feature Elimination (RFE) เป็นอีกหนึ่งเทคนิคของ Wrapper Method ที่ทำการคัดเลือกคุณสมบัติด้วยการคัดเข้าและนำออกคุณลักษณะไปพร้อมกัน ทำให้สามารถพิจารณาความเป็นไปได้ของคุณลักษณะมากขึ้น

3.3.7 แก้ปัญหาข้อมูลไม่สมดุล (Imbalance) เป็นการแก้ไขปัญหาเพื่อให้โมเดลมีผลการทำนายที่ถูกต้องมากขึ้น ปัญหาความไม่สมดุลของข้อมูลพบได้บ่อยครั้ง เช่น การวินิจฉัยการป่วยเป็นโรคมะเร็งหรือการทุจริต เป็นต้น การแก้ปัญหาข้อมูลไม่สมดุลมีด้วยกันหลายวิธีเช่น

3.3.7.1 วิธีสุ่มเกิน (Oversampling) เป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมาก ซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มเลือกจากข้อมูลเดิม วิธีนี้เป็นการหลีกเลี่ยงการสูญเสียข้อมูลในการวิเคราะห์ แต่อาจจะส่งผลต่อการสร้างโมเดลในการใช้ข้อมูลชุดเดียวกันทั้งเทรน (train) และทดสอบ (Test) โมเดล ทำให้ข้อมูลทดสอบไม่ได้เป็นอิสระจากข้อมูลสอน ทำให้ผลการทำนายมีความเอนเอียงได้

3.3.7.2 วิธีสุ่มลด (Undersampling) เป็นการลดจำนวนข้อมูลที่อยู่ในกลุ่มส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อย แต่วิธีนี้จะทำให้สูญสียข้อมูลในการวิเคราะห์จนทำให้โมเดลทำนายผลได้ผิดพลาดมากขึ้น

3.3.7.3 วิธีสังเคราะห์ข้อมูล (Synthetic Minority Oversampling Technique : SMOTE) เป็นเทคนิคการสุ่มตัวอย่างแบบสุ่มเพิ่ม คือ แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิม แต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิม เป็นเทคนิคที่แก้ปัญหาความเอนเอียงและข้อมูลไม่เพียงพอในการสร้างโมเดล เนื่องจากเป็นการสังเคราะห์ข้อมูลของกลุ่มส่วนน้อยให้เพิ่มขึ้นมาจนได้ชุดข้อมูลใหม่ โดยไม่ใช้ข้อมูลเดิมในการเพิ่มจำนวน

ด้วยเหตุผลที่กล่าวมาข้างต้น ผู้วิจัยจะใช้วิธีการสังเคราะห์ข้อมูลในการแก้ปัญหาข้อมูลไม่สมดุล และจะใช้ข้อมูลชุดที่ทำการสังเคราะห์แล้วไปสร้างโมเดลในลำดับถัดไป

3.3.8 แบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดล มีด้วยกัน 3 วิธีคือ

3.3.8.1 วิธี Self consistency test เป็นวิธีการที่ง่ายที่สุด คือ การนำข้อมูลชุดเดียวกันกับการเทรนมาทดสอบประสิทธิภาพของโมเดล ซึ่งวิธีการนี้จะให้ผลการทดสอบประสิทธิภาพที่ค่อนข้างสูง แต่ผลที่ได้อาจจะไม่เหมาะที่จะนำไปรายงานในงานวิจัยอื่น ๆ อาจจะเหมาะกับการพิจารณาแนวโน้มของโมเดลที่สร้างขึ้น

3.3.8.2 วิธี Split test เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น 70% ต่อ 30% หรือ 80% ต่อ 20% โดยข้อมูลส่วนที่หนึ่ง (70% หรือ 80%) ใช้ในการสร้างโมเดลและข้อมูลส่วนที่สอง (30% หรือ 20%) ใช้ใน การทดสอบประสิทธิภาพของโมเดล แต่วิธี Split test นี้ ทำการสุ่มข้อมูลเพียงครั้งเดียว ซึ่งในบางครั้งถ้าสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะคล้ายกับข้อมูลที่ใช้เทรนโมเดล จะทำให้ผลการวัดประสิทธิภาพได้ออกมาดี ในทางตรงข้ามถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้เทรนโมเดลมากทำให้ผลการวัดประสิทธิภาพได้ออกมาแย่

3.3.8.3 วิธี Cross validation test เป็นการแบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ ทำให้วิธีการแบ่งข้อมูลวิธีนี้ให้ผลการวัดประสิทธิภาพที่เหมาะสมกว่า 2 วิธีดังกล่าวข้างต้น ดังนั้น ผู้วิจัยจะใช้วิธีนี้ในการแบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดล

3.3.9 สร้างโมเดล (Modeling) เป็นขั้นตอนการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิงด้วยเทคนิคการจำแนกประเภทข้อมูล เช่น ต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน การวิเคราะห์การถดถอยโลจีสติก หรือโครงข่ายประสาทเทียม เป็นต้น เทคนิคเหล่านี้จะให้ผลลัพธ์ในการจำแนกข้อมูลที่แม่นยำเพียงโมเดลเดียวเท่านั้น ทำให้บางครั้งเกิดปัญหาความโน้มเอียง (Bias) เพื่อลดปัญหาดังกล่าว ดังนั้นผู้วิจัยจะใช้เทคนิคการเรียนรู้แบบกลุ่ม โดยเทคนิคที่ใช้ประกอบด้วย ต้นไม้ตัดสินใจ การวิเคราะห์การถดถอยโลจิสติก และโครงข่ายประสาทเทียม โดยในขั้นตอนสุดท้ายที่เลือกผลลัพธ์เพียงผลลัพธ์เดียวนั้นจะใช้วิธีโหวต (Vote) เพื่อเลือกคำตอบที่ตรงกันมากที่สุด

3.3.10 ทดสอบประสิทธิภาพ (Evaluation) เป็นขั้นตอนการวัดประสิทธิภาพของโมเดลว่ามีความถูกต้องมากน้อยเพียงใด สามารถนำโมเดลไปใช้จริงได้หรือไม่ ในงานวิจัยนี้ ผู้วิจัยจะแสดงผลการวัดประสิทธิภาพของโมเดลในขั้นตอนการสร้าง (Train) และทดสอบ (Test) โมเดล ตัววัดประสิทธิภาพของโมเดล มีดังนี้

3.3.10.1 ตารางแจกแจงผลลัพธ์ (Confusion Matrix)

3.3.10.2 ค่าร้อยละความถูกต้อง (Accuracy)

3.3.10.3 ค่าร้อยละความผิดพลาด (Error)

3.3.10.4 ค่าร้อยละความแม่นยำ (Precision)

3.3.10.5 ค่าร้อยละความระลึก (Recall)

3.3.10.6 ค่าร้อยละเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score)

3.3.10.7 ค่าร้อยละเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ทุกโมเดลจะถูกทดสอบประสิทธิภาพไว้เหมือนกันทั้งหมด

3.3.11 เปรียบเทียบประสิทธิภาพ เป็นขั้นตอนการเปรียบเทียบผลการวัดประสิทธิภาพของโมเดลระหว่างวิธีการที่ 1, 2 และ 3

3.3.12 สรุปผล เป็นขั้นตอนการสรุปผลว่าวิธีการใดให้ผลการวัดประสิทธิภาพของโมเดลที่ดีที่สุด

บทที่ 4

ผลการดำเนินงานวิจัย

4.1 ผลการวิเคราะห์ข้อมูลทั่วไป

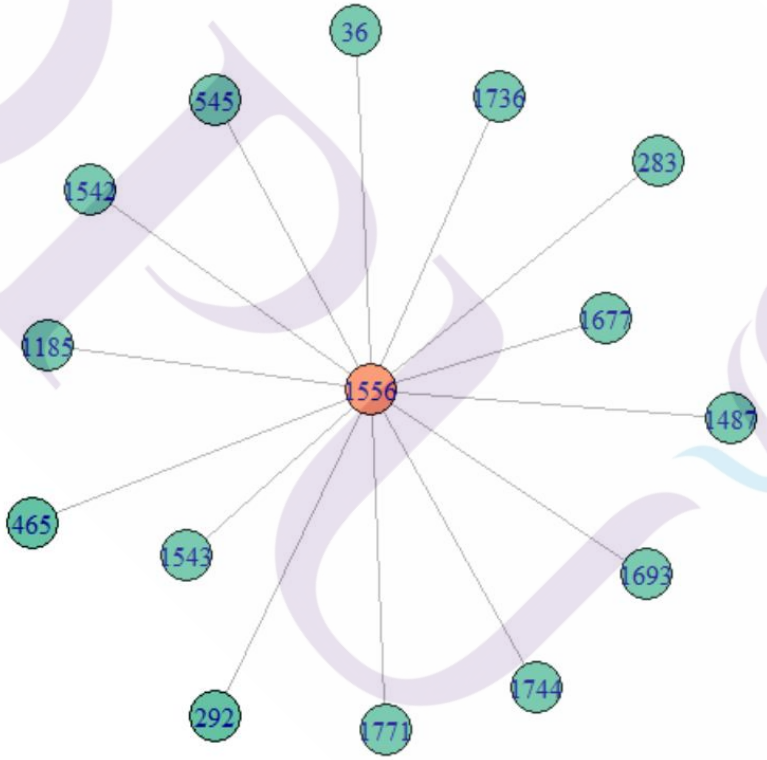
จากข้อมูลการซื้อขาย Bitcoin ของตลาดกลางพาณิชย์อิเล็กทรอนิกส์ชื่อ Bitcoin Alpha พบว่า มีจำนวนผู้ซื้อขายทั้งหมด 9,336 Users แบ่งเป็นผู้ซื้ออย่างเดียว จำนวน 23 Users คิดเป็นร้อยละ 0.2 ผู้ขายอย่างเดียว จำนวน 1,444 Users คิดเป็นร้อยละ 15.5 และเป็นทั้งผู้ซื้อและผู้ขาย จำนวน 7,869 Users คิดเป็นร้อยละ 84.3 และมีจำนวนระเบียนข้อมูลทั้งสิ้น 59,788 ระเบียน ค่าเฉลี่ยในการซื้อขาย Bitcoin เท่ากับ 7.6 ครั้ง และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 22.2 ตัวอย่างจำนวนระเบียนข้อมูลการซื้อขาย Bitcion ราย Users ดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 ตัวอย่างจำนวนระเบียนข้อมูลการซื้อขาย Bitcion ราย Users

UsersID	จำนวนระเบียนข้อมูล (ระเบียน)
1	46
5	3
7	65
8	1
132	24
145	1
149	17
306	1
307	1

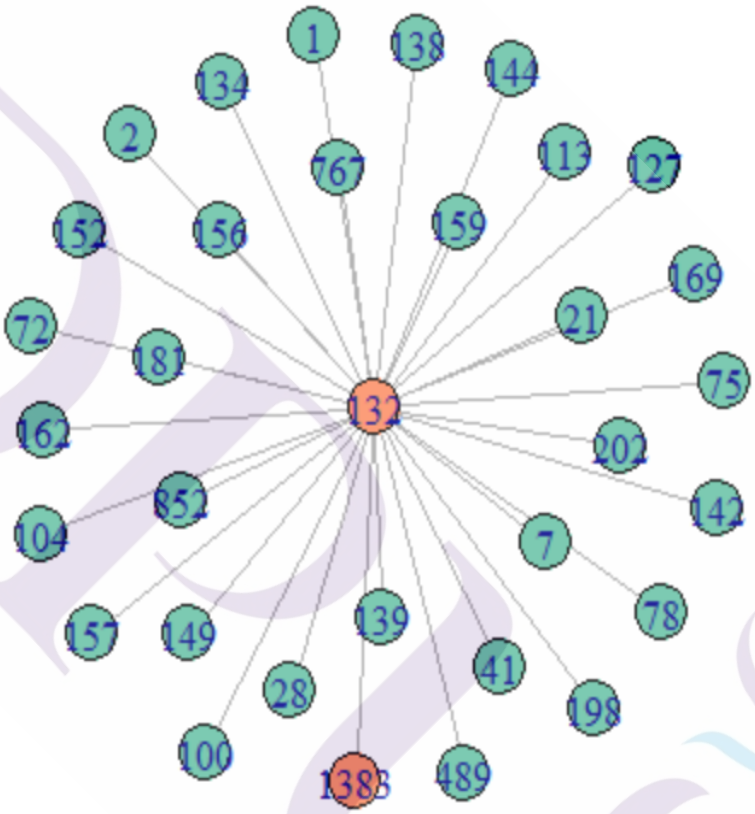
เมื่อพิจารณาข้อมูลการซื้อขาย Bitcion และคะแนนความพึงพอใจราย Users โดยใช้ค่าเฉลี่ยในการซื้อขายเป็นเกณฑ์ในการพิจารณา ถ้าหาก Users รายใด ทำการซื้อขาย Bitcoin มากกว่าค่าเฉลี่ย จะกลายเป็นกลุ่มซื้อขายบ่อย ถ้าต่ำกว่าค่าเฉลี่ยจะกลายเป็นกลุ่มซื้อขายไม่บ่อย ส่วนคะแนนความพึงพอใจ ถ้าหากมีคะแนนความพึงพอใจเป็นบวกทุกการซื้อขาย จะกลายเป็นกลุ่มคะแนนความพึงพอใจเป็นบวก แต่ถ้าหากมีคะแนนความพึงพอใจเป็นลบในบางครั้ง จะกลายเป็นกลุ่มคะแนนความพึงพอใจเป็นลบทันที เมื่อพิจารณาทั้งข้อมูลการซื้อขายและคะแนนความพึงพอใจจะได้กลุ่มลูกค้า 4 กลุ่มดังนี้

- ซื้อขายบ่อย คะแนนเป็นบวก เป็นกลุ่มที่มีการซื้อขาย Bitcoin บ่อย และคะแนนความพึงพอใจเป็นบวกทุกการซื้อขาย เช่น User 1556 มีการซื้อขายทั้งหมด 14 ครั้ง และไม่มีครั้งไหนที่ให้คะแนนความพึงพอใจเป็นลบ ดังแสดงในภาพที่ 4.1



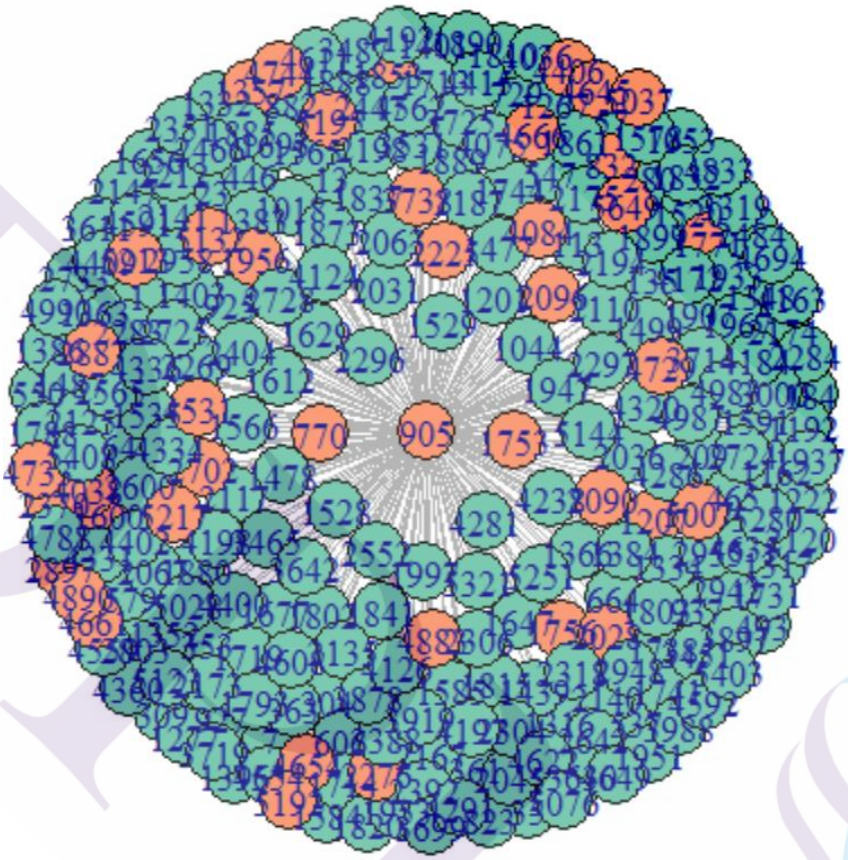
ภาพที่ 4.1 แสดงตัวอย่างระเบียนข้อมูลของ User 1556

- ซื้อขายบ่อย คะแนนเป็นลบ เป็นกลุ่มที่มีการซื้อขาย Bitcoin บ่อย และคะแนนความพึงพอใจเป็นลบเพียง 1 ครั้ง หรือมากกว่าจากการซื้อขายทั้งหมด เช่น User 132 มีการซื้อขายทั้งหมด 24 ครั้ง แต่มีเพียง 1 ครั้งที่มีคะแนนความพึงพอใจเป็นลบ นั้นคือ การซื้อขายกับ User 1383 ดังแสดงในภาพที่ 4.2



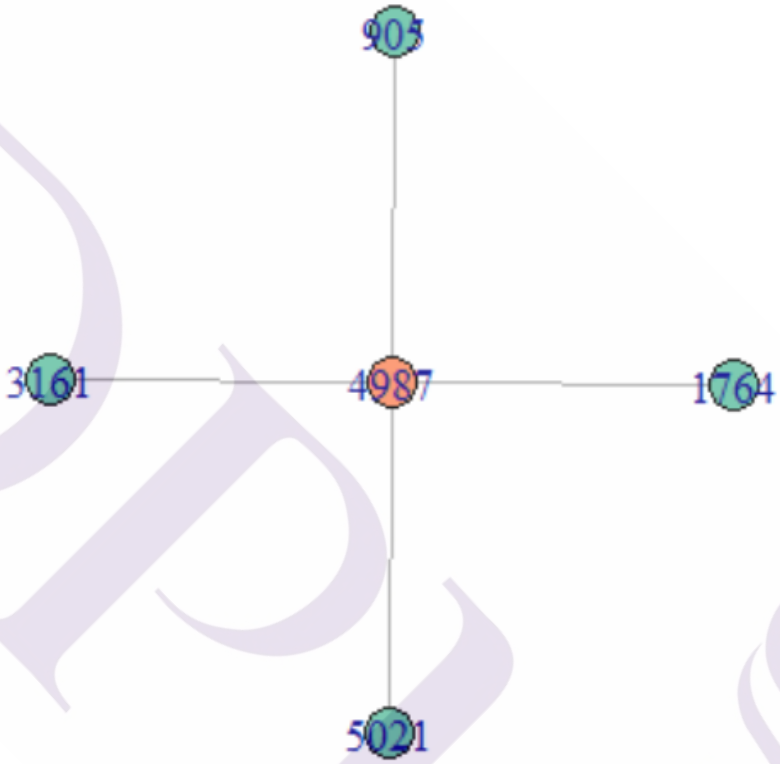
ภาพที่ 4.2 แสดงตัวอย่างระเบียนข้อมูลของ User 132

หรือ User 905 มีการซื้อขายทั้งหมด 264 ครั้ง และมี 41 ครั้งที่มีคะแนนความพึงพอใจเป็นลบ ดังแสดงในภาพที่ 4.3



ภาพที่ 4.3 แสดงตัวอย่างระเบียนข้อมูลของ User 905

- ซื้อขายไม่บ่อย คะแนนเป็นบวก เป็นกลุ่มที่มีการซื้อขาย Bitcoin ไม่บ่อย แต่คะแนนความพึงพอใจเป็นบวกทุกการซื้อขาย เช่น User 4987 มีการซื้อขายทั้งหมด 4 ครั้ง และไม่มีครั้งไหนที่ให้คะแนนความพึงพอใจเป็นลบ ดังแสดงในภาพที่ 4.4

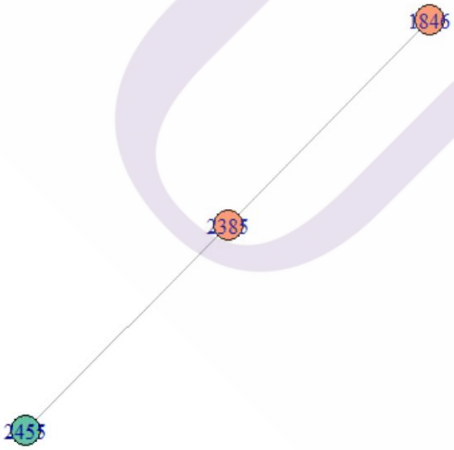


ภาพที่ 4.4 แสดงตัวอย่างระเบียนข้อมูลของ User 4987

- ซื้อขายไม่บ่อย คะแนนเป็นลบ เป็นกลุ่มที่มีการซื้อขาย Bitcoin ไม่บ่อย และคะแนนความพึงพอใจเป็นลบเพียง 1 ครั้ง หรือมากกว่าจากการซื้อขายทั้งหมด เช่น User 1184 มีการซื้อขายเพียง 1 ครั้ง และการซื้อขายนั้นมีคะแนนความพึงพอใจเป็นลบ นั้นคือ การซื้อขายกับ User 1347 ดังแสดงในภาพที่ 4.5

ภาพที่ 4.5 แสดงตัวอย่างระเบียนข้อมูลของ User 1184

หรือ User 2385 มีการซื้อขายทั้งหมด 2 ครั้ง และมี 1 ครั้งที่มีคะแนนความพึงพอใจเป็นลบ ดังแสดงในภาพที่ 4.6



ภาพที่ 4.6 แสดงตัวอย่างระเบียนข้อมูลของ User 2385

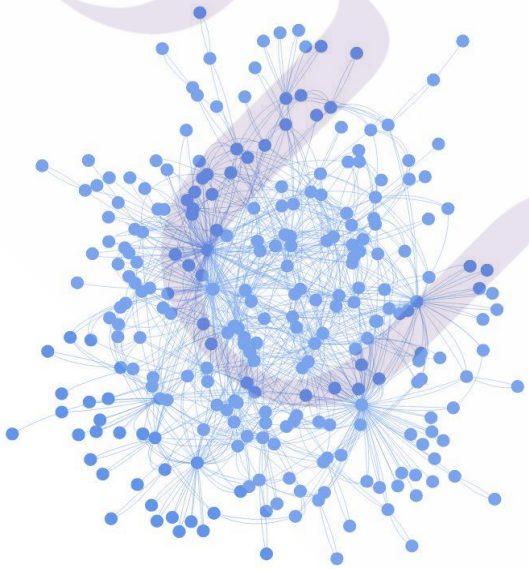
สำหรับลาเบลคำตอบ พบว่า คำตอบที่เป็น Fraud มีจำนวนทั้งสิ้น 5,099 ระเบียน คิดเป็นร้อยละ 8.5 และคำตอบที่เป็น Non Fraud มีจำนวนทั้งสิ้น 54,679 ระเบียน คิดเป็นร้อยละ 91.5 ดังแสดงในตารางที่ 4.2

ตารางที่ 4.2 แสดงความไม่สมดุลของข้อมูล

ลาเบลคำตอบ	จำนวนระเบียนข้อมูล (ระเบียน)	ร้อยละ
Fraud	5,099	8.5
Non-Fraud	54,679	91.5
Total	59,778	100.0

4.2 ผลการวิเคราะห์เครือข่ายทางสังคมของการซื้อขาย Bitcoin

ผลการวิเคราะห์เครือข่ายทางสังคม จะพบว่า ผู้ซื้อสามารถเป็นผู้ขายได้ และผู้ขายสามารถเป็นผู้ซื้อได้ ทำให้ข้อมูลมีการเชื่อมโยงกัน ดังแสดงในภาพที่ 4.7



ภาพที่ 4.7 แสดงเครือข่ายทางสังคมของการซื้อขาย Bitcoin

การวัดความสัมพันธ์ของเครือข่ายด้วยค่าระดับการเป็นศูนย์กลาง ซึ่งจะพิจารณาจากค่า Out-degree ที่แสดงถึงการเชื่อมโยงที่มีทิศทางออกจากโหนดที่สนใจ ค่า In-degree ที่แสดงถึงการเชื่อมโยงจากโหนดอื่น ๆ ที่มีทิศทางเข้าสู่โหนดที่สนใจ ค่า Closeness ที่แสดงถึงโหนดนั้น ๆ มีประสิทธิภาพในการสื่อสาร ข่าวสารข้อมูลหรือข้อความเห็นได้ทั่วถึงตลอดทั้งเครือข่าย ค่า Betweenness ที่แสดงถึงการเป็นตำแหน่งสะพานเชื่อมของโหนดหนึ่งเชื่อมโยงไปยังโหนดอื่นๆ และค่า Eigenvector ที่แสดงถึงโหนดที่มีอิทธิพลต่อโหนดอื่นๆ รายละเอียดดังแสดงในตารางที่ 4.3

ตารางที่ 4.3 ตัวอย่างค่าระดับการเป็นศูนย์กลาง

ลำดับ	Out-degree		In-degree		Closeness		Betweenness		Eigenvector	
	โหนด	ค่า	โหนด	ค่า	โหนด	ค่า	โหนด	ค่า	โหนด	ค่า
1	35	763	35	5535	905	0.2653	35	0.0715	905	0.1886
2	6006	490	2642	412	1	0.2635	2125	0.0287	1810	0.1826
3	2642	406	6006	398	35	0.2601	6006	0.0272	2642	0.1750
4	1810	404	1810	311	2388	0.2549	2642	0.0239	2028	0.1548
5	2125	397	2028	279	1810	0.2500	1810	0.0228	2125	0.1487

4.3 ผลการวัดประสิทธิภาพของโมเดล

จากการเตรียมแอททริบิวต์ด้วยเทคนิค Node2Vec ค่าความเป็นศูนย์กลาง และ ปี เดือน วัน และเวลาในการให้คะแนนการซื้อขาย จะได้แอทริบิวต์ดังแสดงในตารางที่ 4.4

ตารางที่ 4.4 แอททริบิวต์ที่ใช้ในการวิเคราะห์ข้อมูล

แอททริบิวต์	คำอธิบาย	ค่าของข้อมูล
X_1	แอททริบิวต์ที่ได้จาก Node2Vec	x.xxx
⋮	⋮	⋮
X_{14}	แอททริบิวต์ที่ได้จาก Node2Vec	x.xxx
IN-DEGREE	ค่า IN-DEGREE	x.xxx
POS_IN_EDGES	ค่า IN-DEGREE ของโหนดที่ให้คะแนนความพึงพอใจเป็นบวก	x.xxx
NEG_IN_EDGES	ค่า IN-DEGREE ของโหนดที่ให้คะแนนความพึงพอใจเป็นลบ	x.xxx
OUT_DEGREE	ค่า OUT-DEGREE	x.xxx
POS_ OUT _EDGES	ค่า IN-DEGREE ของโหนดที่ให้คะแนนความพึงพอใจเป็นบวก	x.xxx
NEG_ OUT _EDGES	ค่า IN-DEGREE ของโหนดที่ให้คะแนนความพึงพอใจเป็นลบ	x.xxx
CLOSENESS	ค่า CLOSENESS	x.xxx
BETWEENNESS	ค่า BETWEENNESS	x.xxx
EIGENVECTOR	ค่า EIGENVECTOR	x.xxx
DAY_RATE_Monday	ให้คะแนนความพึงพอใจในวันจันทร์	0 = No, 1 = Yes
DAY_RATE_Tuesday	ให้คะแนนความพึงพอใจในวันอังคาร	0 = No, 1 = Yes
DAY_RATE_Wednesday	ให้คะแนนความพึงพอใจในวันพุธ	0 = No, 1 = Yes
DAY_RATE_Thursday	ให้คะแนนความพึงพอใจในวันพฤหัสบดี	0 = No, 1 = Yes
DAY_RATE_Friday	ให้คะแนนความพึงพอใจในวันศุกร์	0 = No, 1 = Yes
DAY_RATE_Saturday	ให้คะแนนความพึงพอใจในวันเสาร์	0 = No, 1 = Yes
DAY_RATE_Sunday	ให้คะแนนความพึงพอใจในวันอาทิตย์	0 = No, 1 = Yes
MONTH_RATE	เดือนที่ให้คะแนนความพึงพอใจ	1, 2, 3, ..., 12

ตารางที่ 4.4 (ต่อ)

แอททริบิวต์	คำอธิบาย	ค่าของข้อมูล
YEAR_RATE_2010	ให้คะแนนความพึงพอใจในปี 2010	0 = No, 1 = Yes
YEAR_RATE_2011	ให้คะแนนความพึงพอใจในปี 2011	0 = No, 1 = Yes
YEAR_RATE_2012	ให้คะแนนความพึงพอใจในปี 2012	0 = No, 1 = Yes
YEAR_RATE_2013	ให้คะแนนความพึงพอใจในปี 2013	0 = No, 1 = Yes
YEAR_RATE_2014	ให้คะแนนความพึงพอใจในปี 2014	0 = No, 1 = Yes
YEAR_RATE_2015	ให้คะแนนความพึงพอใจในปี 2015	0 = No, 1 = Yes
YEAR_RATE_2016	ให้คะแนนความพึงพอใจในปี 2016	0 = No, 1 = Yes
TIME_OF_DAY_MORNING	ให้คะแนนความพึงพอใจในช่วงเช้า	0 = No, 1 = Yes
TIME_OF_DAY_AFTERNOON	ให้คะแนนความพึงพอใจในช่วงกลางวัน	0 = No, 1 = Yes
TIME_OF_DAY_EVENING	ให้คะแนนความพึงพอใจในช่วงกลางคืน	0 = No, 1 = Yes

ผลการคัดเลือกคุณลักษณะเด่น (Features important) ในแต่ละวิธี พบว่า วิธีที่ 1 แอททริบิวต์มีทั้งหมด 64 แอททริบิวต์ และมีคุณลักษณะเด่น 18 แอททริบิวต์ วิธีที่ 2 แอททริบิวต์มีทั้งหมด 46 แอททริบิวต์ และมีคุณลักษณะเด่น 21 แอททริบิวต์ และวิธีที่ 3 แอททริบิวต์มีทั้งหมด 36 แอททริบิวต์ และมีคุณลักษณะเด่น18 แอททริบิวต์ ดังแสดงในตารางที่ 4.5

ตารางที่ 4.5 แสดงผลการคัดเลือกคุณลักษณะเด่น (Features important) ในแต่ละวิธี

แอทริบิวต์	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
SOURCE_X1		✓	
SOURCE_X2		✓	
SOURCE_X3		✓	
SOURCE_X4		✓	
SOURCE_X5		✓	

ตารางที่ 4.5 (ต่อ)

แอทริบิวต์	**วิธีที่ 1**	**วิธีที่ 2**	**วิธีที่ 3**
SOURCE_X6		✓	
SOURCE_X7		✓	
SOURCE_X8		✓	
SOURCE_X9		✓	
SOURCE_X10	✓	✓	
SOURCE_X14		✓	
SOURCE_POS_IN_EDGES	✓		✓
SOURCE_OUT_DEGREE	✓		✓
SOURCE_POS_OUT_DEGREE			✓
SOURCE_NEG_OUT_DEGREE	✓		✓
TARGET_X25		✓	
TARGET_X29		✓	
TARGET_X30		✓	
TARGET_X31		✓	
TARGET_X32		✓	
TARGET_X33		✓	
TARGET_X34	✓		
TARGET_X35	✓		
TARGET_IN_DEGREE	✓		✓
TARGET_POS_IN_DEGREE	✓		✓
TARGET_NEG_IN_DEGREE	✓		✓
TARGET_OUT_DEGREE			✓
TARGET_POS_OUT_DEGREE			✓
TARGET_CLOSNESS			✓

ตารางที่ 4.5 (ต่อ)

แอทริบิวต์	**วิธีที่ 1**	**วิธีที่ 2**	**วิธีที่ 3**
TARGET_ EIGENVECTOR	✓		
DAY_RATE_Tuesday			✓
DAY_RATE_Thursday	✓		
DAY_RATE_Wednesday	✓		
YEAR_RATE_2010	✓	✓	
YEAR_RATE_2011	✓		✓
YEAR_RATE_2012			✓
YEAR_RATE_2013	✓	✓	
YEAR_RATE_2014	✓	✓	✓
YEAR_RATE_2015		✓	✓
TIME_OF_DAY_MORNING			✓
TIME_OF_DAY_AFTERNOON	✓		✓
TIME_OF_DAY_EVENING	✓		✓

หลังจากคัดเลือกคุณลักษณะเด่นในแต่ละวิธีเรียบร้อยแล้ว ผู้วิจัยทำการแก้ปัญหาความไม่สมดุลของข้อมูลด้วยวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) และแบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดลด้วยวิธี 10-fold cross-validation ผลการวัดประสิทธิภาพของโมเดลในขั้นตอนการสร้าง (Train) การตรวจสอบ (Validation) และการทดสอบ (Test) โมเดล ตามลำดับ โดยเรียงผลตามตัวชี้วัดในขั้นตอนที่ 3.3.10 วัดประสิทธิภาพของโมเดล

ตารางที่ 4.6 ผลตารางแจกแจงผลลัพธ์ (Confusion Matrix) ในขั้นตอนการสร้าง (Train) โมเดล

	วิธีที่ 1		**วิธีที่ 2**		**วิธีที่ 3**	
	Actual		**Actual**		**Actual**	
	Fraudulent	**Non-Fraud**	**Fraudulent**	**Non-Fraud**	**Fraudulent**	**Non-Fraud**
รอบที่ 1						
Predict: Fraudulent	34287	564	33763	878	34019	349
Predict: Non-Fraud	116	33839	640	33525	384	34054
รอบที่ 2						
Predict: Fraudulent	34195	408	33799	915	34233	514
Predict: Non-Fraud	196	33983	592	33476	158	33877
รอบที่ 3						
Predict: Fraudulent	34306	525	33707	871	34156	476
Predict: Non-Fraud	87	33868	686	33522	237	33917
รอบที่ 4						
Predict: Fraudulent	34290	488	33732	781	34067	348
Predict: Non-Fraud	119	33921	677	33628	342	34061
รอบที่ 5						
Predict: Fraudulent	34285	488	33873	925	34314	591
Predict: Non-Fraud	141	33938	553	33501	112	33835
รอบที่ 6						
Predict: Fraudulent	34294	507	33967	925	34231	432
Predict: Non-Fraud	151	33938	478	33520	214	34013
รอบที่ 7						
Predict: Fraudulent	34307	577	33886	992	34218	492
Predict: Non-Fraud	105	33835	526	33420	194	33920

ตารางที่ 4.6 (ต่อ)

	วิธีที่ 1		วิธีที่ 2		วิธีที่ 3	
	Actual		Actual		Actual	
	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud
รอบที่ 8						
Predict: Fraudulent	34314	499	33819	896	34223	460
Predict: Non-Fraud	132	33947	627	33550	223	33986
รอบที่ 9						
Predict: Fraudulent	34294	609	33918	904	34133	475
Predict: Non-Fraud	120	33805	496	33510	281	33939
รอบที่ 10						
Predict: Fraudulent	34324	518	33891	871	34146	408
Predict: Non-Fraud	106	33912	539	33559	284	34022
ค่าเฉลี่ย						
Predict: Fraudulent	34289.6	518.3	3383.5	895.8	37174	454.5
Predict: Non-Fraud	127.3	33898.3	581.4	33521.1	242.9	33962.4

ผลการวัดประสิทธิภาพของโมเดลด้วยค่าความถูกต้อง (Accuracy) ค่าความผิดพลาด (Error) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) และค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ในขั้นตอนการสร้าง (Train) โมเดล ดังแสดงในตารางที่ 4.7

ตารางที่ 4.7 ผลการวัดประสิทธิภาพของโมเดล ในขั้นตอนการสร้าง (Train) โมเดล

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
ค่าร้อยละความถูกต้อง (Accuracy)			
รอบที่ 1	**99.01**	97.79	98.93
รอบที่ 2	**99.12**	97.81	99.02
รอบที่ 3	**99.11**	97.73	98.96
รอบที่ 4	**99.11**	97.88	98.99
รอบที่ 5	**99.08**	97.85	98.97
รอบที่ 6	99.04	97.96	**99.06**
รอบที่ 7	**99.00**	97.79	**99.00**
รอบที่ 8	**99.08**	97.78	99.00
รอบที่ 9	**98.94**	97.96	98.90
รอบที่ 10	**99.09**	97.95	98.99
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**99.06±0.05**	97.85±0.08	98.98±0.04
ค่าร้อยละความผิดพลาด (Error)			
รอบที่ 1	**0.99**	2.21	1.07
รอบที่ 2	**0.88**	2.19	0.98
รอบที่ 3	**0.89**	2.27	1.04
รอบที่ 4	**0.89**	2.12	1.01
รอบที่ 5	**0.92**	2.15	1.03
รอบที่ 6	0.96	2.04	**0.94**
รอบที่ 7	**1.00**	2.21	**1.00**
รอบที่ 8	**0.92**	2.22	1.00
รอบที่ 9	**1.06**	2.04	1.10

ตารางที่ 4.7 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
รอบที่ 10	**0.91**	2.05	1.01
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**0.94±0.05**	2.15±0.08	1.02±0.04
ค่าร้อยละความแม่นยำ (Precision)			
รอบที่ 1	98.38	97.46	**98.98**
รอบที่ 2	**98.82**	97.36	98.52
รอบที่ 3	98.49	97.48	**98.62**
รอบที่ 4	98.59	97.73	**98.98**
รอบที่ 5	**98.59**	97.34	98.30
รอบที่ 6	98.54	97.34	**98.75**
รอบที่ 7	98.34	97.15	**98.58**
รอบที่ 8	98.56	97.41	**98.67**
รอบที่ 9	98.25	97.40	**98.62**
รอบที่ 10	98.51	97.49	**98.81**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	98.51±0.15	97.42±0.14	**98.68±0.20**
ค่าร้อยละความระลึก (Recall)			
รอบที่ 1	**99.66**	98.13	98.88
รอบที่ 2	99.43	98.27	**99.54**
รอบที่ 3	**99.74**	98.00	99.31
รอบที่ 4	**99.65**	98.03	99.00
รอบที่ 5	99.59	98.39	**99.67**
รอบที่ 6	**99.56**	98.61	99.37
รอบที่ 7	**99.69**	98.47	99.43

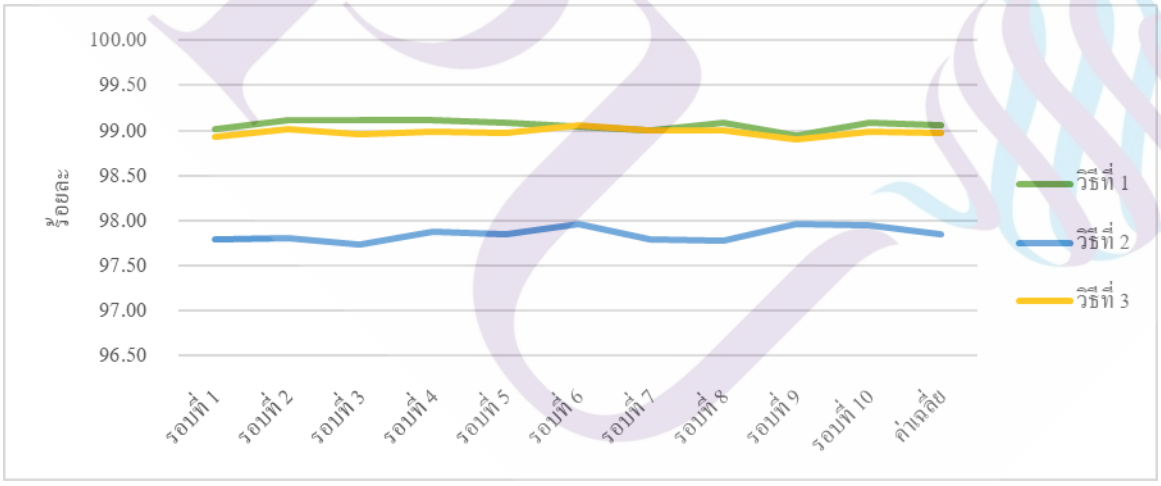
ตารางที่ 4.7 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
รอบที่ 8	**99.61**	98.17	99.35
รอบที่ 9	**99.65**	98.55	99.18
รอบที่ 10	**99.69**	98.43	99.17
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**99.63±0.08**	98.31±0.21	99.29±0.23
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score)			
รอบที่ 1	**99.01**	97.80	98.93
รอบที่ 2	**99.12**	97.81	99.02
รอบที่ 3	**99.11**	97.74	98.96
รอบที่ 4	**99.12**	97.88	98.99
รอบที่ 5	**99.09**	97.86	98.98
รอบที่ 6	99.04	97.97	**99.06**
รอบที่ 7	**99.01**	97.80	99.00
รอบที่ 8	**99.08**	97.79	99.01
รอบที่ 9	**98.94**	97.97	98.90
รอบที่ 10	**99.09**	97.96	98.99
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**99.06±0.06**	97.86±0.08	98.98±0.04
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean)			
รอบที่ 1	**99.02**	97.80	98.93
รอบที่ 2	**99.13**	97.82	99.03
รอบที่ 3	**99.12**	97.74	98.97
รอบที่ 4	**99.12**	97.88	99.00

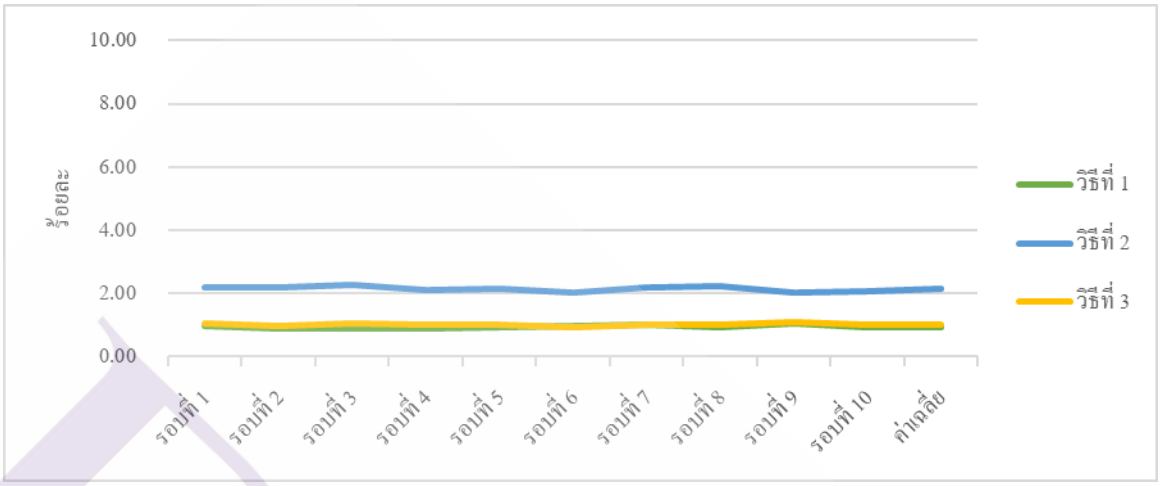
ตารางที่ 4.7 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
รอบที่ 5	**99.09**	97.87	98.99
รอบที่ 6	99.05	97.98	**99.07**
รอบที่ 7	**99.02**	97.81	99.01
รอบที่ 8	**99.09**	97.80	99.01
รอบที่ 9	**98.95**	97.98	98.91
รอบที่ 10	**99.10**	97.96	99.00
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**99.07±0.05**	97.86±0.08	98.99±0.04

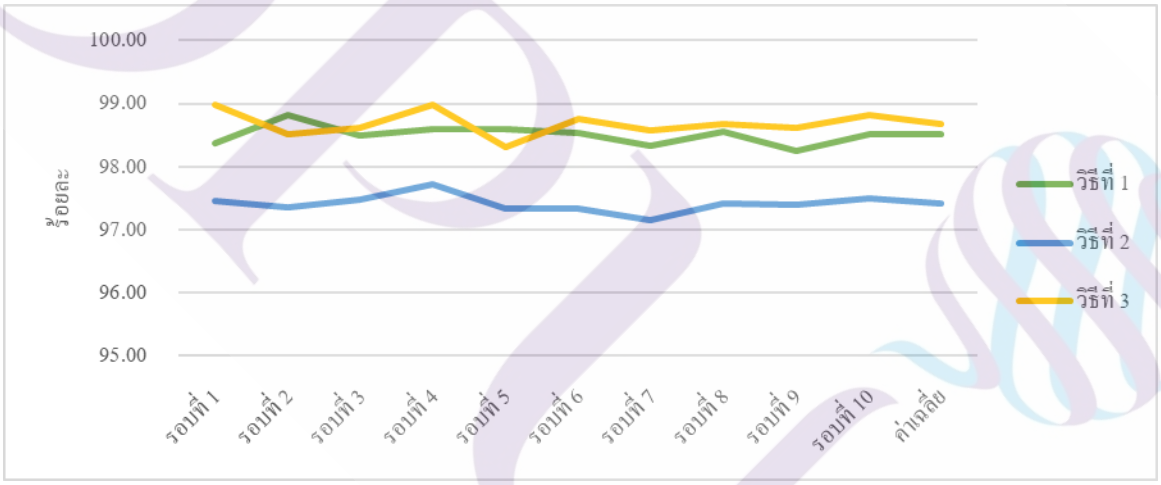
การเปรียบเทียบประสิทธิภาพของโมเดลระหว่างวิธีที่ 1, วิธีที่ 2 และวิธีที่ 3 ในขั้นตอนการสร้าง (Train) โมเดล ดังแสดงในภาพที่ 4.8 – 4.13



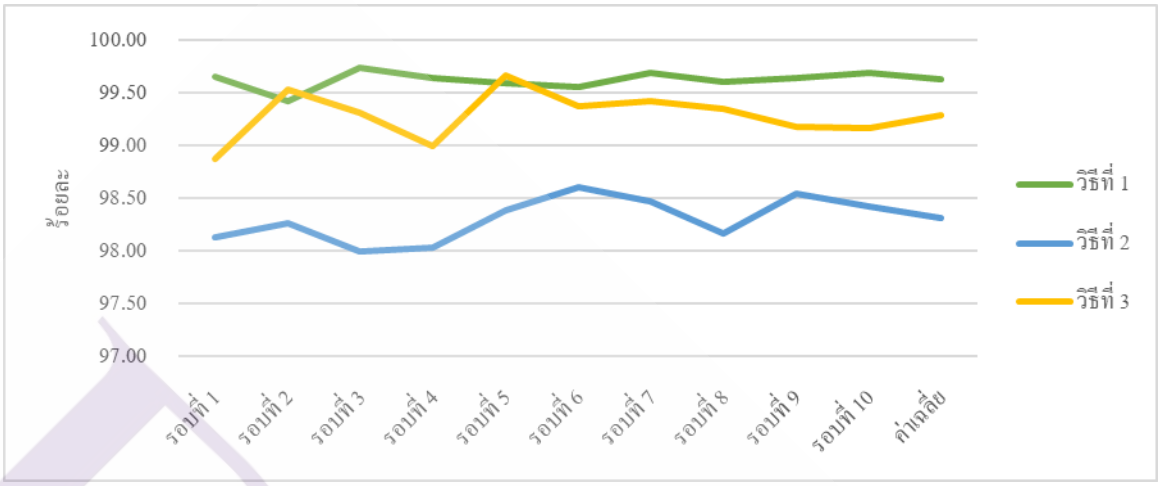
ภาพที่ 4.8 แสดงค่าความถูกต้อง (Accuracy) ในขั้นตอนการสร้าง (Train) โมเดลในแต่ละรอบ



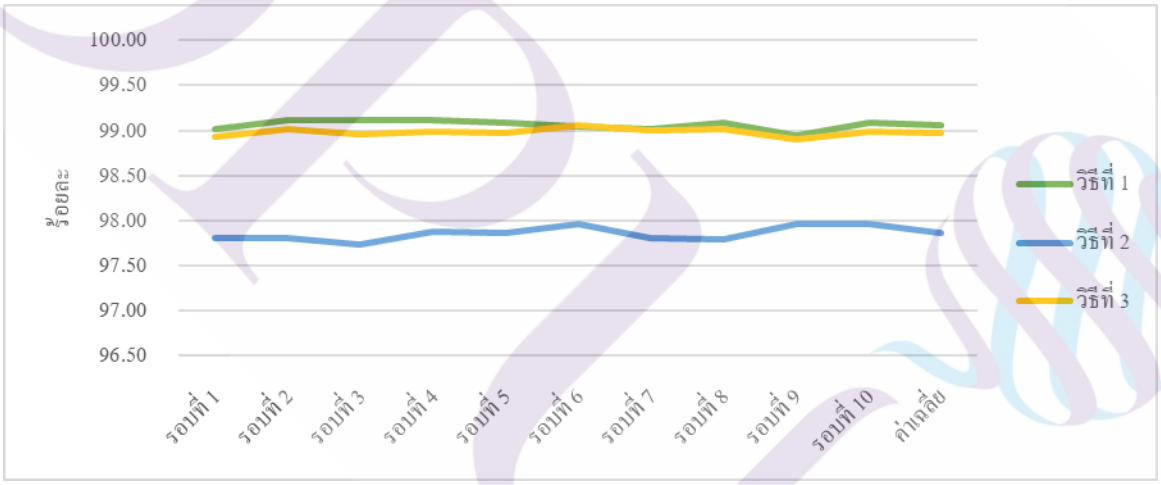
ภาพที่ 4.9 แสดงค่าความผิดพลาด (Error) ในขั้นตอนการสร้าง (Train) โมเดลในแต่ละรอบ



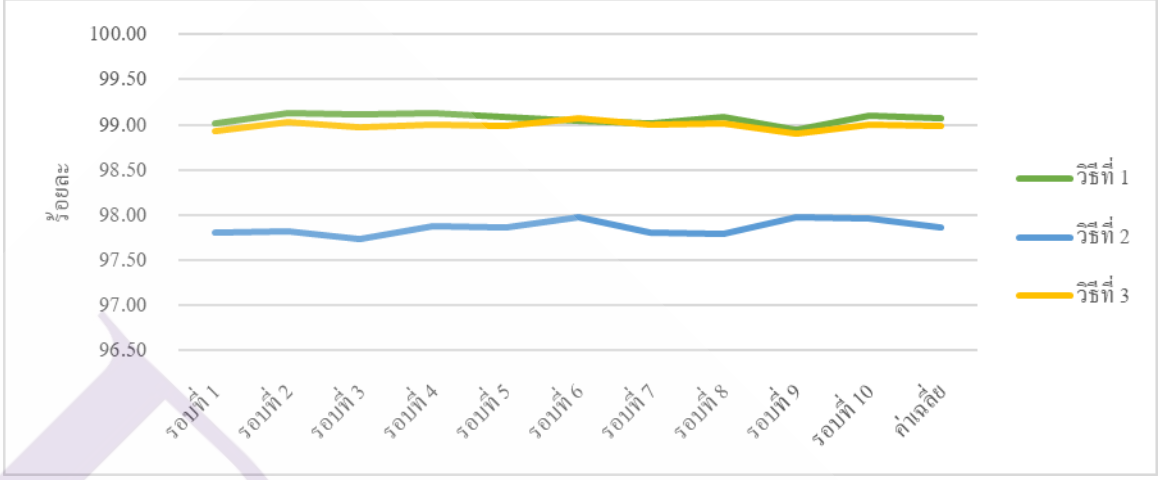
ภาพที่ 4.10 แสดงค่าความแม่นยำ (Precision) ในขั้นตอนการสร้าง (Train) โมเดลในแต่ละรอบ



ภาพที่ 4.11 แสดงค่าความระลึก (Recall) ในขั้นตอนการสร้าง (Train) โมเดลในแต่ละรอบ



ภาพที่ 4.12 แสดงค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) ในขั้นตอนการสร้าง (Train) โมเดลในแต่ละรอบ



ภาพที่ 4.13 แสดงค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ในขั้นตอนการสร้าง (Train) โมเดลในแต่ละรอบ

ผลตารางแจกแจงผลลัพธ์ (Confusion Matrix) ในขั้นตอนการตรวจสอบ (Validation) โมเดล ดังแสดงในตารางที่ 4.8

ตารางที่ 4.8 ผลตารางแจกแจงผลลัพธ์ (Confusion Matrix) ในขั้นตอนการตรวจสอบ (Validation) โมเดล

	วิธีที่ 1		**วิธีที่ 2**		**วิธีที่ 3**	
	Actual		**Actual**		**Actual**	
	Fraudulent	**Non-Fraud**	**Fraudulent**	**Non-Fraud**	**Fraudulent**	**Non-Fraud**
รอบที่ 1						
Predict: Fraudulent	305	123	221	327	294	112
Predict: Non-Fraud	42	3715	126	3511	53	3726

ตารางที่ 4.8 (ต่อ)

	วิธีที่ 1		วิธีที่ 2		วิธีที่ 3	
	Actual		Actual		Actual	
	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud
รอบที่ 2						
Predict: Fraudulent	285	115	196	402	287	124
Predict: Non-Fraud	50	3735	139	3448	48	3726
รอบที่ 3						
Predict: Fraudulent	293	126	201	316	296	101
Predict: Non-Fraud	44	3722	136	3532	41	3747
รอบที่ 4						
Predict: Fraudulent	305	120	200	330	299	86
Predict: Non-Fraud	48	3712	153	3502	54	3746
รอบที่ 5						
Predict: Fraudulent	320	106	230	350	334	108
Predict: Non-Fraud	49	3709	139	3465	35	3707
รอบที่ 6						
Predict: Fraudulent	336	111	221	337	337	111
Predict: Non-Fraud	52	3685	167	3459	51	3685
รอบที่ 7						
Predict: Fraudulent	311	131	236	370	312	119
Predict: Non-Fraud	44	3698	119	3459	43	3710

ตารางที่ 4.8 (ต่อ)

	วิธีที่ 1		วิธีที่ 2		วิธีที่ 3	
	Actual		Actual		Actual	
	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud
รอบที่ 8						
Predict: Fraudulent	350	120	258	338	343	109
Predict: Non-Fraud	39	3675	131	3457	46	3686
รอบที่ 9						
Predict: Fraudulent	313	119	205	354	313	108
Predict: Non-Fraud	44	3708	152	3473	44	3719
รอบที่ 10						
Predict: Fraudulent	3707	104	243	364	322	92
Predict: Non-Fraud	43	3707	130	3447	51	3719
ค่าเฉลี่ย						
Predict: Fraudulent	314.8	117.5	221.1	348.8	313.7	107
Predict: Non-Fraud	45.5	3706.6	139.2	3475.3	46.6	3717.1

ผลการวัดประสิทธิภาพของโมเดลด้วยค่าความถูกต้อง (Accuracy) ค่าความผิดพลาด (Error) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) และค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ในขั้นตอนการตรวจสอบ (Validation) โมเดล ดังแสดงในตารางที่ 4.9

ตารางที่ 4.9 ผลการวัดประสิทธิภาพของโมเดล ในขั้นตอนการตรวจสอบ (Validation) โมเดล

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
ค่าร้อยละความถูกต้อง (Accuracy)			
รอบที่ 1	**96.05**	89.17	**96.05**
รอบที่ 2	**96.05**	87.07	95.89
รอบที่ 3	95.93	89.19	**96.60**
รอบที่ 4	95.98	88.45	**96.65**
รอบที่ 5	96.29	88.31	**96.58**
รอบที่ 6	96.10	87.95	**96.12**
รอบที่ 7	95.81	88.31	**96.12**
รอบที่ 8	96.19	88.79	**96.29**
รอบที่ 9	96.10	87.90	**96.36**
รอบที่ 10	96.48	88.19	**96.58**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	96.10±0.18	88.33±0.63	96.32±0.27
ค่าร้อยละความผิดพลาด (Error)			
รอบที่ 1	**3.95**	10.83	**3.95**
รอบที่ 2	**3.95**	12.93	4.11
รอบที่ 3	4.07	10.81	**3.40**
รอบที่ 4	4.02	11.55	**3.35**
รอบที่ 5	3.71	11.69	**3.42**
รอบที่ 6	3.90	12.05	**3.88**
รอบที่ 7	4.19	11.69	**3.88**
รอบที่ 8	3.81	11.21	**3.71**
รอบที่ 9	3.90	12.10	**3.64**

ตารางที่ 4.9 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
รอบที่ 10	3.52	11.81	**3.42**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	3.90±0.18	11.67±0.63	**3.68±0.27**
ค่าร้อยละความแม่นยำ (Precision)			
รอบที่ 1	71.26	40.32	**72.41**
รอบที่ 2	**71.25**	32.77	69.82
รอบที่ 3	69.92	38.87	**74.55**
รอบที่ 4	71.76	37.73	**77.66**
รอบที่ 5	75.11	39.65	**75.56**
รอบที่ 6	75.16	39.6	**75.22**
รอบที่ 7	70.36	38.94	**72.38**
รอบที่ 8	74.46	43.28	**75.88**
รอบที่ 9	72.45	36.67	**74.34**
รอบที่ 10	76.03	40.03	**77.77**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	72.78±2.21	38.78±2.73	**74.55±2.47**
ค่าร้อยละความระลึก (Recall)			
รอบที่ 1	**87.89**	63.68	84.72
รอบที่ 2	85.07	58.50	**85.67**
รอบที่ 3	86.94	59.64	**87.83**
รอบที่ 4	**86.40**	56.65	84.70
รอบที่ 5	86.72	62.33	**90.51**
รอบที่ 6	86.59	56.95	**86.85**
รอบที่ 7	87.60	66.47	**87.88**

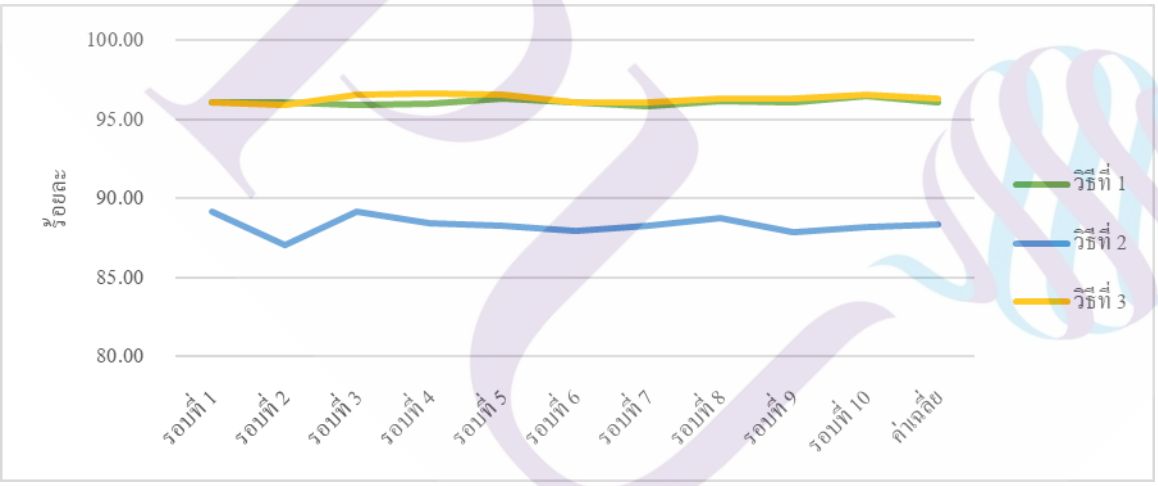
ตารางที่ 4.9 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
รอบที่ 8	**89.97**	66.32	88.17
รอบที่ 9	**87.67**	57.42	**87.67**
รอบที่ 10	**88.47**	65.14	86.32
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**87.33±1.32**	61.31±3.93	87.03±1.77
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score)			
รอบที่ 1	**78.70**	49.38	78.08
รอบที่ 2	**77.55**	42.01	76.94
รอบที่ 3	77.51	47.07	**80.65**
รอบที่ 4	78.40	45.30	**81.02**
รอบที่ 5	80.50	48.47	**82.36**
รอบที่ 6	80.47	46.72	**80.62**
รอบที่ 7	78.04	49.11	**79.38**
รอบที่ 8	81.49	52.38	**81.56**
รอบที่ 9	79.34	44.75	**80.46**
รอบที่ 10	81.78	49.59	**81.82**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	79.38±1.58	47.48±2.95	**80.29±1.69**
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean)			
รอบที่ 1	**79.14**	50.68	78.33
รอบที่ 2	**77.86**	43.79	77.35
รอบที่ 3	77.97	48.15	**80.92**
รอบที่ 4	78.74	46.24	**81.11**
รอบที่ 5	80.71	49.72	**82.70**

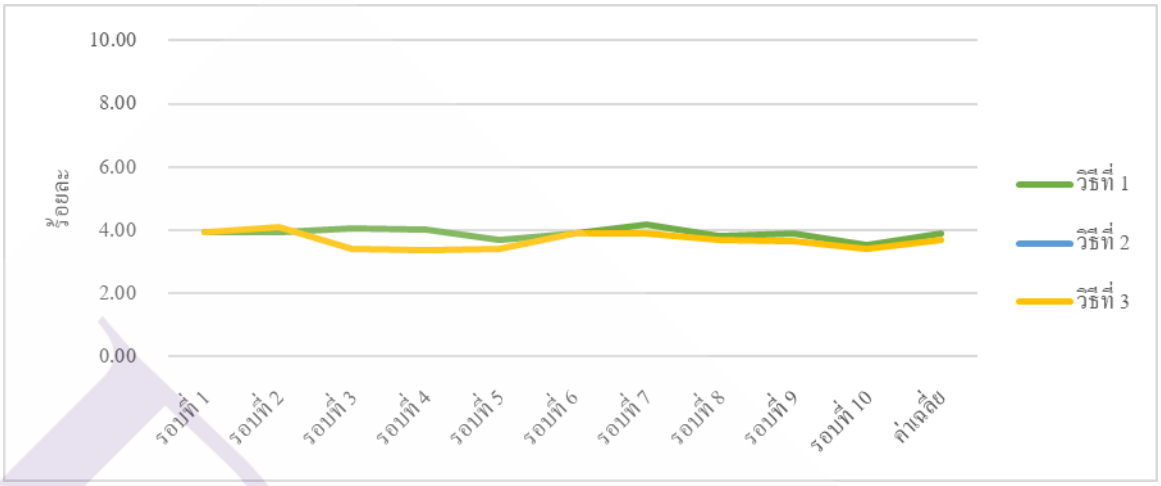
ตารางที่ 4.9 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
รอบที่ 6	80.68	47.50	**80.83**
รอบที่ 7	78.51	50.88	**79.76**
รอบที่ 8	**81.85**	53.58	81.80
รอบที่ 9	79.70	45.89	**80.74**
รอบที่ 10	**98.06**	51.07	81.94
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	**81.32±6.02**	48.75±2.96	80.55±1.65

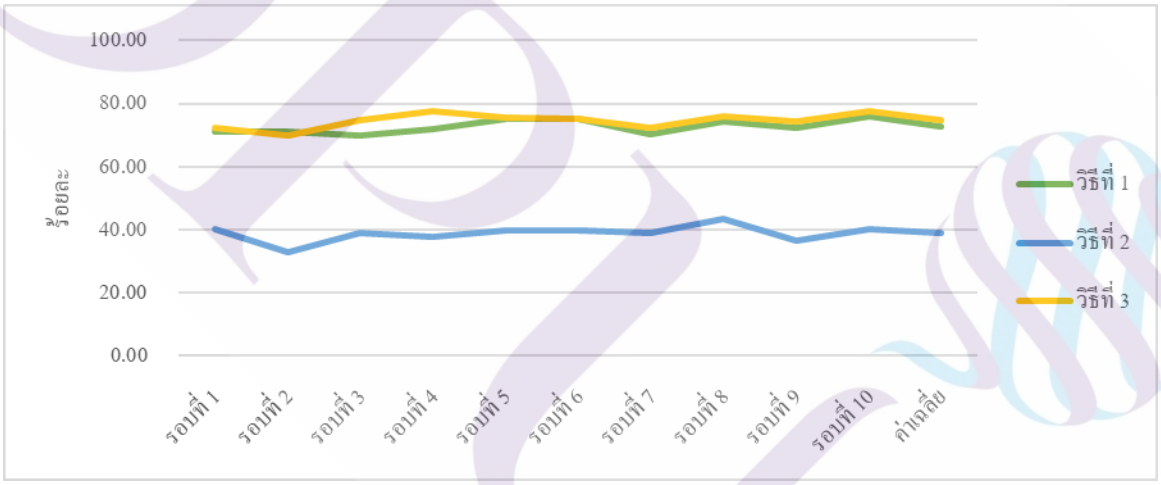
การเปรียบเทียบประสิทธิภาพของโมเดลระหว่างวิธีที่ 1, วิธีที่ 2 และวิธีที่ 3 ในขั้นตอนการตรวจสอบ (Validation) โมเดล ดังแสดงในภาพที่ 4.14 – 4.19



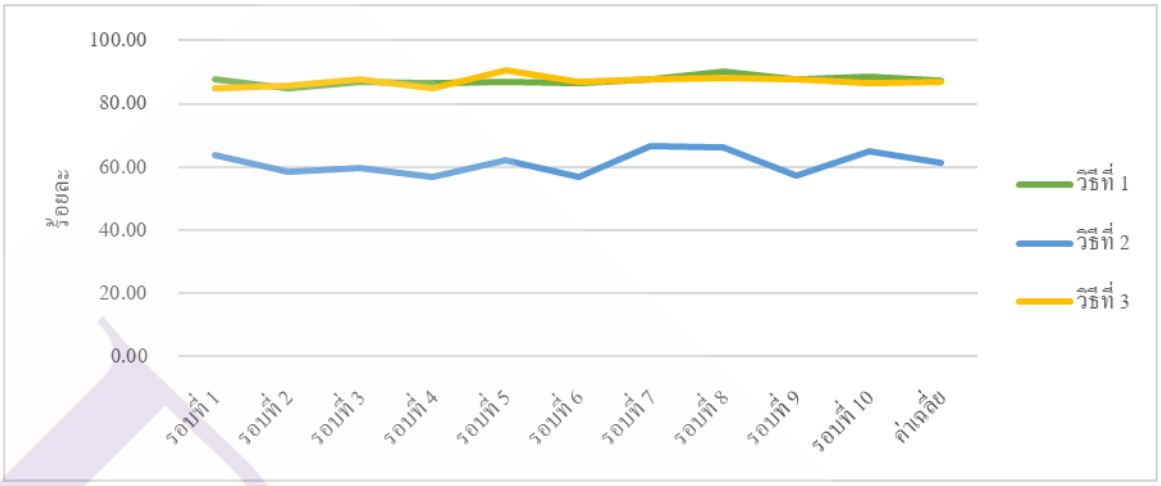
ภาพที่ 4.14 แสดงค่าความถูกต้อง (Accuracy) ในขั้นตอนการตรวจสอบ (Validation) โมเดลในแต่ละรอบ



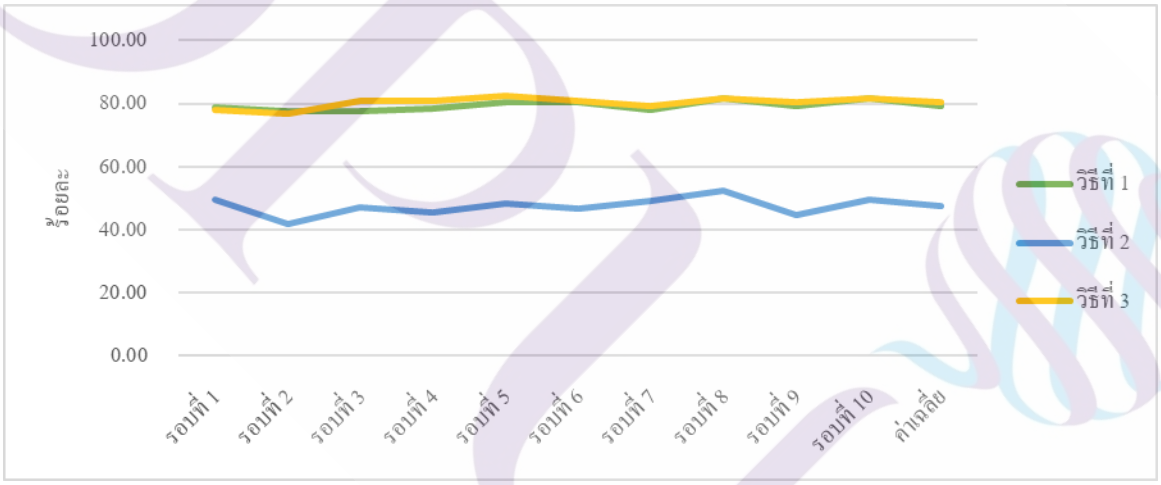
ภาพที่ 4.15 แสดงค่าความผิดพลาด (Error) ในขั้นตอนการตรวจสอบ (Validation) โมเดลในแต่ละรอบ



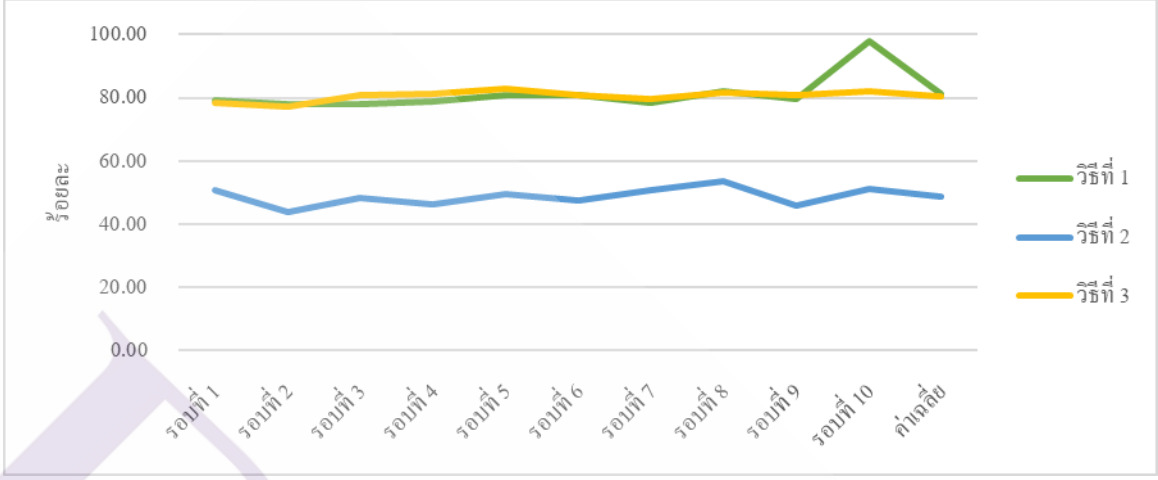
ภาพที่ 4.16 แสดงค่าความแม่นยำ (Precision) ในขั้นตอนการตรวจสอบ (Validation) โมเดลในแต่ละรอบ



ภาพที่ 4.17 แสดงค่าความระลึก (Recall) ในขั้นตอนการตรวจสอบ (Validation) โมเดลในแต่ละรอบ



ภาพที่ 4.18 แสดงค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) ในขั้นตอนการตรวจสอบ (Validation) โมเดลในแต่ละรอบ



ภาพที่ 4.19 แสดงค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ในขั้นตอนการตรวจสอบ (Validation) โมเดลในแต่ละรอบ

ผลตารางแจกแจงผลลัพธ์ (Confusion Matrix) ในขั้นตอนการทดสอบ (Test) โมเดล ดังแสดงในตารางที่ 4.10

ตารางที่ 4.10 ผลตารางแจกแจงผลลัพธ์ (Confusion Matrix) ในขั้นตอนการทดสอบ (Test) โมเดล

	วิธีที่ 1		**วิธีที่ 2**		**วิธีที่ 3**	
	Actual		**Actual**		**Actual**	
	Fraudulent	**Non-Fraud**	**Fraudulent**	**Non-Fraud**	**Fraudulent**	**Non-Fraud**
รอบที่ 1						
Predict: Fraudulent	1284	502	939	1487	1270	378
Predict: Non-Fraud	212	15936	557	14951	226	16060
รอบที่ 2						
Predict: Fraudulent	1259	436	922	1569	1292	450
Predict: Non-Fraud	237	16002	574	14869	204	15988

ตารางที่ 4.10 (ต่อ)

	วิธีที่ 1		วิธีที่ 2		วิธีที่ 3	
	Actual		Actual		Actual	
	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud
รอบที่ 3						
Predict: Fraudulent	1283	461	917	1459	1292	444
Predict: Non-Fraud	213	15977	579	14979	204	15994
รอบที่ 4						
Predict: Fraudulent	1266	466	910	1395	1270	403
Predict: Non-Fraud	230	15972	586	15043	226	16035
รอบที่ 5						
Predict: Fraudulent	1280	461	918	1523	1294	474
Predict: Non-Fraud	216	15977	578	14915	202	15964
รอบที่ 6						
Predict: Fraudulent	1266	449	911	1566	1282	435
Predict: Non-Fraud	230	15989	585	14872	214	16003
รอบที่ 7						
Predict: Fraudulent	1283	478	949	1574	1300	444
Predict: Non-Fraud	213	15960	547	14864	196	15994
รอบที่ 8						
Predict: Fraudulent	1271	461	922	1464	1303	440
Predict: Non-Fraud	225	15977	574	14974	193	15998

ตารางที่ 4.10 (ต่อ)

	วิธีที่ 1		วิธีที่ 2		วิธีที่ 3	
	Actual		Actual		Actual	
	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud	Fraudulent	Non-Fraud
รอบที่ 9						
Predict: Fraudulent	1292	493	933	1526	1284	416
Predict: Non-Fraud	204	15945	563	14912	212	16022
รอบที่ 10						
Predict: Fraudulent	1289	479	906	1455	1277	431
Predict: Non-Fraud	207	15959	590	14983	219	16007
ค่าเฉลี่ย						
Predict: Fraudulent	1277.3	468.6	922.7	1501.8	1286.4	431.5
Predict: Non-Fraud	218.7	15969.4	573.3	14936.2	209.6	16006.5

ผลการวัดประสิทธิภาพของโมเดลด้วยค่าความถูกต้อง (Accuracy) ค่าความผิดพลาด (Error) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) และค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ในขั้นตอนการตรวจสอบ (Test) โมเดล ดังแสดงในตารางที่ 4.11

ตารางที่ 4.11 ผลการวัดประสิทธิภาพของโมเดล ในขั้นตอนการตรวจสอบ (Test) โมเดล

มาตรวัด	**วิธีที่ 1**	**วิธีที่ 2**	**วิธีที่ 3**
ค่าร้อยละความถูกต้อง (Accuracy)			
รอบที่ 1	96.01	88.60	**96.63**
รอบที่ 2	96.24	88.05	**96.35**

ตารางที่ 4.11 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
ค่าร้อยละความถูกต้อง (Accuracy)			
รอบที่ 3	96.24	88.63	**96.38**
รอบที่ 4	96.11	88.95	**96.49**
รอบที่ 5	96.22	88.28	**96.23**
รอบที่ 6	96.21	88.00	**96.38**
รอบที่ 7	96.14	88.17	**96.43**
รอบที่ 8	96.17	88.63	**96.47**
รอบที่ 9	96.11	88.35	**96.49**
รอบที่ 10	96.17	88.59	**96.37**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	96.10±0.18	88.43±0.30	96.42±0.10
ค่าร้อยละความผิดพลาด (Error)			
รอบที่ 1	3.99	11.40	**3.37**
รอบที่ 2	3.76	11.95	**3.65**
รอบที่ 3	3.76	11.37	**3.62**
รอบที่ 4	3.89	11.05	**3.51**
รอบที่ 5	3.78	11.72	**3.77**
รอบที่ 6	3.79	12.00	**3.62**
รอบที่ 7	3.86	11.83	**3.57**
รอบที่ 8	3.83	11.37	**3.53**
รอบที่ 9	3.89	11.65	**3.51**
รอบที่ 10	3.83	11.41	**3.63**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	3.84±0.07	11.58±0.30	**3.58±0.10**

ตารางที่ 4.11 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
ค่าร้อยละความแม่นยำ (Precision)			
รอบที่ 1	71.89	38.70	**77.06**
รอบที่ 2	**74.27**	37.01	74.16
รอบที่ 3	73.56	38.59	**74.42**
รอบที่ 4	73.09	39.47	**75.91**
รอบที่ 5	**73.52**	37.60	73.19
รอบที่ 6	73.81	36.77	**74.66**
รอบที่ 7	72.85	37.61	**74.54**
รอบที่ 8	73.38	38.64	**74.75**
รอบที่ 9	72.38	37.94	**75.52**
รอบที่ 10	72.90	38.37	**74.76**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	73.17±0.69	38.07±0.83	**74.90±1.05**
ค่าร้อยละความระลึก (Recall)			
รอบที่ 1	**85.82**	62.76	84.89
รอบที่ 2	84.15	61.63	**86.36**
รอบที่ 3	85.76	61.29	**86.36**
รอบที่ 4	84.62	60.82	**84.89**
รอบที่ 5	85.56	61.36	**86.49**
รอบที่ 6	84.62	60.89	**85.69**
รอบที่ 7	85.76	63.43	**86.89**
รอบที่ 8	84.95	61.63	**87.09**
รอบที่ 9	**86.36**	62.36	85.82

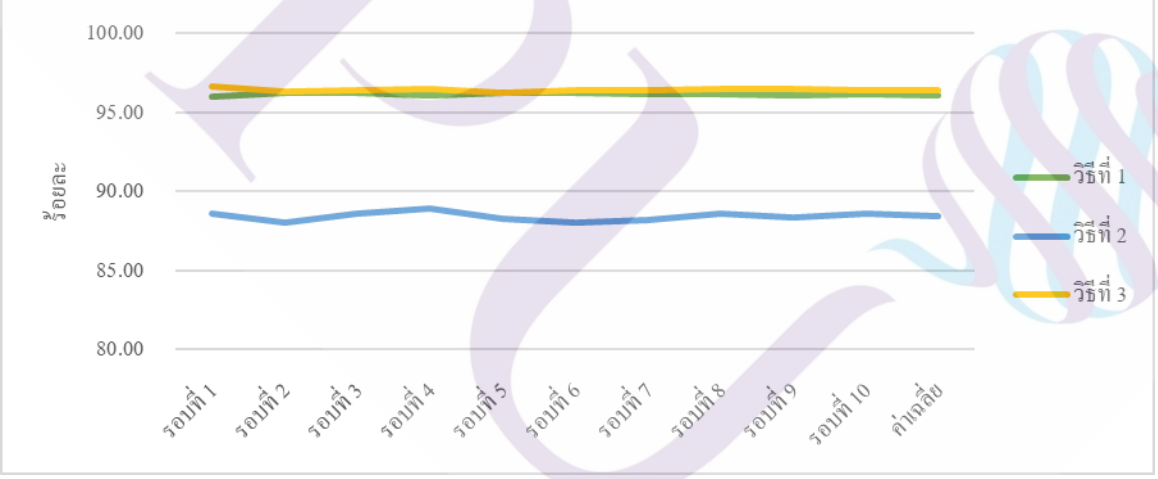
ตารางที่ 4.11 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
ค่าร้อยละความระลึก (Recall)			
รอบที่ 10	**86.16**	60.56	85.36
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	85.37±0.74	61.67±0.91	**85.98±0.78**
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score)			
รอบที่ 1	78.24	47.88	**80.78**
รอบที่ 2	78.90	46.25	**79.80**
รอบที่ 3	79.19	47.36	**79.95**
รอบที่ 4	78.43	47.88	**80.15**
รอบที่ 5	79.08	46.63	**79.28**
รอบที่ 6	78.85	45.85	**79.80**
รอบที่ 7	78.78	47.22	**80.24**
รอบที่ 8	78.74	47.50	**80.45**
รอบที่ 9	78.75	47.18	**80.35**
รอบที่ 10	78.98	46.97	**79.71**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	78.79±0.28	47.07±0.66	**80.05±0.43**
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean)			
รอบที่ 1	78.55	49.29	**80.88**
รอบที่ 2	79.06	47.76	**80.03**
รอบที่ 3	79.43	48.64	**80.17**
รอบที่ 4	78.65	49.00	**80.28**
รอบที่ 5	79.31	48.04	**79.57**
รอบที่ 6	79.04	47.32	**79.99**

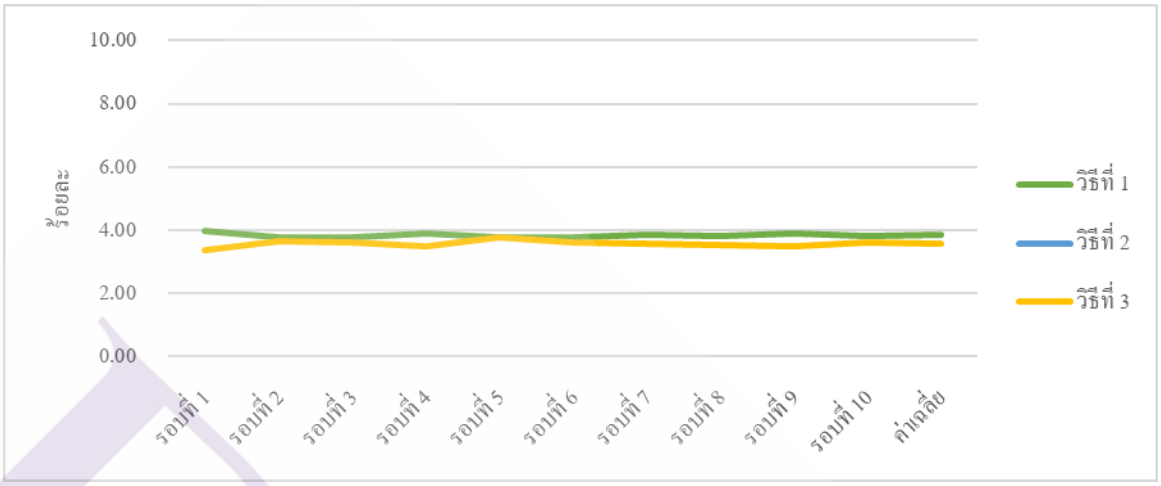
ตารางที่ 4.11 (ต่อ)

มาตรวัด	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3
ค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean)			
รอบที่ 7	79.05	48.85	**80.48**
รอบที่ 8	78.96	48.80	**80.69**
รอบที่ 9	79.06	48.64	**80.51**
รอบที่ 10	79.26	48.21	**79.89**
ค่าเฉลี่ย±ส่วนเบี่ยงเบนมาตรฐาน	79.04±0.27	48.46±0.60	**80.25±0.40**

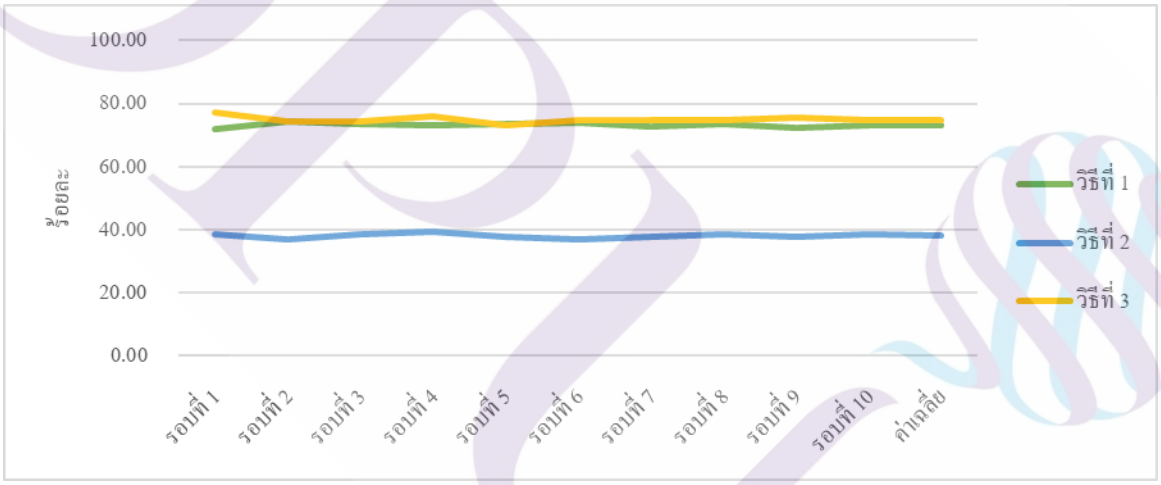
การเปรียบเทียบประสิทธิภาพของโมเดลระหว่างวิธีที่ 1, วิธีที่ 2 และวิธีที่ 3 ในขั้นตอนการตรวจสอบ (Test) โมเดล ดังแสดงในภาพที่ 4.20 – 4.25



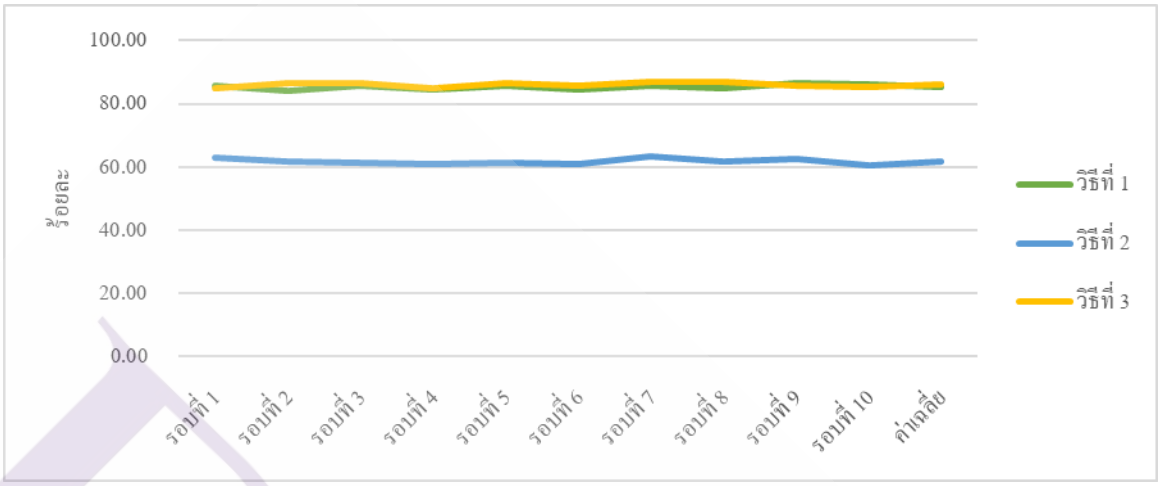
ภาพที่ 4.20 แสดงค่าความถูกต้อง (Accuracy) ในขั้นตอนการตรวจสอบ (Test) โมเดลในแต่ละรอบ



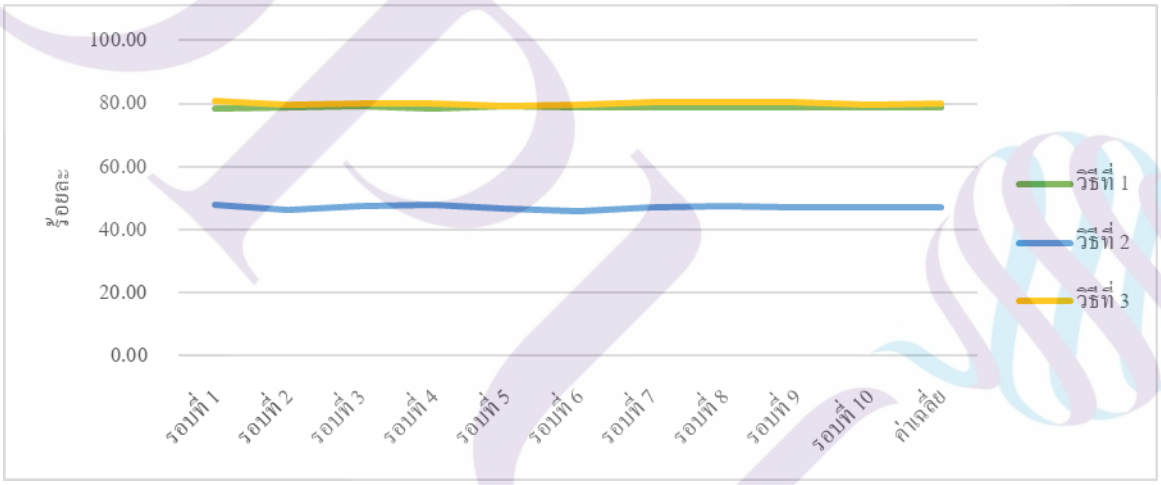
ภาพที่ 4.21 แสดงค่าความผิดพลาด (Error) ในขั้นตอนการตรวจสอบ (Test) โมเดลในแต่ละรอบ



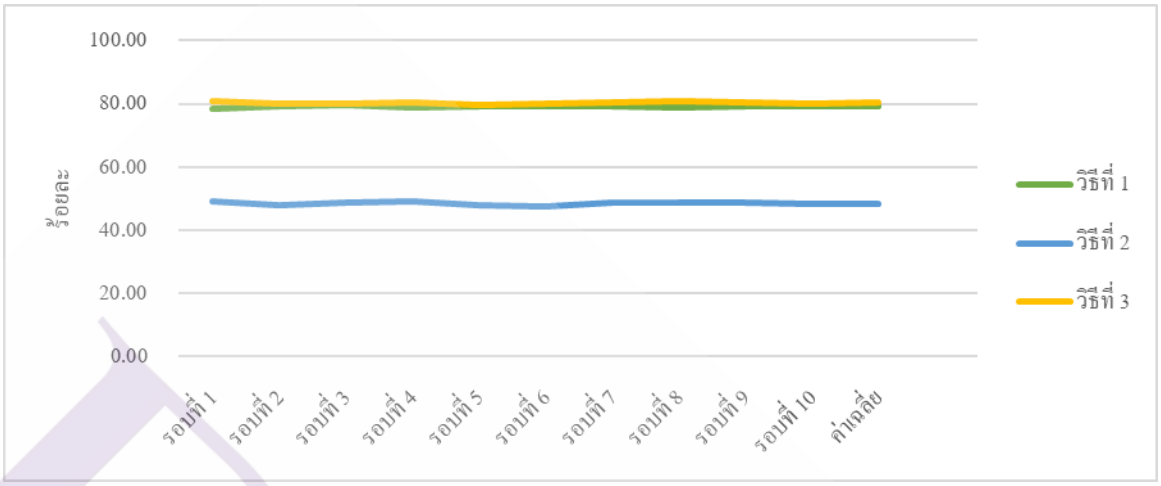
ภาพที่ 4.22 แสดงค่าความแม่นยำ (Precision) ในขั้นตอนการตรวจสอบ (Test) โมเดลในแต่ละรอบ



ภาพที่ 4.23 แสดงค่าความระลึก (Recall) ในขั้นตอนการตรวจสอบ (Test) โมเดลในแต่ละรอบ



ภาพที่ 4.24 แสดงค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (F1 Score) ในขั้นตอนการตรวจสอบ (Test) โมเดลในแต่ละรอบ



ภาพที่ 4.25 แสดงค่าเฉลี่ยความแม่นยำของผลลัพธ์ข้อมูลทุกกลุ่ม (Geometric Mean) ในขั้นตอนการตรวจสอบ (Test)โมเดลในแต่ละรอบ

บทที่ 5
บทสรุปและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอวิธีการทำนายธุรกรรมที่หลอกลวง โดยใช้เทคนิคการเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วยวิธี Node2Vec และค่าความเป็นศูนย์กลาง มาช่วยทำนายธุรกรรมที่มีแนวโน้มจะหลอกลวง โดยใช้ข้อมูลการซื้อขาย Bitcoin ของตลาดกลางพาณิชย์อิเล็กทรอนิกส์ชื่อ Bitcoin Alpha ซึ่งเป็นข้อมูลที่เผยแพร่ให้บุคคลทั่วไปสามารถนำข้อมูลไปใช้ประโยชน์ได้ ข้อมูลชุดนี้ มีแอททริบิวต์ทั้งหมด 4 แอทริบิวต์ ได้แก่ Source, Target, Rating, TimeStamp มีจำนวนระเบียนข้อมูลทั้งสิ้น 59,788 ระเบียน ทำการสร้างลาเบลคำตอบจากคะแนนความพึงพอใจ (Rating) และแปลงรูปข้อมูลวันและเวลาของการให้คะแนนความพึงพอใจ แต่เนื่องจากข้อมูลเป็นข้อมูลรายธุรกรรม การสกัดคุณลักษณะจึงมีความสำคัญ ผู้วิจัยจึงแบ่งวิธีการทำนายออกเป็น 3 วิธี ตามการสกัดคุณลักษณะ ได้แก่ วิธีที่ 1 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และค่าความเป็นศูนย์กลาง วิธีที่ 2 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วย Node2Vec และวิธีที่ 3 การเรียนรู้แบบกลุ่มร่วมกับการสกัดคุณลักษณะด้วยค่าความเป็นศูนย์กลาง ทั้ง 3 วิธีจะทำการคัดเลือกคุณลักษณะด้วยวิธี Recursive Feature Elimination (RFE) หลังจากนั้นแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธี SMOTE แบ่งข้อมูลเพื่อนำไปทดสอบประสิทธิภาพของโมเดลด้วยวิธี 10-fold cross validation และสร้างโมเดลด้วยการเรียนรู้แบบกลุ่ม จากนั้นทำการเปรียบเทียบประสิทธิภาพของโมเดลในแต่ละวิธี

5.1 สรุปผลการศึกษา

ในการเปรียบเทียบผลในการทำนายธุรกรรมที่หลอกลวง ในด้านตารางแจกแจงผลลัพธ์ (Confusion Matrix) พบว่า วิธีที่ 1 และ 3 สามารถทำนายว่าทุจริต แล้วทุจริตจริง (True Positive) ได้สูงขึ้น ส่วนผลการวัดประสิทธิภาพของโมเดล พบว่า ค่าเฉลี่ยความถูกต้องของทั้ง 3 วิธีเท่ากับ 96.10%,

88.43% และ 96.42% ตามลำดับ รองลงมาคือค่าเฉลี่ยความระลึกของทั้ง 3 วิธีเท่ากับ 85.37%, 61.67% และ 85.98% ตามลำดับ

5.2 ข้อเสนอแนะ

ประสิทธิภาพของโมเดลในแต่ละวิธีสามารถทำนายธุรกรรมที่หลอกลวงได้ดีในระดับหนึ่ง แต่ยังสามารถปรับปรุงเพื่อให้ผลลัพธ์ที่ดียิ่งขึ้นดังนี้

5.2.1 ลาเบลคำตอบ

ในงานวิจัยนี้ใช้คะแนนความพึงพอใจมาประยุกต์ใช้เป็นลาเบลคำตอบ ผลการทำนายมีความแม่นยำในระดับหนึ่ง สำหรับงานวิจัยครั้งถัดไปอาจจะใช้เทคนิค Semi-Supervise โดยการนำข้อมูลที่ทราบว่าทุจริตจริงมาประยุกต์ใช้ร่วมกับข้อมูลที่ยังไม่ทราบว่าทุจริตหรือไม่ เพื่อช่วยสร้างลาเบลคำตอบได้อีกวิธีหนึ่ง

5.2.2 การสกัดคุณลักษณะ

ในงานวิจัยนี้ได้นำเสนอวิธี Node2Vec เป็นเทคนิคหนึ่งในการทำ Node embedding ที่ใช้ความน่าจะเป็นเข้ามาช่วยในการสุ่มเดิน เพื่อช่วยจัดกลุ่มโหนดที่มีความคล้ายคลึงกัน (Homophily) เข้ามาอยู่ในกลุ่มเดียวกัน หากทำการสกัดคุณลักษณะวิธีอื่น เช่น graph2vec หรือ X2vec ซึ่งเป็นเทคนิคทางด้าน Graph embedding อาจจะได้คุณลักษณะเพิ่มเติมที่หลากหลาย และประสิทธิภาพในการทำนายดียิ่งขึ้น

5.2.3 ค่าพารามิเตอร์

ในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เป็นค่าเริ่มต้น (Default) ซึ่งอาจจะไม่ใช่ค่าพารามิเตอร์ที่เหมาะสมที่สุด ดังนั้นหากสามารถปรับเปลี่ยนค่าพารามิเตอร์ได้ ก็อาจจะเพิ่มประสิทธิภาพของโมเดลได้แม่นยำขึ้น

5.2.4 การพิจารณาข้อมูลทีละช่วงเวลา

ในงานวิจัยนี้ได้นำข้อมูลทั้งหมดทุกช่วงเวลามาสร้างโมเดล ในงานวิจัยครั้งต่อไป อาจจะแบ่งข้อมูลออกเป็นแต่ละปี จากนั้นเปรียบเทียบผลการคัดเลือกคุณลักษณะในแต่ละปีว่ามีความสอดคล้องกันมากน้อยเท่าใด รวมถึงประสิทธิภาพของโมเดลในแต่ละช่วงเวลาแตกต่างกันหรือไม่

5.2.5 การเพิ่มข้อมูลย้อนหลัง (Historical Data)

ในงานวิจัยนี้ได้ใช้ข้อมูลที่เผยแพร่ให้บุคคลทั่วไปสามารถนำข้อมูลไปใช้ประโยชน์ได้ โดยชุดข้อมูลนี้เป็นข้อมูลรายธุรกรรมในปีค.ศ. 2010 – 2015 ซึ่งถ้าสามารถนำข้อมูลย้อนหลังเพิ่มขึ้นอีกอาจจะช่วยเพิ่มประสิทธิภาพของโมเดลได้แม่นยำขึ้น

บรรณานุกรม

บรรณานุกรม

เชษฐพงศ์ ปัญญาชนกุล และ อานนท์ ศักดิ์วรวิชญ์. (2559). “การพยากรณ์การสูญเสียลูกค้าด้วยเทคนิคการวิเคราะห์เครือข่าย,” สถาบันบัณฑิตพัฒนบริหารศาสตร์.

สรีพร โพธิ์งาม. (2560). “อิทธิพลของภาพลักษณ์ตราสินค้า ความเชื่อถือ และความพึงพอใจต่อความภักดีต่อลาซาด้าของลูกค้าในกรุงเทพมหานคร,” บัณฑิตวิทยาลัย มหาวิทยาลัยกรุงเทพ.

สำนักงานตำรวจแห่งชาติ. (2562). “กองปราบปรามเตือนกลโกงรูปแบบใหม่ หลอกลวงผู้ซื้อขายสินค้าบนโลกออนไลน์,” สืบค้นจาก https://www.onlinenewstime.com/กองปราบปรามเตือนกลกองป/news-update/

สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. (2558). “รายงานผลการสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทย ปี 2557,” กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม.

สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. (2561). “รายงานผลการสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทย ปี 2560,” กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม.

สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. (2562). “เอกสารการแถลงผลการสำรวจมูลค่าพาณิชย์อิเล็กทรอนิกส์ ปี 2561,” Thailand e-Commerce Week 2019.

สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. (2562). “รายงานผลการสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทย ปี 2561,” กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม.

Aditya Grover and Jure Leskovec. (2016). “Node2vec: Scalable Feature Learning for Networks,” KDD. 13-17 August 2016.

Efraim Turban, Jon Outland, David King, Jae Kyu Lee, Ting-Peng Liang and Deborrah C. Turban. (2018). “Electronic Commerce 2018: A Managerial and Social Networks Perspective.” Springer.

David W Hosmer and Lemeshow Stanley. (1989). “Applied Logistic Regression,” New York: John Wiley.

Jayesh Soni and Himanshu Upadhyay. (2019). “Feature Extraction through Deepwalk on Weighed Graph,” International Conference on Data Science. ICDATA’19.

Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos and V.S. Subrahmanian. (2018). "REV2: Fraudulent User Prediction in Rating Platforms," In Proceedings of 11th ACM International Conf. on Web Search and Data Mining (WSDM 2018).

ประวัติผู้เขียน

ชื่อ-นามสกุล	วราภรณ์ พิมา
ประวัติการศึกษา	พ.ศ. 2548 สำเร็จการศึกษาระดับปริญญาตรี สาขาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏอุบลราชธานี พ.ศ. 2551 สำเร็จการศึกษาระดับปริญญาโท สาขาสถิติประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่
ตำแหน่งและสถานที่ทำงานปัจจุบัน	นักวิทยาศาสตร์ข้อมูล บริษัท กรุงศรี ออโต้