การพัฒนาแบบจำลองการเรียนรู้ของเครื่องด้วยเทคนิค Stacking โดยใช้กลยุทธ์ปรับสมดุลข้อมูล สำหรับการประเมินความเสี่ยงสินเชื่อ SME

นายสรธัญ อินทรสมพงค์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์และวิทยาการวิเคราะห์ข้อมูล
บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2568

การพัฒนาแบบจำลองการเรียนรู้ของเครื่องด้วยเทคนิค Stacking โดยใช้กลยุทธ์ปรับสมดุลข้อมูล สำหรับการประเมินความเสี่ยงสินเชื่อ SME

นายสรธัญ อินทรสมพงค์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติประยุกต์และวิทยาการวิเคราะห์ข้อมูล
บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2568

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

เรื่อง การพัฒนาแบบจำลองการเรียนรู้ของเครื่องด้วยเทคนิค Stacking โดยใช้กลยุทธ์ปรับสมดุลข้อมูลสำหรับการประเมินความเสี่ยงสินเชื่อ SME

โดย นายสรธัญ อินทรสมพงค์

ได้รับอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์และวิทยาการวิเคราะห์ข้อมูล

คณบดีบัณฑิตวิทยาลัย / หัวหน้าภาควิชา

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(พัทธ์ชนก ศรีสุรเดชชัย)

..... อาจารย์ที่ปรึกษา

(รองศาสตราจารย์ ดร.วิกานดา ผาพันธ์)

..... กรรมการ

(ศิรประภา มโนมัธย์)

ชื่อ : นายสรธัญ อินทรสมพงค์
ชื่อวิทยานิพนธ์ : การพัฒนาแบบจำลองการเรียนรู้ของเครื่องด้วยเทคนิค Stacking โดยใช้กลยุทธ์ปรับสมดุลข้อมูลสำหรับการประเมินความเสี่ยงสินเชื่อ SME
สาขาวิชา : สถิติประยุกต์และวิทยาการวิเคราะห์ข้อมูล มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : รองศาสตราจารย์ ดร.วิกานดา ผาพันธ์
ปีการศึกษา : 2568

บทคัดย่อ

วิสาหกิจขนาดกลางและขนาดย่อม (Small and Medium Enterprises: SMEs) มักเผชิญอุปสรรคในการขอสินเชื่อ เนื่องจากถูกมองว่ามีความเสี่ยงด้านเครดิตสูง งานวิจัยนี้มุ่งพัฒนาแบบจำลองการเรียนรู้เชิงลึกด้วยเทคนิคการรวมแบบจำลอง (Stacking) เพื่อเพิ่มประสิทธิภาพในการประเมินความเสี่ยงสินเชื่อของวิสาหกิจขนาดกลางและขนาดย่อม โดยใช้ชุดข้อมูลจากกระทรวงอุตสาหกรรม ซึ่งประกอบด้วยตัวแปรเชิงปริมาณและเชิงคุณภาพ รวม 14 ตัวแปร เนื่องจากข้อมูลมีความไม่สมดุล จึงใช้กลยุทธ์ปรับสมดุลข้อมูล 4 วิธี ได้แก่ เทคนิคการสุ่มตัวอย่างสังเคราะห์สำหรับกลุ่มตัวอย่างน้อย (Synthetic Minority Over-sampling Technique: SMOTE), เทคนิคการสุ่มตัวอย่างแบบปรับตัวเอง (Adaptive Synthetic Sampling: ADASYN), การรวมกันของการสุ่มตัวอย่างสังเคราะห์และเทคนิคการตัดตัวอย่างที่มีเสียงรบกวน (SMOTE and Edited Nearest Neighbors: SMOTE + ENN) และการรวมกันของการสุ่มตัวอย่างสังเคราะห์และการกำจัดตัวอย่างที่ซ้ำซ้อน (SMOTE and Tomek Links: SMOTE + Tomek) จากนั้นทำการเปรียบเทียบประสิทธิภาพของแบบจำลอง 9 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree), วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines), วิธีบูสต์ติงไล่ระดับ (Gradient Boosting), วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), วิธีการจำแนกแบบเบส์นาอีฟ (Naïve Bayes), วิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (Extreme Gradient Boosting with Meta-Learning) และวิธีเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning) โดยวัดผลด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าความจำ (Recall), ค่ามาตราวัดเอฟ (F1-score)

และตัววัดประสิทธิภาพโดยใช้พื้นที่ใต้กราฟ (Area Under an ROC Curve : AUCs) ผลการศึกษาพบว่า เทคนิคการรวมแบบจำลอง โดยเฉพาะแบบจำลองเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning) ให้ประสิทธิภาพสูงสุด โดยมีค่ามาตราวัดเอฟ 0.953 และค่าพื้นที่ใต้กราฟเส้นโค้งลักษณะการทำงานของตัวจำแนกประเภท 0.990 ในขณะที่แบบจำลองวิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning) และแบบจำลองเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning) ให้ผลลัพธ์รองลงมา เมื่อเทียบกับแบบจำลองพื้นฐาน เช่น แบบจำลองบูสต์ติงไล่ระดับ (Gradient Boosting) ซึ่งมีประสิทธิภาพต่ำกว่า นอกจากนี้ การใช้กระบวนการคัดเลือกตัวแปรแบบเป็นขั้นตอน (Stepwise Feature Selection) ช่วยลดจำนวนตัวแปรโดยไม่กระทบต่อประสิทธิภาพของแบบจำลอง เทคนิคการรวมแบบจำลอง ร่วมกับการปรับสมดุลข้อมูลและการเลือกตัวแปรที่เหมาะสม สามารถเพิ่มประสิทธิภาพในการประเมินความเสี่ยงสินเชื่อของวิสาหกิจขนาดกลางและขนาดย่อมได้อย่างมีนัยสำคัญ โดยการปรับสมดุลข้อมูลด้วยวิธี SMOTE + ENN ให้ผลลัพธ์ดีที่สุด

(มีจำนวนทั้งสิ้น 67 หน้า)

คำสำคัญ : แบบจำลองความเสี่ยงด้านเครดิต สถานการณ์ผ่อนชำระ ความน่าเชื่อถือทางเครดิต เทคนิคการรวมแบบจำลอง วิธีต้นไม้ตัดสินใจ, วิธีซัพพอร์ตเวกเตอร์แมชชีน, วิธีบูสต์ติงไล่ระดับ, วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด, วิธีการจำแนกแบบเบส์นาอีฟ

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Name : Mr. Sornthun Intharasompong
Thesis Title : Development of a Machine Learning Model Using Stacking Techniques with Data Balancing Strategies for SME Credit Risk Assessment
Major Field : Applied Statistics and Data Analytics
King Mongkut's University of Technology North Bangkok
Thesis Advisor : Associate Professor Wikanda Phaphan, Ph.D.
Academic Year : 2025

ABSTRACT

Small and Medium Enterprises (SMEs) often face challenges in obtaining credit due to their perceived high credit risk. This study aims to develop a deep learning model using the stacking ensemble technique to enhance the accuracy of credit risk assessment for SMEs. The research utilizes a dataset from the Ministry of Industry, consisting of 14 quantitative and qualitative variables. Due to data imbalance, four data balancing techniques are applied: Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), a combination of SMOTE and Edited Nearest Neighbors (SMOTE + ENN), and a combination of SMOTE and Tomek Links (SMOTE + Tomek). The study compares the performance of nine machine learning models: Decision Tree, Support Vector Machines (SVM), Gradient Boosting, K-Nearest Neighbors (KNN), Naïve Bayes, Logistic Regression with Meta-Learning, Gradient Boosting with Meta-Learning, Extreme Gradient Boosting with Meta-Learning, and Multi-layer Perceptron Neural Network with Meta-Learning. Model performance is evaluated based on Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUCs). Results indicate that the stacking ensemble technique, particularly the Multi-layer Perceptron Neural Network with Meta-Learning, achieves the highest performance, with an F1-score of 0.953 and an AUCs of 0.990. Logistic Regression with Meta-Learning and Gradient Boosting with Meta-Learning also yield strong results, outperforming baseline models such as standard Gradient Boosting. Furthermore, applying Stepwise Feature Selection reduces the number of variables without compromising model performance. Overall, the

combination of stacking ensemble models, data balancing techniques, and optimal feature selection significantly enhances the accuracy of credit risk assessment for SMEs. Through data balancing using the SMOTE + ENN technique, which yields superior performance compared to other methods.

(Total 67 Pages)

Keywords: Credit Risk Model, Loan Repayment Status, Creditworthiness, Stacking Ensemble Technique, Decision Tree, Support Vector Machines, Gradient Boosting, K-Nearest Neighbors, Naïve Bayes

_____ Advisor

กิตติกรรมประกาศ

การศึกษาค้นคว้าครั้งนี้ สามารถสำเร็จลุล่วงได้ด้วยความอนุเคราะห์และความช่วยเหลือจาก รองศาสตราจารย์ ดร.วิกานดา ผาพันธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่กรุณาช่วยเหลือให้คำปรึกษาและแนะนำตลอดจนการตรวจทานและแก้ไขข้อบกพร่องต่างๆ เป็นอย่างดีมาโดยตลอด จนกระทั่งโครงงานพิเศษฉบับนี้เสร็จสมบูรณ์

ผู้วิจัยขอขอบพระคุณคณาจารย์ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือทุกท่านที่ประสิทธิ์ประสาทวิชาความรู้แก่คณะผู้วิจัยให้ได้มีความรู้ ประสบการณ์ทางด้านสถิติอันเป็นประโยชน์ต่อผู้วิจัยจนสำเร็จการศึกษา

ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา ผู้สนับสนุนทุนทรัพย์ส่งเสริมการศึกษาที่ให้แก่ผู้วิจัยตลอดจนสำเร็จการศึกษา และขอขอบคุณรุ่นพี่ เพื่อน และรุ่นน้อง สาขาวิชาสถิติประยุกต์ที่คอยให้คำปรึกษา และความช่วยเหลือแก่ผู้วิจัยเสมอมา กระทั่งวิทยานิพนธ์สำเร็จลุล่วงไปได้ด้วยดี

สรธัญ อินทรสมพงค์

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1

บทนำ

1.1. ที่มาและความสำคัญ

วิสาหกิจขนาดกลางและขนาดย่อม (SMEs) เป็นกลไกสำคัญในการขับเคลื่อนเศรษฐกิจโลก โดยมีบทบาทในการสร้างงานและกระตุ้นการเติบโตทางเศรษฐกิจ อย่างไรก็ตาม SMEs มักเผชิญอุปสรรคในการเข้าถึงสินเชื่อจากสถาบันการเงิน เนื่องจากถูกมองว่ามีความเสี่ยงด้านเครดิตสูง โดยเฉพาะในช่วงภาวะเศรษฐกิจถดถอยหรือการแพร่ระบาดของโควิด-19 ซึ่งส่งผลให้รายได้ลดลงขณะที่ต้นทุนดำเนินงานยังคงที่ ความท้าทายเหล่านี้เพิ่มความเสี่ยงต่อการปิดกิจการและทำให้ SMEs ต้องพึ่งพาแหล่งเงินทุนภายนอกมากขึ้น เพื่อบริหารความเสี่ยงของหนี้ที่ไม่ก่อให้เกิดรายได้ (NPLs) สถาบันการเงินจึงปรับมาตรฐานการปล่อยสินเชื่อให้เข้มงวดขึ้น สิ่งนี้เน้นย้ำถึงความจำเป็นในการพัฒนาแนวทางการประเมินความเสี่ยงด้านเครดิต (Credit Risk Assessment) ที่แม่นยำและเป็นกลางมากขึ้น เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) จึงเป็นทางเลือกที่มีศักยภาพในการวิเคราะห์ข้อมูลผู้กู้และช่วยให้สถาบันการเงินสามารถประเมินความสามารถในการชำระหนี้ของ SMEs ได้อย่างมีประสิทธิภาพและยุติธรรมมากขึ้น นำไปสู่การสร้างสมดุลระหว่างการบริหารความเสี่ยงและการส่งเสริมการเข้าถึงสินเชื่อของธุรกิจที่มีศักยภาพ ในการวิจัยครั้งนี้ ผู้วิจัยสนใจศึกษาการพัฒนาแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking เพื่อวิเคราะห์ปัจจัยสำคัญที่เกี่ยวข้องกับความเสี่ยงด้านเครดิต 14 ปัจจัย ซึ่งเป็นข้อมูลทุติยภูมิจากกระทรวงอุตสาหกรรม ได้แก่ อายุของผู้ขอสินเชื่อ (Age), จำนวนสมาชิกในครอบครัวหรือธุรกิจที่เกี่ยวข้องกับการขอสินเชื่อ (Member), หนี้สินที่ได้รับเข้ามาของธุรกิจ (IN_debt), หนี้สินที่ชำระออกไปของธุรกิจ (OUT_debt), จำนวนพนักงานของธุรกิจ (Num_Emp), ค่าใช้จ่ายรายปี (Expenses_Y), อัตราส่วนของลูกค้ารายใหม่ (ratio_CCus-TCus), อัตราการเติบโตของธุรกิจในช่วง 3 ปี (Growth_rate_3Y), รายได้สุทธิของปีที่ผ่านมา (Net-income_LY), หนี้สินที่ได้รับเข้ามาโดยเฉพาะจากธุรกิจ (IN_debt_Bus), หนี้สินที่ชำระออกไปโดยเฉพาะจากธุรกิจ (OUT_debt_Bus), จำนวนเงินกู้ที่ขอจากสถาบันการเงิน (Loan_Amount), ระยะเวลาการกู้ยืม (period), ตัวแปรเป้าหมาย (Target 0 = ไม่เป็น NPL หรือ ไม่ผิดนัดชำระ, 1 = เป็น NPL หรือ ผิด นัดชำระ) และใช้กลยุทธ์ปรับสมดุลข้อมูล 4 วิธี ได้แก่ SMOTE, ADASYN,

SMOTE + ENN และ SMOTE + Tomek เพื่อปรับสมดุลของชุดข้อมูลให้มีข้อมูลสถานะหนี้เสียและสถานะปรกติเท่ากันซึ่งจะช่วยเพิ่มประสิทธิภาพ ในการจำแนกความเสี่ยงด้านเครดิต พร้อมทั้งเปรียบเทียบประสิทธิภาพตัวแบบจำนวน 9 ตัวแบบ ได้แก่ วิธีต้นไม้ตัดสินใจ (Decision Tree Model) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) วิธีบูสต์ติงไล่ระดับ (Gradient Boosting Model) วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbours Model) วิธีการจำแนกแบบเบส์นาอีฟ (Naïve Bayesian Model) วิธีเมตาการถดถอยลอจิสติก (META in Logistic Regression) วิธีเมตาบูสต์ติงไล่ระดับ (META in Gradient Boosting Model) วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (META in Extreme Gradient Boosting Model) วิธีเมตาโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้น (META in Multi-layer Perceptron Neural Network) โดยใช้ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความจำ (Recall) ค่ามาตราวัดเอฟ (F-measure) และ ตัววัดประสิทธิภาพโดยใช้พื้นที่ใต้กราฟ (Area Under an ROC Curve : AUCs) เป็นเกณฑ์ในการวัดประสิทธิภาพ และใช้ Google Colab ในการประมวลผล

1.2. วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อพัฒนาแบบจำลองการเรียนรู้เชิงลึกด้วยเทคนิค Stacking สำหรับการประเมินความเสี่ยงสินเชื่อ SME

1.2.2 เพื่อศึกษาผลกระทบของกลยุทธ์การปรับสมดุลข้อมูลที่แตกต่างกันต่อประสิทธิภาพของแบบจำลอง

1.2.3 เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้เทคนิค Stacking กับแบบจำลองพื้นฐานอื่นๆ

1.3. ขอบเขตการวิจัย

1.3.1. ด้านข้อมูล

การวิจัยนี้มุ่งเน้นการวิเคราะห์ฐานข้อมูลผู้กู้ของกองทุนหมุนเวียน โดยใช้ข้อมูลจากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทย ซึ่งอยู่ภายใต้กรมส่งเสริมอุตสาหกรรม กระทรวงอุตสาหกรรม ประเทศไทย การศึกษาจะเน้นไปที่ผู้กู้รายบุคคล ซึ่งเป็น

กลุ่มลูกค้าหลักของกองทุน และเป็นกลุ่มที่มีสัดส่วนสูงในหนี้ที่ไม่ก่อให้เกิดรายได้ (Non-Performing Loans: NPLs) ขอบเขตของการวิเคราะห์จะครอบคลุมโครงสร้างของฐานข้อมูลผู้กู้ ปัจจัยที่อาจส่งผลต่อสถานะหนี้ และแนวโน้มของหนี้ที่ไม่ก่อให้เกิดรายได้ในกลุ่มผู้กู้รายบุคคล

1.3.2. ด้านตัวแบบและวิธีการที่ใช้

1.3.2.1. ตัวแบบการเรียนรู้ของเครื่องพื้นฐาน 5 วิธี

1.3.2.2.1. วิธีต้นไม้ตัดสินใจ (Decision Tree Model)

1.3.2.2.2. ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines)

1.3.2.2.3. วิธีบูสต์ติงไล่ระดับ (Gradient Boosting Model)

1.3.2.2.4. วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbours Model)

1.3.2.2.5. วิธีการจำแนกแบบเบส์นาอีฟ (Naïve Bayesian Model)

1.3.2.2. การปรับปรุงตัวแบบการเรียนรู้ของเครื่อง งานวิจัยนี้ศึกษาการปรับปรุงตัวแบบการเรียนรู้ของเครื่องทั้งหมด 4 วิธี คือ

1.3.2.2.1. วิธีเมตาการถดถอยลอจิสติก (META in Logistic Regression)

1.3.2.2.2. วิธีเมตาบูสต์ติงไล่ระดับ (META in Gradient Boosting Model)

1.3.2.2.3. วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (META in Extreme Gradient Boosting Model)

1.3.2.2.4. วิธีเมตาโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้น (META in Multi-layer Perceptron Neural Network)

1.3.3. การปรับสมดุลข้อมูลมีดังนี้

1.3.3.1. SMOTE (Synthetic Minority Over-sampling Technique)

1.3.3.2. ADASYN (Adaptive Synthetic Sampling Approach)

1.3.3.3. SMOTE + ENN (SMOTE + Edited Nearest Neighbors)

1.3.3.4. SMOTE + Tomek (SMOTE + Tomek Links)

1.3.4. ด้านการคำนวณ

งานวิจัยนี้มุ่งเน้นการพัฒนาตัวแบบโดยใช้ภาษา ไพทอน (Python) ซึ่งเป็นภาษาระดับสูงที่พัฒนาโดยคีโด ฟัน โรสซึม (Guido van Rossum) ตั้งแต่ปี พ.ศ. 2533 การออกแบบของไพทอน มุ่งเน้นให้โค้ดสามารถอ่านและทำความเข้าใจได้ง่าย โดยใช้ การเว้นวรรค (whitespace) เป็นส่วนหนึ่งของโครงสร้างภาษา ซึ่งช่วยให้โปรแกรมมีความเป็นระเบียบและสะดวกต่อการปรับปรุงแก้ไข ด้วยคุณลักษณะดังกล่าวไพทอน จึงได้รับความนิยมอย่างแพร่หลายในงานพัฒนาโปรแกรมและสามารถนำไปประยุกต์ใช้ในการพัฒนาต่อยอดในอนาคต

1.3.5. ข้อจำกัดของการวิจัย

งานวิจัยนี้อาศัยการรวบรวมข้อมูลทุติยภูมิ ซึ่งเป็นข้อมูลที่ถูกจัดเก็บและรวบรวมไว้จากกระทรวงอุตสาหกรรม โดยที่ผู้วิจัยไม่ได้ดำเนินการเก็บรวบรวมข้อมูลโดยตรง

1.4. คำนิยามศัพท์

1.4.1. วิสาหกิจขนาดกลางและขนาดย่อม (Small and Medium Enterprises - SMEs) หมายถึง ธุรกิจที่มีขนาดกลางหรือเล็กที่มีรายได้ สินทรัพย์และพนักงานจำนวนน้อย ดำเนินธุรกิจโดยผู้ประกอบการรายย่อย ไม่ขึ้นอยู่กับกลุ่มธุรกิจใด ใช้เงินลงทุนต่ำ เป็นเงินทุนของเจ้าของธุรกิจเองหรือเงินทุนจากการกู้ยืมสินเชื่อจากธนาคาร

1.4.2. ความเสี่ยงด้านเครดิต (Credit Risk) หมายถึง ความเสี่ยงที่ลูกหนี้ไม่สามารถชำระหนี้คืนได้ตามเงื่อนไขที่กำหนด

1.4.3. สินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-Performing Loan: NPL) หมายถึง สินเชื่อที่ลูกหนี้ผิดนัดชำระหรือมีการค้างชำระเกินระยะเวลาที่กำหนด (เช่น 90 วันขึ้นไป)

1.4.4. หนี้สินที่ได้รับเข้ามาของธุรกิจ (IN_debt) หมายถึง จำนวนหนี้สินที่ธุรกิจได้รับเข้ามาเพื่อใช้ในการดำเนินงานหรือขยายกิจการ

1.4.5. หนี้สินที่ชำระออกไปของธุรกิจ (OUT_debt) หมายถึง จำนวนหนี้สินที่ธุรกิจได้ชำระออกไปเพื่อลดภาระหนี้สิน

1.4.6. จำนวนพนักงานของธุรกิจ (Num_Emp) หมายถึง จำนวนพนักงานทั้งหมดที่ทำงานในธุรกิจ

1.4.7. ค่าใช้จ่ายรายปี (Expenses_Y) หมายถึง ค่าใช้จ่ายรวมของธุรกิจในรอบปีที่ผ่านมา

1.4.8. อัตราส่วนของลูกค้ารายใหม่ (ratio_CCus-TCus) หมายถึง อัตราส่วนของลูกค้าใหม่ที่เข้ามาเทียบกับจำนวนลูกค้าทั้งหมด

1.4.9. อัตราการเติบโตของธุรกิจในช่วง 3 ปี (Growth_rate_3Y) หมายถึง อัตราการเติบโตของธุรกิจในช่วงเวลา 3 ปีที่ผ่านมา

1.4.10. รายได้สุทธิของปีที่ผ่านมา (Net-income_LY) หมายถึง รายได้สุทธิของธุรกิจในปีที่ผ่านมาหลังหักค่าใช้จ่ายทั้งหมด

1.4.11. หนี้สินที่ได้รับเข้ามาโดยเฉพาะจากธุรกิจ (IN_debt_Bus) หมายถึง จำนวนหนี้สินที่ธุรกิจได้รับเข้ามาโดยเฉพาะสำหรับการดำเนินธุรกิจ

1.4.12. หนี้สินที่ชำระออกไปโดยเฉพาะจากธุรกิจ (OUT_debt_Bus) หมายถึง จำนวนหนี้สินที่ธุรกิจได้ชำระออกไปโดยเฉพาะสำหรับภาระหนี้ของธุรกิจ

1.4.13. จำนวนเงินกู้ (Loan Amount) หมายถึง จำนวนเงินที่ผู้กู้ได้รับจากสถาบันการเงิน

1.4.14. ระยะเวลาการกู้ยืม (Period) หมายถึง ช่วงเวลาที่ผู้กู้ต้องชำระคืนเงินกู้

1.4.15. การเรียนรู้ของเครื่อง (Machine Learning) หมายถึง กระบวนการที่คอมพิวเตอร์เรียนรู้จากข้อมูลเพื่อสร้างแบบจำลองและคาดการณ์ผลลัพธ์

1.4.16. แบบจำลองการจำแนกประเภท (Classification Model) หมายถึง ตัวแบบที่ใช้ในการจำแนกข้อมูลออกเป็นกลุ่ม เช่น จำแนกผู้กู้เป็น เสี่ยงต่ำ หรือเสี่ยงสูง

1.4.17. การเรียนรู้แบบ Stacking (Stacking Learning) หมายถึง เทคนิคที่ใช้โมเดลหลายตัวมาผสมกันเป็น Meta-Model เพื่อให้ได้ผลลัพธ์ที่แม่นยำขึ้น

1.4.18. เทคนิคการปรับสมดุลข้อมูล (Data Balancing Techniques) หมายถึง วิธีการทำให้ข้อมูลกลุ่มที่มีจำนวนน้อยมีสัดส่วนใกล้เคียงกับกลุ่มที่มีจำนวนมากขึ้น

1.4.19. Synthetic Minority Over-sampling Technique (SMOTE) หมายถึง เทคนิคการเพิ่มข้อมูลของกลุ่มที่มีจำนวนตัวอย่างน้อยโดยการสร้างตัวอย่างใหม่แบบสังเคราะห์

1.4.20. Adaptive Synthetic Sampling (ADASYN) หมายถึง เทคนิคปรับปรุง SMOTE โดยเน้นการเพิ่มตัวอย่างที่อยู่ใกล้ขอบเขตการจำแนก ซึ่งช่วยให้แบบจำลองเรียนรู้ข้อมูลที่มีความสำคัญมากขึ้น และลดปัญหาการแบ่งประเภทที่ผิดพลาดในข้อมูลที่ไม่สมดุล

1.4.21. SMOTE + ENN (SMOTE + Edited Nearest Neighbors) หรือ SMOTEENN หมายถึง เทคนิคการผสมผสานระหว่าง SMOTE และ Edited Nearest Neighbors (ENN) ซึ่งนอกจากจะเพิ่มตัวอย่างข้อมูลกลุ่มที่มีจำนวนน้อยแล้ว ยังช่วยลบตัวอย่างที่มีความผิดปกติออกไป

1.4.22. SMOTE + Tomek (SMOTE + Tomek Links) หรือ SMOTETomek หมายถึง เทคนิคการผสมผสาน SMOTE กับ Tomek Links ซึ่งช่วยลดปัญหาการซ้อนทับกันของข้อมูลระหว่างคลาส โดยการลบตัวอย่างที่อยู่ใกล้เส้นแบ่งของกลุ่มข้อมูลออกไป

1.4.23. การแบ่งชุดข้อมูลแบบ Cross-Validation หมายถึง เทคนิคที่ใช้แบ่งข้อมูลออกเป็นหลายชุด

1.4.24. Akaike Information Criterion (AIC) หมายถึง ตัวชี้วัดที่ใช้เลือกตัวแบบที่ดีที่สุดโดยพิจารณาจากค่าความคลาดเคลื่อนและความซับซ้อนของตัวแบบ

1.4.25. Multicollinearity หมายถึง ปัญหาที่เกิดขึ้นเมื่อตัวแปรอิสระในตัวแบบมีความสัมพันธ์กันมากเกินไป

1.4.26. Overfitting หมายถึง สถานการณ์ที่ตัวแบบเรียนรู้จากข้อมูลฝึกซ้อมมากเกินไปจนไม่สามารถทำนายข้อมูลใหม่ได้

1.4.27. Underfitting หมายถึง ตัวแบบที่มีความสามารถในการเรียนรู้ต่ำเกินไป ทำให้ไม่สามารถจับรูปแบบของข้อมูลได้

1.5. ประโยชน์ของงานวิจัย

1.5.1. พัฒนาแนวทางในการประเมินความเสี่ยงสินเชื่อ SME ที่มีประสิทธิภาพสูงขึ้น

1.5.2. นำเสนอแนวทางการใช้ เทคนิคปรับสมดุลข้อมูล เพื่อเพิ่มความแม่นยำของแบบจำลองการเรียนรู้

1.5.3. ช่วยให้สถาบันการเงินสามารถใช้ แบบจำลองที่มีประสิทธิภาพสูงขึ้น เพื่อลดความเสี่ยงในการให้สินเชื่อ

1.5.4. สนับสนุนการตัดสินใจด้านสินเชื่อโดยใช้ Machine Learning และ Data Science

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการพัฒนาแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking เพื่อวิเคราะห์ปัจจัยสำคัญที่เกี่ยวข้องกับความเสี่ยงด้านเครดิต 14 ปัจจัย ซึ่งเป็นข้อมูลทุติยภูมิจากกระทรวงอุตสาหกรรม และใช้กลยุทธ์ปรับสมดุลข้อมูล 4 วิธี พร้อมทั้งเปรียบเทียบประสิทธิภาพตัวแบบหรือวิธีเชิงทำนาย 9 ตัวแบบ ได้แก่ วิธีต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน วิธีบูสต์ติงไล่ระดับ วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด วิธีการจำแนกแบบเบส์นาอีฟ วิธีเมตาการถดถอยลอจิสติก วิธีเมตาบูสต์ติงไล่ระดับ วิธีเมตาบูสต์ติงไล่ระดับสุดขีด วิธีเมตาโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้น ซึ่งในบทนี้จะกล่าวถึงเอกสารและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

2.1. แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.2. งานวิจัยที่เกี่ยวข้อง

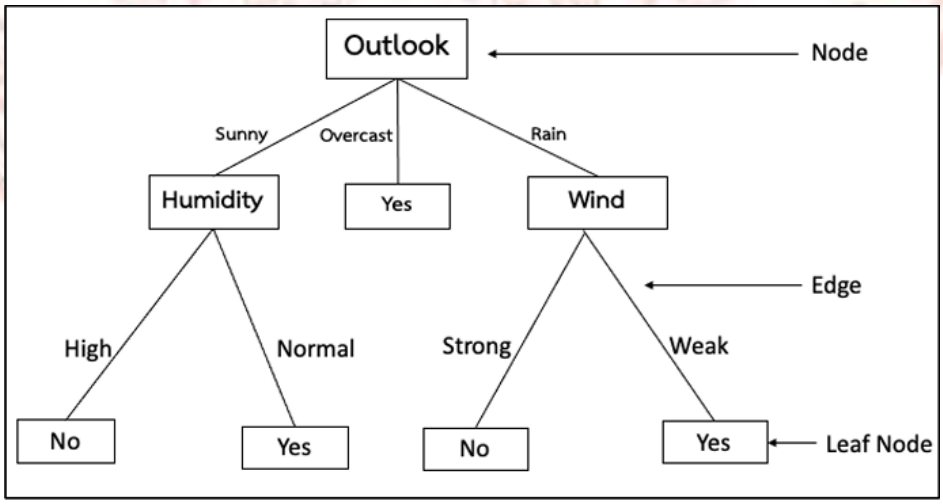
2.1. แนวคิดและทฤษฎีที่เกี่ยวข้อง

ในการดำเนินการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาและประยุกต์ใช้แนวคิดและทฤษฎีที่เกี่ยวข้องกับการพัฒนาแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ทั้งในระดับพื้นฐานและระดับที่ซับซ้อนมากขึ้น เพื่อประเมินความเสี่ยงด้านเครดิตของวิสาหกิจขนาดกลางและขนาดย่อม (SMEs) โดยเน้นการใช้เทคนิคการรวมโมเดลแบบ Stacking ร่วมกับวิธีการปรับสมดุลข้อมูล (Data Balancing) เพื่อเพิ่มประสิทธิภาพในการทำนายและความแม่นยำของตัวแบบ ดังนั้น ในส่วนนี้จึงได้นำเสนอทฤษฎีและหลักการที่เกี่ยวข้องกับเทคนิคที่ใช้ในการวิจัย ซึ่งประกอบด้วยรายละเอียดของโมเดลการเรียนรู้ของเครื่องแต่ละประเภท เทคนิคการรวมโมเดล และวิธีการปรับสมดุลข้อมูล ดังต่อไปนี้

2.1.1. วิธีต้นไม้ตัดสินใจ (Decision Tree Model)

วิธีต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้ในการแบ่งชั้นข้อมูลหรือจัดกลุ่มตัวแปรอิสระให้อยู่ในอาณาบริเวณต่าง ๆ อย่างเป็นระบบ เพื่อนำมาใช้ในการทำนายค่าของข้อมูลใหม่ โดยอาศัยค่าเฉลี่ยหรือค่าฐานนิยมของข้อมูลที่อยู่ในแต่ละบริเวณนั้นเป็นเกณฑ์ในการคาดการณ์ ซึ่งโครงสร้างของต้นไม้ตัดสินใจสามารถเขียนอยู่ในรูปแบบ “ถ้า แล้ว ผลลัพธ์” ที่ช่วยอธิบายการตัดสินใจของแบบจำลองได้อย่างชัดเจน (James, G., Witten, D., Hastie, T., & Tibshirani, R., 2021)

การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ (Classification Tree) เป็นวิธีทางคณิตศาสตร์ที่ใช้ในการเลือกแนวทางที่เหมาะสมที่สุดในการจำแนกข้อมูล โดยอาศัยกระบวนการเรียนรู้แบบมีผู้สอนซึ่งแบบจำลองจะถูกฝึกด้วยชุดข้อมูลที่มีตัวแปรตาม และตัวแปรอิสระ จากนั้นต้นไม้ตัดสินใจจะสร้างกฎเกณฑ์จากข้อมูลชุดฝึกเพื่อนำไปใช้ทำนายข้อมูลใหม่ โครงสร้างต้นไม้ได้แสดงให้เห็นดังภาพที่ 2-1 โดยมี 3 องค์ประกอบ ได้แก่ 1) จุดต่อ (Node) ใช้แสดงถึงคุณสมบัติหรือตัวแปรอิสระที่ใช้ในการแบ่งข้อมูล 2) เส้นเชื่อมโหนด (Edge) ใช้เชื่อมกับโหนดชั้นถัดไปจากโหนดรากซึ่งเป็นโหนดที่ใช้ระบุค่าของคุณลักษณะของโหนดลำดับชั้นก่อนหน้า 3) ใบ (Leaf) ใบแต่ละใบแทนกลุ่มหรือหมวดหมู่ของตัวแปรตาม



ภาพที่ 2-1 โครงสร้างต้นไม้ตัดสินใจ (Decision Tree)

2.1.2. วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines)

วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) เป็นเทคนิคในการเรียนรู้ของเครื่องแบบมีผู้สอน ซึ่งมีรากฐานมาจากทฤษฎีทางคณิตศาสตร์และการเพิ่มประสิทธิภาพ (Optimization Theory) จุดประสงค์หลักของ SVM คือการหาขอบเขตการจำแนกที่เหมาะสมที่สุด (Optimal Separating Hyperplane) ซึ่งเป็นเส้นหรือระนาบในเชิงเรขาคณิตที่สามารถแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ได้โดยมีระยะขอบ (Margin) ระหว่างกลุ่มข้อมูลมากที่สุด กระบวนการของ SVM เริ่มจากการพิจารณาคลาสข้อมูลที่สามารถแยกได้อย่างเป็นเส้นตรง (Linearly Separable Data) โดยการหาค่าพารามิเตอร์ของไฮเปอร์เพลนที่สามารถเพิ่มระยะห่างจากจุดข้อมูลที่อยู่ใกล้ที่สุดในแต่ละคลาส ซึ่งจุดเหล่านี้เรียกว่า Support Vectors สมการของไฮเปอร์เพลนสามารถกำหนดได้เป็น

$$wx + b = 0$$

โดยที่ w แทน เวกเตอร์ของน้ำหนัก

b แทน คือค่าคงที่ (bias)

จุดมุ่งหมายหลักของซัพพอร์ตเวกเตอร์แมชชีน คือการสร้างแบบจำแนก (classifier) ที่มีความสามารถในการแยกข้อมูลออกเป็นคลาสต่าง ๆ ได้อย่างมีประสิทธิภาพสูงสุด โดยเฉพาะในกรณีที่ข้อมูลมีมิติสูง หรือจุดข้อมูลมีลักษณะซับซ้อนในกรณีของข้อมูลสองคลาสที่สามารถแยกกันได้เชิงเส้น (linearly separable), ซัพพอร์ตเวกเตอร์แมชชีนจะค้นหา ไฮเปอร์เพลนที่สามารถแยกข้อมูลทั้งสองคลาสออกจากกัน โดยให้มีระยะห่าง (margin) ระหว่างขอบเขตกับข้อมูลที่อยู่ใกล้ที่สุดมากที่สุด จุดข้อมูลที่อยู่ชิด margin เรียกว่า support vectors และเป็นตัวกำหนดตำแหน่งของไฮเปอร์เพลนโดยตรง

2.1.3. วิธีบูสต์ติงไล่ระดับ (Gradient Boosting Model)

ตัวแบบบูสต์ติงไล่ระดับ เป็นตัวแบบที่พัฒนาจาก ตัวแบบบูสต์ติงปรับได้ (Adaptive Boosting Model) เพื่อลดความคลาดเคลื่อนของการทำนาย และเพิ่มความแม่นยำ

การหาค่าการทำนายของตัวแบบบูสต์ติงไล่ระดับ ประกอบด้วย ผลลัพธ์ หรือ ตัวแปรตอบสนอง แทนด้วย y และ เซตของตัวแปรต้น หรือ ตัวแปรอธิบาย แทนด้วย $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ โดยใช้ training set $\{y_i, x_i\}_1^N$ โดยที่ทราบค่า (y,x)

ตัวแบบบูสต์ติงไล่ระดับ ประกอบด้วย 3 ส่วน คือ

1. ฟังก์ชันการสูญเสีย (Loss function) ใช้เพื่อปรับตัวแบบให้เหมาะสม
2. Weak learner ใช้เพื่อสร้างการทำนาย
3. Additive model ใช้เพิ่ม Weak learner เพื่อฟังก์ชันการสูญเสียต่ำสุด

2.1.4. วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbours Model)

วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด (KNN) เป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่องที่ใช้สำหรับงานจำแนกประเภทและการถดถอย โดยอาศัยหลักการเปรียบเทียบความคล้ายคลึงของข้อมูลใหม่กับข้อมูลตัวอย่างที่มีอยู่ มีกระบวนการดังนี้

1. กำหนดค่า k (มักเลือกให้เป็นจำนวนคี่)
2. คำนวณระยะห่างระหว่างข้อมูลใหม่กับข้อมูลตัวอย่าง
3. เรียงลำดับข้อมูลตัวอย่างตามระยะห่างที่คำนวณได้
4. เลือกข้อมูลที่ใกล้ที่สุดจำนวน k ตัว
5. กำหนดกลุ่มให้กับข้อมูลใหม่ ใช้หลักเสียงข้างมากจาก k ตัวอย่างที่เลือกมา

2.1.5. วิธีการจำแนกแบบเบส์นาอีฟ (Naïve Bayesian Model)

การจำแนกประเภทด้วยวิธีเบส์นาอีฟเป็นวิธีการเรียนรู้ของเครื่องที่อาศัยหลักการของความน่าจะเป็น โดยมีพื้นฐานมาจาก ทฤษฎีของเบส์ (Bayes' Theorem) จุดมุ่งหมายหลักของวิธีนี้คือการสร้างตัวแบบเชิงทำนายที่อยู่ในรูปของความน่าจะเป็น ซึ่งสามารถคำนวณได้จากข้อมูลที่สังเกตพบ เมื่อได้รับข้อมูลใหม่ วิธีนี้จะใช้ความน่าจะเป็นของข้อมูลที่มีอยู่ร่วมกับความรู้ก่อนหน้า (Prior Knowledge) เพื่อปรับปรุงการคาดการณ์ให้มีความแม่นยำมากขึ้น กระบวนการนี้ช่วยให้สามารถปรับปรุงสมมติฐานเกี่ยวกับข้อมูลที่ต้องการจำแนกได้อย่างมีประสิทธิภาพ เมื่อกล่าวถึงความน่าจะ

เป็นของเหตุการณ์ที่เกิดขึ้น (A) ถ้ามีเหตุการณ์อีกเหตุการณ์หนึ่งเกิดมาแล้ว (B) สามารถเขียนให้อยู่ในรูปความสัมพันธ์ ดังสมการที่ 2-1

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2\text{-}1)$$

โดยที่ $P(A|B)$ แทน ความน่าจะเป็นที่เหตุการณ์ A จะเกิดขึ้นเมื่อเหตุการณ์ B เกิดขึ้นแล้ว

$P(B|A)$ แทน ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้นเมื่อเหตุการณ์ A เกิดขึ้นแล้ว

$P(A)$ แทน ความน่าจะเป็นที่จะเกิดเหตุการณ์ A

$P(B)$ แทน ความน่าจะเป็นที่จะเกิดเหตุการณ์ B

เมื่อใช้ทฤษฎีนี้กับปัญหาการจำแนกประเภท ให้กำหนดให้ c_j เป็นกลุ่มของข้อมูลที่ต้องการจำแนกที่มีคุณสมบัติ n ตัว และกำหนดให้ $x = \{x_1, x_2, \dots, x_n\}$ เมื่อพิจารณาข้อมูลชุดฝึกเพื่อต้องการจำแนกลักษณะที่มีคุณสมบัติ X จะมีค่าความน่าจะเป็นอยู่ในหมวดใดสามารถคำนวณได้จากสมการที่ 2-2

$$P(c_j|x) = \frac{P(x|c_j)P(c_j)}{P(x)} \quad (2\text{-}2)$$

กระบวนการจำแนกประเภทใช้หลักการเลือกค่าที่มีความน่าจะเป็นสูงสุดพิจารณาเพียงค่า $P(x|c_j)P(c_j)$ ดังนั้นซึ่งสามารถคำนวณได้ดังสมการที่ 2-3

$$C_{NB} = argmax(P(c_j) \cdot \prod_{i=1}^{n} P(x_i|c_j) \quad (2\text{-}3)$$

โดยที่ $P(x_i|c_j) = \frac{1}{\sqrt{2\pi\sigma_{c_j}^2}} exp\left(-\frac{(x_i-\mu_{C_j})^2}{2\sigma_{c_j}^2}\right)$

2.1.6. วิธีเมตาการถดถอยลอจิสติก (META in Logistic Regression)

การศึกษาครั้งนี้ใช้ในการตรวจจับแบบสองขั้นตอนโดยการนำผลการทำนายจากขั้นตอนที่ หนึ่งมี 5 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน วิธีบูสต์ติงไล่ระดับ วิธีการจำแนกเพื่อนบ้าน

ใกล้ที่สุด วิธีการจำแนกแบบเบส์นาอีฟใช้เป็นตัวแปรที่สนใจในขั้นตอนที่สองโดยใช้การถดถอยลอจิสติกในการหาค่าการทำนายใหม่

2.1.7. วิธีเมตาบูสต์ติงไล่ระดับ (META in Gradient Boosting Model)

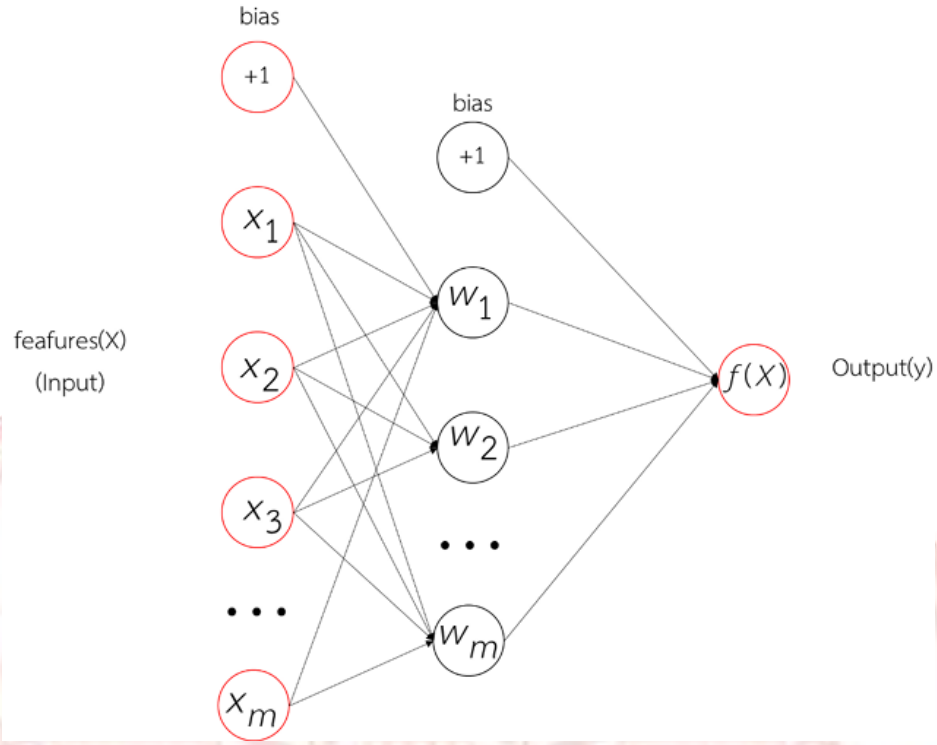
การศึกษาครั้งนี้ใช้ในการตรวจจับแบบสองขั้นตอนโดยการนำผลการทำนายจากขั้นตอนที่ หนึ่งมี 5 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน วิธีบูสต์ติงไล่ระดับ วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด วิธีการจำแนกแบบเบส์นาอีฟใช้เป็นตัวแปรที่สนใจในขั้นตอนที่สองโดยใช้วิธีบูสต์ติงไล่ระดับในการหาค่าการทำนายใหม่

2.1.8. วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (META in Extreme Gradient Boosting Model)

การศึกษาครั้งนี้ใช้ในการตรวจจับแบบสองขั้นตอนโดยการนำผลการทำนายจากขั้นตอนที่ หนึ่งมี 5 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน วิธีบูสต์ติงไล่ระดับ วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด วิธีการจำแนกแบบเบส์นาอีฟใช้เป็นตัวแปรที่สนใจในขั้นตอนที่สองโดยใช้วิธีบูสต์ติงไล่ระดับสุดขีดในการหาค่าการทำนายใหม่

2.1.9. วิธีเมตาโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้น (META in Multi-layer Perceptron Neural Network)

วิธีเมตาโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้น เป็นวิธีการเรียนรู้แบบมี ผู้สอน (supervised learning) โดยมีฟังก์ชันสำหรับชุดฝึกสอนคือ $f(\cdot) = R^m \rightarrow R^0$ เมื่อ m คือจำนวนมิติของตัวแปรอธิบายและ o คือจำนวนมิติของตัวแปรสอบสนอง กำหนดตัวแปรสอบสนอง y และตัวแปรอธิบาย $X = x_1, x_2 \ldots, x_m$ สามารถใช้กับข้อมูลที่ไม่เป็นเชิงเส้น (non-linear) ถ้ามีตัวแปรสอบสนองและตัวแปรอธิบายที่ไม่เป็นเชิงเส้นจะเรียกว่า ชั้นที่ซ่อน (hidden layers)



ภาพที่ 2-2 โครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้นที่มี 1 ชั้นที่ซ่อน

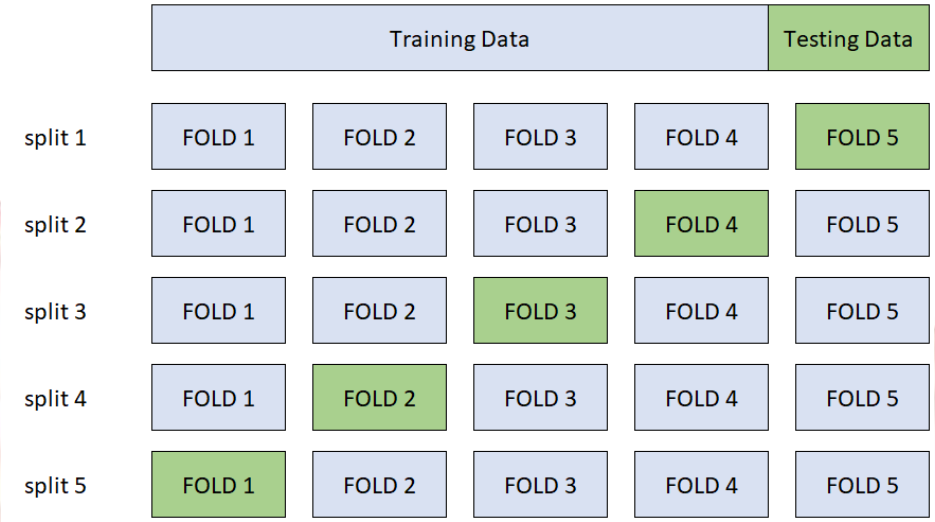
เซลล์ประสาทแต่ละเซลล์ในชั้นที่ซ่อนจะแปลงค่าจากชั้นก่อนหน้าด้วยผลรวมเชิงเส้นแบบถ่วงน้ำหนัก $w_1x_1 + w_2x_2 + \cdots, w_mx_m$ ตามด้วยฟังก์ชัน $g(\cdot) = R \to R -$ เช่นเดียวกับฟังก์ชันแทนไฮเปอร์โบลิก (Hyperbolic an function) ตัวแปรตอบสนองรับค่าจากชั้นที่ซ่อนอยู่สุดท้ายและแปลงเป็นตัวแปรอธิบาย

การศึกษาครั้งนี้ใช้ในการตรวจจับแบบสองขั้นตอนโดยการนำผลการทำนายจากขั้นตอนที่หนึ่ง มี 5 วิธี ได้แก่ วิธีต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน วิธีบูสต์ติงไล่ระดับ วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด วิธีการจำแนกแบบเบส์นาอีฟ ใช้เป็นตัวแปรที่สนใจในขั้นตอนที่สองโดยใช้วิธีโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายชั้นในการหาค่าการทำนายใหม่

2.1.10. วิธีการตรวจสอบไขว้ (Cross-Validation Test)

วิธีการตรวจสอบไขว้ (Cross-Validation Test) ใช้วัดประสิทธิภาพของตัวแบบหรือวิธีเชิงทำนายที่สร้างจากชุดข้อมูลฝึกสอนโดยจะแบ่งข้อมูลทั้งหมดออกเป็นหลายส่วน ซึ่งเรียกว่า K-Fold Cross Validation โดยจะสุ่มแบ่งข้อมูลทั้งหมดออกเป็น k ส่วน ส่วนละเท่า ๆ กัน โดยจะนำข้อมูลที่แบ่งแล้ว 1 ส่วน ใช้เป็นข้อมูลชุดทดสอบ (Test Set) และใช้ข้อมูลที่เหลืออีก k-1 เป็นข้อมูลชุดฝึกสอน

(Training Set) ใช้วิเคราะห์ข้อมูล ในการวิเคราะห์ข้อมูลแต่ละครั้งจะใช้สัดส่วนระหว่างข้อมูลชุดฝึกสอนและข้อมูลชุดทดสอบ เท่ากับ k-1 ส่วน ต่อ 1 ส่วน ซึ่งจะทำการวิเคราะห์ข้อมูลทั้งหมด k รอบ โดยในแต่ละครั้งจะเปลี่ยนข้อมูลชุดทดสอบและข้อมูลชุดฝึกสอนไปเรื่อย ๆ จนครบทั้งหมด k รอบ



ภาพที่ 2-3 การแบ่งข้อมูลด้วยวิธีไขว้ 5-Fold Cross-Validation

2.1.11. การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Selection)

เป็นวิธีการคัดเลือกผสมผสานระหว่างวิธีการคัดเลือกตัวแปรอิสระทั้งแบบการเพิ่มตัวแปรและการลดตัวแปรเข้าด้วยกัน วิธีการถดถอยแบบขั้นตอนสามารถทำได้ ดังนี้

1. วิธีคัดเลือกตัวแปรโดยเพิ่มตัวแปร คือคัดเลือกตัวแปรอิสระ 1 ตัวเข้าไว้ในตัวแบบการถดถอย โดยเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม Y สูงสุด หรือที่ทำให้ตัวแบบการถดถอยมีค่า SSE ต่ำสุดไว้ในตัวแบบการถดถอยเป็นตัวแปรแรก จากนั้นคำนวณค่าสถิติ Partial F หรือค่าสถิติ t เพื่อทดสอบสมมุติฐานว่าตัวแปรอิสระมีความสำคัญต่อการทำนายค่าของ Y หรือไม่ โดยถ้า $F_0 \leq F_{IN} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| \leq t_{IN} = t_{\frac{\alpha}{2},(n-p)}$ หรือ $p - value \geq p_{IN}$ เมื่อ p_{IN} คือระดับนัยสำคัญสำหรับการคัดตัวแปรอิสระเข้า แสดงว่าตัวแปรอิสระนั้นไม่มีความสำคัญต่อการทำนายค่าของ Y จะสิ้นสุดการคัดเลือกตัวแปรอิสระเข้าไว้ในตัวแบบการถดถอย แต่ถ้า $F_0 > F_{IN}$ หรือ $|t_0| > t_{IN}$ หรือ $p - value < p_{IN}$ แสดงว่าตัวแปรอิสระนั้นมีความสำคัญต่อการทำนายค่าของ Y ซึ่งจะคัดเข้าสู่ตัวแบบการถดถอยเป็นตัวแรก

2. วิธี Backward elimination คัดตัวแปรอิสระนั้นออกจากตัวแบบการถดถอย หากเป็นไปได้ โดยจะสามารถคัดตัวแปรอิสระที่อยู่ในตัวแบบออกได้ หาก $F_0 < F_{OUT} = F_{\alpha,(1,n-2)}$

หรือ $|t_0| < t_{OUT} = t_{\frac{\alpha}{2},(n-p)}$ หรือ $p - value > p_{OUT}$ คือระดับนัยสำคัญสำหรับการคัดตัวแปรอิสระออก การดำเนินการคัดเลือกตัวแปรอิสระจะสิ้นสุดลงและสรุปว่าไม่มีตัวแปรอิสระใดเลยที่มีความสำคัญต่อการทำนายค่าของตัวแปรตาม Y แต่ถ้า $F_0 \geq F_{OUT} = F_{\alpha,(1,n-2)}$หรือ $|t_0| \geq t_{OUT} = t_{\frac{\alpha}{2},(n-p)}$ หรือ $p - value \leq p_{OUT}$ แสดงว่าตัวแปรอิสระนั้นมีความสำคัญต่อการทำนายจะไม่สามารถคัดออกจากตัวแบบการถดถอยได้

3. พิจารณาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระกับ Y โดยจะเลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนกับตัวแปรตามสูงที่สุด โดยถ้า $F_0 > F_{IN}$ หรือ $|t_0| > t_{IN}$ หรือ $p - value < p_{IN}$ แสดงว่าตัวแปรอิสระที่ถูกพิจารณาตัวต่อมามีความสำคัญต่อการทำนายค่าของ Y เมื่อมีตัวแปรอิสระก่อนหน้าอยู่ในตัวแบบการถดถอยก่อนแล้ว จะคัดตัวแปรอิสระที่ถูกพิจารณาเข้ามาอยู่ในตัวแบบเป็นตัวต่อมา

4. หลังจากนั้นจึงทำการทดสอบความมีนัยสำคัญของตัวแปรที่เข้ามาเป็นตัวสุดท้ายก่อน โดยพิจารณาจากค่าสถิติเอฟบางส่วน ถ้ามีค่าน้อยกว่าเกณฑ์ที่กำหนดไว้จะต้องนำตัวแปรนั้นออกจากตัวแบบ แต่ถ้าค่าสถิติเอฟบางส่วนมีค่ามากกว่าเกณฑ์ที่กำหนดไว้ ก็จะย้อนกลับไปทดสอบตัวแปรอิสระตัวก่อนหน้าที่เข้ามาอยู่ในตัวแบบ โดยพิจารณาจากค่าสถิติเอฟบางส่วน เช่นเดียวกัน ทำซ้ำในขั้นตอนที่สามและสี่ต่อไป จนกระทั่งไม่สามารถนำตัวแปรอิสระใดออกจากตัวแบบได้ หรือไม่สามารถนำตัวแปรอิสระใดเข้าสู่ตัวแบบได้อีก ข้อดีของวิธีการนี้ คือ สามารถแก้จุดบกพร่องของวิธีการเลือกตัวแปรโดยวิธีเพิ่มตัวแปร (Forward Selection) และวิธีการเลือกตัวแปรโดยวิธีลดตัวแปร (Backward Elimination) ได้ (Montgomery, Peck and Vining, 2012)

เกณฑ์การประเมินและตรวจสอบความเหมาะสมของตัวแบบ

1. ค่ารากที่สองของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE) โดยจะเป็นการวัดค่าความคลาดเคลื่อนระหว่างค่าจริงและค่าที่ประมาณจากแบบจำลองมีความแตกต่างกันมากน้อยเพียงใด ซึ่งหากค่า RMSE มีค่าเท่ากับ 0 จะหมายถึงแบบจำลองที่ประมาณได้มีค่าเท่ากันกับค่าจริงพอดีดังนั้นหากว่าค่า RMSE ยิ่งมีค่าน้อยแบบจำลองนั้นสามารถเป็นตัวแทนค่าจริงได้ดี ดังสมการที่ 2-3

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2\text{-}4)$$

โดยที่ y_i แทน ค่าความคลาดเคลื่อนระหว่างค่าจริง

$\hat{y}_i$ แทน ค่าที่ประมาณจากแบบจำลอง

2. ค่า R-Squared (R^2) หรือ ค่าสัมประสิทธิ์แสดงการตัดสินใจ คือการวัดค่าตัวแปรอิสระสามารถอธิบายตัวแปรตามได้ดีเพียงใด หากค่านี้เท่ากับ 1 ก็หมายความว่า ตัวแปรอิสระสามารถอธิบายตัวแปรตามได้ 100% ในทางกลับกัน หากค่านี้มีค่า เท่ากับ 0 แปลความหมายว่าตัวแปรอิสระไม่สามารถอธิบายตัวแปรตามได้ แต่อย่างไรก็ตาม พบว่าหากมีการเพิ่มตัวแปรอิสระเข้าไปในสมการมาก ก็จะทำให้ค่า R-Squared มากขึ้นด้วย ซึ่งนับเป็น ข้อจำกัดของค่าสถิตนี้ ดังสมการที่ 2-5

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2\text{-}5)$$

โดยที่ $\bar{y}$ แทน ค่าเฉลี่ยของค่าจริง

3. ค่า adjusted R-Squared (adjusted R^2) นำค่า R–squared มาปรับเพื่อให้สอดคล้องกับจำนวนตัวแปรที่อยู่ในตัวแบบ ค่า adjusted R–squared จะมีค่าเพิ่มขึ้นจากเดิมก็ต่อเมื่อพจน์ที่เพิ่มเข้ามาใหม่นั้นทำให้ตัวแบบอธิบายความได้ดีขึ้นจากค่าความคลาดเคลื่อน และค่า adjusted R – squared จะมีค่าลดลงถ้าพจน์ที่เพิ่มเข้ามาในตัวแบบนั้นทำให้ตัวแบบอธิบายความได้น้อยกว่าค่าความคลาดเคลื่อน ดังสมการที่ 2-6

$$adjusted\ R^2 = 1 - \frac{(1-R)^2(n-1)}{n-p-1} \quad (2\text{-}6)$$

โดยที่ n แทน จำนวนข้อมูลทั้งหมด

p แทน จำนวนตัวแปรอิสระ

4. Akaike Information Criterion (AIC) คือค่าสถิติที่คล้ายกับค่า R^2 แต่ใช้รูปแบบการใส่ค่าลอกาลิทึมฐานธรรมชาติ (natural logarithm) โดยหากค่าสถิตินี้ยิ่งมีค่าน้อย หมายความว่าแบบจำลองที่ได้สามารถเป็นตัวแทนข้อมูลจริงได้ดี Akaike Information Criterion (AIC) (Akaike,

1973; Akaike, 1974) เกณฑ์นี้เป็นตัวประมาณไม่เอนเอียงเชิงเส้นกำกับ (Asymptotically Unbiased Estimator) ของความผันแปรของ Kullback's Directed Divergence ในการเลือกตัวแบบคือ ตัวแบบที่มีค่า AIC ต่ำที่สุด จะถือว่าเป็นตัวแบบที่เหมาะสมที่สุด เนื่องจากค่า AIC สะท้อนถึงสมดุลระหว่างความแม่นยำของตัวแบบและจำนวนพารามิเตอร์ที่ใช้ (Burnham & Anderson, 2002) เนื่องจากสะท้อนถึงความสมดุลระหว่างคุณภาพของการพยากรณ์และจำนวนพารามิเตอร์ที่ใช้

$$AIC = 2p - 2ln(L) \quad (2\text{-}7)$$

โดยที่ p แทน จำนวนพารามิเตอร์ในแบบจำลอง

L แทน ค่าความน่าจะเป็นสูงสุด (Likelihood) ของแบบจำลอง

5. มัลติคอลิเนียริตี (Multicollinearity) เป็นปัญหาที่เกิดขึ้นเมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง ส่งผลให้ค่าประมาณของสัมประสิทธิ์ถดถอยมีความไม่แน่นอนสูง และทำให้ตัวแบบไม่สามารถตีความได้อย่างถูกต้อง วิธีการตรวจสอบมัลติคอลิเนียริตีสามารถทำได้โดยใช้ค่า Variance Inflation Factor (VIF) ซึ่งคำนวณจากสมการ 2-4

$$VIF_j = \frac{1}{1-R_j^2} \quad (2\text{-}8)$$

โดยที่ R_j^2 คือค่าสัมประสิทธิ์การตัดสินใจของตัวแปรอิสระตัวที่ เมื่อทำการถดถอยกับตัวแปรอิสระอื่นๆ หากค่า VIF สูงกว่า 10 มักบ่งชี้ว่ามีปัญหามัลติคอลิเนียริตี (Kutner et al., 2005)

2.1.12. การวัดประสิทธิภาพของตัวแบบด้วยเมตริกซ์วัดประสิทธิภาพ (Confusion Matrix)

เมตริกซ์วัดประสิทธิภาพถือเป็นเครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนายที่ได้จากตัวแบบหรือวิธีเชิงทำนาย โดยใช้การเปรียบเทียบผลลัพธ์ของการทำนาย (Predicted) กับสิ่งที่เกิดขึ้นจริง(Actually) ซึ่งสามารถประเมินผลได้ 4 รูปแบบ ได้แก่

True Positive (TP) = การทำนายมีผลลัพธ์ว่า จริง และสิ่งที่เกิดขึ้นจริง คือ จริง

True Negative (TN) = การทำนายมีผลลัพธ์ว่า ไม่จริง และสิ่งที่เกิดขึ้นจริง คือ ไม่จริง

False Positive (FP) = การทำนายมีผลลัพธ์ว่า จริง แต่สิ่งที่เกิดขึ้นจริง คือ ไม่จริง

False Negative (FN) = การทำนายมีผลลัพธ์ว่า ไม่จริง แต่สิ่งที่เกิดขึ้นจริง คือ จริง

		Actually	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

ภาพที่ 2-4 Confusion Matrix

ในการวัดประสิทธิภาพของตัวแบบหรือวิธีเชิงทำนายในการจำแนกประเภทข้อมูล สำหรับงานวิจัยฉบับนี้ใช้เกณฑ์ในการวัดประสิทธิภาพที่คำนวณได้จากสัดส่วนของค่า TP, FP, TN และ FN ซึ่งจะใช้เกณฑ์การวัดประสิทธิภาพ 5 เกณฑ์ ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความจำ (Recall) ค่ามาตราวัดเอฟ (F-measure) และ ตัววัดประสิทธิภาพโดยใช้พื้นที่ใต้กราฟ (Area Under an ROC Curve : AUCs)

ค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้องของตัวแบบหรือวิธีเชิงทำนายในภาพรวม (พิจารณาหมวดหมู่ จริง และหมวดหมู่ ไม่จริง)

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2\text{-}9)$$

ค่าความแม่นยำ (Precision) เป็นการวัดความถูกต้องของตัวแบบหรือวิธีเชิงทำนายมีผลลัพธ์ว่าจริง

$$Precision = \frac{TP}{TP+FP} \quad (2\text{-}10)$$

ค่าความจำ (Recall) เป็นการวัดความถูกต้องของตัวแบบหรือวิธีเชิงทำนายในการหาตัวอย่างในสิ่งที่เกิดขึ้นจริงว่าจริง

$$Recall = \frac{TP}{TP+FN} \quad (2\text{-}11)$$

ค่ามาตราวัดเอฟ (F-measure) เป็นวัดค่าความแม่นยำและค่าความจำพร้อมกันในสิ่งที่เกิดขึ้นจริงและผลลัพธ์ว่าจริง

$$F - measure = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (2\text{-}12)$$

ROC เป็นวิธีพื้นฐานสำหรับการสรุปผลการดำเนินงานของตัวจำแนกประเภท ซึ่งเป็นค่าความสัมพันธ์ระหว่าง Recall และ False Positive (1-Specificity) โดยเส้นโค้งของ ROC มีแกน X เป็น False Positive (1 – Specificity) และแกน Y เป็น Recall และ AUCs คือ ค่าพื้นที่ใต้โค้ง ROC

$$\text{False Positive } (1 - \text{ Specificity}) = \frac{FP}{TN+FP} \quad (2\text{-}13)$$

AUC (Area Under the Curve) คือ ค่าพื้นที่ใต้เส้นโค้ง ROC มีค่าอยู่ในช่วง [0, 1] ซึ่งใช้วัดความสามารถของโมเดลในการแยกแยะระหว่างคลาสทั้งสองได้อย่างถูกต้อง ซึ่งยิ่งค่ามีการเข้าใกล้ 1 หมายถึงสามารถจำแนกได้สมบูรณ์ หากมีค่าน้อยกว่า 0.5 หมายถึงจำแนกผิด หรือโมเดลกลับทิศทางการจำแนก ค่า AUC ยังสามารถตีความได้ว่า เป็น ความน่าจะเป็นที่โมเดลจะจัดอันดับตัวอย่างบวก (positive instance) ได้สูงกว่าตัวอย่างลบ (negative instance) โดยการสุ่มเลือกตัวอย่างบวกและลบหนึ่งคู่

2.1.13. Synthetic Minority Oversampling Technique (SMOTE)

เป็นวิธีการสร้างข้อมูลสังเคราะห์ที่เพิ่มตัวอย่างใหม่ให้กับคลาสที่มีจำนวนน้อยโดยใช้เทคนิคการประมาณค่าระหว่างจุดข้อมูลที่อยู่ใกล้เคียงกันในเชิงคุณลักษณะ (Feature Space) วิธีนี้ช่วยลด

ปัญหาการเกิด Overfitting ที่มักพบในกรณีของการทำ Random Oversampling ซึ่งเป็นการสุ่มเพิ่มข้อมูลเดิมที่มีอยู่แล้ว ตัวอย่างข้อมูลใหม่ของ SMOTE จะถูกสร้างขึ้นโดยใช้การ Interpolation ระหว่างข้อมูลของคลาสที่มีจำนวนน้อยที่ใกล้เคียงกัน วิธีนี้ทำให้แบบจำลองสามารถเรียนรู้รูปแบบของข้อมูลได้ดีขึ้น และเพิ่มความสามารถในการแยกแยะข้อมูลของทั้งสองคลาสได้อย่างมีประสิทธิภาพ

2.1.14. Adaptive Synthetic Sampling Approach (ADASYN)

เป็นการพัฒนาต่อยอดจาก SMOTE ซึ่งมีแนวคิดหลักคล้ายกัน แต่มีจุดเด่นที่แตกต่างกันตรงที่ ADASYN จะให้ความสำคัญกับตัวอย่างข้อมูลที่ยากต่อการเรียนรู้มากกว่า กล่าวคือ จำนวนตัวอย่างสังเคราะห์ที่ถูกสร้างขึ้นจะขึ้นอยู่กับความซับซ้อนของข้อมูล ซึ่งกำหนดโดยค่าความหนาแน่นของข้อมูล (Density Distribution) ทำให้แบบจำลองสามารถปรับขอบเขตการตัดสินใจให้เหมาะสมกับข้อมูลที่ยากต่อการเรียนรู้ได้อย่างมีประสิทธิภาพ

2.1.15. Hybridization: SMOTE + ENN (Edited Nearest Neighbors) หรือ SMOTEENN

เป็นอีกหนึ่งเทคนิคการผสมผสานระหว่าง Oversampling และ Undersampling โดยใช้ ENN (Edited Nearest Neighbors) ซึ่งเป็นเทคนิคที่ใช้ตรวจสอบความผิดพลาดของข้อมูลโดยใช้วิธี K-Nearest Neighbors (KNN) หากข้อมูลของคลาสที่มีจำนวนมากถูกล้อมรอบด้วยข้อมูลของคลาสที่มีจำนวนน้อย และถูก KNN จัดว่าอยู่ในคลาสผิด ข้อมูลเหล่านั้นจะถูกลบออก การใช้ ENN ร่วมกับ SMOTE ทำให้สามารถทำความสะอาดข้อมูลได้อย่างมีประสิทธิภาพ โดยไม่เพียงแต่ลบข้อมูลจากคลาสที่มีจำนวนมากเท่านั้น แต่ยังลบข้อมูลของคลาสที่มีจำนวนน้อยบางส่วนที่อาจก่อให้เกิดความซับซ้อนในการเรียนรู้ของแบบจำลองด้วย ซึ่งช่วยให้ข้อมูลมีความชัดเจนมากขึ้น และทำให้แบบจำลองสามารถแยกแยะคลาสได้ดีขึ้น

2.1.16. Hybridization: SMOTE + Tomek Links (SMOTE + Tomek) หรือ SMOTETomek

เป็นเทคนิคที่รวมระหว่างการ Oversampling และ Undersampling เข้าด้วยกันเพื่อเพิ่มประสิทธิภาพของแบบจำลอง ในขั้นแรกจะทำการเพิ่มข้อมูลคลาสที่มีจำนวนน้อยโดยใช้ SMOTE แต่ปัญหาที่อาจเกิดขึ้นคือ ข้อมูลที่ถูกเพิ่มเข้ามาอาจรุกล้ำพื้นที่ของคลาสที่มีจำนวนมาก ส่งผลให้แบบจำลองเกิด Overfitting ได้ง่าย ดังนั้น Tomek Links จึงถูกนำมาใช้เพื่อลบข้อมูลบางส่วนที่ทำให้

เกิดความซ้อนทับกัน Tomek Links เป็นคู่ของตัวอย่างข้อมูลที่อยู่ในคลาสตรงข้ามกันและเป็นเพื่อนบ้านที่ใกล้ที่สุดกัน หากมี Tomek Links ในชุดข้อมูลหลังจากที่ทำ SMOTE แล้ว ข้อมูลจากคลาสที่มีจำนวนมากกว่าจะถูกลบออกเพื่อลดความซับซ้อนของพื้นที่ข้อมูล วิธีนี้ช่วยทำให้ขอบเขตการตัดสินใจของแบบจำลองชัดเจนขึ้น และลดโอกาสในการเกิด Overfitting

2.1.17. Google Colab

Google Colab โดยมีชื่อเต็มว่า Google Colaboratory เป็นบริการ Software as a Service (Saas) ที่พัฒนามากจาก Jupyter notebook โดยสามารถทำงานบนคลาวด์โดยที่ไม่ต้องลงโปรแกรม โดยใช้ภาษาไพทอน (Python) มีไลเบอรี่ที่เพื่ออำนวยความสะดวกสำหรับผู้ใช้ที่สำคัญเช่น NumPy, Padas, Scikit Learn เป็นต้น

2.2. งานวิจัยที่เกี่ยวข้อง

ศราวุฒิ ทองเชื้อ (2566). การทำนายสัญญาณการซื้อขายสกุลเงินดิจิทัลโดยใช้การเรียนรู้ของเครื่องและการวิเคราะห์ทางเทคนิค. คณะพาณิชยศาสตร์และการบัญชี. การศึกษาการลงทุนในตลาดสกุลเงินดิจิทัลได้รับความสนใจจากนักลงทุนจำนวนมาก เนื่องจากเป็นสินทรัพย์ที่สามารถให้ผลตอบแทนสูง แต่ในขณะเดียวกันก็มีความเสี่ยงที่จะเกิดความสูญเสียสูง การลงทุนในตลาดนี้จึงจำเป็นต้องอาศัยเครื่องมือวิเคราะห์ที่มีประสิทธิภาพ งานวิจัยนี้ศึกษาการทำนายสัญญาณการซื้อขายสกุลเงินดิจิทัลโดยเปรียบเทียบประสิทธิภาพของตัวแบบป่าสุ่ม (Random Forest), ตัวแบบการถดถอยโลจิสติก (Logistic Regression) และตัวแบบการเรียนรู้แบบกลุ่มด้วยเทคนิค Stacking โดยใช้ตัวบ่งชี้ทางเทคนิค 50 ตัว สำหรับสกุลเงินดิจิทัลที่มีมูลค่าตามราคาตลาดสูงสุด 20 อันดับแรก วิเคราะห์จากข้อมูลความถี่ 1 วัน และ 1 สัปดาห์ ผลการศึกษา พบว่าตัวแบบป่าสุ่มให้ค่าความถูกต้องระหว่าง 0.4423 - 0.7115, ตัวแบบการถดถอยโลจิสติกให้ค่าความถูกต้องระหว่าง 0.4615 - 0.6346 และตัวแบบการเรียนรู้แบบกลุ่มด้วยเทคนิค Stacking ให้ค่าความถูกต้องระหว่าง 0.4423 - 0.6538 โดยสรุปพบว่าตัวแบบการถดถอยโลจิสติกเหมาะสมสำหรับสกุลเงินดิจิทัลส่วนใหญ่ และการใช้เทคนิค Stacking สามารถช่วยเพิ่มประสิทธิภาพการทำนายได้ นอกจากนี้ ข้อมูลความถี่ 1 สัปดาห์ให้ผลลัพธ์ดีกว่าข้อมูลความถี่ 1 วัน

ณรงค์ศักดิ์ โค้ววิลัยแสง (2564). การพัฒนากรอบแนวคิดการจำแนกประเภทซัพพลายเออร์สำหรับปัญหาการประเมินซัพพลายเออร์ในระบบ SAP ERP. การพัฒนาแนวคิดการจำแนกประเภทซัพพลายเออร์สำหรับปัญหาการประเมินซัพพลายเออร์ในระบบ SAP ERP ที่มีความไม่แน่นอนจากการวิเคราะห์ค่าคะแนนเชิงเส้น (Linear scoring model) โดยนำรายการสั่งซื้อ (Purchase orders) และข้อมูลใบขอซื้อ (Purchase requisition) มาใช้เป็นข้อมูลนำเข้าเพื่อจำแนกประสิทธิภาพซัพพลายเออร์ โดยพิจารณาปัจจัยด้าน ปริมาณสินค้า (Quantity), คุณภาพสินค้า (Quality), ความน่าเชื่อถือในการจัดส่ง (Delivery commitment) รวมถึงกระบวนการ ดึงข้อมูล (Extract), แปลงข้อมูล (Transform) และนำมาใช้ในการวิเคราะห์ โมเดลการจำแนกประเภทที่ใช้เปรียบเทียบ ได้แก่ Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Adaboost, Decision Tree, Random Forest และโมเดลแบบ Stacking ซึ่งรวม SVM+DT และ SVM+LR เพื่อเพิ่มประสิทธิภาพในการจำแนก ผลการศึกษาพบว่า Adaboost เป็นโมเดลที่ให้ความแม่นยำสูงสุดอยู่ที่ 81.6% รองลงมาคือ 67.6%

วิชญ์วิสิฐ และคณะ (2561) ศึกษาการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนก ผู้ป่วยโรคเบาหวาน โดยใช้ข้อมูลการกลับมารักษาซ้ำในโรงพยาบาลของผู้ป่วยโรคเบาหวาน เมื่อ ข้อมูลไม่สมดุลใช้ข้อมูลจาก 130 โรงพยาบาลในประเทศสหรัฐอเมริกา โดยการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูลจำนวน 4 วิธีคือ วิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) โดยใช้ เทคนิคการจำแนกคือ วิธีการถดถอยโลจิสติกแบบมัลติโนเมียล และวิธีต้นไม้การตัดสินใจในการจำแนกผู้ป่วยโรคเบาหวาน จากการเปรียบเทียบประสิทธิภาพของสถิติและอัลกอริทึมในการจำแนก พบว่าข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีวิธีสังเคราะห์ข้อมูลใหม่สามารถจำแนกผู้ป่วยโรคเบาหวานด้วยวิธีต้นไม้การตัดสินใจมีผลลัพธ์ดีที่สุด

Veronesi (2017) ศึกษาเกณฑ์การประเมินความแม่นยำของแบบจำลองโดยใช้ตัวชี้วัดทางสถิติ ได้แก่ ค่า R-squared, Adjusted R-squared, RMSE, MAE และ AIC เพื่อวิเคราะห์ประสิทธิภาพของแบบจำลองเชิงพยากรณ์ โดยงานวิจัยนี้ได้อธิบายข้อดีและข้อจำกัดของแต่ละเกณฑ์พร้อมทั้งเปรียบเทียบผลลัพธ์ที่ได้จากการใช้แบบจำลองต่าง ๆ พบว่า การใช้ค่า AIC เป็นแนวทางที่เหมาะสมในการเลือกแบบจำลองที่มีความสมดุลระหว่างความแม่นยำและจำนวนพารามิเตอร์ที่

บทที่ 3
วิธีดำเนินการวิจัย

ในการวิจัยครั้งนี้ ได้ดำเนินการพัฒนาแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking เพื่อวิเคราะห์ปัจจัยสำคัญที่เกี่ยวข้องกับความเสี่ยงด้านเครดิต รวม 14 ปัจจัย โดยใช้ข้อมูลทุติยภูมิจากกระทรวงอุตสาหกรรม ข้อมูลถูกปรับสมดุลด้วย 4 วิธี ได้แก่ SMOTE, ADASYN, SMOTEEN และ SMOTE + Tomek จากนั้นได้เปรียบเทียบประสิทธิภาพของ 9 วิธีเชิงทำนาย ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine), บูสต์ติงไล่ระดับ (Gradient Boosting), การจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), การจำแนกแบบเบส์นาอีฟ (Naïve Bayes), วิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (Extreme Gradient Boosting with Meta-Learning) และวิธีเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning) กระบวนการวิเคราะห์ประสิทธิภาพของตัวแบบใช้ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความจำ (Recall) ค่ามาตราวัดเอฟ (F-measure) และ ตัววัดประสิทธิภาพโดยใช้พื้นที่ใต้กราฟ (Area Under an ROC Curve : AUCs) เป็นตัวชี้วัด ซึ่งในบทนี้จะกล่าวถึงขั้นตอนการดำเนินการวิจัยโดยมีรายละเอียด ดังนี้

3.1 การจัดการข้อมูล
3.2 ขั้นตอนการสร้างตัวแบบ Stacking
3.3 การเปรียบเทียบประสิทธิภาพการทำนายทั้ง 9 วิธีเชิงทำนาย
3.4 สรุปผล

3.1 การจัดการข้อมูล

ในการวิจัยนี้มุ่งเน้นการวิเคราะห์ฐานข้อมูลผู้กู้ของกองทุนหมุนเวียน โดยใช้ข้อมูลจากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทย ซึ่งอยู่ภายใต้กรมส่งเสริมอุตสาหกรรม กระทรวงอุตสาหกรรม ประเทศไทย การศึกษาจะเน้นไปที่ผู้กู้รายบุคคล ซึ่งเป็นกลุ่มลูกค้าหลักของกองทุน และเป็นกลุ่มที่มีสัดส่วนสูงในหนี้ที่ไม่ก่อให้เกิดรายได้ (Non-Performing Loans: NPLs) ขอบเขตของการวิเคราะห์จะครอบคลุมโครงสร้างของฐานข้อมูลผู้กู้ ปัจจัยที่อาจส่งผลต่อสถานะหนี้ และแนวโน้มของหนี้ที่ไม่ก่อให้เกิดรายได้ในกลุ่มผู้กู้รายบุคคล ปัจจัยสำคัญที่เกี่ยวข้องกับความเสี่ยงด้านเครดิต 14 ปัจจัย ซึ่งเป็นข้อมูลทุติยภูมิจากกระทรวงอุตสาหกรรม ได้แก่ อายุของผู้ขอสินเชื่อ (Age), จำนวนสมาชิกในครอบครัวหรือธุรกิจที่เกี่ยวข้องกับการขอสินเชื่อ (Member), หนี้สินที่ได้รับเข้ามาของธุรกิจ (IN_debt), หนี้สินที่ชำระออกไปของธุรกิจ (OUT_debt), จำนวนพนักงานของธุรกิจ (Num_Emp), ค่าใช้จ่ายรายปี (Expenses_Y), อัตราส่วนของลูกค้ารายใหม่ (ratio_CCus-TCus), อัตราการเติบโตของธุรกิจในช่วง 3 ปี (Growth_rate_3Y), รายได้สุทธิของปีที่ผ่านมา (Net-income_LY), หนี้สินที่ได้รับเข้ามาโดยเฉพาะจากธุรกิจ (IN_debt_Bus), หนี้สินที่ชำระออกไปโดยเฉพาะจากธุรกิจ (OUT_debt_Bus), จำนวนเงินกู้ที่ขอจากสถาบันการเงิน (Loan_Amount), ระยะเวลาการกู้ยืม (period), ตัวแปรเป้าหมาย (Target 0 = ไม่เป็น NPL, 1 = เป็น NPL)

ตารางที่ 3-1 ชื่อตัวแปรของข้อมูลจากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทย

ชื่อตัวแปร	ความหมาย
Age	อายุของผู้ขอสินเชื่อ (หน่วย: ปี)
Member	จำนวนสมาชิกในครอบครัวหรือธุรกิจที่เกี่ยวข้องกับการขอสินเชื่อ
IN_debt	หนี้สินที่ได้รับเข้ามาของธุรกิจ
OUT_debt	หนี้สินที่ชำระออกไปของธุรกิจ
Num_Emp	จำนวนพนักงานของธุรกิจ
Expenses_Y	ค่าใช้จ่ายรายปี
ratio_CCus-TCus	อัตราส่วนของลูกค้ารายใหม่
Growth_rate_3Y	อัตราการเติบโตของธุรกิจในช่วง 3 ปี

ตารางที่ 3-1 ชื่อตัวแปรของข้อมูลจากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทย

Net-income_LY	รายได้สุทธิของปีที่ผ่านมา
IN_debt_Bus	หนี้สินที่ได้รับเข้ามาโดยเฉพาะจากธุรกิจ
OUT_debt_Bus	หนี้สินที่ชำระออกไปโดยเฉพาะจากธุรกิจ
Loan_Amount	จำนวนเงินกู้ที่ขอจากสถาบันการเงิน
period	ระยะเวลาการกู้ยืม
Target	ตัวแปรเป้าหมาย (Target 0 = ไม่เป็น NPL, 1 = เป็น NPL)

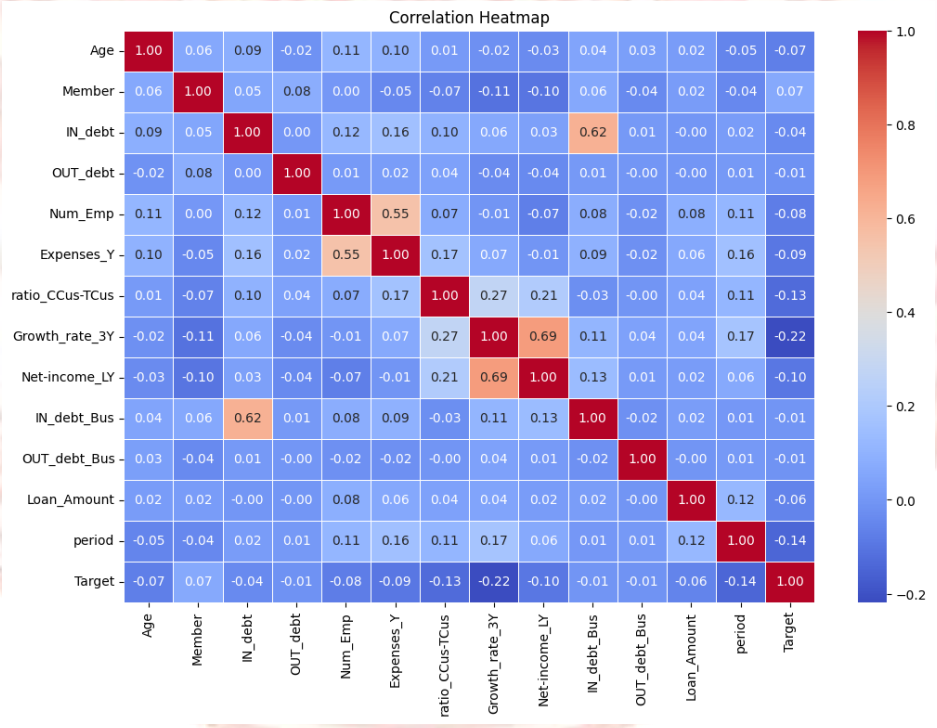
ตารางที่ 3-2 ค่าสถิติเบื้องต้น

index	min	mean	std	max
Age	20	44.62	9.87	74
Member	0	1.21	1.30	10
IN_debt	0	587794.03	1280887.69	19,834,276
OUT_debt	0	16617.84	257638.82	7,214,728
Num_Emp	0	6.56	9.31	201
Expenses_Y	0	708148.10	904281.69	18,540,000
ratio_CCus-TCus	0	52.36	23.39	225
Growth_rate_3Y	-10	21.38	18.41	100
Net-income_LY	-60	23.51	17.15	100
IN_debt_Bus	0	407685.90	1033702.13	19,834,276
OUT_debt_Bus	0	2393.02	43475.00	1,300,000
Loan_Amount	0	253024.65	650243.44	20,000,000

ตารางที่ 3-2 ค่าสถิติเบื้องต้น

Target	0	0.27	0.44	1

วิเคราะห์ข้อมูลเบื้องต้นว่าปัจจัยใดมีความสำคัญและส่งผลกระทบต่อความเสี่ยงด้านเครดิต จึงเลือกใช้ Heatmap เป็นเครื่องมือในการวิเคราะห์ความสัมพันธ์ของตัวแปรทางการเงินและปัจจัยที่เกี่ยวข้อง Heatmap เป็นการแสดงผลในรูปแบบของตารางที่ใช้ ค่าสหสัมพันธ์ (Correlation Coefficient) ในการวัดระดับความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ซึ่งมีค่าตั้งแต่ -1 ถึง 1 โดยค่าที่เข้าใกล้ 1 หรือ -1 บ่งบอกถึงความสัมพันธ์ที่สูง ขณะที่ค่าที่ใกล้ 0 หมายถึงไม่มีความสัมพันธ์ที่ชัดเจน ดังภาพที่ 3-1



ภาพที่ 3-1 Correlation Heatmap

จากภาพที่ 3-1 แสดงให้เห็นถึงระดับความสัมพันธ์ของตัวแปรทางการเงินกับตัวแปรเป้าหมาย (Target: NPL หรือไม่เป็น NPL) และความสัมพันธ์ระหว่างตัวแปรอื่น ๆ ที่อาจมีผลต่อการพิจารณาความเสี่ยงด้านเครดิต ค่าสหสัมพันธ์ที่เป็น บวก (สีแดงเข้ม) หมายถึง เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้นอีกตัวแปรหนึ่งมีแนวโน้มเพิ่มขึ้นตามไปด้วย ในขณะที่ค่าสหสัมพันธ์ที่เป็น ลบ (สีน้ำเงินเข้ม) หมายถึง

ตัวแปรสองตัวมีแนวโน้มเปลี่ยนแปลงไปในทิศทางตรงกันข้าม และค่าสหสัมพันธ์ที่ใกล้ ศูนย์ (สีฟ้าอ่อน) หมายถึงไม่มีความสัมพันธ์ที่ชัดเจน สามารถกล่าวได้ว่าตัวแปรที่มีความสัมพันธ์สูงกับความเสี่ยงด้านเครดิต (Target) เช่น Growth_rate_3Y มีค่าอยู่ที่ -0.22 กล่าวคือ ธุรกิจที่เติบโตสูงในช่วง 3 ปี มีแนวโน้มเป็น NPL น้อยลง, ratio_CCus-TCus มีค่าอยู่ที่ -0.13 กล่าวคือ อัตราส่วนลูกค้ารายใหม่ที่สูงขึ้น อาจช่วยลดความเสี่ยง NPL, period มีค่าอยู่ที่ -0.14 กล่าวคือ ระยะเวลากู้ยืมที่ยาวขึ้น มีแนวโน้มลดความเสี่ยง NPL และปัจจัยอื่นๆ จากผลการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร พบว่าบางปัจจัยมีผลต่อความเสี่ยงด้านเครดิตอย่างมีนัยสำคัญ ขณะที่บางปัจจัยไม่มีความสัมพันธ์ที่ชัดเจน ดังนั้นผู้วิจัยจึงเห็นควรคัดเลือกตัวแปรที่เหมาะสมเพื่อใช้ในการพัฒนาแบบจำลองในขั้นตอนถัดไป

3.1.1. การคัดเลือกตัวแปร

หลังจากวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรทางการเงินและปัจจัยต่าง ๆ ที่เกี่ยวข้องกับความเสี่ยงด้านเครดิตแล้ว ในขั้นตอนนี้คือการคัดเลือกตัวแปรที่เหมาะสมที่สุดเพื่อนำไปสร้างแบบจำลองเชิงพยากรณ์ ซึ่งมีความสำคัญอย่างยิ่งต่อประสิทธิภาพของแบบจำลอง ในการวิจัยนี้ ใช้วิธี Stepwise Selection ซึ่งเป็นเทคนิคการคัดเลือกตัวแปรเชิงสถิติที่ได้รับความนิยม โดยอาศัยหลักการเพิ่ม (Forward Selection) และลด (Backward Elimination) ตัวแปรอย่างเป็นระบบ โดยพิจารณาจากค่าทางสถิติเพื่อให้ได้ชุดตัวแปรที่เหมาะสมที่สุด ลดความซับซ้อนของแบบจำลอง และเพิ่มความแม่นยำในการพยากรณ์ การดำเนินการนี้ทำผ่านโปรแกรม R ซึ่งช่วยคัดเลือกตัวแปรที่มีนัยสำคัญ และตัดตัวแปรที่ไม่ส่งผลกระทบต่อโมเดลออก

ตารางที่ 3-3 ตารางผลการวิเคราะห์แบบจำลองถดถอยเชิงเส้นทั่วไป

ตัวแปร	ค่าสัมประสิทธิ์	S.D.	t	P
(Intercept)	0.7049	0.0853	8.263	<0.001***
Age	-0.0035	0.0014	-2.527	0.0117*
Member	0.0151	0.0104	1.45	0.1474
Num_Emp	-0.0022	0.0015	-1.491	0.1362
ratio_CCus.TCus	0.0006	0.0006	-2.123	0.034*
Growth_rate_3Y	-0.0058	0.001	-5.551	<0.001***
Net.income_LY	0.0023	0.0011	2.062	0.0394*
period	-0.0025	0.0008	-3.126	0.0018

***มีนัยสำคัญทางสถิติที่ระดับ .001

**มีนัยสำคัญทางสถิติที่ระดับ .01

*มีนัยสำคัญทางสถิติที่ระดับ .05

จากตารางแสดงการวิเคราะห์แบบจำลองถดถอยเชิงเส้นทั่วไป โดยใช้ตัวแปรตามคือ Target และตัวแปรอิสระที่ผ่านกระบวนการคัดเลือกตัวแปร ได้แก่ Age, Member, Num_Emp, ratio_CCus.TCus, Growth_rate_3Y, Net.income_LY และ period พบว่าผลลัพธ์ของโมเดลเป็นดังตารางที่ 3-3 และทำการวิเคราะห์ค่าความสัมพันธ์เชิงพหุคอลิเนียร์ (Multicollinearity) ด้วย Variance Inflation Factor (VIF) ในการวิเคราะห์โมเดลถดถอย ค่าความสัมพันธ์เชิงพหุคอลิเนียร์ (Multicollinearity) เป็นปัญหาที่อาจส่งผลต่อความถูกต้องของตัวประมาณค่าสัมประสิทธิ์ (Coefficient Estimates) ซึ่งสามารถตรวจสอบได้โดยใช้ค่า VIF

ตารางที่ 3-4 ตารางผล Variance Inflation Factor (VIF)

ตัวแปร	Age	Member	Num_ Emp	ratio_CCus .TCus	Growth_rate _3Y	Net.income _LY	period
VIF	**1.021**	**1.019**	**1.040**	**1.095**	**2.037**	**1.947**	**1.060**

จากตารางที่ 3-4 แสดงให้เห็นว่าจากการวิเคราะห์ Variance Inflation Factor (VIF) ค่า VIF ต่ำกว่า 5 ทุกตัวแปร หมายความว่าไม่มีปัญหา Multicollinearity ในโมเดลนี้ ซึ่งหมายความว่าตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์ซ้ำซ้อนกันมากจนส่งผลต่อเสถียรภาพของโมเดล

3.1.2. การจัดการปัญหาความไม่สมดุลของคลาส

ในงานวิจัยนี้ ตัวแปรเป้าหมาย (Target) ถูกกำหนดให้มีค่า 0 หมายถึง "ไม่เป็น NPL หรือ ไม่ผิดนัดชำระ" และค่า 1 หมายถึง "เป็น NPL หรือ ผิดนัดชำระ" โดยจากข้อมูลต้นฉบับพบว่า มีจำนวนตัวอย่างที่อยู่ในกลุ่ม ไม่เป็น NPL (Target = 0) มากกว่ากลุ่ม เป็น NPL (Target = 1) อย่างมีนัยสำคัญ ซึ่งทำให้เกิดปัญหาความไม่สมดุลของข้อมูล (Imbalanced Data) ที่อาจส่งผลกระทบต่อความแม่นยำของแบบจำลองการพยากรณ์ เพื่อแก้ไขปัญหานี้ งานวิจัยได้นำ 4 เทคนิคการปรับสมดุลข้อมูล มาใช้ ได้แก่:

SMOTE (Synthetic Minority Oversampling Technique) ใช้วิธีการสร้างข้อมูลสังเคราะห์ขึ้นมาเพื่อเพิ่มจำนวนตัวอย่างของกลุ่มที่เป็น NPL (Target = 1) ให้มีจำนวนใกล้เคียงกับกลุ่มที่ไม่เป็น NPL (Target = 0)

ADASYN (Adaptive Synthetic Sampling Approach) เป็นการขยายข้อมูลแบบสังเคราะห์โดยเน้นไปที่ตัวอย่างที่อยู่ใกล้ขอบเขตการจำแนกมากที่สุด

SMOTE + Tomek Links หรือ SMOTETomek ผสมผสานการสร้างข้อมูลสังเคราะห์ (SMOTE) ร่วมกับการลบตัวอย่างที่อาจทำให้เกิดความซับซ้อนของการจำแนกด้วยวิธี Tomek

SMOTE + ENN (Edited Nearest Neighbors) หรือ SMOTEENN ใช้ SMOTE ในการเพิ่มข้อมูลสังเคราะห์ และใช้ ENN เพื่อลบข้อมูลที่อาจก่อให้เกิดความผิดพลาดในการจำแนก

3.1.3. การแบ่งชุดข้อมูล

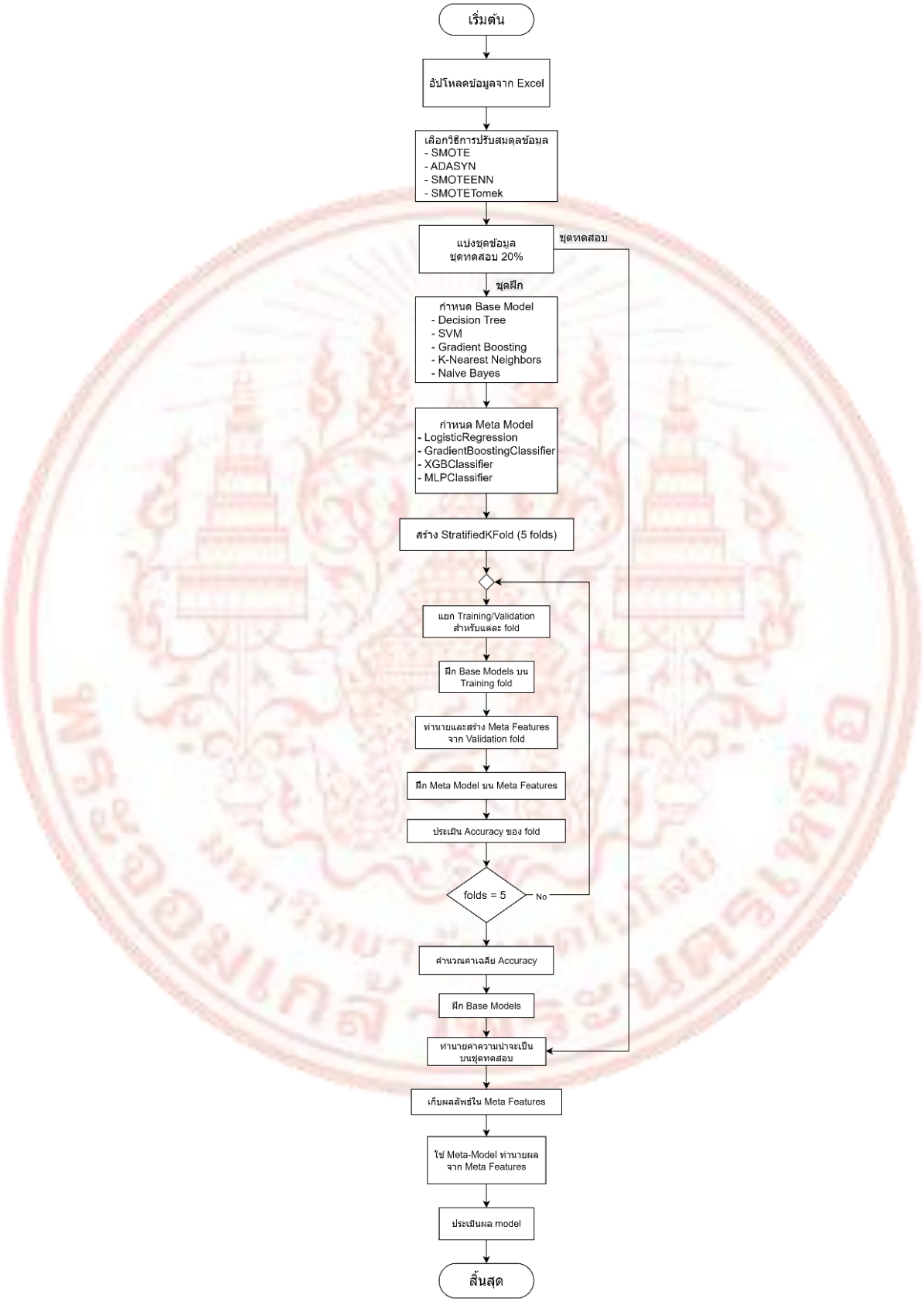
จากข้อมูลจากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทยได้ดำเนินการแบ่งชุดข้อมูลโดยใช้ K-Fold Cross-Validation (K=5) ซึ่งเป็นวิธีที่แบ่งข้อมูลออกเป็น 5 ส่วน โดยใช้ 4 ส่วนเป็นข้อมูลฝึกสอน (Training Set) และอีก 1 ส่วนเป็นข้อมูลทดสอบ (Testing Set) สลับไปเรื่อย ๆ จนครบทุกส่วนจากนั้นได้นำข้อมูลที่ผ่านการปรับสมดุลมาใช้ในการสร้างแบบจำลอง

3.2 ขั้นตอนการสร้างตัวแบบ Stacking

3.2.1. ขั้นตอนมีดังนี้

กระบวนการพัฒนาตัวแบบ Stacking ในการประเมินความเสี่ยงด้านเครดิตประกอบด้วยขั้นตอนสำคัญดังต่อไปนี้ เริ่มต้นด้วยการนำเข้าชุดข้อมูลจากไฟล์ Excel และดำเนินการแยกตัวแปรอิสระ (Features) และตัวแปรตาม (Target) ออกจากกันอย่างชัดเจน จากนั้นทำการตรวจสอบความไม่สมดุลของข้อมูล และปรับสมดุลข้อมูลหากจำเป็น โดยใช้เทคนิคที่เหมาะสม ทั้งนี้เพื่อเพิ่มประสิทธิภาพของโมเดลในการเรียนรู้ข้อมูลกลุ่มน้อย ภายหลังการเตรียมข้อมูล ได้ทำการแบ่งข้อมูลออกเป็นชุดฝึกสอน (Training set) และชุดทดสอบ (Test set) ตามสัดส่วนที่กำหนด โดยชุดฝึกสอนถูกใช้สำหรับการฝึกโมเดลในกระบวนการ Stacking ขณะที่ชุดทดสอบจะถูกสงวนไว้สำหรับการประเมินผลขั้นสุดท้ายของตัวแบบ ในกระบวนการสร้างตัวแบบ Stacking ได้กำหนดโมเดลพื้นฐาน (Base Models) จำนวน 5 วิธี ได้แก่ Decision Tree, Support Vector Machine, Gradient Boosting, K-Nearest Neighbors และ Naïve Bayes โดยโมเดลเหล่านี้จะทำหน้าที่เรียนรู้จากชุดข้อมูลฝึกสอน และสร้างค่าทำนายเบื้องต้นค่าทำนายจากโมเดลพื้นฐานทั้งห้าถูกนำมาใช้เป็นคุณลักษณะใหม่ (Meta Features) สำหรับการฝึกโมเดลระดับเมตา (Meta Model) ซึ่งประกอบด้วยตัวเลือก 4 วิธี ได้แก่ วิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (Extreme Gradient Boosting with Meta-Learning) และวิธีเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning) เพื่อให้กระบวนการฝึกโมเดลมีความน่าเชื่อถือและลดความเอนเอียงจากข้อมูล ได้ประยุกต์ใช้เทคนิค Stratified K-Fold Cross Validation โดยกำหนดจำนวนพับ (fold) เท่ากับ 5 ในแต่ละพับ ชุดข้อมูลฝึกสอนจะถูกแบ่ง

ออกเป็น Training fold และ Validation fold โดยโมเดลพื้นฐานจะถูกฝึกด้วย Training fold และทำนาย Validation fold เพื่อนำค่าทำนายดังกล่าวไปสร้าง Meta Features ซึ่งจะถูกใช้ในการฝึก Meta Model ต่อไป เมื่อฝึกโมเดลครบทั้ง 5 folds แล้ว จึงทำการคำนวณค่าเฉลี่ยของตัวชี้วัด Accuracy เพื่อประเมินประสิทธิภาพเบื้องต้นของโครงสร้างตัวแบบ Stacking หลังจากนั้น โมเดลพื้นฐานทั้งหมดจะถูกฝึกใหม่โดยใช้ชุดข้อมูล Training set เต็ม และทำการสร้าง Meta Features จากชุด Test set ด้วยโมเดลที่ได้ จากนั้นนำ Meta Features ดังกล่าวมาทำนายด้วย Meta Model ที่ได้ฝึกไว้ ผลลัพธ์ที่ได้จะถูกนำมาประเมินด้วยตัวชี้วัดด้านประสิทธิภาพ ได้แก่ Accuracy, Precision, Recall และ F1-score รวมถึงการวิเคราะห์ประสิทธิภาพเชิงกราฟด้วยค่า AUC จาก ROC Curve ดังภาพที่ 3-2



ภาพที่ 3-2 แผนภาพ flow chart แสดงขั้นตอนการดำเนินงาน

3.2.2. การวัดประสิทธิภาพของโมเดล

เพื่อประเมินประสิทธิภาพของวิธีเชิงทำนายแต่ละวิธี งานวิจัยนี้ใช้การแบ่งชุดข้อมูลฝึกสอนและชุดทดสอบในอัตราส่วนที่แตกต่างกัน และวัดผลลัพธ์โดยเกณฑ์การวัดประสิทธิภาพของโมเดล ได้แก่ Accuracy (ค่าความถูกต้อง) เพื่อวัดสัดส่วนของการทำนายที่ถูกต้องเมื่อเทียบกับข้อมูลทั้งหมด, Precision (ค่าความแม่นยำ) เพื่อวัดอัตราส่วนของตัวอย่างที่ถูกทำนายว่าเป็น NPL อย่างถูกต้อง เมื่อเทียบกับตัวอย่างทั้งหมดที่ถูกทำนายว่าเป็น NPL, Recall (ค่าความจำ) เพื่อวัดความสามารถของโมเดลในการตรวจจับตัวอย่างที่เป็น NPL ทั้งหมดในข้อมูล, F1-Score (ค่ามาตราวัดเอฟ) เป็นค่าที่คำนวณจาก Precision และ Recall เพื่อหาสมดุลระหว่างสองค่าดังกล่าวและ AUCs-ROC Curve (ค่าพื้นที่ใต้โค้ง ROC) ใช้ประเมินความสามารถของโมเดลในการแยกแยะระหว่างกลุ่ม NPL และไม่เป็น NPL โดยค่าที่สูงกว่าหมายถึงโมเดลมีประสิทธิภาพในการจำแนกได้ดียิ่งขึ้น โมเดลทั้งหมดถูกประมวลผลและวิเคราะห์ผลลัพธ์โดยใช้ ภาษา Python บน Google Colab ไลบรารีที่ใช้ imblearn.over_sampling และimblearn.combine พารามิเตอร์ที่สำคัญ ได้แก่

sampling_strategy พารามิเตอร์นี้จะกำหนดจำนวนของข้อมูลที่เพิ่มขึ้นเพื่อทำให้คลาสที่มีจำนวนน้อยมีจำนวนเท่ากับคลาสที่มีจำนวนมาก หรือสามารถตั้งค่าเป็นค่าเฉพาะเพื่อกำหนดอัตราส่วนของข้อมูลที่ต้องการเพิ่มขึ้น หรือใช้ค่า auto เพื่อปรับสมดุลให้เหมาะสม

k_neighbors จำนวนเพื่อนบ้านที่ใช้ในการคำนวณการสร้างข้อมูลสังเคราะห์ ค่าเริ่มต้นคือ 5 แต่สามารถปรับให้เหมาะสมกับลักษณะของข้อมูล

random_state เพื่อใช้ในการตั้งค่าให้การสุ่มในการสร้างข้อมูลสังเคราะห์เกิดขึ้นได้ซ้ำๆ ในกรณีที่ต้องการผลลัพธ์ที่เหมือนเดิมทุกครั้ง

n_splits เป็นการแบ่งข้อมูลในการทำ Cross-Validation (ในที่นี้ใช้ 5-fold cross-validation) ซึ่งหมายถึงการแบ่งข้อมูลออกเป็น 5 ชุดและฝึกโมเดล 5 ครั้งเพื่อหาค่าประเมินผลที่ดีขึ้น

test_size กำหนดอัตราส่วนของข้อมูลที่จะใช้สำหรับชุดทดสอบ

shuffle ใช้ในการสุ่มลำดับของข้อมูลก่อนการแบ่งเพื่อให้ผลการแบ่งข้อมูลไม่เกิดความลำเอียง

MLPClassifier เป็นโมเดล Multi-Layer Perceptron (MLP) ซึ่งเป็น Neural Network แบบ fully connected ที่มีหลายชั้น (layers) โดยมีการเชื่อมโยงระหว่างนิวรอนในแต่ละชั้น

hidden_layer_sizes=(64, 32) กำหนดขนาดของชั้นซ่อน (hidden layers) โดยในที่นี้ชั้นแรกมี 64 นิวรอนและชั้นที่สองมี 32 นิวรอน

activation='relu' ใช้ฟังก์ชัน ReLU (Rectified Linear Unit) เป็นฟังก์ชันการเปิดใช้งาน (activation function) ซึ่งเป็นที่นิยมใน Neural Network เพราะมีความง่ายและประสิทธิภาพที่ดีในการลดการหายไปของ gradient

solver='adam' ใช้ Adam (Adaptive Moment Estimation) ซึ่งเป็นวิธีการปรับค่าของ weights ที่ดีสำหรับ Neural Network โดยการใช้ momentum และ adaptive learning rate

max_iter=500 กำหนดจำนวนรอบ (iterations) ที่โมเดลจะทำการฝึก โดยในที่นี้กำหนดให้ฝึกสูงสุด 500 รอบ

3.3 การเปรียบเทียบประสิทธิภาพการทำนายทั้ง 9 วิธีเชิงทำนาย

งานวิจัยนี้ใช้ข้อมูลจากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทยเกี่ยวกับการชำระหนี้ของลูกค้าสินเชื่อ เพื่อนำมาวิเคราะห์และเปรียบเทียบประสิทธิภาพของการทำนายระหว่างโมเดลพื้นฐาน ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine), บูสต์ติงไล่ระดับ (Gradient Boosting), การจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), การจำแนกแบบเบส์นาอีฟ (Naïve Bayes), วิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (Extreme Gradient Boosting with Meta-Learning) และวิธีเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning) การเปรียบเทียบประสิทธิภาพของแต่ละโมเดลดำเนินการภายใต้อัตราส่วนที่แตกต่างกันระหว่าง ข้อมูลชุดฝึกสอน (Training Set) และข้อมูลชุดทดสอบ (Test Set) โดยใช้ตัวชี้วัด

ประสิทธิภาพ ได้แก่ ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าความจำ (Recall), ค่ามาตราวัดเอฟ (F-measure) และพื้นที่ใต้โค้ง AUCs-ROC (AUCs-ROC Curve) เพื่อประเมินผลลัพธ์ของแต่ละวิธีและระบุโมเดลที่เหมาะสมที่สุดสำหรับการทำนายสถานะของลูกค้าสินเชื่อ

3.4 สรุปผล

สรุปผลโดยการสร้างตารางเปรียบเทียบ โดยใช้ค่าความถูกต้อง ค่าความแม่นยำ ค่าความจำ ค่ามาตราวัดเอฟ และ และพื้นที่ใต้โค้ง โดยพิจารณาจากเปอร์เซ็นต์ที่มาตรวัดแต่ละค่านั้นเป็นเท่าไหร่ ค่ายิ่งสูงหมายถึง วิธีการที่ใช้ยิ่งมีประสิทธิภาพเท่านั้นและเพื่อเปรียบเทียบประสิทธิภาพระหว่าง วิธีต้นไม้ตัดสินใจ (Decision Tree), วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine), วิธีบูสต์ติงไล่ระดับ (Gradient Boosting), วิธีการจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), วิธีการจำแนกแบบเบส์นาอีฟ (Naïve Bayes), วิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (Extreme Gradient Boosting with Meta-Learning) และวิธีเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning)

บทที่ 4
ผลการดำเนินงาน

งานวิจัยครั้งนี้ได้ดำเนินการพัฒนาแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking เพื่อวิเคราะห์ปัจจัยสำคัญที่เกี่ยวข้องกับความเสี่ยงด้านเครดิต รวม 14 ปัจจัย โดยใช้ข้อมูลทุติยภูมิจากกระทรวงอุตสาหกรรม ข้อมูลถูกปรับสมดุลด้วย 4 วิธี ได้แก่ SMOTE, ADASYN, SMOTEEN และ SMOTE + Tomek และเปรียบเทียบประสิทธิภาพของ 9 วิธีเชิงทำนาย ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine), บูสต์ติงไล่ระดับ (Gradient Boosting), การจำแนกเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors), การจำแนกแบบเบส์นาอีฟ (Naïve Bayes), วิธีเมตาการถดถอยลอจิสติก (Logistic Regression with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับ (Gradient Boosting with Meta-Learning), วิธีเมตาบูสต์ติงไล่ระดับสุดขีด (Extreme Gradient Boosting with Meta-Learning) และวิธีเมตาโครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer Perceptron Neural Network with Meta-Learning) กระบวนการวิเคราะห์ประสิทธิภาพของตัวแบบใช้ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความจำ (Recall) ค่ามาตราวัดเอฟ (F-measure) และตัววัดประสิทธิภาพโดยใช้พื้นที่ใต้กราฟ (Area Under an ROC Curve : AUCs) เป็นตัวชี้วัด ซึ่งในบทนี้จะกล่าวถึงผลการดำเนินการวิจัยโดยมีรายละเอียด ดังนี้

4.1 ปัจจัยที่มีอิทธิพลต่อความเสี่ยงด้านเครดิต

4.2 ผลการศึกษาของวิธีเชิงทำนายหรือตัวแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking

4.3 การเปรียบเทียบประสิทธิภาพของการตรวจจับความผิดปกติด้วยเทคนิค Stacking สำหรับความเสี่ยงด้านเครดิต

4.1 ปัจจัยที่มีอิทธิพลต่อความเสี่ยงด้านเครดิต

การศึกษาปัจจัยที่มีผลต่อความเสี่ยงด้านเครดิตมีความสำคัญอย่างยิ่ง เนื่องจากช่วยให้สามารถพัฒนาแบบจำลองพยากรณ์ที่แม่นยำ ลดความซับซ้อน และเพิ่มประสิทธิภาพของการวิเคราะห์ข้อมูลทางการเงิน การศึกษานี้ใช้วิธีการคัดเลือกตัวแปรเชิงสถิติ โดยอาศัยหลักการของ Stepwise Selection ซึ่งเป็นเทคนิคที่อาศัยการเลือกตัวแปรเชิงระบบผ่านกระบวนการเพิ่ม (Forward Selection) และลด (Backward Elimination) ตัวแปรตามค่าทางสถิติดังตารางที่ 4-1

ตารางที่ 4-1 ผลลัพธ์ของแบบจำลองถอดถอยเชิงเส้น และค่าความสัมพันธ์เชิงพหุคอลิเนียร์

ตัวแปร	ค่าสัมประสิทธิ์	S.D.	t	P	VIF
(Intercept)	0.7049	0.0853	8.263	<0.001***	
Age	-0.0035	0.0014	-2.527	0.0117*	1.021294
Member	0.0151	0.0104	1.45	0.1474	1.019204
Num_Emp	-0.0022	0.0015	-1.491	0.1362	1.040343
ratio_CCus.TCus	0.0006	0.0006	-2.123	0.034*	1.094899
Growth_rate_3Y	-0.0058	0.001	-5.551	<0.001***	2.036554
Net.income_LY	0.0023	0.0011	2.062	0.0394*	1.947140
period	-0.0025	0.0008	-3.126	0.0018	1.059931

***มีนัยสำคัญทางสถิติที่ระดับ .001

**มีนัยสำคัญทางสถิติที่ระดับ .01

*มีนัยสำคัญทางสถิติที่ระดับ .05

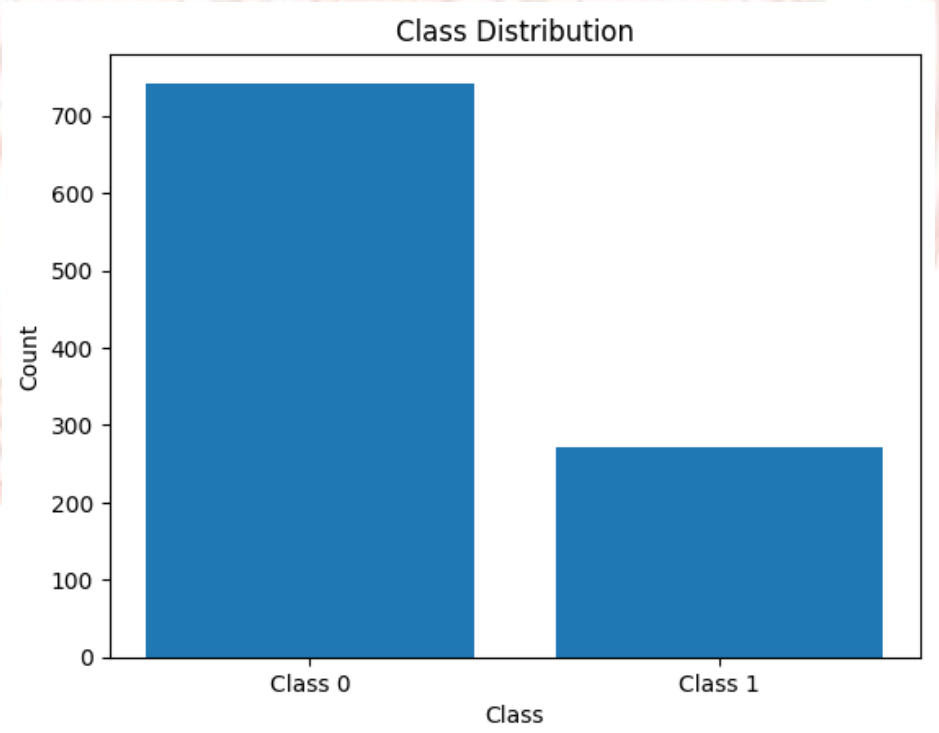
จากตารางที่ 4-1 พบว่าตัวแปร Age, Growth_rate_3Y, Net.income_LY และ period มีอิทธิพลต่อ Target อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และค่า VIF ของตัวแปรอิสระทั้งหมดพบว่าไม่มีตัวแปรใดที่มีค่า VIF สูงเกิน 5 แสดงว่าแบบจำลองไม่มีปัญหาความสัมพันธ์เชิงพหุคอลิเนียร์

4.2 ผลการศึกษาของวิธีเชิงทำนายหรือตัวแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking

การศึกษาประสิทธิภาพของตัวแบบจำลองการเรียนรู้ด้วยเทคนิค Stacking ดำเนินการโดยเปรียบเทียบผลลัพธ์จากข้อมูลเดิมและข้อมูลที่ผ่านการปรับสมดุลด้วยวิธีต่างๆ เพื่อให้เข้าใจผลกระทบของการปรับสมดุลข้อมูลต่อความแม่นยำของตัวแบบ

4.2.1 ข้อมูลเดิม

ข้อมูลที่ใช้ในการศึกษาเป็นชุดข้อมูลที่ยังไม่ได้มีการปรับแต่งเพื่อแก้ไขปัญหาความไม่สมดุลของคลาส โดยตัวแปรเป้าหมาย (Target) แบ่งออกเป็น 2 คลาส ได้แก่ Class 0 (กลุ่มที่ไม่เป็น NPL หรือไม่ผิดนัดชำระ) มีจำนวน 742 รายการ และ Class 1 (กลุ่มที่เป็น NPL หรือผิดนัดชำระ) มีจำนวน 272 รายการ ซึ่งจากการสำรวจพบว่าข้อมูลอยู่ในสภาวะ ไม่สมดุล (Imbalanced Data) โดยมีสัดส่วนของ Class 1 ต่ำกว่า Class 0 อย่างมีนัยสำคัญ ส่งผลต่อประสิทธิภาพของโมเดลที่อาจเกิดปัญหา Overclassification ไปยัง Class ที่มีจำนวนมากกว่า ดังภาพที่ 4-1



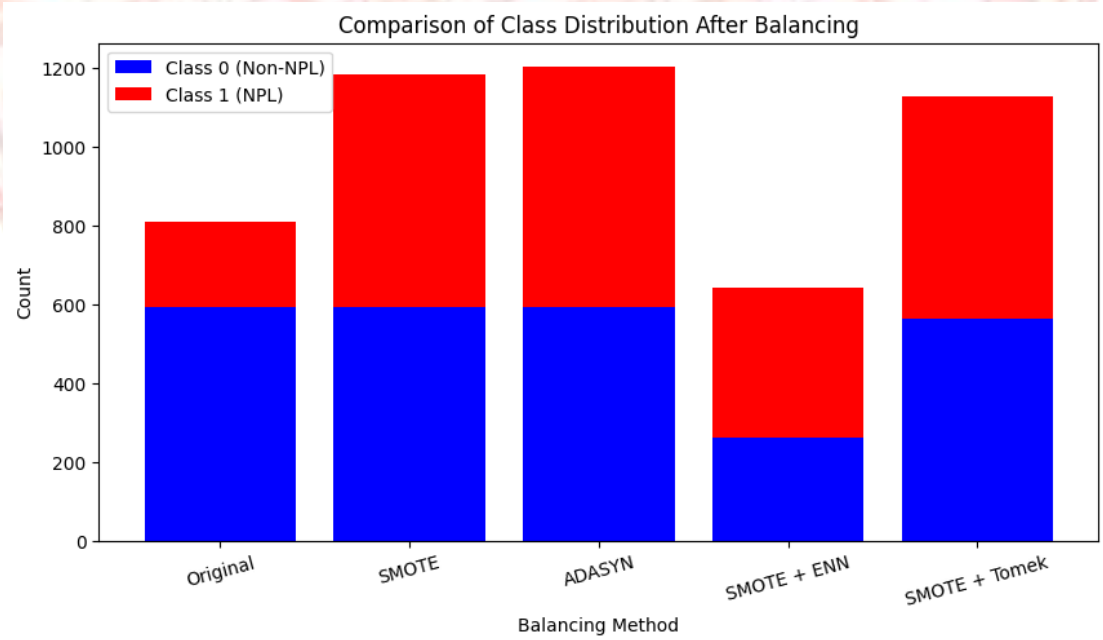
ภาพที่ 4-1 ปัญหาความไม่สมดุลของคลาส โดยตัวแปรเป้าหมาย (Target)

4.2.2 การปรับชุดข้อมูลให้สมดุลด้วยวิธีที่แตกต่างกัน

เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล มีการใช้เทคนิคต่างๆ ในการปรับสมดุลข้อมูลก่อนนำไปฝึกสอนโมเดล ในงานวิจัยนี้ได้ใช้เทคนิคการจัดการปัญหาความไม่สมดุลของข้อมูล 4 วิธี ได้แก่ SMOTE, ADASYN, SMOTE + ENN และ SMOTE + Tomek โดยผลลัพธ์ที่ได้แสดงใน ตารางที่ 4-2 และ ภาพที่ 4-2 ซึ่งแสดงให้เห็นถึงการกระจายตัวของข้อมูลในแต่ละคลาสหลังจากการปรับสมดุล

ตารางที่ 4-2 แสดงจำนวนตัวอย่างข้อมูลในแต่ละคลาส

Method	Class 0	Class 1	Total
Original	593	272	1014
SMOTE	593	593	1186
ADASYN	593	611	1204
SMOTE + ENN	262	381	643
SMOTE + Tomek	564	564	1128



ภาพที่ 4-2 เปรียบเทียบปัญหาความไม่สมดุลของคลาส หลังจากการปรับสมดุล

จากตารางที่ 4-2 แสดงให้เห็นจำนวนตัวอย่างข้อมูลในแต่ละคลาส (**Class 0: ไม่เป็น NPL และ Class 1: เป็น NPL**) ก่อนและหลังการปรับสมดุลด้วยเทคนิคต่างๆ ดังภาพที่ 4-2 ในปัญหาการเรียนรู้ของเครื่องที่มีข้อมูลไม่สมดุล (imbalanced data) ซึ่งพบว่าคลาสหนึ่งมีจำนวนมากกว่าคลาสอื่นอย่างมีนัยสำคัญ อาจทำให้แบบจำลองที่พัฒนาเกิด bias ต่อคลาสที่มีมากกว่า และไม่สามารถจำแนกคลาสที่มีน้อยได้อย่างมีประสิทธิภาพ ดังนั้นจึงมีการใช้เทคนิค Oversampling เพื่อเพิ่มจำนวนตัวอย่างของคลาสที่มีจำนวนน้อยเพื่อให้ข้อมูลมีความสมดุลมากขึ้น เดิมชุดข้อมูลที่มีการกระจายของคลาสเดิมคือ Class 0 มีจำนวน 593 รายการ และ Class 1 มีจำนวน 272 รายการ และทำการเปรียบเทียบผลลัพธ์หลังจากใช้แต่ละเทคนิค พบว่าการใช้ SMOTE + ENN นั้นไม่เพียงแต่ช่วยเพิ่มจำนวนข้อมูลของคลาสที่มีน้อย แต่ยังช่วยลบตัวอย่างที่อาจสร้างความสับสนให้กับแบบจำลอง ซึ่งสามารถนำไปสู่การพัฒนาโมเดลที่สามารถจำแนกข้อมูลได้ดีขึ้น อย่างไรก็ตาม การเลือกใช้เทคนิคการปรับสมดุลควรพิจารณาตามลักษณะของข้อมูลและเป้าหมายของการวิเคราะห์

4.2.3 ผลลัพธ์แบบจำลอง

ในการศึกษาครั้งนี้ ได้มีการใช้เทคนิค Stacking เพื่อพัฒนาและปรับปรุงประสิทธิภาพของตัวแบบจำลองการเรียนรู้ โดยทำการเปรียบเทียบผลลัพธ์ของตัวแบบที่แตกต่างกันผ่านการใช้วิธีการเพิ่มตัวอย่างของกลุ่มข้อมูลที่มีสัดส่วนไม่สมดุล (Imbalanced Data) ด้วยเทคนิคการสุ่มตัวอย่าง เช่น SMOTE, ADASYN, SMOTE + ENN และ SMOTE + Tomek ทั้งนี้ ผลลัพธ์ที่ได้รับจากตัวแบบแต่ละตัวได้รับการประเมินผ่านตัวชี้วัดที่สำคัญ ได้แก่ Accuracy, Precision, Recall, F-measure และ AUCs

เพื่อประเมินประสิทธิภาพของโมเดลในการพยากรณ์ความเสี่ยงด้านเครดิต งานวิจัยนี้ได้เปรียบเทียบโมเดลหลายประเภทภายใต้การจัดการปัญหาคลาสไม่สมดุล (Imbalanced Classification) ด้วยเทคนิคต่าง ๆ ได้แก่ SMOTE, ADASYN, SMOTE + ENN และ SMOTE + Tomek โดยประเมินจากค่าความแม่นยำ (Accuracy), ความแม่นยำเฉพาะคลาส (Precision), ค่าความครอบคลุม (Recall), ค่า F1 Score และค่า AUCs รวมถึงนำค่า Baseline Accuracy จากการ

ทำนายด้วยค่าที่พบมากที่สุด (Mode) มาใช้เป็นตัวเปรียบเทียบเบื้องต้นเพื่อประเมินว่าโมเดลสามารถให้ผลลัพธ์ที่ดีกว่าแบบจำลองพื้นฐานหรือไม่

ผลการวิเคราะห์พบว่า ค่า Baseline Accuracy อยู่ที่ 0.734 ซึ่งเป็นเกณฑ์ขั้นต่ำที่โมเดลควรจะสามารถทำได้ดีกว่า อย่างไรก็ตาม มีโมเดลหลายตัวที่ให้ค่า Accuracy ต่ำกว่าค่านี้ โดยเฉพาะในกรณีที่ไม่ได้มีการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งแสดงให้เห็นว่าแม้โมเดลจะมีความซับซ้อนมากขึ้น แต่ก็ไม่ได้หมายความว่าจะมีประสิทธิภาพดีขึ้นเสมอไป ดังตารางที่ 4-3

ตารางที่ 4-3 ผลลัพธ์ของตัวแบบพื้นฐาน

Model		SMOTE					ADASYN				
		Accuracy	Precision	Recall	F-measure	AUCs	Accuracy	Precision	Recall	F-measure	AUCs
Decision Tree	Full Model	0.680	0.343	0.222	0.270	0.610	0.680	0.343	0.222	0.270	0.610
	Stepwise	0.522	0.319	0.704	0.439	0.640	0.522	0.319	0.704	0.439	0.640
Support Vector Machine	Full Model	0.424	0.294	0.833	0.435	0.570	0.424	0.294	0.833	0.435	0.570
	Stepwise	0.557	0.339	0.704	0.458	0.670	0.557	0.339	0.704	0.458	0.670
Gradient Boosting	Full Model	0.709	0.456	0.481	0.468	0.720	0.709	0.456	0.481	0.468	0.720
	Stepwise	0.665	0.397	0.500	0.443	0.690	0.665	0.397	0.500	0.443	0.690
K-Nearest Neighbor	Full Model	0.606	0.345	0.537	0.420	0.610	0.606	0.345	0.537	0.420	0.610
	Stepwise	0.631	0.376	0.593	0.460	0.630	0.631	0.376	0.593	0.460	0.630
Naïve Bayes	Full Model	0.365	0.291	0.963	0.446	0.620	0.365	0.291	0.963	0.446	0.620
	Stepwise	0.507	0.331	0.833	0.474	0.710	0.507	0.331	0.833	0.474	0.710

ตารางที่ 4-3 ผลลัพธ์ของตัวแบบพื้นฐาน **(ต่อ)**

Model		SMOTE+ENN					SMOTE + Tomek				
		Accuracy	Precision	Recall	F-measure	AUCs	Accuracy	Precision	Recall	F-measure	AUCs
Decision Tree	Full Model	0.542	0.321	0.648	0.429	0.620	0.665	0.375	0.389	0.382	0.610
	Stepwise	0.488	0.305	0.722	0.429	0.610	0.502	0.303	0.667	0.416	0.620
Support Vector Machine	Full Model	0.414	0.296	0.870	0.441	0.590	0.429	0.299	0.852	0.442	0.570
	Stepwise	0.512	0.338	0.870	0.487	0.680	0.562	0.348	0.741	0.473	0.670
Gradient Boosting	Full Model	0.635	0.381	0.593	0.464	0.690	0.724	0.481	0.463	0.472	0.720
	Stepwise	0.635	0.396	0.704	0.507	0.710	0.645	0.355	0.407	0.379	0.680
K-Nearest Neighbor	Full Model	0.557	0.324	0.611	0.423	0.580	0.635	0.375	0.556	0.448	0.620
	Stepwise	0.571	0.343	0.667	0.453	0.660	0.626	0.375	0.611	0.465	0.620
Naïve Bayes	Full Model	0.360	0.291	0.981	0.449	0.630	0.365	0.291	0.963	0.446	0.620
	Stepwise	0.498	0.333	0.889	0.485	0.700	0.507	0.331	0.833	0.474	0.710

จากผลลัพธ์ของตัวแบบพื้นฐานเมื่อใช้วิธีการคัดเลือกตัวแปรแบบ Stepwise Selection ร่วมกับเทคนิคการปรับสมดุลข้อมูล พบว่าหลายโมเดลสามารถเพิ่มค่า Recall และ F1 Score ได้อย่างมีนัยสำคัญ โดยเฉพาะโมเดล Naïve Bayes, Support Vector Machine และ K-Nearest Neighbor ที่สามารถให้ค่า Recall สูงเกินกว่า 0.80 ในหลายกรณี ซึ่งบ่งชี้ว่าเหมาะกับสถานการณ์ที่การตรวจจับกลุ่มเสี่ยงมีความสำคัญมากกว่าการลด False Positive และ ในบางกรณี โมเดลพื้นฐานกลับให้ค่า Accuracy ต่ำกว่าค่า Baseline Accuracy ซึ่งได้จากการทำนายด้วยคลาสที่พบบ่อยที่สุด (mode) แสดงให้เห็นว่า โมเดลบางตัวไม่มีประสิทธิภาพเพียงพอ แม้จะผ่านกระบวนการเรียนรู้แบบมีผู้สอน (supervised learning) และใช้เทคนิคขั้นสูงในการปรับสมดุลข้อมูล จึงมีความจำเป็นต้องพัฒนาแนวทางที่สามารถ รวมจุดแข็งของแต่ละโมเดลพื้นฐานและลดจุดอ่อนของแต่ละตัว ซึ่งนำไปสู่การประยุกต์ใช้เทคนิค Stacking Ensemble Learning ที่สามารถนำผลลัพธ์ของโมเดลพื้นฐานหลายตัวมาใช้เป็นข้อมูลนำเข้าของโมเดลระดับบน (Meta-learner) เพื่อสร้างการตัดสินใจที่แม่นยำและรอบด้านมากยิ่งขึ้น และผลลัพธ์ที่ได้แสดงดังตารางที่ 4-4

ตารางที่ 4-4 ผลลัพธ์ของตัวแบบ Stacking (META Models)

Model		SMOTE					ADASYN				
		Accuracy	Precision	Recall	F-measure	AUCs	Accuracy	Precision	Recall	F-measure	AUCs
META-LR	Full Model	0.805	0.789	0.831	0.809	0.890	0.814	0.839	0.791	0.814	0.900
	Stepwise	0.791	0.769	0.831	0.799	0.860	0.763	0.753	0.788	0.770	0.870
META-GB	Full Model	0.748	0.759	0.723	0.741	0.790	0.769	0.774	0.779	0.776	0.860
	Stepwise	0.795	0.796	0.791	0.793	0.850	0.777	0.800	0.742	0.770	0.840
META-XGB	Full Model	0.761	0.769	0.743	0.756	0.830	0.733	0.744	0.734	0.739	0.820
	Stepwise	0.734	0.738	0.723	0.730	0.820	0.743	0.761	0.715	0.737	0.830
META-MLP	Full Model	0.808	0.794	0.831	0.812	0.890	0.831	0.831	0.842	0.837	0.910
	Stepwise	0.795	0.764	0.851	0.805	0.860	0.800	0.773	0.854	0.811	0.880

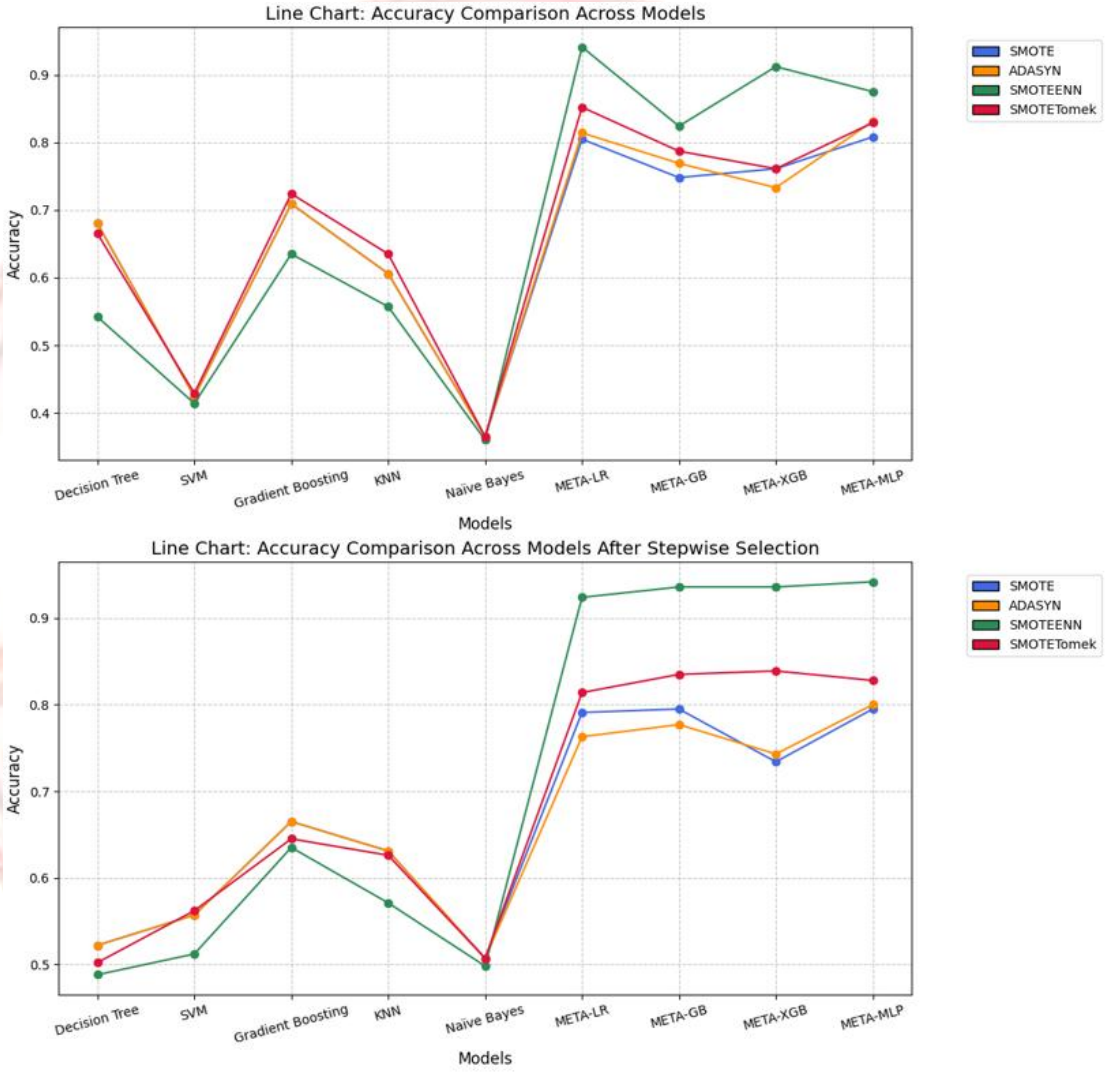
ตารางที่ 4-4 (ต่อ)

Model		SMOTE + ENN					SMOTE + Tomek				
		Accuracy	Precision	Recall	F-measure	AUCs	Accuracy	Precision	Recall	F-measure	AUCs
META-LR	Full Model	0.941	0.922	0.973	0.947	0.980	0.852	0.833	0.878	0.855	0.920
	Stepwise	0.924	0.894	0.990	0.940	0.980	0.814	0.795	0.845	0.819	0.900
META-GB	Full Model	0.824	0.827	0.849	0.838	0.920	0.787	0.791	0.779	0.785	0.850
	Stepwise	0.936	0.910	0.990	0.948	0.990	0.835	0.819	0.859	0.839	0.890
META-XGB	Full Model	0.912	0.868	0.986	0.923	0.940	0.761	0.770	0.741	0.755	0.810
	Stepwise	0.936	0.910	0.990	0.948	0.990	0.839	0.833	0.845	0.839	0.880
META-MLP	Full Model	0.875	0.850	0.932	0.889	0.900	0.829	0.791	0.893	0.839	0.920
	Stepwise	0.942	0.918	0.990	0.953	0.990	0.828	0.804	0.866	0.834	0.900

จากตารางที่ 4-4 แสดงให้เห็นผลลัพธ์จากค่าความแม่นยำ (Accuracy), ค่าความถูกต้องของการทำนาย (Precision), ค่าความสามารถในการตรวจจับกลุ่มเป้าหมาย (Recall), ค่าความสมดุลระหว่าง Precision และ Recall (F-measure) และค่าพื้นที่ใต้โค้ง ROC (AUCs) เมื่อใช้เทคนิค Stacking ในการรวมตัวแบบพื้นฐานเข้าด้วยกัน พบว่าผลลัพธ์ของตัวแบบ META-learners ได้แก่ META-LR (Logistic Regression), META-GB (Gradient Boosting), META-XGB (Extreme Gradient Boosting) และ META-MLP (Multi-layer Perceptron) มีค่าประสิทธิภาพที่ดีขึ้นในทุกตัวชี้วัด ตัวแบบ META-LR, META-GB, META-XGB และ META-MLP มีค่าความแม่นยำที่สูงขึ้นเมื่อเทียบกับตัวแบบพื้นฐาน และยังสามารถรักษาสมดุลระหว่างค่าตัวชี้วัดต่างๆ ได้ดีขึ้น เช่น ค่าความสามารถในการตรวจจับกลุ่มเป้าหมาย (Recall) และค่าความถูกต้องของการทำนาย (Precision) ทำให้ตัวแบบเหล่านี้มีประสิทธิภาพในการจำแนกประเภทข้อมูลมากขึ้น การใช้เทคนิคการเพิ่มข้อมูล เช่น SMOTE, ADASYN, SMOTE +ENN และ SMOTE + Tomek มีผลต่อค่าชี้วัดของตัวแบบ โดยบางเทคนิคช่วยเพิ่มค่า Recall ในขณะที่บางเทคนิคช่วยเพิ่มค่า Precision หรือ Accuracy โดยรวม ซึ่งแสดงให้เห็นถึงผลกระทบที่แตกต่างกันของแต่ละวิธีต่อผลลัพธ์ของตัวแบบ

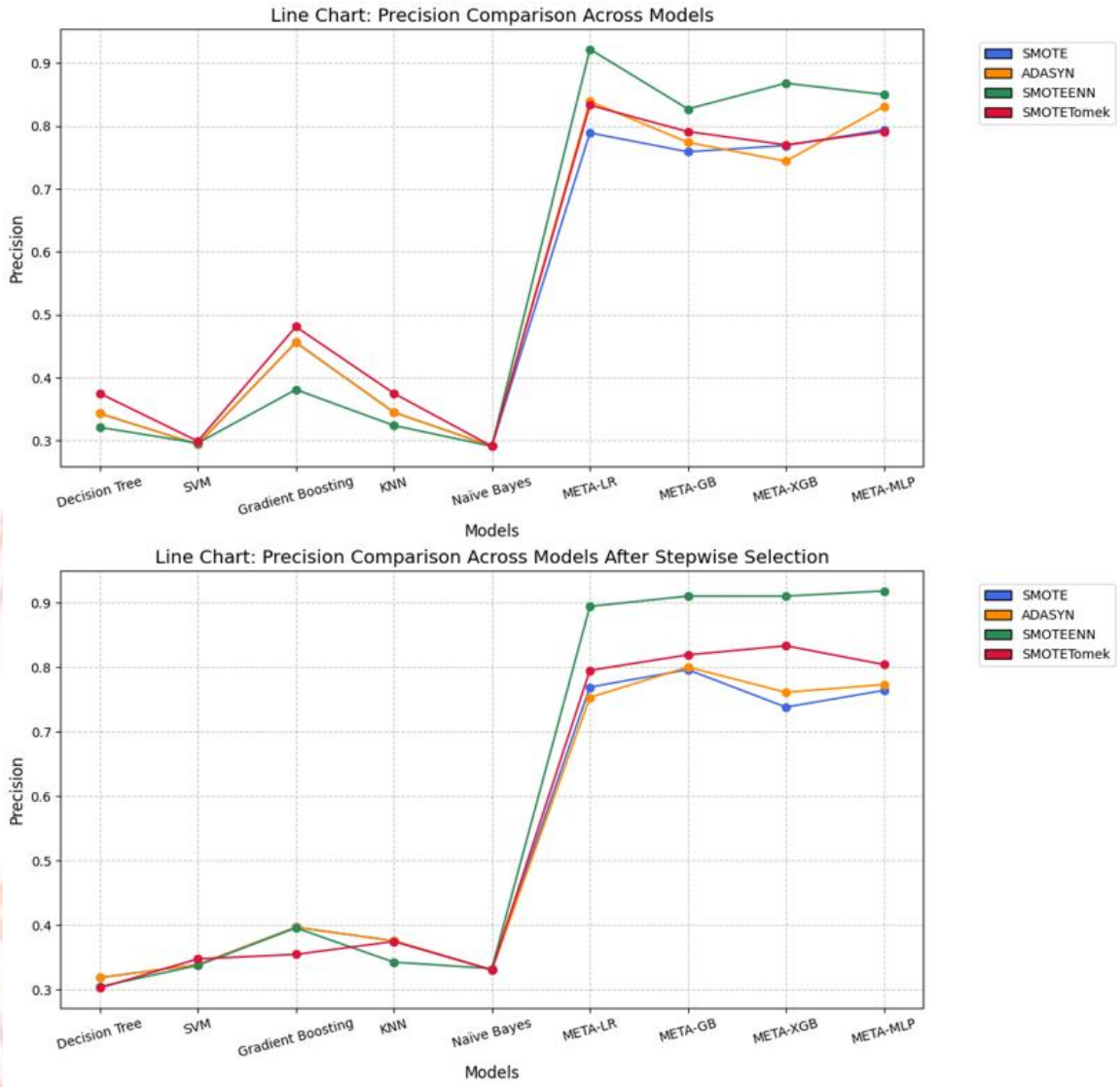
4.3 การเปรียบเทียบประสิทธิภาพของการตรวจจับความผิดปกติด้วยเทคนิค Stacking สำหรับความเสี่ยงด้านเครดิต

4.3.1 การเปรียบเทียบประสิทธิภาพระหว่างข้อมูลดั้งเดิมกับการคัดเลือกตัวแบบภายใต้เทคนิคการปรับสมดุลข้อมูลที่แตกต่างกัน



ภาพที่ 4-3 เปรียบเทียบค่าความถูกต้องของแบบจำลองพื้นฐานและแบบจำลอง Stacking ภายใต้กลยุทธ์การปรับสมดุลข้อมูล ระหว่างแบบจำลอง Full Model และแบบจำลองหลังการคัดเลือกตัวแปร

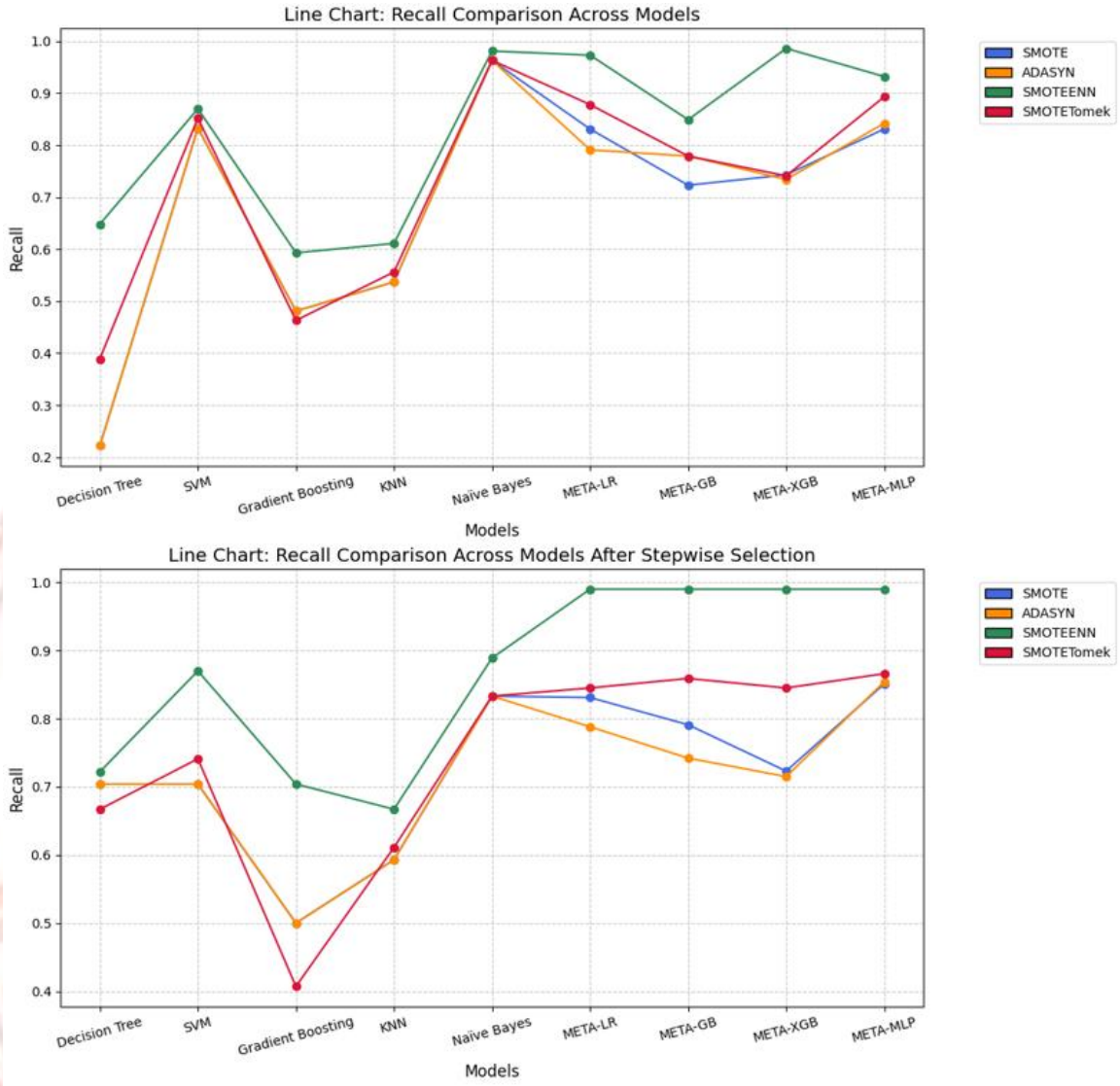
จากภาพที่ 4-3 กราฟทั้งสองแสดงให้เห็นถึงผลของการคัดเลือกตัวแปรด้วยวิธี Stepwise Selection ต่อค่าความถูกต้องของแบบจำลองประเภทต่าง ๆ โดยแบ่งออกเป็นสองช่วง คือ แบบจำลองที่ใช้ตัวแปรทั้งหมด (Full Model) และแบบจำลองหลังผ่านการคัดเลือกตัวแปร (Stepwise Model) จะเห็นได้ว่าแบบจำลอง Stacking (กลุ่ม META-Model) ยังคงให้ค่าความแม่นยำสูงกว่ากลุ่มแบบจำลองพื้นฐาน (Base Models) อย่างสม่ำเสมอในทั้งสองกรณี โดยเฉพาะ META-XGB และ META-MLP ซึ่งแสดงค่าความแม่นยำเฉลี่ยสูงกว่า 0.8 ขึ้นไปภายใต้หลายกลยุทธ์การปรับสมดุลข้อมูล และสามารถเห็นได้ชัดว่ามีลักษณะของเส้นกราฟที่ ทับซ้อนกันในบริเวณกลุ่ม Base Models เช่น Decision Tree, Support Vector Machine, Gradient Boosting, K-Nearest Neighbor ซึ่งสามารถอธิบายได้ว่า การปรับสมดุลข้อมูลมีผลต่อประสิทธิภาพของแบบจำลองพื้นฐานในทิศทางที่คล้ายคลึงกัน แม้จะใช้ตัวแปรจำนวนมากหรือผ่านการคัดเลือกแล้วก็ตาม เนื่องจากแบบจำลองพื้นฐานหลายตัว เช่น Decision Tree และ K-Nearest Neighbor มีความไวต่อการกระจายตัวของข้อมูลในคลาสย่อย (class imbalance) มากกว่าการเปลี่ยนแปลงในคุณลักษณะของตัวแปร (feature selection) ทำให้เส้นกราฟของกลยุทธ์ SMOTE, ADASYN, และ SMOTE + ENN มีแนวโน้มที่ทับซ้อนกันอยู่ในช่วงค่า Accuracy ใกล้เคียงกัน การคัดเลือกตัวแปรช่วยปรับความเสถียรของผลลัพธ์ในบางโมเดลพื้นฐาน และยังคงเสริมความโดดเด่นของแบบจำลอง Stacking



ภาพที่ 4-4 เปรียบเทียบค่าความแม่นยำของแบบจำลองพื้นฐานและแบบจำลอง Stacking ภายใต้กลยุทธ์การปรับสมดุลข้อมูล ระหว่างแบบจำลอง Full Model และแบบจำลองหลังการคัดเลือกตัวแปร

จากภาพที่ 4-4 กราฟทั้งสองกราฟทั้งสองแสดงให้เห็นถึงผลของการคัดเลือกตัวแปรด้วยวิธี Stepwise Selection ต่อ ค่าความแม่นยำ (Precision) ของแบบจำลองประเภทต่าง ๆ โดยแบ่งออกเป็นสองช่วง ได้แก่ แบบจำลองที่ใช้ตัวแปรทั้งหมด (Full Model) และแบบจำลองหลังผ่านการคัดเลือกตัวแปร (Stepwise Model) จะเห็นได้ว่าแบบจำลอง Stacking (กลุ่ม META-Model) ยังคงให้ค่าความแม่นยำสูงกว่ากลุ่มแบบจำลองพื้นฐาน (Base Models) อย่างสม่ำเสมอในทั้งสองกรณี โดยเฉพาะ META-XGB และ META-MLP ซึ่งมีค่า Precision สูงกว่า 0.8 อย่างต่อเนื่องภายใต้ทุกกลยุทธ์การปรับสมดุลข้อมูล ในกลุ่มแบบจำลองพื้นฐาน (เช่น Decision Tree, Support Vector

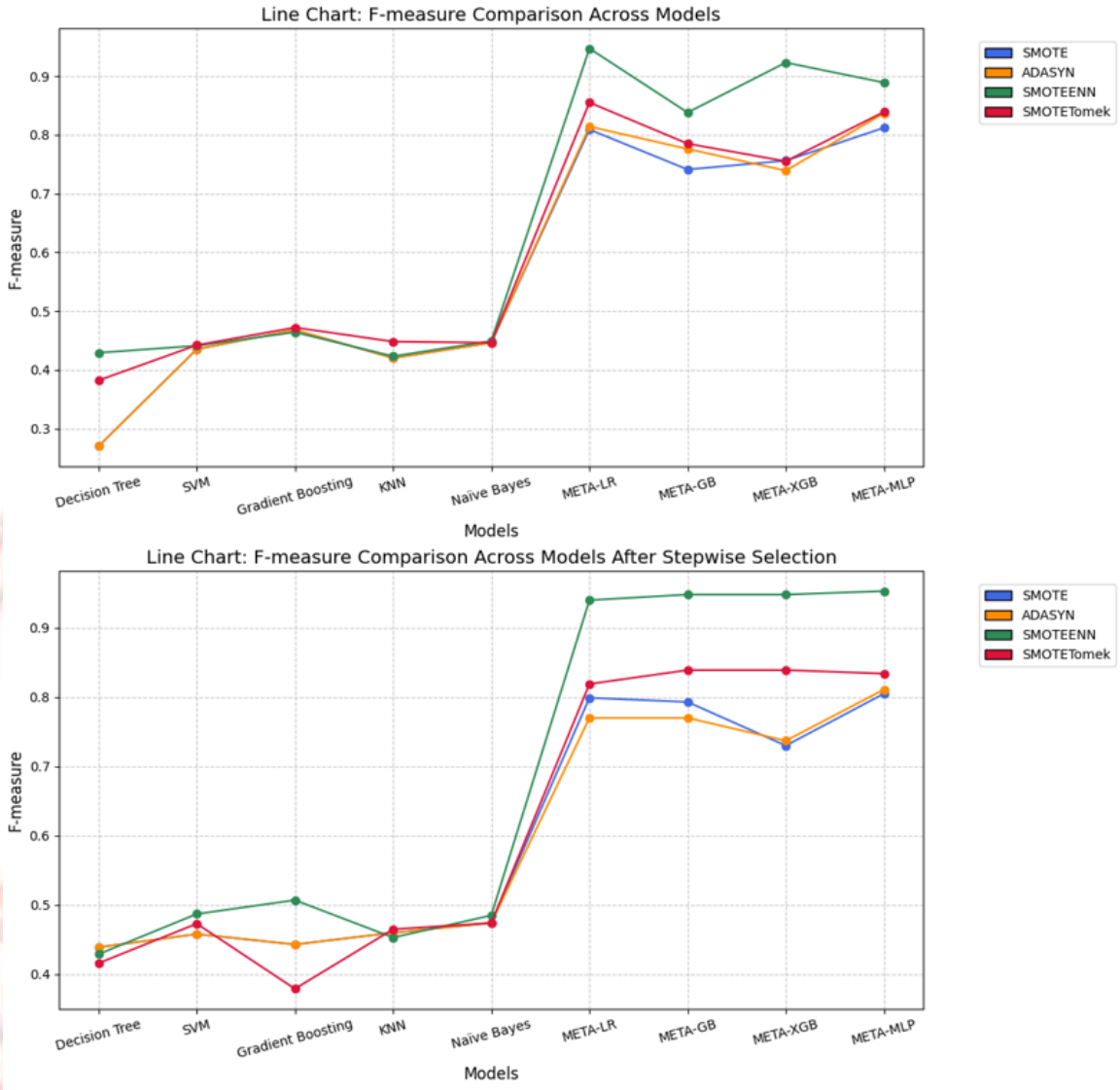
Machine, Gradient Boosting และ K-Nearest Neighbor) สังเกตได้ว่าค่า Precision มีแนวโน้มใกล้เคียงกันระหว่างแต่ละวิธีการปรับสมดุลข้อมูล โดยเฉพาะในช่วงค่าที่ต่ำกว่า 0.5 ซึ่งเห็นได้ชัดว่าลักษณะของเส้นกราฟมีการทับซ้อนกัน ในกลุ่ม Base Models อย่างชัดเจน สามารถอธิบายได้ว่า กลยุทธ์การปรับสมดุลข้อมูลมีผลต่อ Precision ของแบบจำลองพื้นฐานในทิศทางที่คล้ายคลึงกัน โดยเฉพาะในกรณีที่แบบจำลองมีความไวต่อการเปลี่ยนแปลงของอัตราส่วนข้อมูลในแต่ละคลาสมากกว่าคุณลักษณะของตัวแปร เช่น Support Vector Machine และ Naive Bayes ซึ่งไม่ได้รับประโยชน์อย่างชัดเจนจากการคัดเลือกตัวแปร จึงมีค่า Precision ที่ไม่แตกต่างกันมากนักไม่ว่าจะเป็น Full Model หรือ Stepwise Model ในขณะเดียวกัน การคัดเลือกตัวแปรช่วยลด Noise และเพิ่มเสถียรภาพให้กับแบบจำลอง Stacking



ภาพที่ 4-5 เปรียบเทียบค่าความจำของแบบจำลองพื้นฐานและแบบจำลอง Stacking ภายใต้กลยุทธ์การปรับสมดุลข้อมูล ระหว่างแบบจำลอง Full Model และแบบจำลองหลังการคัดเลือกตัวแปร

จากภาพที่ 4-5 กราฟทั้งสองแสดงให้เห็นถึงผลของการคัดเลือกตัวแปรด้วยวิธี Stepwise Selection ต่อค่าความสามารถในการจำ (Recall) ของแบบจำลองประเภทต่าง ๆ โดยแบ่งออกเป็นสองช่วง คือแบบจำลองที่ใช้ตัวแปรทั้งหมด (Full Model) และแบบจำลองหลังผ่านการคัดเลือกตัวแปร (Stepwise Model) จะเห็นได้ว่าแบบจำลอง Stacking (กลุ่ม META-Model) ยังคงให้ค่า Recall ที่สูงกว่ากลุ่มแบบจำลองพื้นฐาน (Base Models) อย่างสม่ำเสมอในทั้งสองกรณี โดยเฉพาะ META-XGB และ META-MLP ซึ่งแสดงค่า Recall เฉลี่ยใกล้เคียง 1.0 ภายใต้หลายกลยุทธ์การปรับสมดุลข้อมูล แสดงให้เห็นถึงศักยภาพของการผสมแบบจำลองหลายตัวร่วมกัน (Stacking) ที่สามารถ

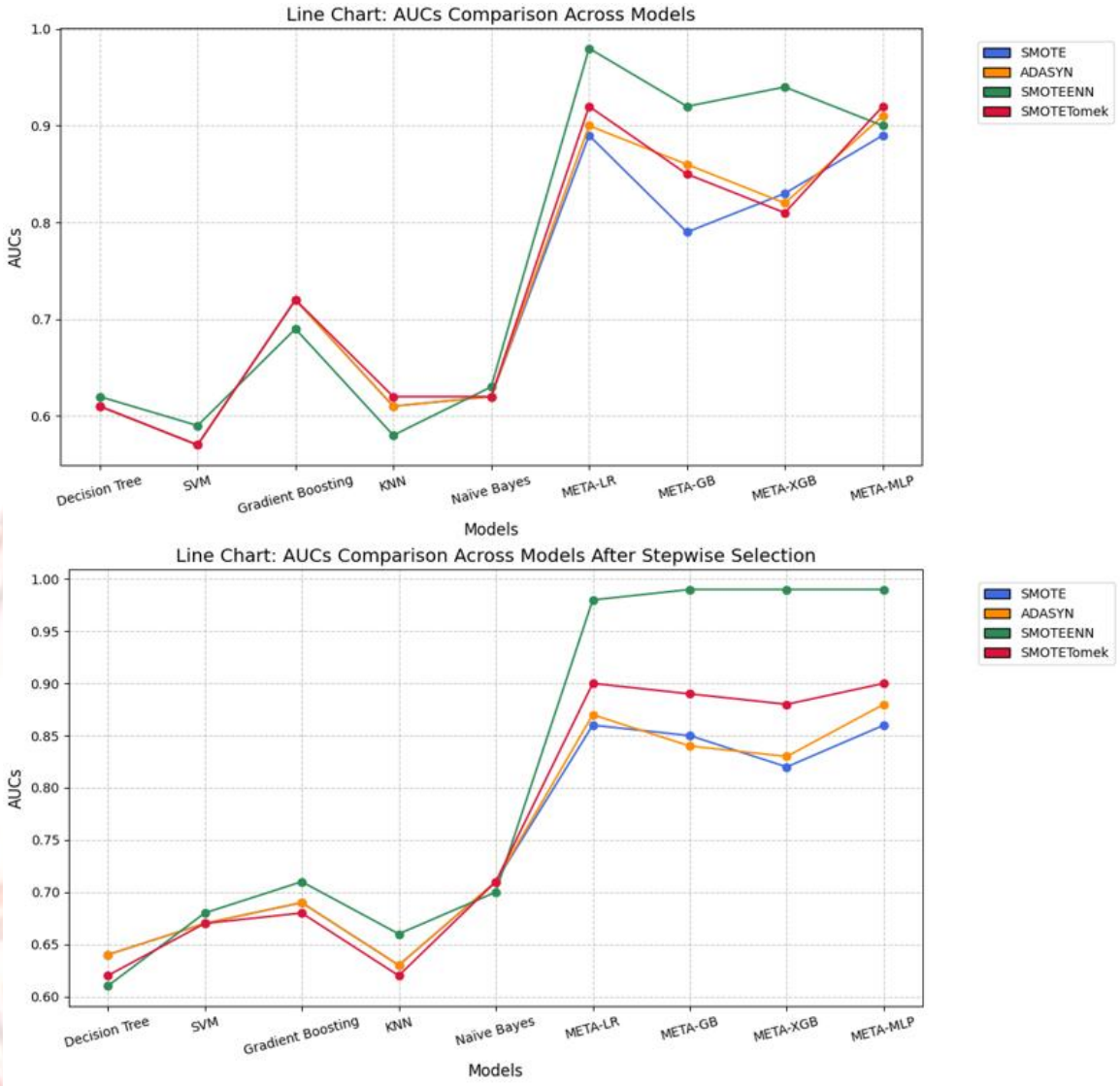
จับข้อมูลคลาสเป้าหมายได้อย่างมีประสิทธิภาพ แม้จะอยู่ภายใต้ชุดข้อมูลที่ไม่สมดุลในขณะที่แบบจำลองพื้นฐาน เช่น Decision Tree, Support Vector Machine, Gradient Boosting และ K-Nearest Neighbor มีแนวโน้มของกราฟ Recall ที่ ทับซ้อนกันอย่างชัดเจน ระหว่างกลยุทธ์การปรับสมดุลข้อมูล โดยเฉพาะในช่วง Full Model การทับซ้อนของเส้นกราฟเหล่านี้สะท้อนให้เห็นว่า กลยุทธ์การปรับสมดุลข้อมูลมีผลต่อการเพิ่มค่า Recall ของแบบจำลองพื้นฐานในทิศทางที่คล้ายคลึงกัน ซึ่งสามารถอธิบายได้ว่า แบบจำลองพื้นฐานเหล่านี้มีความไวต่อการกระจายตัวของข้อมูลในคลาสย่อย (Class Imbalance) มากกว่าความเปลี่ยนแปลงในชุดคุณลักษณะ (Feature Subset) ทำให้การคัดเลือกตัวแปรมีผลกระทบเพียงเล็กน้อยต่อค่า Recall เมื่อเทียบกับผลของการใช้เทคนิคอย่าง SMOTE, ADASYN, หรือ SMOTE + ENN อย่างไรก็ตาม หลังจากทำ Stepwise Selection พบว่าค่า Recall ของบางแบบจำลองมีความเสถียรและดีขึ้น โดยเฉพาะแบบจำลองในกลุ่ม META-Learners ที่ยังคงรักษาค่า Recall ให้อยู่ในระดับสูงได้อย่างต่อเนื่อง และบางกลยุทธ์ เช่น SMOTE + ENN ยังให้ผลลัพธ์ที่โดดเด่นกว่ากลยุทธ์อื่นในหลายโมเดลพื้นฐาน ซึ่งสะท้อนถึงประสิทธิภาพในการจัดการกับข้อมูล Noise และ Outlier ได้อย่างเหมาะสม กราฟนี้เน้นย้ำว่า การปรับสมดุลข้อมูลมีบทบาทสำคัญมากกว่าการเลือกตัวแปรเพียงอย่างเดียว ในการยกระดับค่า Recall ของแบบจำลองพื้นฐาน และเมื่อผสานเข้ากับเทคนิคการเรียนรู้แบบ Stacking จะสามารถส่งเสริมให้โมเดลมีความสามารถในการตรวจจับข้อมูลกลุ่มเป้าหมายได้อย่างมีประสิทธิภาพสูงสุด



ภาพที่ 4-6 เปรียบเทียบมาตรวัดเอฟของแบบจำลองพื้นฐานและแบบจำลอง Stacking ภายใต้กลยุทธ์การปรับสมดุลข้อมูล ระหว่างแบบจำลอง Full Model และแบบจำลองหลังการคัดเลือกตัวแปร

จากภาพที่ 4-6 กราฟทั้งสองสามารถเห็นได้ชัดเจนว่า การคัดเลือกตัวแปรด้วยวิธี Stepwise Selection ส่งผลต่อค่าประสิทธิภาพของแบบจำลองในแง่ของ F-measure ซึ่งเป็นตัวชี้วัดที่ผสานความสมดุลระหว่าง Precision และ Recall จากการเปรียบเทียบพบว่า แบบจำลองกลุ่ม META-Models ซึ่งใช้เทคนิค Stacking ยังคงให้ค่า F-measure ที่สูงกว่าแบบจำลองพื้นฐานอย่างสม่ำเสมอ ไม่ว่าจะใช้กลยุทธ์การปรับสมดุลข้อมูลแบบใด โดยเฉพาะ META-LR และ META-XGB แสดงผลลัพธ์ที่โดดเด่นในทุกกลยุทธ์ ทั้งใน Full Model และ Stepwise Model ซึ่งสะท้อนถึงความสามารถของการผสานโมเดลหลายตัวในการลดผลกระทบจากปัญหาความไม่สมดุลของข้อมูล และยังรักษาความ

แม่นยำและความสามารถในการตรวจจับข้อมูลเป้าหมายได้อย่างสมดุล ในขณะที่แบบจำลองพื้นฐาน เช่น Decision Tree, Support Vector Machine, KNN และ Gradient Boosting มีค่า F-measure ที่ค่อนข้างใกล้เคียงกัน โดยเฉพาะในช่วง Full Model แสดงให้เห็นว่า กลยุทธ์การปรับสมดุลข้อมูลมีผลต่อการเพิ่มค่า F-measure ของแบบจำลองพื้นฐานในทิศทางที่คล้ายคลึงกัน นอกจากนี้ การคัดเลือกตัวแปรยังไม่ส่งผลอย่างมีนัยสำคัญต่อแบบจำลองพื้นฐานเหล่านี้เมื่อเทียบกับผลของการใช้เทคนิคปรับสมดุลข้อมูล อย่างไรก็ตาม เทคนิค SMOTE + ENN แสดงผลลัพธ์ที่โดดเด่นในหลายโมเดลพื้นฐาน โดยเฉพาะหลังจาก Stepwise Selection ซึ่งสะท้อนถึงความสามารถในการจัดการกับข้อมูลผิดปกติและ outlier ได้ดี ทำให้ค่า F-measure มีแนวโน้มเพิ่มขึ้นหรือมีความเสถียรมากขึ้น กราฟทั้งสองแสดงให้เห็นว่า แม้การคัดเลือกตัวแปรจะมีบทบาทในการปรับปรุงผลลัพธ์ในบางกรณี แต่ กลยุทธ์การปรับสมดุลข้อมูลยังคงเป็นปัจจัยสำคัญที่ส่งผลต่อ F-measure ของแบบจำลองอย่างมีนัยสำคัญ และเมื่อผสานกับเทคนิคการเรียนรู้แบบ Stacking ก็จะยิ่งส่งเสริมให้โมเดลสามารถจำแนกข้อมูลกลุ่มเป้าหมายได้อย่างมีประสิทธิภาพและสมดุลยิ่งขึ้น



ภาพที่ 4-7 เปรียบเทียบค่าพื้นที่ใต้โค้งของแบบจำลองพื้นฐานและแบบจำลอง Stacking ภายใต้กลยุทธ์การปรับสมดุลข้อมูล ระหว่างแบบจำลอง Full Model และแบบจำลองหลังการคัดเลือกตัวแปร

จากภาพที่ 4-7 กราฟทั้งสองแสดงให้เห็นถึงผลของการคัดเลือกตัวแปรด้วยวิธี Stepwise Selection ต่อค่าความสามารถในการจำแนกกลุ่มข้อมูลด้วยค่า AUC (Area Under the Curve) ของแบบจำลองประเภทต่าง ๆ โดยเปรียบเทียบระหว่างแบบจำลองที่ใช้ตัวแปรทั้งหมด (Full Model) กับแบบจำลองหลังผ่านการคัดเลือกตัวแปร (Stepwise Model) จะเห็นได้ว่าแบบจำลองกลุ่ม Stacking (META-Model) ยังคงแสดงค่า AUC ที่สูงกว่าแบบจำลองพื้นฐาน (Base Models) อย่างต่อเนื่องในทั้งสองกรณีโดยเฉพาะโมเดล META-LR, META-XGB และ META-MLP ที่แสดงผลลัพธ์ AUC ใกล้เคียง 1.0 ในหลายกลยุทธ์การปรับสมดุลข้อมูล สะท้อนถึงศักยภาพของเทคนิคการรวม

แบบจำลองหลายตัว (Stacking) ในการเพิ่มความแม่นยำในการจำแนกกลุ่มเป้าหมาย แม้อยู่ภายใต้ข้อมูลที่ไม่สมดุล ในขณะที่แบบจำลองพื้นฐาน เช่น Decision Tree, Support Vector Machine, Gradient Boosting และ K-Nearest Neighbor มีแนวโน้มของกราฟ AUC ที่ใกล้เคียงและทับซ้อนกันในหลายกลยุทธ์ โดยเฉพาะในช่วง Full Model สะท้อนให้เห็นว่า กลยุทธ์การปรับสมดุลข้อมูลส่งผลต่อค่า AUC ของแบบจำลองพื้นฐานในลักษณะที่คล้ายคลึงกัน นั่นแสดงว่าแบบจำลองพื้นฐานตอบสนองต่อการปรับสมดุลข้อมูลได้ดี แต่ยังมีข้อจำกัดในการเพิ่มศักยภาพโดยลำพัง หากไม่มีการรวมผลจากโมเดลอื่นหรือการปรับแต่งที่ลึกขึ้น เมื่อผ่านกระบวนการ Stepwise Selection พบว่า ค่า AUC ของบางแบบจำลองมีความเสถียรมากขึ้น โดยเฉพาะกลุ่ม META-Models ที่ยังสามารถรักษาระดับค่า AUC ให้อยู่ในระดับสูงได้อย่างสม่ำเสมอ แสดงถึงความยืดหยุ่นของโมเดลกลุ่มนี้ต่อการลดจำนวนตัวแปร ขณะเดียวกันบางกลยุทธ์ เช่น SMOTE + ENN และ SMOTE + Tomek ยังให้ผลลัพธ์ที่เด่นชัดในหลายแบบจำลอง ทั้ง Base และ META ซึ่งแสดงถึงความสามารถในการจัดการกับข้อมูลไม่สมดุลและ Noise ได้อย่างมีประสิทธิภาพ กราฟชุดนี้เน้นให้เห็นว่า การปรับสมดุลข้อมูลยังคงมีอิทธิพลสำคัญต่อประสิทธิภาพของแบบจำลองมากกว่าการเลือกตัวแปรเพียงอย่างเดียว และเมื่อนำมาผสานกับเทคนิค Stacking จะสามารถเพิ่มศักยภาพในการจำแนกข้อมูลกลุ่มเป้าหมายได้อย่างมีประสิทธิภาพ

4.3.2 การเปรียบเทียบประสิทธิภาพของแบบจำลองภายใต้เทคนิคการปรับสมดุลข้อมูลที่เหมาะสมที่สุดร่วมกับการคัดเลือกตัวแปรแบบ Stepwise

ตารางที่ 4-5 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองภายใต้การปรับสมดุลข้อมูลและการคัดเลือกตัวแปร

Model	Accuracy	Precision	Recall	F-measure	AUCs
Baseline (Mode)	0.734	-	-	-	-
Decision Tree	0.488	0.305	0.722	0.429	0.610
Support Vector Machine	0.512	0.338	0.870	0.487	0.680
Gradient Boosting	0.635	0.396	0.704	0.507	0.710
K-Nearest Neighbor	0.571	0.343	0.667	0.453	0.660
Naïve Bayes	0.498	0.333	0.889	0.485	0.700
META-LR	0.924	0.894	0.990	0.940	0.980
META-GB	0.936	0.910	0.990	0.948	0.990
META-XGB	0.936	0.910	0.990	0.948	0.990
META-MLP	0.942	0.918	0.990	0.953	0.990

จากตารางที่ 4-5 ตารางนี้แสดงผลการประเมินประสิทธิภาพของแบบจำลองพื้นฐานและแบบจำลอง Stacking ภายใต้การปรับสมดุลข้อมูลด้วยเทคนิค SMOTE + ENN ซึ่งจากการทดลองพบว่าให้ผลลัพธ์ดีที่สุดในภาพรวม และได้ทำการคัดเลือกตัวแปรด้วยวิธี Stepwise เพื่อเพิ่มความกระชับของข้อมูลป้อนเข้า โดยมีการเปรียบเทียบผลลัพธ์ด้วยตัวชี้วัด ได้แก่ Accuracy, Precision, Recall, F-measure และ AUC

จากตารางพบว่าแบบจำลองพื้นฐาน (Base Models) ทั้งหมดมีค่า Accuracy ต่ำกว่าค่า baseline (0.734) อย่างมีนัยสำคัญ โดย baseline ในที่นี้คือโมเดลที่ทำนายเฉพาะคลาสที่มีจำนวนมากที่สุด (Mode Classifier) ซึ่งแม้จะให้ค่า Accuracy สูง แต่ไม่สามารถสะท้อนความสามารถในการจำแนกคลาสเป้าหมายได้จริง การที่แบบจำลองพื้นฐานมีค่า Accuracy ต่ำกว่า baseline สะท้อนถึงปัญหา Class Imbalance ที่รุนแรง และความสามารถจำแนกคลาสส่วนน้อยที่ยังจำกัด แม้บางโมเดล เช่น

Support Vector Machine และ Naïve Bayes จะมีค่า Recall ค่อนข้างสูง 0.870 และ 0.889 ตามลำดับ แต่กลับมีค่า Precision และ F-measure ต่ำมาก แสดงให้เห็นถึงการเกิด False Positive จำนวนมาก ซึ่งส่งผลให้สมรรถนะโดยรวมของโมเดลยังไม่สมดุล

ในทางกลับกัน แบบจำลองในกลุ่ม META-Learners ที่ใช้เทคนิค Stacking ได้แก่ META-LR, META-GB, META-XGB และ META-MLP สามารถยกระดับค่าชี้วัดทุกตัวได้อย่างมีนัยสำคัญ โดยเฉพาะค่า F-measure และ AUC ซึ่งอยู่ในช่วง 0.940–0.953 และ 0.980–0.990 ตามลำดับ แสดงให้เห็นถึงความสามารถในการจำแนกทั้งสองคลาสได้อย่างครอบคลุมและแม่นยำ โดยเฉพาะ META-MLP ที่ให้ผลลัพธ์สูงสุดเกือบทุกมิติ

ผลการทดลองนี้เน้นย้ำถึงประสิทธิภาพของแนวทางการพัฒนาโมเดลด้วยการผสมผสานเทคนิคต่าง ๆ อย่างเป็นระบบ ทั้งการปรับสมดุลข้อมูล การเลือกตัวแปรที่เหมาะสม และการเรียนรู้แบบ Stacking ซึ่งช่วยยกระดับประสิทธิภาพของโมเดลให้สามารถรับมือกับข้อมูลที่ไม่สมดุลและมีโครงสร้างซับซ้อนได้อย่างมีประสิทธิภาพ

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องด้วยเทคนิค Stacking สำหรับการประเมินความเสี่ยงสินเชื่อ SME โดยเน้นศึกษาผลกระทบของกลยุทธ์การปรับสมดุลข้อมูลที่แตกต่างกัน และเปรียบเทียบประสิทธิภาพของแบบจำลองที่พัฒนาขึ้นกับแบบจำลองพื้นฐานอื่น ๆ ทั้งนี้ การวิเคราะห์ดำเนินการบนฐานข้อมูลผู้กู้จากกองทุนส่งเสริมอุตสาหกรรมในครอบครัวและหัตถกรรมไทย ซึ่งอยู่ภายใต้กระทรวงอุตสาหกรรม โดยข้อมูลที่ใช้มีลักษณะไม่สมดุลระหว่างกลุ่มผู้กู้ที่ผิดนัดชำระหนี้ (NPL) และไม่ผิดนัดชำระหนี้ ซึ่งเป็นปัจจัยสำคัญที่ส่งผลต่อการประเมินความแม่นยำของแบบจำแนกประเภท (Classification Models)

จากผลการทดลองพบว่า ค่าความแม่นยำพื้นฐาน (Baseline Accuracy) ของการทำนายด้วยวิธี Mode Classifier อยู่ที่ 73.4% ซึ่งแม้จะดูเหมือนมีค่าสูง แต่เป็นเพียงการสะท้อนสัดส่วนของกลุ่มข้อมูลที่มากกว่า (Class 0) โดยไม่ได้สะท้อนความสามารถที่แท้จริงของโมเดลในการจำแนกข้อมูลกลุ่มเป้าหมาย ดังนั้น โมเดลใด ๆ ที่มีค่าความแม่นยำต่ำกว่าค่านี้จึงถือว่าขาดศักยภาพในการประเมินความเสี่ยงเชิงคุณภาพ ซึ่งจากผลการทดลองพบว่า โมเดลพื้นฐานหลายตัว เช่น Decision Tree, Gradient Boosting, K-Nearest Neighbor มีค่า Accuracy ต่ำกว่า baseline อย่างมีนัยสำคัญ แม้ว่าบางโมเดล เช่น Naïve Bayes และ SVM จะมีค่า Recall สูงก็ตาม แต่ยังมีค่า Precision และ F1-score ที่ต่ำ ซึ่งชี้ถึงปัญหา False Positive ที่สูง และประสิทธิภาพโดยรวมที่ไม่สมดุล

เมื่อมีการประยุกต์ใช้เทคนิค Stacking ผ่านกลุ่ม META-Models ได้แก่ META-LR, META-GB, META-XGB และ META-MLP ร่วมกับการปรับสมดุลข้อมูลโดยใช้เทคนิค SMOTE, ADASYN, SMOTE + ENN และ SMOTE + Tomek พบว่าโมเดลกลุ่มนี้สามารถยกระดับค่าประสิทธิภาพทุกตัวชี้วัดอย่างชัดเจน

โดยเฉพาะในกลุ่ม META-XGB และ META-MLP ที่สามารถสร้างผลลัพธ์ F1-score ได้ระหว่าง 0.940–0.953 และ AUC ระหว่าง 0.980–0.990 ซึ่งแสดงถึงความสามารถในการจำแนกข้อมูลทั้งสองคลาสได้อย่างสมดุลและแม่นยำมากยิ่งขึ้นเมื่อเทียบกับโมเดลพื้นฐานเดี่ยว นอกจากนี้ ยังพบว่าเทคนิค Stepwise Selection ในการคัดเลือกตัวแปรส่งผลให้โมเดลบางประเภท โดยเฉพาะ META-LR และ META-MLP มีค่าความเสถียรสูงขึ้นในเชิงของ F1-score และ AUC แสดงถึงการลดความซ้ำซ้อนและ noise ของข้อมูลที่ไม่จำเป็นได้อย่างมีประสิทธิภาพ

การทดลองยังแสดงให้เห็นว่ากลยุทธ์การปรับสมดุลข้อมูล เช่น SMOTE + ENN และ SMOTE + Tomek ส่งผลต่อการเพิ่มค่าประสิทธิภาพของโมเดลได้อย่างเด่นชัด โดยเฉพาะ SMOTE + ENN ซึ่งสามารถช่วยลดปัญหาข้อมูล outlier และ noise ได้ดีกว่ากลยุทธ์อื่น ๆ ส่งผลให้โมเดลมีค่าความแม่นยำและ F1-score ที่สูงขึ้นอย่างต่อเนื่อง โดยเฉพาะเมื่อผสานกับเทคนิคการเรียนรู้แบบ Stacking จะยิ่งส่งเสริมให้โมเดลสามารถจำแนกข้อมูลกลุ่มเป้าหมายที่มีความไม่สมดุลได้อย่างมีประสิทธิภาพยิ่งขึ้น

จากการวิเคราะห์ภาพรวมทั้งหมด สรุปได้ว่าแนวทางการพัฒนาแบบจำลองโดยใช้เทคนิค Stacking ร่วมกับการเลือกตัวแปรอย่างเหมาะสม และการปรับสมดุลข้อมูลอย่างมีประสิทธิภาพ ส่งผลให้โมเดลสามารถจำแนกกลุ่มผู้กู้ที่มีความเสี่ยงผิดนัดชำระหนี้ (NPL) ได้อย่างแม่นยำ และสามารถรับมือกับปัญหาข้อมูลไม่สมดุลได้ดีกว่าแบบจำลองพื้นฐานอย่างมีนัยสำคัญ โดยเฉพาะในบริบทของการประเมินสินเชื่อภาครัฐ ซึ่งต้องการความสามารถในการตรวจจับความเสี่ยงที่แม่นยำเพื่อสนับสนุนการตัดสินใจเชิงนโยบายที่มีประสิทธิภาพ

5.2 ข้อเสนอแนะ

จากผลการวิจัยเกี่ยวกับการพัฒนาแบบจำลองการเรียนรู้เชิงลึกด้วยเทคนิค Stacking สำหรับการประเมินความเสี่ยงสินเชื่อ SME มีข้อเสนอแนะที่สามารถนำไปใช้ปรับปรุงและขยายขอบเขตของการศึกษาต่อไปได้ ด้านการพัฒนาแบบจำลองเพิ่มเติมใช้เทคนิคการเลือกตัวแปรที่ลึกซึ้งขึ้นเพื่อคัดเลือกตัวแปรที่มีอิทธิพลสูงสุดต่อการทำนายความเสี่ยง การใช้เทคนิคการปรับสมดุลข้อมูลอื่น ๆ เช่น Balanced Bagging การผสานโมเดลแบบอื่น เช่น Blending หรือ Voting Classifiers เพื่อตรวจสอบว่าให้ประสิทธิภาพดีกว่า Stacking หรือไม่ การประยุกต์ใช้เทคนิค Deep Learning อื่น ๆ ในการวิเคราะห์ความสัมพันธ์เชิงซับซ้อนของตัวแปรต่าง ๆ คาดว่าจะสามารถช่วยพัฒนาแบบจำลองการเรียนรู้ของเครื่องให้มีประสิทธิภาพสูงขึ้น และสามารถนำไปใช้ในเชิงปฏิบัติได้อย่างมีประสิทธิผลมากขึ้นในอนาคต

บรรณานุกรม

ภาษาไทย

ณรงค์ศักดิ์ โค้ววิลัยแสง. การพัฒนากรอบแนวคิดการจำแนกประเภทซัพพลายเออร์สำหรับปัญหาการประเมินซัพพลายเออร์ในระบบ SAP ERP. วิทยานิพนธ์ปริญญามหาบัณฑิต, มหาวิทยาลัยเกษตรศาสตร์, 2564.

วิชญ์วิสิฐ เกสรสิทธิ์, วิชิต หล่อจีระชุณห์กุล และจิราวัลย์ จิตรถเวช. “การแก้ปัญหาข้อมูลไม่สมดุล ของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน.” วารสารวิจัย มข. [วารสารออนไลน์] 2561. [สืบค้นวันที่ 12 มกราคม 2566]. จาก https://ph02.tci-thaijo.org/index.php/ gskku/ article/view/145226/107320

ศราวุฒิ ทองเชื้อ. การทำนายสัญญาณการซื้อขายสกุลเงินดิจิทัลโดยใช้การเรียนรู้ของเครื่องและการวิเคราะห์ทางทคนิค. ปริญญานิพนธ์ปริญญามหาบัณฑิต, คณะพาณิชยศาสตร์และการบัญชี, มหาวิทยาลัยธรรมศาสตร์, 2566.

ภาษาอังกฤษ

Abdelmoula, A.K. "Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks." *Accounting and Management Information Systems*. 14(1): 79, 2015.

Ahmed, A.N. "Credit-Risk Assessment of Small Business Loans using Naïve Bayes, Decision Tree and Random Forest." PhD dissertation, National College of Ireland, Dublin, 2019.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. (1974).

Akaike, H. Information theory and an extension of the maximum likelihood principle. *In B. N. Petrov & F. Csaki (Eds.), Second International Symposium on Information Theory* (pp. 267–281). Akademiai Kiado. (1973).

Bhowmik, R. "Detecting Auto Insurance Fraud by Data Mining Techniques." *CIS Journal*. 2(4): 156-162, 2011.

Botchey, F.E., Qin, Z., Hughes-Lartey, K. "Mobile Money Fraud Prediction—A Cross-Case Analysis on the Efficiency of Support Vector Machines, Gradient Boosted Decision Trees, and Naïve Bayes Algorithms." *Information*. 11: 383, 2020.

Burnham, K. P., & Anderson, D. R. *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer. (2002).

Chen, N., Ribeiro, B., Chen, A. "Financial credit risk assessment: a recent review." *Artificial Intelligence Review*. 45: 1-23, 2016.

Cortes, C., & Vapnik, V. Support-vector networks. *Machine Learning, 20*(3), 273–297. (1995).

CoŞkun, S.B., Turanli, M. "Credit risk analysis using boosting methods." *Journal of Applied Mathematics, Statistics and Informatics*. 19(1): 5-18, 2023.

Cristianini, N., & Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press. (2000).

Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. (2006).

Friedman, J.H. "Stochastic Gradient Boosting." *CSIRO CMIS*. 1-10, 1999.

Galindo, J., Tamayo, P. "Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications." *Computational Economics*. 15: 107-143, 2000.

Ginting, S.L.B., Adler, J., Ginting, Y.R., Kurniadi, A.H. "The development of bank application for debtors' selection by using Naïve Bayes Classifier technique." *In IOP Conference Series: Materials Science and Engineering*. Vol. 407, No. 1: p. 012177, 2018.

Hanley, J. A., & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. (1982).

Henley, W.E., Hand, D.J. "A k-nearest-neighbour classifier for assessing consumer credit risk." *Journal of the Royal Statistical Society Series D: The Statistician*. 45(1): 77-95, 1996.

James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer. (2021).

KhowwilaiSaeng, N. "Supplier classification framework for SAP ERP evaluation using machine learning models." 2021.

Krichene, A. "Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank." *Journal of Economics, Finance and Administrative Science*. 22(42): 3-24, 2017.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. *Applied linear statistical models* (5th ed.). (2005).

Mittal, L., Gupta, T., Sangaiah, A.K. "Prediction of credit risk evaluation using naive bayes, artificial neural network and support vector machine." *The IIOAB Journal*. 7(2): 33-42, 2016.

Montgomery, D. C., Peck, E. A., & Vining, G. G. *Introduction to linear regression analysis* (5th ed.). Wiley. (2012).

Mukid, M.A., Widiharih, T., Rusgiyono, A., Prahutama, A. "Credit scoring analysis using weighted k nearest neighbor." *Journal of Physics: Conference Series*. Vol. 1025, No. 1: p.012114, 2018.

Patchanok, S., Wararit, P. "Using Ensemble Machine Learning Methods to Forecast Particulate Matter (2.5 PM) in Bangkok, Thailand." *In International Conference on Multi-disciplinary Trends in Artificial Intelligence*, pp. 204-215, 2022.

Schmitt, M. "Deep learning vs. gradient boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring." *arXiv preprint arXiv*. 2205.10535, 2022.

Schölkopf, B., & Smola, A. J. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press. (2002).

Teles, G., Rodrigues, J.J.P.C., Rabê, R.A., Kozlov, S.A. "Artificial neural network and Bayesian network models for credit risk prediction." *Journal of Artificial Intelligence and Systems*. 2(1): 118-132, 2020.

Thongchuea, S. "Cryptocurrency trading signal prediction using ensemble learning and technical indicators." Faculty of Commerce and Accountancy, 2023.

Tian, Z., Xiao, J., Feng, H., Wei, Y. "Credit risk assessment based on gradient boosting decision tree." *Procedia Computer Science*. 174: 150-160, 2020.

Torvekar, N., Game, P.S. "Predictive analysis of credit score for credit card defaulters." *International Journal of Recent Technology and Engineering*. 7(1): 4, 2019.

Vapnik, V.N. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1998.

Veronesi, F. "Assessing the Accuracy of our models (R Squared, Adjusted R Squared, RMSE, MAE, AIC)." *R-bloggers*, 2017.

Veronesi, F. *Assessing the accuracy of our models (R Squared, Adjusted R Squared, RMSE, MAE, AIC)*. (2017).

Wang, Y., Zhang, Y., Lu, Y., Yu, X. "A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data." *Procedia Computer Science*. 174: 141-149, 2020.

Zan, H., Hsinchun, C., Chia-Jung, H., Wun-Hwa, C., Soushan, W. "Credit rating analysis with support vector machines and neural networks: a market comparative study." *Decision Support Systems*. 37(4): 543-558, 2004.

Zhang, L., Hu, H., Zhang, D. "A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance." *Financial Innovation*. 1(1): 1-21, 2015.

Zhang, T. "An introduction to support vector machines and other kernel-based learning methods." *Ai Magazine*. 22(2): 103, 2001.

ประวัติผู้เขียน

ชื่อ นายสรธัญ อินทรสมพงค์

ชื่อวิทยานิพนธ์ การพัฒนาการพัฒนาแบบจำลองการเรียนรู้ของเครื่องด้วยเทคนิค Stacking โดยใช้กลยุทธ์ปรับสมดุลข้อมูลสำหรับการประเมินความเสี่ยงสินเชื่อ SME

สาขาวิชา สถิติประยุกต์

ประวัติ

สำเร็จการศึกษาระดับปริญญาตรี หลักสูตรวิทยาศาสตร์บัณฑิต สาขาสถิติธุรกิจและการประกันภัย ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ในปีการศึกษา 2565

ศึกษาต่อในระดับปริญญาโท หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาสถิติประยุกต์และวิทยาการวิเคราะห์ข้อมูล ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ในปีการศึกษา 2566