ทำนายการผิดนัดชำระหนี้งวดแรกสำหรับสินเชื่อเพื่อธุรกิจขนาดเล็ก : กรณีศึกษาธนาคารแห่งหนึ่งในประเทศไทย

PREDICTING FIRST PAYMENT DEFAULT FOR SMALL BUSINESS LOANS: A CASE STUDY OF A BANK IN THAILAND

ภรภัทร ชินคำวงค์

PHORNRAPHAT CHINKHAMWONG

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติและการวิเคราะห์ธุรกิจ
ภาควิชาสถิติ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2568

KMITL-2025-SC-M-050-053

PREDICTING FIRST PAYMENT DEFAULT FOR SMALL BUSINESS LOANS: A CASE STUDY OF A BANK IN THAILAND

PHORNRAPHAT CHINKHAMWONG

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS AND BUSINESS ANALYTICS

DEPARTMENT OF STATISTICS SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2025

KMITL-2025-SC-M-050-053

หัวข้อการค้นคว้าอิสระ	ทำนายการผิดนัดชำระหนี้งวดแรกสำหรับสินเชื่อเพื่อธุรกิจขนาดเล็ก : กรณีศึกษาธนาคารกรณีศึกษาธนาคารแห่งหนึ่งในประเทศไทย
ชื่อนักศึกษา	ภรภัทร ชินคำวงค์
รหัสประจำตัว	65056108
ปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติและการวิเคราะห์ธุรกิจ)
ภาควิชา	สถิติ
พ.ศ.	2568
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ	ผู้ช่วยศาสตราจารย์ ดร.กนกกรรณ์ ลี้โรจนาประภา

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่มีผลต่อการเกิดการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มสินเชื่อธุรกิจขนาดเล็กและสร้างโมเดลการพยากรณ์ที่เหมาะสมสำหรับระบุลูกค้าที่มีความเสี่ยง FPD ภายใต้สภาวะข้อมูลที่ไม่สมดุล โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ได้แก่ การถดถอยโลจิสติก (Logistic Regression), ป่าสุ่ม (Random Forest) และ XGBoost ร่วมกับวิธีการจัดการข้อมูลไม่สมดุล เช่น SMOTE, SMOTEENN และ Random Undersampling การประเมินประสิทธิภาพของโมเดลใช้ตัวชี้วัดหลัก คือ พื้นที่ใต้กราฟ ROC (ROC-AUC) และค่า KS (Kolmogorov-Smirnov) ผลการวิจัยพบว่า เทคนิค SMOTE ร่วมกับ XGBoost ให้ประสิทธิภาพสูงสุด โดยได้ค่า ROC-AUC เท่ากับร้อยละ 81.42 และค่า KS เท่ากับร้อยละ 50.12 ซึ่งอยู่ในระดับที่สูงกว่ามาตรฐานที่สถาบันการเงินไทยยอมรับ รองลงมาคือ SMOTEENN ร่วมกับ Random Forest ให้ค่า ROC-AUC ร้อยละ 75.56 และ KS ร้อยละ 37.96 และ Random Undersampling ร่วมกับ Random Forest ให้ค่า ROC-AUC ร้อยละ 72.18 และ KS ร้อยละ 37.03 ปัจจัยที่มีอิทธิพลสำคัญต่อความเสี่ยงต่อการผิดนัดชำระหนี้งวดแรก ได้แก่ อัตราส่วนหนี้ต่อรายได้ ประเภทธุรกิจ ระยะเวลากู้ และจำนวนครั้งขอสินเชื่อย้อนหลัง ทั้งนี้ สามารถนำโมเดลที่พัฒนาขึ้นไปใช้ในการประเมินและแบ่งกลุ่มความเสี่ยงของลูกค้า (Credit Scoring) เพื่อเป็นแนวทางในการกำหนดนโยบายอนุมัติสินเชื่อและบริหารจัดการความเสี่ยงของสถาบันการเงินได้อย่างมีประสิทธิภาพ

คำสำคัญ: การถดถอยโลจิสติก, การผิดนัดชำระหนี้งวดแรก, เครดิตสกอริ่ง, ข้อมูลไม่สมดุล, ป่าสุ่ม, สินเชื่อธุรกิจขนาดเล็ก, SMOTE, SMOTEENN, XGBoost, Random Undersampling

Independent Study Title	Predicting First Payment Default in Micro Business Loans: Case Study of a Bank in Thailand
Student Name	Phornraphat Chinkhamwong
Student ID	65056108
Degree	Master of Science (Statistics and Business Analytics)
Department	Statistics
Year	2025
Independent Study Advisor	Assistant Professor Dr. Kanogkan Leerojanaprapa

Abstract

This study aims to analyze the key factors influencing the first payment default (FPD) in micro business loan portfolios and develop an appropriate predictive model for identifying borrowers with a high risk of FPD under imbalanced data conditions. Machine learning techniques, including Logistic Regression, Random Forest, and XGBoost, were employed in conjunction with resampling methods such as SMOTE, SMOTEENN, and Random Undersampling to address class imbalance. Model performance was primarily evaluated using the area under the ROC curve (ROC-AUC) and the Kolmogorov-Smirnov (KS) statistic. The findings revealed that the combination of SMOTE and XGBoost yielded the highest predictive performance, with a ROC-AUC of 81.42% and a KS statistic of 50.12%, both exceeding the industry benchmark for credit risk modeling in Thai financial institutions. The next best results were achieved by SMOTEENN with Random Forest (ROC-AUC = 75.56%, KS = 37.96%) and Random Undersampling with Random Forest (ROC-AUC = 72.18%, KS = 37.03%). The most influential predictors of FPD risk included the debt-to-income ratio, business type, loan tenure, and the number of past loan applications. The developed models can be applied to credit scoring systems for micro business loans, providing practical guidelines for loan approval policies and risk management within financial institutions.

Keywords: Logistic Regression, First Payment Default, Credit Scoring, Imbalanced Data, Random Forest, Micro Business Loans, SMOTE, SMOTEENN, XGBoost, Random Undersampling

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบคุณอาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.กนกกรรณ์ ลี้โรจนาประภา ที่ได้กรุณาให้คำแนะนำ ชี้แนะแนวทาง ตลอดจนให้ความช่วยเหลือในทุกขั้นตอนของการดำเนินงานวิจัยฉบับนี้ด้วยความเอาใจใส่และอดทนเสมอมา

ขอขอบคุณคณะกรรมการสอบผู้ช่วยศาสตราจารย์ ดร.พรพิมล ชัยวุฒิศักดิ์ และผู้ช่วยศาสตราจารย์ ดร.สิทธิชัย เจริญเศรษฐศิลป์ ที่ได้ให้ข้อเสนอแนะและคำแนะนำอันมีคุณค่า ทำให้งานวิจัยฉบับนี้มีความสมบูรณ์ยิ่งขึ้น

ขอขอบคุณสถาบันการเงินที่สนับสนุนด้านข้อมูลและสนับสนุนทรัพยากรอันสำคัญในการดำเนินการวิจัยนี้ รวมถึงเพื่อนร่วมงานและผู้ที่เกี่ยวข้องทุกท่านที่ให้กำลังใจและความร่วมมือด้วยดีตลอดมา

ท้ายที่สุดนี้ขอขอบคุณ คุณเปรมสุดา ปัดสาโย เพื่อนที่ได้เข้ามาศึกษาร่วมกันและคอยช่วยเหลือให้คำปรึกษามาโดยตลอด ขอขอบคุณ คุณนพพร แย้มสุวรรณ และคุณกุลปาลิกา ปรรณารม รวมถึงครอบครัวที่เป็นกำลังใจและส่งแรงสนับสนุนสำคัญในทุกช่วงเวลาของการศึกษาและการดำเนินงานวิจัยฉบับนี้ให้สำเร็จลุล่วงไปด้วยดี ขอขอบพระคุณทุกท่านมา ณ โอกาสนี้

นางสาวภรภัทร ชินคำวงค์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

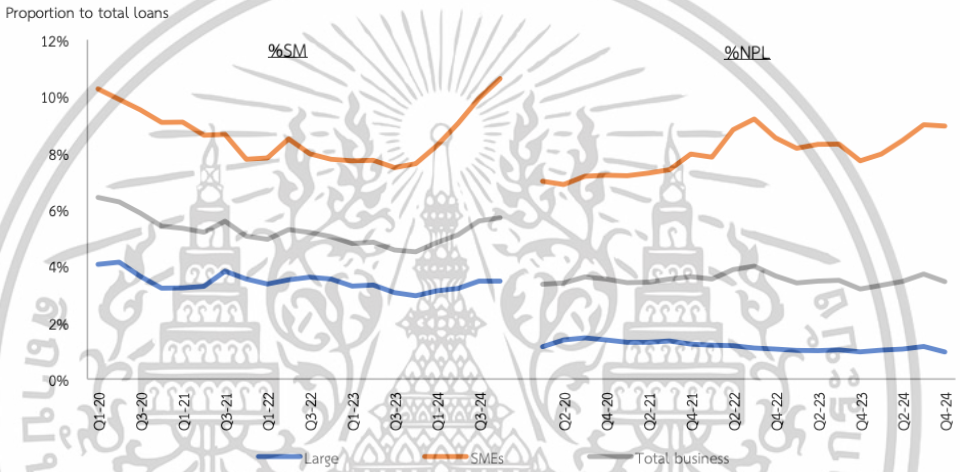
สารบัญรูป

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

ในยุคปัจจุบัน ธุรกิจขนาดเล็ก (Small Business) เป็นฟันเฟืองสำคัญในการขับเคลื่อนเศรษฐกิจของประเทศ สถาบันการเงินและธนาคารพาณิชย์ต่างให้ความสำคัญกับการสนับสนุนสินเชื่อแก่ผู้ประกอบการรายย่อย เพื่อเสริมสร้างศักยภาพในการแข่งขัน เพิ่มโอกาสการเข้าถึงแหล่งเงินทุน และช่วยกระตุ้นการจ้างงานในระดับชุมชน (ธนาคารแห่งประเทศไทย, 2566)



รูปที่ 1.1 คุณภาพสินเชื่อของระบบสถาบันการเงิน
ที่มา ธนาคารแห่งประเทศไทย (2567)

จากรูปที่ 1.1 ซึ่งแสดงอัตราส่วนสินเชื่อด้อยคุณภาพ (SM) และหนี้ที่ไม่ก่อให้เกิดรายได้ (NPL) ของระบบสถาบันการเงินในประเทศไทย จะเห็นได้ว่า กลุ่มธุรกิจขนาดกลางและขนาดย่อม (SMEs) ซึ่งรวมถึงธุรกิจขนาดเล็ก มีสัดส่วน SM และ NPL สูงกว่ากลุ่มธุรกิจขนาดใหญ่ตลอดช่วงปี 2020-2024 และยังคงมีแนวโน้มเพิ่มสูงขึ้นอย่างต่อเนื่อง โดยเฉพาะ %SM ของ SMEs ที่พุ่งสูงอย่างมีนัยสำคัญในช่วงปลายปี 2567 สะท้อนให้เห็นถึงปัญหาคุณภาพสินเชื่อและความเปราะบางของผู้ประกอบการรายเล็กมากขึ้นในระบบการเงิน นอกจากนี้ NPL ในกลุ่ม SMEs อยู่ในระดับสูงกว่า 4% ตลอดช่วงเวลาดังกล่าว ในขณะที่กลุ่มธุรกิจขนาดใหญ่มีอัตรา NPL ต่ำกว่าอย่างมีนัยสำคัญ สถานการณ์ดังกล่าวจึงเป็นข้อท้าทายสำคัญต่อสถาบันการเงินและธนาคารพาณิชย์ในด้านการบริหารความเสี่ยงและเสถียรภาพของพอร์ตสินเชื่อ ข้อมูลจากงานวิจัยสากล อาทิ Brown & Mues (2012), Altman & Sabato (2007), Agarwal et al. (2020) ยังยืนยันว่า ลูกหนี้ที่ผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) มีโอกาสผิดนัดต่อเนื่องและกลายเป็น NPL ในอนาคตโดยเฉพาะในกลุ่มธุรกิจขนาดเล็ก เนื่องจากกลุ่มนี้มีความเปราะบางทางกระแสเงินสดและขาดกลไกป้องกันความ

เสี่ยงที่มีประสิทธิภาพดังนั้น การลดสัดส่วน FPD ในกลุ่มธุรกิจขนาดเล็กจะส่งผลเชิงบวกทั้งต่อคุณภาพพอร์ตสินเชื่อของสถาบันการเงินและต่อเสถียรภาพของระบบเศรษฐกิจโดยรวม อันจะนำไปสู่การลดต้นทุน ความสูญเสีย และสร้างโอกาสในการปล่อยสินเชื่อใหม่ให้กับกลุ่มธุรกิจที่มีศักยภาพมากยิ่งขึ้น ดังนั้นการพัฒนา Credit Scoring Model ที่แม่นยำและสามารถจัดกลุ่มความเสี่ยงของลูกค้าจึงมีความสำคัญต่อธนาคารและสถาบันการเงิน เพราะโมเดลเหล่านี้สามารถนำผลลัพธ์ไปใช้ในกระบวนการอนุมัติหรือปฏิเสธสินเชื่อ การกำหนดวงเงินและอัตราดอกเบี้ย อีกประเด็นสำคัญที่มักพบในข้อมูลสินเชื่อกลุ่มธุรกิจขนาดเล็ก คือ ปัญหาความไม่สมดุลของข้อมูล (Class Imbalance) กล่าวคือ จำนวนตัวอย่างของลูกค้าที่ไม่ผิดนัดชำระ (Non-FPD) มีมากกว่าจำนวนตัวอย่างของลูกค้าที่ผิดนัดชำระหนี้งวดแรก (FPD) อย่างมาก เช่น สัดส่วน FPD ต่ำกว่า 10% ของทั้งหมด ปัญหานี้ส่งผลให้โมเดล Machine Learning แบบดั้งเดิมมีแนวโน้มที่จะทำนายกลุ่มเสียงต่ำ (Non-FPD) ได้แม่นยำแต่ขาดประสิทธิภาพในการระบุลูกค้าที่มีความเสี่ยงสูง (FPD) ซึ่งเป็นกลุ่มที่ธุรกิจและธนาคารต้องการโฟกัสมากที่สุด (He & Garcia, 2009; Chawla et al., 2002) ผลกระทบจากปัญหาข้อมูลไม่สมดุลได้แก่ ความแม่นยำโดยรวมของโมเดลอาจดูสูงแต่จริงๆ แล้วโมเดลอาจทำนายกลุ่มเสียงต่ำถูกต้องแต่ขาดความสามารถในการจับกลุ่มเสียงสูง ทำให้เกิดความเสี่ยงในการปล่อยสินเชื่อผิดพลาด

จากเหตุผลข้างต้นการศึกษาครั้งนี้จึงมุ่งเน้นการสร้างและเปรียบเทียบประสิทธิภาพของโมเดล Machine Learning ที่ได้รับการยอมรับในงานวิจัยได้แก่ ป่าสุ่ม (Random Forest), XGBoost และการถดถอยโลจิสติก (Logistic Regression) ร่วมกับเทคนิคการจัดการข้อมูลไม่สมดุลต่างๆ โดยงานวิจัยนี้เลือกใช้เทคนิคการปรับสมดุลข้อมูล ได้แก่ SMOTE และ Random Undersampling ซึ่งเป็นวิธีการที่ได้รับการยอมรับและพิสูจน์ทางวิชาการแล้วว่าช่วยเพิ่มประสิทธิภาพโมเดลในการทำนายกรณีข้อมูลไม่สมดุล (Brown & Mues, 2012; Fernández et al., 2018) ร่วมด้วยเทคนิคผสม SMOTEENN เพื่อลด Bias และเพิ่มความน่าเชื่อถือของ Credit Scoring Model สำหรับกลุ่มลูกค้าที่มีความเสี่ยงสูงสำหรับทำนายโอกาสการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มสินเชื่อธุรกิจขนาดเล็ก ผลของโมเดลจะประเมินจากตัวชี้วัดได้แก่ค่าพื้นที่ใต้กราฟ ROC และ KS (Kolmogorov-Smirnov) ซึ่งเป็นตัวชี้วัดมาตรฐานที่ใช้ในการประเมินคุณภาพของโมเดลเครดิตสกอริ่งและการบริหารความเสี่ยงสินเชื่อ (ธนาคารแห่งประเทศไทย, 2565) เพื่อประเมินโมเดลที่เหมาะสมที่สุดเพื่อนำไปประยุกต์ใช้สำหรับการสนับสนุนการตัดสินใจทางธุรกิจตามนโยบายหรือลดความเสี่ยงการเกิด NPL และเพิ่มประสิทธิภาพการตัดสินใจอนุมัติสินเชื่อของธนาคารไทยในอนาคต

1.2 วัตถุประสงค์ของงานวิจัย

1) ศึกษาปัจจัยสำคัญที่มีผลต่อการเกิดการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD)

2) สร้างโมเดลที่เหมาะสมสำหรับลูกค้าที่มีโอกาสการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มสินเชื่อธุรกิจขนาดเล็กภายใต้สภาวะข้อมูลที่ไม่สมดุล

1.3 ขอบเขตการวิจัย

1) ขอบเขตของแผนงาน

งานวิจัยนี้มุ่งเน้นศึกษาปัจจัยที่มีอิทธิพลต่อการผิดนัดชำระหนี้งวดแรก (FPD) โดยนิยามการผิดนัดชำระหนี้งวดแรกว่า เป็นกรณีที่ลูกค้าไม่ชำระค่างวดแรกเกินกว่า 30 วันหลังจากวันครบกำหนด โดยใช้ข้อมูลจากการสมัครสินเชื่อธุรกิจขนาดเล็กของสถาบันการเงินแห่งหนึ่ง ซึ่งเป็นข้อมูลคำขอสินเชื่อและถูกอนุมัติระหว่างวันที่ 1 กันยายน พ.ศ. 2562 ถึงวันที่ 1 มีนาคม พ.ศ. 2567

2) ขอบเขตของข้อมูล

ข้อมูลที่ใช้วิเคราะห์เป็นข้อมูลการสมัครสินเชื่อธุรกิจขนาดเล็กจากสถาบันการเงินแห่งหนึ่ง ครอบคลุมลูกค้าที่สมัครเข้ามาตั้งแต่วันที่ 1 กันยายน พ.ศ. 2562 ถึงวันที่ 1 มีนาคม พ.ศ. 2567 ทั้งนี้ โมเดลถูกพัฒนาเพื่อทำนายสถานะการผิดนัดชำระหนี้งวดแรก โดย 0 หมายถึง ลูกหนี้ไม่ผิดนัดชำระหนี้ (Non-First Payment Default: Non-FPD) และ 1 หมายถึง ลูกหนี้ผิดนัดชำระหนี้ (First Payment Default: FPD) หลังจากการสมัครสินเชื่อ

3) ขอบเขตด้านเวลา

นำข้อมูลคำขอสินเชื่อธุรกิจขนาดเล็กและถูกอนุมัติระหว่างวันที่ 1 กันยายน พ.ศ. 2562 ถึงวันที่ 1 มีนาคม พ.ศ. 2567 เพื่อสร้างตัวแปรอิสระที่เกี่ยวข้องกับพฤติกรรมการชำระหนี้ในงวดแรกที่ถึงกำหนดชำระหลังจากอนุมัติสินเชื่อ

4) เครื่องมือที่ใช้ในการสร้างโมเดล

การพัฒนาโมเดลการทำนายการผิดนัดชำระหนี้งวดแรกดำเนินการผ่านโปรแกรมภาษาไพธอน (Python Programming) โดยใช้ Jupyter Notebook เป็นแพลตฟอร์มหลักในการพัฒนาโมเดล และใช้ SAS Program สำหรับการจัดเก็บและเตรียมข้อมูลที่เกี่ยวข้อง

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) สร้างแบบโมเดลที่ช่วยให้สถาบันการเงินใช้ในการวางแผนและกำหนดกลยุทธ์ให้สอดคล้องกับนโยบายของสถาบันการเงิน

2) สถาบันการเงินสามารถนำผลจากโมเดลไปใช้ในการประเมินความเสี่ยงและคัดกรองลูกค้าได้อย่างมีประสิทธิภาพมากขึ้น

1.5 กรอบแนวคิดการวิจัย

จากการศึกษาข้อมูลสินเชื่อธุรกิจขนาดเล็ก ทบทวนวรรณกรรม และงานวิจัยที่เกี่ยวข้อง พบว่าปัจจัยที่อาจมีอิทธิพลต่อการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ประกอบด้วยปัจจัยส่วนบุคคล ปัจจัยด้านธุรกิจ ปัจจัยด้านสินเชื่อ ปัจจัยด้านรายได้และการเงิน ปัจจัยด้านภาระหนี้ ปัจจัยด้านประวัติสินเชื่อ และปัจจัยด้านดัชนีชี้วัดทางการเงิน ซึ่งสามารถนำมาใช้เป็น

ตัวแปรอิสระ (Independent Variables) โดยใช้สถานะการผิดนัดชำระหนี้งวดแรกเป็นตัวแปรตามในการวิเคราะห์ดังรูปต่อไปนี้

ตัวแปรต้น (Independent Variables)

รูปที่ 1.2 กรอบแนวคิดในการวิจัย

1.6 นิยามศัพท์เฉพาะ

การผิดนัดชำระหนี้งวดแรกและไม่ผิดนัดชำระนี้งวดแรก ในงานวิจัยนี้ **“การผิดนัดชำระหนี้งวดแรก”** (1) หมายถึง สถานการณ์ที่ลูกค้าไม่สามารถชำระหนี้ในงวดแรกของสัญญาเงินกู้ได้ตรงตามกำหนดเวลาเกินกว่า 30 วัน ในทางกลับกัน **“ไม่ผิดนัดชำระหนี้งวดแรก”** (0) หมายถึง ลูกค้าที่ชำระหนี้ได้ครบเมื่อถึงวันที่ครบกำหนดชำระในงวดแรกของสัญญาเงินกู้

สินเชื่อธุรกิจขนาดเล็กภายใต้การกำกับ หมายถึง สินเชื่อที่ธนาคารหรือสถาบันการเงินปล่อยให้แก่ผู้ประกอบการธุรกิจที่มีขนาดเล็ก ใช้เพื่อเสริมสภาพคล่องทางธุรกิจ ขยายกิจการ หรือเป็นเงินทุนหมุนเวียน สินเชื่อประเภทนี้จะเน้นกลุ่มผู้ประกอบการรายย่อยที่มีรายได้หรือยอดขายต่อปีในระดับต่ำตามที่แต่ละองค์กรหรือสถาบันการเงินกำหนด ตัวอย่างเช่น ธนาคารแห่งประเทศไทยได้ให้คำนิยาม "ธุรกิจขนาดเล็ก" หรือ "วิสาหกิจขนาดกลางและขนาดย่อม (SMEs)" โดยใช้เกณฑ์มูลค่ายอดขายหรือรายได้รวมต่อปี และจำนวนพนักงาน (ธนาคารแห่งประเทศไทย, 2566)

ข้อมูลไม่สมดุล (Imbalanced Data) หมายถึง คือชุดข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสแตกต่างกันมาก เช่น กลุ่มที่ผิดนัดชำระหนี้มีจำนวนน้อยกว่ากลุ่มที่ไม่ผิดนัดอย่างชัดเจน (He & Garcia, 2009)

คะแนนเครดิต (Credit Scoring) หมายถึง กระบวนการให้คะแนนหรือจัดอันดับความน่าเชื่อถือทางการเงินของผู้กู้ โดยใช้โมเดลทางสถิติหรืออัลกอริทึมในการวิเคราะห์ข้อมูลประวัติทางการเงิน ข้อมูลส่วนบุคคล และพฤติกรรมการชำระหนี้ เพื่อประเมินความเสี่ยงของผู้ขอกู้ว่ามีแนวโน้มจะผิดนัดชำระหนี้หรือไม่ โดยคะแนนที่ได้จะใช้ประกอบการตัดสินใจอนุมัติสินเชื่อ กำหนดวงเงิน หรืออัตราดอกเบี้ย (ธนาคารแห่งประเทศไทย, 2566; Hand & Henley, 1997)

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาเรื่อง การทำนายการผิดนัดชำระหนี้งวดแรกของสินเชื่อธุรกิจขนาดเล็กผู้วิจัยได้ศึกษาค้นคว้าแนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้องต่างๆ ดังต่อไปนี้

2.1 แนวคิดเกี่ยวกับสินเชื่อธุรกิจขนาดเล็ก

2.2 แนวคิดและทฤษฎีเกี่ยวกับความเสี่ยงสินเชื่อและการบริหารความเสี่ยง

2.3 แนวคิดเกี่ยวกับการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD)

2.4 แนวคิดเกี่ยวกับกระบวนการ CRISP-DM

2.5 แนวคิดและทฤษฎีเกี่ยวกับการจัดการข้อมูลไม่สมดุล (Imbalanced Data)

2.6 แนวคิดและทฤษฎีเกี่ยวกับอัลกอริทึม Machine Learning สำหรับการจำแนกประเภท

2.7 แนวคิดและทฤษฎีเกี่ยวข้องกับการเลือกไฮเปอร์พารามิเตอร์แบบกริด (Grid Search Cross-Validation)

2.8 แนวคิดและทฤษฎีเกี่ยวกับการประเมินประสิทธิภาพโมเดล

2.9 แนวคิดและทฤษฎีเกี่ยวกับ Feature Importance

2.10 แนวคิดและทฤษฎีเกี่ยวกับ Credit Scoring

2.11 งานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎีเกี่ยวกับสินเชื่อธุรกิจขนาดเล็ก

ธุรกิจขนาดเล็ก หมายถึง บริการสินเชื่อที่ออกแบบเฉพาะสำหรับธุรกิจขนาดเล็กหรือผู้ประกอบการรายย่อย (SME) ซึ่งมักประสบปัญหาเข้าถึงสินเชื่อธุรกิจแบบดั้งเดิมได้ยาก สินเชื่อประเภทนี้มีลักษณะเฉพาะแตกต่างจากสินเชื่อธุรกิจทั่วไปหลายประการ ได้แก่ วงเงินกู้ที่มักอยู่ในระดับหลักพันถึงหลักแสนบาท (ต่ำกว่าสินเชื่อธุรกิจขนาดใหญ่ที่อาจสูงถึงหลักล้านบาท) ระยะเวลาชำระคืนสั้นกว่า (ประมาณ 6 เดือนถึง 2 ปี เทียบกับสินเชื่อธุรกิจปกติที่ผ่อน 5–10 ปี) และเงื่อนไขอนุมัติที่ยืดหยุ่นกว่า เช่น อาจไม่ต้องมีหลักประกัน แต่มุ่งเน้นพิจารณาศักยภาพธุรกิจและความสามารถชำระหนี้แทน ขณะเดียวกัน อัตราดอกเบี้ย ของสินเชื่อธุรกิจขนาดเล็กมักจะสูงกว่าสินเชื่อธุรกิจทั่วไป เพราะความเสี่ยงที่สูงกว่าและต้นทุนการดำเนินงานต่อสัญญาที่สูงกว่า และผู้ให้บริการบางแห่งยังมีการให้คำปรึกษาหรือสนับสนุนเพิ่มเติมแก่ผู้กู้รายย่อยเพื่อเพิ่มโอกาสความสำเร็จของธุรกิจอีกด้วย (Funding Societies, 2024) สำหรับในประเทศไทย สินเชื่อธุรกิจขนาดเล็กอาจครอบคลุมตั้งแต่สินเชื่อเพื่อผู้ประกอบการรายย่อย (ไมโครเครดิต) เช่น นาโนไฟแนนซ์หรือพิโกไฟแนนซ์ ซึ่งวงเงินไม่เกิน 100,000 บาท และไม่ต้องมีหลักทรัพย์ค้ำประกันไปจนถึงสินเชื่อ SMEs ขนาดเล็กที่ธนาคารพาณิชย์ให้แก่ธุรกิจมียอดขายไม่เกินเกณฑ์ที่กำหนด ภาครัฐและธนาคารแห่ง

ประเทศไทยสนับสนุนสินเชื่อกลุ่มนี้เพื่อกระตุ้นเศรษฐกิจฐานรากและเพิ่มการเข้าถึงแหล่งเงินทุนของผู้ประกอบการรายเล็ก โดยผู้ให้บริการสินเชื่อธุรกิจขนาดเล็กมักใช้ข้อมูลทางเลือกในการประเมินความเสี่ยง เช่น ข้อมูลธุรกรรม หรือ พฤติกรรมการชำระเงิน แทนหลักทรัพย์ค้ำประกัน เพื่อลดอุปสรรคในการเข้าถึงสินเชื่อของ SMEs (ธนาคารแห่งประเทศไทย, 2566)

2.2 แนวคิดและทฤษฎีเกี่ยวกับความเสี่ยงสินเชื่อและการบริหารความเสี่ยง

ความเสี่ยงสินเชื่อ (Credit Risk) หมายถึง ความเสี่ยงที่เกิดขึ้นจากการที่ผู้กู้ไม่สามารถชำระหนี้คืนได้ตามข้อตกลง ซึ่งส่งผลให้สถาบันการเงินสูญเสียรายได้ที่ควรจะได้รับและอาจเกิดความเสียหายทางการเงิน การบริหารความเสี่ยงสินเชื่อ (Credit Risk Management) จึงมีบทบาทสำคัญในการควบคุมและลดความเสี่ยงดังกล่าว โดยการประเมินความสามารถในการชำระหนี้ของลูกค้า ความมั่นคงของรายได้ ประวัติการชำระหนี้ และพฤติกรรมทางการเงินเป็นหลักที่ใช้ในการตัดสินใจ (Basel Committee on Banking Supervision, 2000)

2.2.1 ประเภทของความเสี่ยงสินเชื่อ

ความเสี่ยงสินเชื่อสามารถจำแนกออกได้เป็น 3 ประเภทหลัก ดังนี้

1) ความเสี่ยงจากลูกค้า (Borrower Risk) ความเสี่ยงประเภทนี้เกิดจากลักษณะเฉพาะของผู้กู้ เช่น รายได้ อาชีพ การศึกษา และประวัติทางการเงิน หากผู้กู้มีรายได้ไม่แน่นอน หรือมีประวัติผิดนัดชำระหนี้ในอดีต จะมีแนวโน้มที่จะผิดนัดชำระหนี้สูงกว่ากลุ่มที่มีความมั่นคงทางการเงิน

2) ความเสี่ยงจากผลิตภัณฑ์สินเชื่อ (Product Risk) เป็นความเสี่ยงที่เกิดจากลักษณะของผลิตภัณฑ์สินเชื่อเอง เช่น ประเภทของสินเชื่อ ระยะเวลาการผ่อนชำระ และอัตราดอกเบี้ย หากมีการกำหนดเงื่อนไขที่ไม่เหมาะสม เช่น ดอกเบี้ยสูงหรือระยะเวลาการผ่อนชำระสั้น อาจส่งผลต่อความสามารถในการชำระหนี้ของลูกค้า

3) ความเสี่ยงจากสภาพแวดล้อม (Environmental Risk) เป็นความเสี่ยงที่เกิดจากปัจจัยภายนอกซึ่งอยู่นอกเหนือการควบคุมของลูกค้าและสถาบันการเงิน เช่น ภาวะเศรษฐกิจถดถอย ความผันผวนของตลาดแรงงาน ภัยธรรมชาติ หรือสถานการณ์เฉพาะอย่าง (Basel Committee on Banking Supervision, 2000)

2.2.2 มาตรการสำคัญในการบริหารความเสี่ยงสินเชื่อ

การบริหารความเสี่ยงสินเชื่ออย่างมีประสิทธิภาพจำเป็นต้องอาศัยมาตรการหลายด้าน โดยสามารถสรุปแนวทางสำคัญได้ดังนี้

1) การใช้โมเดลให้คะแนนเครดิต (Credit Scoring Models) โมเดลนี้มีบทบาทในการประเมินความเสี่ยงของผู้กู้แต่ละราย โดยใช้ข้อมูลเชิงประวัติและข้อมูลพฤติกรรมมาคำนวณเป็นคะแนนความเสี่ยง ซึ่งช่วยให้สามารถตัดสินใจในการปล่อยสินเชื่อได้อย่างแม่นยำ (Thomas, 2009)

2) การตั้งเกณฑ์อนุมัติที่เหมาะสม การกำหนดเกณฑ์หรือ Threshold ในการอนุมัติสินเชื่อตามระดับความเสี่ยงที่ยอมรับได้ ช่วยให้สถาบันการเงินสามารถควบคุมระดับความเสี่ยงรวมของพอร์ตสินเชื่อได้อย่างมีประสิทธิภาพ และสามารถปรับเงื่อนไขให้สอดคล้องกับลักษณะของลูกค้าแต่ละกลุ่ม (Basel Committee on Banking Supervision, 2000)

3) การติดตามและประเมินพฤติกรรมลูกค้าหลังการปล่อยกู้ หลังจากปล่อยสินเชื่อแล้ว การติดตามพฤติกรรมการชำระหนี้ของลูกค้าอย่างต่อเนื่อง เช่น การตรวจสอบความถี่ในการจ่ายชำระหรือการเปลี่ยนแปลงรายได้ จะช่วยให้สามารถวางมาตรการเชิงป้องกันล่วงหน้า เช่น การโทรแจ้งเตือนหรือการปรับเงื่อนไขสินเชื่อได้ทันเวลา

2.2.3 ตัวชี้วัดความเสี่ยงที่สำคัญ

ในการบริหารความเสี่ยงสินเชื่อ จำเป็นต้องใช้ตัวชี้วัดที่สะท้อนสถานะความเสี่ยงของลูกหนี้ในระยะต่างๆ ดังนี้

1) First Payment Default (FPD) คือการผิดนัดชำระตั้งแต่งวดแรกหลังจากได้รับการอนุมัติสินเชื่อ ซึ่งถือเป็นสัญญาณเตือน (Early Warning Sign) ที่สำคัญ เพราะสะท้อนให้เห็นถึงปัญหาในการคัดกรองความสามารถของผู้กู้ (Anderson, 2007)

2) Non-Performing Loan (NPL) หมายถึงหนี้ที่ผิดนัดชำระมาแล้วเกิน 90 วัน ซึ่งจัดเป็นหนี้ที่มีความเสี่ยงสูงและอาจก่อให้เกิดผลกระทบต่อสภาพคล่องและฐานะทางการเงินของสถาบันการเงินในระยะยาว การติดตามอัตรา NPL อย่างใกล้ชิดจึงเป็นกลไกสำคัญในการประเมินประสิทธิภาพของการบริหารความเสี่ยง (Bank of Thailand, 2022)

2.3 แนวคิดเกี่ยวกับการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD)

การผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) หมายถึงสถานการณ์ที่ผู้กู้ไม่สามารถชำระค่างวดแรกของสินเชื่อตามกำหนดในสัญญาเงินกู้ภายในระยะเวลาที่สถาบันการเงินกำหนด โดยทั่วไปจะนิยามว่าเป็น “การไม่ชำระค่างวดแรกภายใน 30 วันหลังวันครบกำหนด” (FICO, 2022; Liu, Wang, & Lin, 2020) ถือเป็นหนึ่งในตัวชี้วัดสำคัญของระบบบริหารความเสี่ยงด้านสินเชื่อ เนื่องจากลูกหนี้ที่ผิดนัดในงวดแรกมักมีแนวโน้มสูงที่จะผิดนัดซ้ำในอนาคต หรือกลายเป็นหนี้ที่ไม่ก่อให้เกิดรายได้ (Non-Performing Loan: NPL) ในที่สุด ดังนั้น FPD จึงได้รับการยอมรับอย่างกว้างขวางในภาคการเงินว่าเป็นตัวบ่งชี้ความเสี่ยงในระยะเริ่มต้น (Early Warning Indicator) ที่มีประสิทธิภาพ

2.3.1 บทบาทและความสำคัญของ FPD ในระบบสินเชื่อ

การใช้ตัวชี้วัด FPD มีบทบาทหลายด้านในระบบการปล่อยสินเชื่อและการบริหารความเสี่ยงของสถาบันการเงิน ดังนี้

1) ตัวชี้วัดคุณภาพการปล่อยสินเชื่อ อัตรา FPD เป็นหนึ่งในตัวชี้วัดหลัก (Key Performance Indicator: KPI) ที่ใช้วัดคุณภาพของกระบวนการปล่อยสินเชื่อ หากมีอัตรา FPD สูง แสดงให้เห็นถึงปัญหาในการคัดกรองลูกค้าหรือความไม่เหมาะสมของเกณฑ์การอนุมัติที่ใช้ในช่วงเวลาดังกล่าว (FICO, 2022; ธนาคารแห่งประเทศไทย, 2566)

2) การควบคุมคุณภาพลูกค้า ข้อมูลจาก FPD สามารถนำไปใช้ในการวิเคราะห์เชิงลึกเพื่อปรับเกณฑ์การอนุมัติสินเชื่อให้เหมาะสมยิ่งขึ้นกับกลุ่มลูกค้าที่มีความเสี่ยงแตกต่างกัน เช่น การเพิ่มข้อกำหนดในการตรวจสอบรายได้ หรือการลดวงเงินในกลุ่มลูกค้าที่มีพฤติกรรมความเสี่ยงสูง

3) เครื่องมือในการกำหนดวงเงินและดอกเบี้ย การวิเคราะห์แนวโน้มการเกิด FPD สามารถนำมาใช้เป็นพื้นฐานในการกำหนดเงื่อนไขของวงเงิน อัตราดอกเบี้ย หรือหลักประกันเพิ่มเติม สำหรับกลุ่มลูกค้าที่มีความเสี่ยงสูง ทั้งนี้เพื่อชดเชยความเสี่ยงและป้องกันการเกิด NPL ในระยะยาว (ธนาคารแห่งประเทศไทย, 2566)

2.3.2 วิธีนิยาม FPD ในบริบทสากลและระดับประเทศ

แนวทางการนิยาม FPD มีความแตกต่างกันไปตามสถาบันและบริบทของแต่ละภูมิภาค โดยมีรายละเอียดดังนี้

1) สหรัฐอเมริกาและยุโรป มาตรฐานของบริษัท FICO และหน่วยงานสินเชื่อในสหรัฐอเมริกาและยุโรปมักใช้เกณฑ์ว่า FPD คือ “การไม่ชำระค่างวดแรกภายใน 30 วันหลังครบกำหนด” ซึ่งเป็นที่ยอมรับอย่างกว้างขวางในระบบเครดิตของประเทศที่มีโครงสร้างสินเชื่อพัฒนาแล้ว (FICO, 2022)

2) เอเชียและตลาดเกิดใหม่ ในประเทศกำลังพัฒนา เช่น เวียดนาม จีน และไทย นิยมใช้เกณฑ์เดียวกับมาตรฐานสากล แต่บางบริบทอาจขยายระยะเวลาการพิจารณา FPD ออกไปเป็น 60 วัน ขึ้นอยู่กับรูปแบบผลิตภัณฑ์สินเชื่อ (Liu, Wang, & Lin, 2020; Khandani, Kim, & Lo, 2010)

3) ประเทศไทย ธนาคารแห่งประเทศไทยและหน่วยงานวิจัยด้านสินเชื่อของไทย นิยมใช้เกณฑ์ “เกิน 30 วัน” ในการนิยาม FPD โดยยึดเป็นมาตรฐานในการติดตามคุณภาพของสินเชื่อรายใหม่ และใช้ประกอบการกำหนดเกณฑ์สินเชื่ออย่างต่อเนื่อง (ธนาคารแห่งประเทศไทย, 2566) โดยสูตรการคำนวณอัตราการผิดนัดชำระหนี้งวดแรก (FPD Rate) คือ

$$FPD\ Rate = \frac{\textit{จำนวนสินเชื่อที่ผิดนัดชำระหนี้งวดแรก}}{\textit{จำนวนสินเชื่อที่ปล่อยใหม่ในช่วงเวลาเดียวกัน}} \times 100 \qquad (2.1)$$

2.3.3 ปัจจัยเสี่ยงสำคัญที่สัมพันธ์กับการผิดนัดชำระหนี้งวดแรก (FPD)

จากการศึกษางานวิจัยทั้งในระดับสากลและภายในประเทศ พบว่ามีหลายปัจจัยที่สัมพันธ์อย่างมีนัยสำคัญกับความเสี่ยงในการเกิด FPD ซึ่งสามารถจำแนกได้เป็น 5 กลุ่มหลัก ดังนี้

1) ข้อมูลส่วนบุคคล เช่น อายุ เพศ และสถานภาพสมรส โดยเฉพาะลูกหนี้ที่มีอายุน้อย หรือยังไม่มีสถานภาพครอบครัวที่มั่นคง มักมีความเสี่ยงทางการเงินที่สูงกว่ากลุ่มอื่น (Rodríguez, 2024)

2) ข้อมูลทางธุรกิจ เช่น ประเภทกิจการและระยะเวลาประสบการณ์ในการดำเนินงาน โดยกิจการที่เพิ่งเริ่มต้นหรืออยู่ในภาคธุรกิจที่มีความไม่แน่นอน เช่น ค้าขายรายย่อย จะมีแนวโน้มผิดนัดสูงขึ้น (Liu et al., 2020)

3) ข้อมูลทางการเงิน โดยเฉพาะระดับรายได้สุทธิ อัตราส่วนความสามารถในการชำระหนี้ (Debt Service Coverage Ratio: DSCR) และภาระหนี้สินรวม หากลูกหนี้มีรายได้สุทธิต่อเดือนต่ำหรือมี DSCR ต่ำกว่าเกณฑ์มาตรฐาน มักไม่สามารถรับภาระค่างวดในระยะยาวได้อย่างต่อเนื่อง (Koç & Sevgili, 2020)

4) ประวัติเครดิต เช่น จำนวนครั้งที่เคยขอสินเชื่อในอดีต และระดับความเสี่ยง (Risk Grade) ซึ่งถือเป็นสัญญาณเตือนที่สำคัญหากลูกค้ามีพฤติกรรมการยื่นขอสินเชื่อถี่โดยไม่มีความสามารถในการชำระหนี้รองรับ (Liu et al., 2020)

5) ภูมิภาคที่อยู่อาศัย ซึ่งสะท้อนถึงความมั่นคงทางเศรษฐกิจในพื้นที่ หรือระดับการจ้างงานในภูมิภาคนั้นๆ โดยลูกหนี้ที่อาศัยในพื้นที่ชนบทหรือพื้นที่ที่มีความไม่แน่นอนทางเศรษฐกิจมักมีแนวโน้มเกิด FPD สูงกว่าลูกค้าในเขตเมือง (Koç & Sevgili, 2020)

2.3.4 กรณีศึกษาในระดับสากลและประเทศไทย

เพื่อให้เห็นภาพการประยุกต์ใช้ FPD ในบริบทที่แตกต่างกัน งานวิจัยและรายงานจากหลายประเทศสามารถสรุปได้เป็นกรณีศึกษาสำคัญ ดังนี้

ประเทศสหรัฐอเมริกา รายงานโดย FICO (2022) ระบุว่า อัตรา FPD เฉลี่ยในสินเชื่อบุคคลอยู่ในช่วง 2–5% โดยสถาบันการเงินได้นำข้อมูลจาก FPD ไปใช้เป็นข้อมูลตั้งต้นในการสร้างโมเดลให้คะแนนเครดิต (Credit Scoring Models) ที่มีความแม่นยำสูงขึ้น และสามารถจำแนกกลุ่มลูกค้าที่มีความเสี่ยงตั้งแต่ก่อนอนุมัติสินเชื่อได้อย่างมีประสิทธิภาพ

เวียดนามและจีน Liu, Wang และ Lin (2020) ศึกษา FPD ในกลุ่มสินเชื่อไมโครไฟแนนซ์ โดยพบว่า FPD มีความสัมพันธ์อย่างชัดเจนกับความเสี่ยงที่จะกลายเป็น NPL โดยเฉพาะในกลุ่มธุรกิจที่มีรายได้ไม่แน่นอน เช่น กลุ่มแรงงานอิสระ หรือผู้ค้ารายย่อย และในกลุ่มลูกค้าที่มีประวัติเครดิตไม่สมบูรณ์ ซึ่งบ่งชี้ว่าความไม่แน่นอนของรายได้และการขาดข้อมูลเครดิตที่น่าเชื่อถือเป็นปัจจัยเสี่ยงสำคัญ

นอกจากนี้ ธนาคารแห่งประเทศไทย (2566) ได้กำหนดให้ FPD เป็นหนึ่งในตัวชี้วัดคุณภาพของพอร์ตสินเชื่อใหม่ และใช้ในการกำกับดูแลระดับระบบเพื่อป้องกันความเสี่ยงต่อเสถียรภาพของระบบการเงินไทย

2.4 แนวคิดเกี่ยวกับกระบวนการ CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) เป็นกรอบแนวคิดมาตรฐานสากลสำหรับการทำเหมืองข้อมูล (Data Mining) และงานวิเคราะห์ข้อมูลเชิงลึก ได้รับการ

ยอมรับอย่างกว้างขวางในอุตสาหกรรมและสถาบันวิจัยทั่วโลก เนื่องจากมีโครงสร้างที่ชัดเจน ยืดหยุ่นต่อทุกประเภทข้อมูล และสามารถประยุกต์ใช้ได้กับปัญหาหลากหลายประเภท CRISP-DM เป็นวิธีที่เน้น “ความเข้าใจธุรกิจและการทำซ้ำ (Iterative)” ช่วยให้การพัฒนาโมเดลมีความถูกต้องและสอดคล้องกับเป้าหมายขององค์กรหรือโครงการ (Wirth & Hipp, 2000; Shearer, 2000)

2.4.1 ขั้นตอนหลักของ CRISP-DM

ขั้นตอนหลักของ CRISP-DM แบ่งออกเป็น 6 ขั้นตอนหลักที่วนเป็นวัฏจักร (Cyclical Process) ดังนี้

1) การทำความเข้าใจธุรกิจ (Business Understanding) เริ่มต้นด้วยการศึกษาวัตถุประสงค์ทางธุรกิจ ปัญหาที่ต้องการแก้ไข และความต้องการของผู้มีส่วนได้ส่วนเสีย จากนั้นแปลงปัญหาเหล่านั้นเป็นรูปแบบที่สามารถแก้ด้วยข้อมูลและวางแผนโครงการโดยคร่าวๆ

2) การทำความเข้าใจข้อมูล (Data Understanding) รวบรวมข้อมูลที่มีอยู่ที่เกี่ยวข้องกับปัญหา แล้วทำการสำรวจเบื้องต้น ตรวจสอบคุณภาพข้อมูล เช่น ความสมบูรณ์ ความถูกต้อง ความสอดคล้อง รวมถึงทำความเข้าใจลักษณะการกระจายตัวของข้อมูล เพื่อให้ทราบว่าข้อมูลใดบ้างที่สามารถนำมาใช้ในขั้นตอนถัดไปได้

3) การเตรียมข้อมูล (Data Preparation) ทำการปรับปรุงและแปลงรูปแบบข้อมูลดิบให้พร้อมสำหรับการสร้างโมเดลในขั้นตอนต่อไป งานส่วนนี้อาจรวมถึงการล้างข้อมูล (เช่น ลบหรือแก้ไขข้อมูลที่ผิดปกติ), การเลือกคุณลักษณะ (Feature Selection), การสร้างตัวแปรใหม่ รวมถึงการแปลงข้อมูลให้อยู่ในสเกลหรือรูปแบบที่เหมาะสมต่ออัลกอริทึม เช่น การทำ Log Transform กับข้อมูลที่มีการกระจายแบบเบ้มาก หรือการเข้ารหัสข้อมูลประเภทหมวดหมู่ด้วย One-Hot Encoding เป็นต้น โดยขั้นตอนเตรียมข้อมูลนับว่ามีความสำคัญมากเพราะคุณภาพของข้อมูลจะส่งผลโดยตรงต่อคุณภาพของโมเดล (“Garbage In, Garbage Out”)

4) การสร้างโมเดล (Modeling) ทำการเลือกเทคนิคและอัลกอริทึมการเรียนรู้ที่เหมาะสม เช่น การถดถอยโลจิสติก, การตัดสินใจเชิงต้นไม้, หรือโครงข่ายประสาทเทียม จากนั้นทำการปรับแต่ง (Train) โมเดลเหล่านั้นด้วยชุดข้อมูลฝึกอบรม ผู้วิเคราะห์อาจทดลองสร้างโมเดลหลายรูปแบบและปรับค่า Hyperparameters ของแต่ละโมเดลเพื่อค้นหารุ่นที่ให้ผลลัพธ์ดีที่สุด ทั้งนี้ หากยังไม่ได้โมเดลที่น่าพอใจ ก็สามารถย้อนกลับไปขั้นเตรียมข้อมูลเพื่อปรับปรุงข้อมูลเพิ่มเติมได้

5) การประเมินโมเดล (Evaluation) เมื่อได้โมเดลมาแล้ว จะต้องทำการประเมินประสิทธิภาพของโมเดลบนชุดข้อมูลทดสอบหรือตรวจสอบ โดยใช้ตัวชี้วัดที่เหมาะสม เช่น Accuracy, Precision, Recall, F1-Score, ROC-AUC เป็นต้น เพื่อดูว่าโมเดลมีความแม่นยำเพียงพอต่อการนำไปใช้งานหรือไม่ หากโมเดลยังไม่เป็นที่น่าพอใจ อาจต้องย้อนกลับไปขั้นตอนก่อนหน้าเพื่อปรับปรุง ไม่ว่าจะเป็นการปรับคุณลักษณะข้อมูลหรือเลือกโมเดลใหม่

6) การนำโมเดลไปใช้งานจริง (Deployment) ขั้นตอนสุดท้ายคือการนำโมเดลที่ผ่านการทดสอบแล้วไปใช้จริงในสภาพแวดล้อมการปฏิบัติงาน อาจอยู่ในรูปแบบของระบบสารสนเทศหรือแดชบอร์ดที่ให้ผู้ใช้ธุรกิจเรียกใช้งานโมเดลเพื่อพยากรณ์หรือช่วยตัดสินใจได้ โมเดลที่นำไปใช้จริงควรมีการติดตามผลและบำรุงรักษา เช่น การวัดประสิทธิผลอย่างต่อเนื่องและการปรับปรุงโมเดลเมื่อสภาพแวดล้อมเปลี่ยนแปลงไป

กระบวนการ CRISP-DM มีลักษณะ วนรอบ (Iterative) และ ยืดหยุ่น ขั้นตอนต่างๆ สามารถย้อนกลับไปมาได้ เช่น หลังประเมินโมเดลแล้วอาจพบจุดบกพร่องที่ต้องกลับไปแก้ที่ขั้นตอนการเตรียมข้อมูลหรือทำความเข้าใจธุรกิจเพิ่มเติม ทั้งนี้ CRISP-DM ได้รับการยอมรับอย่างกว้างขวางในวงการวิทยาการข้อมูลเพราะช่วยจัดระบบความคิดในการวิเคราะห์ข้อมูลได้เป็นขั้นเป็นตอน ลดความสับสนฟุ้งซ่านในโครงการที่ซับซ้อน และสามารถประยุกต์ใช้ได้ในหลากหลายอุตสาหกรรมและปัญหาทางธุรกิจ

2.4.2 ลักษณะสำคัญของ CRISP-DM

กระบวนการ CRISP-DM (Cross Industry Standard Process for Data Mining) เป็นกรอบแนวทางที่ได้รับความนิยมสูงสุดสำหรับการพัฒนาโครงการเหมืองข้อมูล (Data Mining Projects) โดยมีจุดเด่นที่สำคัญหลายประการ ลักษณะสำคัญของ CRISP-DM มีดังนี้

1) กระบวนการที่มีลักษณะการวนซ้ำ (Iterative Process) กล่าวคือ แม้ว่าจะประกอบด้วย 6 ขั้นตอน ได้แก่ การเข้าใจธุรกิจ (Business Understanding) การเข้าใจข้อมูล (Data Understanding) การเตรียมข้อมูล (Data Preparation) การสร้างโมเดล (Modeling) การประเมินผล (Evaluation) และการนำไปใช้งาน (Deployment) แต่ผู้พัฒนาสามารถย้อนกลับไปปรับปรุงขั้นตอนก่อนหน้าได้ตามความจำเป็น เช่น หากพบว่าโมเดลมีประสิทธิภาพต่ำ อาจย้อนกลับไปที่ขั้นการเตรียมข้อมูลหรือเข้าใจธุรกิจเพื่อหาสาเหตุ (Chapman et al., 2000)

2) CRISP-DM มีจุดแข็งในการ สร้างความเชื่อมโยงระหว่างนักวิทยาศาสตร์ข้อมูล (Data Scientists) กับผู้มีส่วนได้เสียทางธุรกิจ โดยการเริ่มต้นจากการวิเคราะห์ความต้องการเชิงกลยุทธ์ของธุรกิจ ช่วยให้โมเดลที่พัฒนาขึ้นตอบสนองเป้าหมายที่แท้จริงขององค์กร

3) แนวคิดที่ เน้นการเข้าใจข้อมูล (Data Understanding) ก่อนกระบวนการสร้างโมเดล (Modeling) โดยเฉพาะการสำรวจโครงสร้างของข้อมูล คุณภาพของข้อมูล และความสัมพันธ์ของตัวแปรต่างๆ ซึ่งช่วยให้สามารถออกแบบโมเดลได้อย่างมีประสิทธิภาพ (Shearer, 2000)

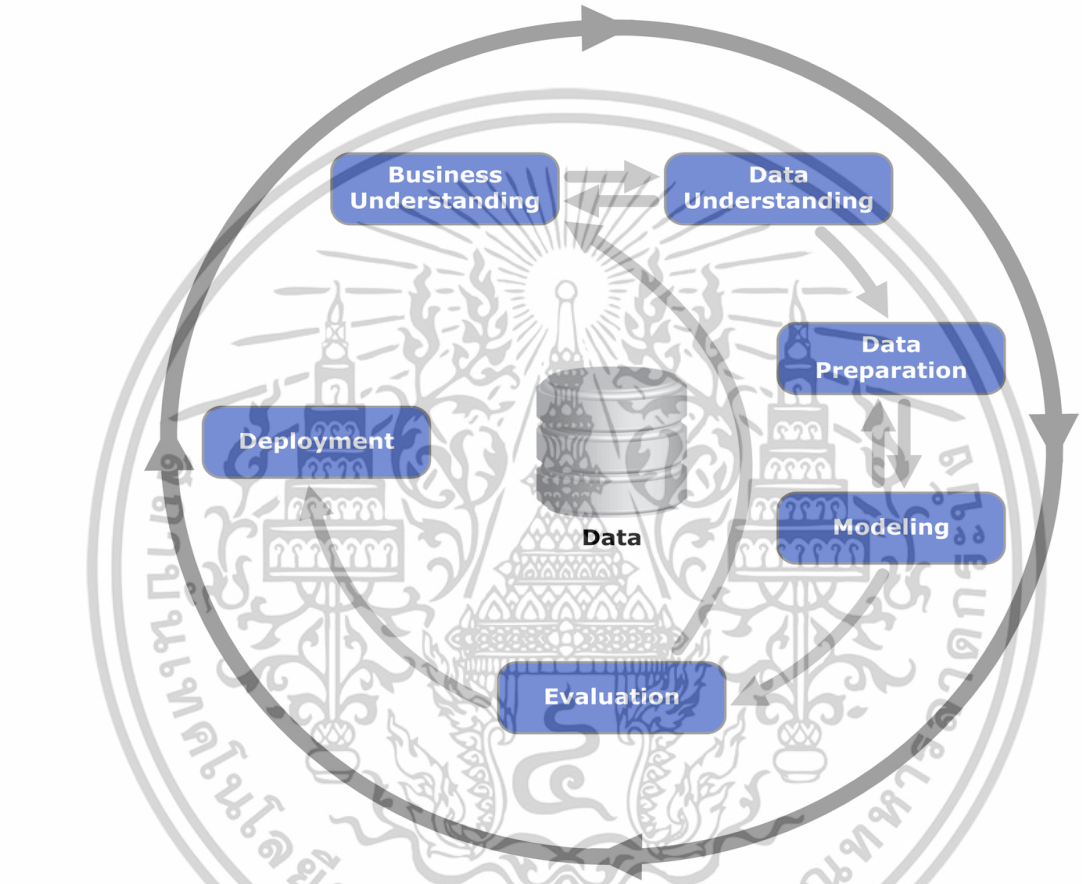
2.4.3 ข้อดีของกระบวนการ CRISP-DM

CRISP-DM ได้รับการยอมรับในวงกว้างว่าเป็นมาตรฐานสำหรับการวิเคราะห์ข้อมูลเชิงลึก เนื่องจากมีข้อดีหลายประการที่ส่งเสริมให้เกิดการพัฒนาโครงการข้อมูลได้อย่างมีประสิทธิผล ดังนี้

1) ความเป็นมาตรฐานที่เข้าใจง่ายและสามารถประยุกต์ใช้ได้กับทุกอุตสาหกรรม เนื่องจากโครงสร้างของกระบวนการไม่ได้ยึดติดกับรูปแบบเฉพาะด้านใดด้านหนึ่ง (Wirth & Hipp, 2000)

2) CRISP-DM ช่วยให้แต่ละขั้นตอนมีความเป็นระบบและสามารถตรวจสอบย้อนหลังได้ เช่น การเก็บบันทึกข้อสมมติฐานในการเลือกตัวแปร หรือการเลือกวิธีประเมินผลโมเดล ช่วยเพิ่มความโปร่งใสและทำให้สามารถตรวจสอบคุณภาพของงานได้อย่างชัดเจน

3) ความยืดหยุ่นต่อการเปลี่ยนแปลงของปัญหาหรือความต้องการทางธุรกิจ โดยเฉพาะในกรณีที่องค์กรต้องปรับกลยุทธ์หรือมีการเปลี่ยนข้อมูลที่ใช้ การมีโครงสร้างที่สามารถวนซ้ำและย้อนกลับได้ ทำให้ทีมงานสามารถปรับแนวทางได้อย่างทันท่วงทีโดยไม่ต้องเริ่มใหม่ทั้งหมด



รูปที่ 2.1 แผนภาพแสดงวัฏจักร CRISP-DM
ที่มา: Wikimedia Commons (2022)

2.5 แนวคิดและทฤษฎีเกี่ยวกับการจัดการข้อมูลไม่สมดุล (Imbalanced Data)

ปัญหาข้อมูลไม่สมดุล (Imbalanced Data) คือสถานการณ์ที่ชุดข้อมูลการฝึกมีสัดส่วนกลุ่มตัวอย่างของคลาสเป้าหมายไม่เท่ากันอย่างมาก เช่น ข้อมูลที่มีตัวอย่างของกลุ่ม “เกิดเหตุการณ์” (เช่น ผิดนัดชำระหนี้) น้อยมากเมื่อเทียบกับกลุ่ม “ไม่เกิดเหตุการณ์” ปัญหานี้ส่งผลให้โมเดลแมชชีนเลิร์นนิงที่ฝึกจากข้อมูลดังกล่าวมักเกิด Bias เข้าข้างกลุ่มใหญ่ (Majority Class) และมองข้ามการทำนายกลุ่มส่วนน้อยที่สำคัญ เพราะการทำนายทุกอย่างเป็นกลุ่มใหญ่ก็ยังได้ความแม่นยำโดยรวมสูง การแก้ไขจึงต้องอาศัยเทคนิคการปรับสมดุลข้อมูล

2.5.1 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE หรือเทคนิคการเพิ่มข้อมูลสังเคราะห์ในกลุ่มข้อมูลส่วนน้อย (Minority Class) เป็นแนวทางที่เสนอโดย Chawla et al. (2002) ซึ่งมีวัตถุประสงค์เพื่อแก้ไขปัญหาความไม่สมดุลของข้อมูล (Class Imbalance) โดยไม่ต้องลดจำนวนของกลุ่มข้อมูลส่วนใหญ่ เทคนิคนี้ทำงานโดยสร้างตัวอย่างใหม่ระหว่างจุดข้อมูลจริงในกลุ่มส่วนน้อยและเพื่อนบ้านที่ใกล้ที่สุดจำนวน k ตัว (โดยทั่วไป k=5) ที่เลือกมาจากกลุ่มเดียวกัน สำหรับแต่ละตัวอย่าง x_i ในกลุ่มส่วนน้อย SMOTE จะสุ่มเลือกเพื่อนบ้านที่ใกล้ที่สุด x_{NN} และคำนวณจุดสังเคราะห์ x_{new} ตามสมการด้านล่าง

$$x_{\text{new}} = x_i + rand(0,1) \times (x_{NN} - x_i) \tag{2.2}$$

โดยที่

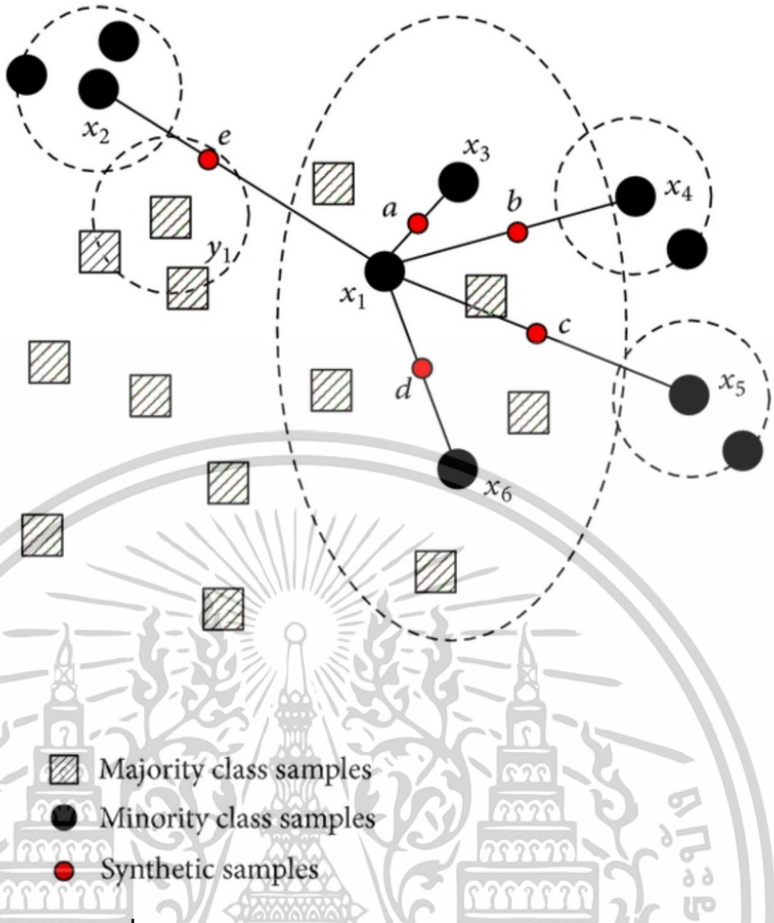
x_i คือตัวอย่างจริงใน Minority Class

x_{NN} คือเพื่อนบ้านที่ใกล้ที่สุดของ x_i

$rand(0,1)$ คือค่าสุ่มระหว่าง 0 ถึง 1

การสร้างข้อมูลในลักษณะนี้จะทำให้ตัวอย่างสังเคราะห์มีลักษณะใกล้เคียงกับข้อมูลจริงและกระจายตัวใน Feature Space ของกลุ่มส่วนน้อยได้ดีขึ้น ช่วยลดปัญหาการกระจุกตัวของข้อมูลและเพิ่มความหลากหลายให้กับข้อมูลที่สร้างขึ้น ส่งผลให้ขอบเขตการจำแนกของโมเดลมีความชัดเจนขึ้นและช่วยหลีกเลี่ยงการสูญเสียข้อมูลที่อาจเกิดจากวิธีการลดขนาดกลุ่มใหญ่ (Undersampling)

อย่างไรก็ตาม SMOTE มีข้อจำกัดที่สำคัญ เช่น หากกลุ่มข้อมูลส่วนน้อยอยู่ใกล้กลุ่มใหญ่ อาจเกิดการซ้อนทับ (Overlap) ระหว่างข้อมูลทั้งสองกลุ่ม และการเลือกค่า k ที่ไม่เหมาะสมอาจนำไปสู่การสร้างข้อมูลที่ไม่สมจริง นอกจากนี้ SMOTE ไม่สามารถจัดการกับ Noise ได้โดยตรง หากข้อมูลต้นฉบับมี Noise อยู่ ตัวอย่างใหม่ที่สร้างขึ้นก็อาจสะท้อนและแพร่กระจาย Noise เหล่านั้นออกไปด้วย (Han, Kamber, & Pei, 2011)



รูปที่ 2.2 ตัวอย่างการสร้างข้อมูลด้วย SMOTE
ที่มา: Hu and Li (2013)

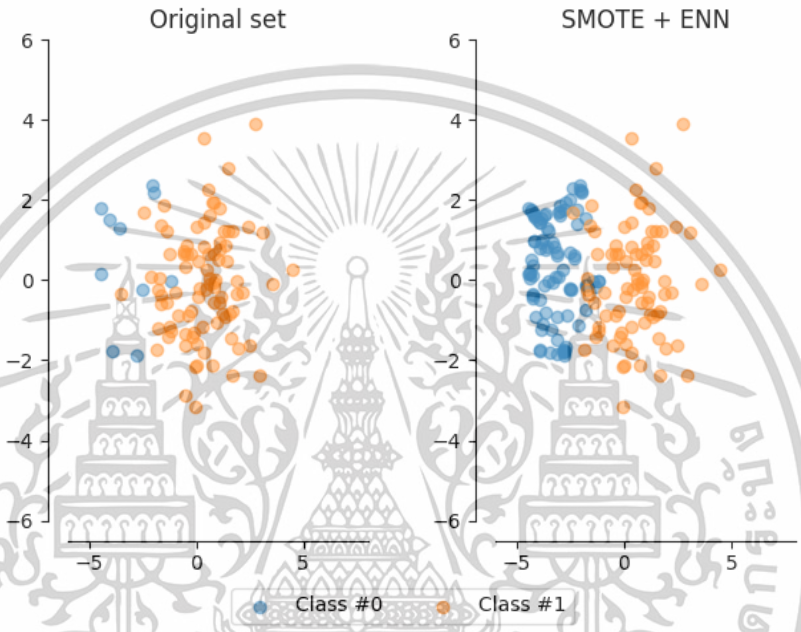
2.5.2 SMOTEENN

SMOTEENN เป็นเทคนิคการปรับสมดุลของข้อมูลที่ผสานระหว่างการสุ่มเพิ่มข้อมูลสังเคราะห์ด้วยเทคนิค SMOTE (Synthetic Minority Over-Sampling Technique) และการลบข้อมูลรบกวนด้วยวิธี ENN (Edited Nearest Neighbors) โดยมีวัตถุประสงค์เพื่อเพิ่มคุณภาพของชุดข้อมูลสำหรับการฝึกโมเดลในปัญหาการจำแนกประเภทที่มีข้อมูลไม่สมดุล กระบวนการของ SMOTEENN ประกอบด้วยสองขั้นตอนหลัก ได้แก่

1) การใช้ SMOTE เพื่อเพิ่มจำนวนตัวอย่างในกลุ่มส่วนน้อย โดยการสร้างตัวอย่างสังเคราะห์ระหว่างข้อมูลจริงกับเพื่อนบ้านที่ใกล้เคียงในกลุ่มเดียวกัน
2) การใช้เทคนิค ENN เพื่อลบข้อมูลที่อยู่ใกล้ขอบเขตการจำแนกซึ่งอาจก่อให้เกิดความสับสน โดยเฉพาะตัวอย่างในกลุ่มใหญ่ที่มีลักษณะคล้ายกับกลุ่มน้อยหรือถือเป็น Noise

การทำงานร่วมกันของสองเทคนิคนี้ก่อให้เกิดผลลัพธ์ที่ดีกว่าการใช้ SMOTE เพียงอย่างเดียว โดย SMOTE ช่วยเพิ่มความสมดุลในข้อมูล ส่วน ENN ช่วยขจัดข้อมูลรบกวนที่อาจทำให้โมเดลเกิด Overfitting หรือมีขอบเขตการจำแนกไม่ชัดเจน (Batista et al., 2004) ผลลัพธ์ที่ได้คือชุดข้อมูลที่ไม่

เพียงแต่มีความสมดุลของกลุ่มเป้าหมาย แต่ยังผ่านกระบวนการทำความสะอาดที่ช่วยให้โมเดลเรียนรู้ได้มีประสิทธิภาพมากขึ้น ทั้งในด้านความแม่นยำและความสามารถในการตรวจจับกลุ่มส่วนน้อยที่มักมีความสำคัญทางธุรกิจ เช่น กลุ่มที่มีความเสี่ยงสูงในการผิดนัดชำระหนี้ อย่างไรก็ตาม เช่นเดียวกับเทคนิคการสุ่มตัวอย่างอื่นๆ การใช้ SMOTEENN ควรปรับให้เหมาะสมกับลักษณะเฉพาะของข้อมูล เนื่องจากประสิทธิภาพของเทคนิคนี้อาจแตกต่างกันไปตามสัดส่วนของความไม่สมดุล รูปแบบของข้อมูล และโครงสร้างของ Feature Space



รูปที่ 2.3 กระบวนการทำงานของ SMOTEENN

ที่มา: Lemaître, N., et al. (2017)

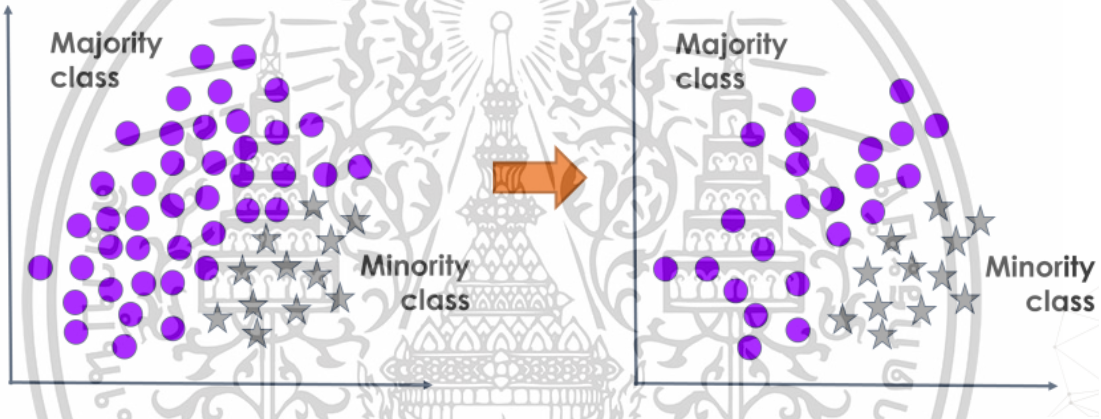
จากรูปที่ 2.3 แสดงการเปรียบเทียบลักษณะของข้อมูลก่อนและหลังการประยุกต์ใช้เทคนิค SMOTEENN โดยฝั่งซ้ายแสดงชุดข้อมูลต้นฉบับ (Original Set) ที่มีความไม่สมดุลระหว่างคลาสอย่างชัดเจน กลุ่มส่วนน้อยมีจำนวนตัวอย่างต่ำและกระจายตัวแบบไม่สม่ำเสมอ ส่งผลให้ขอบเขตการจำแนกระหว่างคลาสไม่แน่นอน อาจทำให้โมเดลเรียนรู้ได้ไม่เต็มประสิทธิภาพ ฝั่งขวาแสดงผลลัพธ์หลังจากใช้ SMOTEENN โดยขั้นตอนแรก SMOTE จะเพิ่มตัวอย่างในกลุ่มส่วนน้อยผ่านการสร้างจุดข้อมูลใหม่ระหว่างข้อมูลจริงและเพื่อนบ้าน จากนั้น ENN จะทำหน้าที่ลบข้อมูลที่อยู่ใกล้ขอบเขตการจำแนกหรือข้อมูลรบกวน โดยเฉพาะจากกลุ่มส่วนใหญ่ ทำให้โมเดลสามารถเรียนรู้ได้จากข้อมูลที่สมดุลและชัดเจนยิ่งขึ้น

2.5.3 เทคนิคการสุ่มลดข้อมูล (Random Undersampling)

เทคนิคการสุ่มลดข้อมูล (Random Undersampling) เป็นแนวทางพื้นฐานในกระบวนการปรับสมดุลของชุดข้อมูลที่ไม่สมดุล (Imbalanced Data) มักเกิดขึ้นเมื่อกลุ่มข้อมูลส่วนใหญ่ (Majority

Class) มีจำนวนตัวอย่างมากกว่ากลุ่มข้อมูลส่วนน้อย (Minority Class) อย่างมีนัยสำคัญ ส่งผลให้โมเดลการเรียนรู้ของเครื่องมีแนวโน้มเรียนรู้จากข้อมูลในกลุ่มใหญ่เป็นหลัก และละเลยการเรียนรู้จากกลุ่มส่วนน้อย ซึ่งมักเป็นกลุ่มเป้าหมายที่สำคัญ เช่น กลุ่มผู้ป่วยโรคหายาก หรือกลุ่มที่มีแนวโน้มผิดนัดชำระหนี้ (He & Garcia, 2009) เพื่อแก้ปัญหานี้ เทคนิคการสุ่มลดข้อมูลจะดำเนินการโดยการสุ่มลบตัวอย่างจากกลุ่มใหญ่ให้เหลือในสัดส่วนที่สมดุลกับกลุ่มเล็ก วิธีการนี้ช่วยลดอคติในการเรียนรู้ของโมเดล และเพิ่มโอกาสในการตรวจจับกลุ่มเป้าหมายที่มีจำนวนตัวอย่างน้อยได้แม่นยำมากขึ้น

Random UnderSampling มีข้อดีในด้านความเรียบง่าย ประหยัดเวลา และลดภาระในการประมวลผล แต่ก็มีข้อจำกัดสำคัญ คือความเสี่ยงในการลบตัวอย่างที่มีความสำคัญต่อการเรียนรู้ ซึ่งอาจนำไปสู่ปัญหา Underfitting หรือการเรียนรู้ที่ไม่สมบูรณ์ โดยเฉพาะในชุดข้อมูลที่มีลักษณะเชิงมิติสูง (High-dimensional space) หรือมีความซับซ้อนของโครงสร้าง (Dal Pozzolo et al., 2015)



รูปที่ 2.4 กระบวนการของเทคนิคการสุ่มลดข้อมูล (Random UnderSampling)
ที่มา: Train in Data (2022)

จากรูปที่ 2.4 แสดงให้เห็นหลักการของเทคนิคการสุ่มลดข้อมูล (Random UnderSampling) ซึ่งเป็นหนึ่งในวิธีการปรับสมดุลของชุดข้อมูลที่ไม่สมดุล (Imbalanced Data) โดยภาพดังกล่าวแบ่งออกเป็นสองส่วนเพื่อเปรียบเทียบสถานะของข้อมูลก่อนและหลังการประยุกต์ใช้เทคนิคดังกล่าว ในฝั่งซ้ายของภาพ แสดงชุดข้อมูลก่อนการปรับสมดุล ซึ่งประกอบด้วยจุดข้อมูลของกลุ่มส่วนใหญ่ (Majority Class) ในสีม่วงจำนวนมาก และจุดข้อมูลของกลุ่มส่วนน้อย (Minority Class) ในสีเทาจำนวนจำกัด การกระจายตัวของข้อมูลในลักษณะนี้ทำให้โมเดลการเรียนรู้มีแนวโน้มที่จะโฟกัสกับกลุ่มใหญ่ และละเลยข้อมูลจากกลุ่มเล็ก ซึ่งส่งผลต่อประสิทธิภาพของโมเดลโดยเฉพาะในด้านการตรวจจับกลุ่มเป้าหมายที่มีความสำคัญ ในขณะที่ฝั่งขวาของภาพ แสดงผลลัพธ์หลังจากการประยุกต์ใช้ Random Undersampling โดยมีการลบจุดข้อมูลบางส่วนจากกลุ่มใหญ่แบบสุ่ม เพื่อให้จำนวนของข้อมูลทั้งสองกลุ่มมีความใกล้เคียงกันมากขึ้น ส่งผลให้ชุดข้อมูลโดยรวมมีความสมดุลยิ่งขึ้น ซึ่งเอื้อต่อการเรียนรู้ของโมเดลในเชิงเท่าเทียม

2.6 แนวคิดและทฤษฎีเกี่ยวกับอัลกอริธึม Machine Learning สำหรับการจำแนกประเภท

ในงานวิจัยนี้ได้ใช้แนวคิดและทฤษฎี Machine Learning สำหรับการจำแนกประเภทดังนี้

2.6.1 Logistic Regression

Logistic Regression ถือเป็นหนึ่งในโมเดลพื้นฐานของสถิติและ Machine Learning สำหรับปัญหาการจำแนกประเภทแบบสองกลุ่ม (Binary Classification) ที่ได้รับความนิยมอย่างสูงในงานประยุกต์ด้านการเงิน โมเดลนี้ไม่ได้เป็นเพียงการถดถอยเชิงเส้นธรรมดา แต่ใช้ฟังก์ชันโลจิสติก (Logistic Function หรือ Sigmoid Function) เพื่อจำกัดค่าทำนายให้อยู่ในช่วง 0 ถึง 1 อันเป็นขอบเขตของ “ความน่าจะเป็น” (Probability) ในเชิงสถิติ สมมติฐานสำคัญของ Logistic Regression คือ ความสัมพันธ์ระหว่าง Log-Odds ของการอยู่ในกลุ่มเป้าหมายกับตัวแปรอิสระ (Feature) เป็นเส้นตรง ดังสมการด้านล่าง

$$\log\left(\frac{P(y=1\mid X)}{1-P(y=1\mid X)}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p \tag{2.3}$$

ฟังก์ชันโลจิสติกจะปรับค่าทำนายให้เป็น “ความน่าจะเป็น” ที่เข้าใจง่าย ดังสมการนี้

$$P(y=1\mid X)=\frac{1}{1+e^{-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_j\right)}} \tag{2.4}$$

จุดเด่นด้านทฤษฎีคือตีความง่าย ค่าสัมประสิทธิ์ β บ่งบอกว่า ตัวแปรอิสระ x_j มีผลต่อ Log-Odds หรือโอกาสที่ข้อมูลจะเป็นกลุ่มเป้าหมายอย่างไร Probabilistic Output โมเดลนี้สามารถประเมิน “ความมั่นใจ” ในการจำแนกกลุ่มของแต่ละตัวอย่าง ไม่ใช่แค่ระบุคลาส เหมาะกับงานที่ต้องการเหตุผล ใช้บ่อยในงานธนาคาร (Credit Scoring) โดยขั้นตอนการฝึกโมเดลจะใช้ Maximum Likelihood Estimation (MLE) เพื่อประมาณค่าสัมประสิทธิ์ที่เหมาะสม จากนั้นใช้ Gradient Descent หรือ Newton-Raphson Method ในการหาค่าที่เหมาะสมที่สุด ข้อดีคือ ตีความค่าพารามิเตอร์ตรงไปตรงมา ใช้เวลาในการฝึกไม่นาน ตอบโจทย์ปัญหาที่ความสัมพันธ์เป็นเชิงเส้น ข้อจำกัดคือ ประสิทธิภาพต่ำเมื่อความสัมพันธ์ข้อมูลไม่เชิงเส้น หรือมี Interaction สูง อ่อนไหวต่อ Outlier และ Multicollinearity

2.6.2 Random Forest

Random Forest เป็นอัลกอริธึมในกลุ่มของเทคนิค **Ensemble Learning** ซึ่งใช้แนวคิดการรวมการตัดสินใจของโมเดลหลายต้นไม้ตัดสินใจ (Decision Trees) เข้าด้วยกัน เพื่อเพิ่มความ

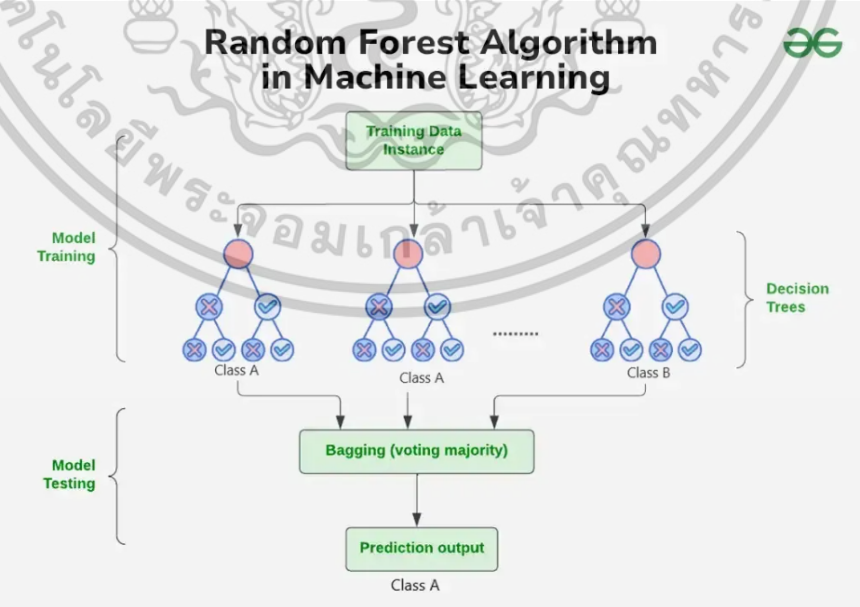
แม่นยำและลดความเอนเอียง (Bias) ของโมเดลเดี่ยว โดยเฉพาะอย่างยิ่ง Random Forest ได้รับความนิยมอย่างกว้างขวางในปัญหาการจำแนกประเภท (Classification) และการถดถอย (Regression) เนื่องจากมีความสามารถในการจัดการกับข้อมูลที่ซับซ้อนและมีมิติสูงได้ดี (Breiman, 2001) หัวใจของการทำงานของ Random Forest คือการสร้าง “ความหลากหลาย” (Diversity) ให้กับต้นไม้แต่ละต้นใน “ป่า” โดยใช้เทคนิคหลัก 2 ประการดังนี้

1) Bootstrap Sampling: แต่ละต้นไม้ในป่าจะได้รับชุดข้อมูลสำหรับฝึกที่ถูกสุ่มเลือกจากข้อมูลทั้งหมดโดยสามารถสุ่มซ้ำได้ (Sampling With Replacement) เทคนิคนี้เรียกว่า Bagging (Bootstrap Aggregating) ซึ่งช่วยลดความแปรปรวน (Variance) ของโมเดล (Breiman, 1996)

2) Random Feature Selection: ที่แต่ละจุดแยก (Split) ของต้นไม้ จะพิจารณาเพียงชุดย่อยของฟีเจอร์ทั้งหมด (Random Subset of Features) เพื่อใช้ในการตัดสินใจเท่านั้น วิธีนี้ช่วยลดความสัมพันธ์กันระหว่างต้นไม้ต่างๆ และเพิ่มความหลากหลายของโมเดลภายใน Random Forest

โดยกระบวนการของ Random Forest สามารถสรุปได้ดังนี้ สร้างต้นไม้จำนวน n ต้น โดยแต่ละต้นฝึกด้วยชุดข้อมูลที่สุ่มด้วยการทำ Bootstrap Sampling และใช้ฟีเจอร์แบบสุ่มในการแบ่งข้อมูล เมื่อมีอินพุตใหม่เข้ามา แต่ละต้นไม้จะให้การทำนาย และนำผลลัพธ์ที่ได้มาใช้ในการโหวต (ในกรณี Classification) หรือเฉลี่ยผล

แนวทางนี้ทำให้ Random Forest มีความสามารถในการลด Overfitting เมื่อเปรียบเทียบกับการใช้ต้นไม้เพียงต้นเดียว และยังสามารถประเมินความสำคัญของตัวแปร (Feature Importance) ได้อีกด้วย โดยคำนวณจากการลดค่าความไม่บริสุทธิ์ (Impurity Reduction) เช่น Gini Importance หรือ Mean Decrease in Impurity



รูปที่ 2.5 กระบวนการของ Random Forest ใน Machine Learning

ที่มา: GeeksforGeeks (2025)

Random Forest เป็นเทคนิคการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) ที่มีจุดเด่นในการเพิ่มความแม่นยำของการจำแนกและการถดถอยผ่านการรวมผลการตัดสินใจจากต้นไม้หลายต้น (Decision Trees) เข้าด้วยกัน โดยมีคุณสมบัติที่สำคัญซึ่งช่วยให้โมเดลนี้เป็นที่นิยมใช้อย่างแพร่หลายในการวิเคราะห์ข้อมูลเชิงตาราง (Tabular Data) ดังนี้

1) Random Forest มีความสามารถในการลดปัญหา Overfitting ได้อย่างมีประสิทธิภาพ โดยใช้กระบวนการสุ่มข้อมูลฝึก (Bootstrap Sampling) และการสุ่มเลือกคุณลักษณะ (Feature Subsets) ในการสร้างต้นไม้แต่ละต้น ทำให้ต้นไม้แต่ละต้นมีโครงสร้างที่แตกต่างกัน

2) Random Forest มีความสามารถในการประเมินความสำคัญของตัวแปร (Feature Importance) ได้อย่างเป็นระบบ โดยอิงจากค่าการลดลงของความไม่บริสุทธิ์ (Impurity Reduction) เช่น ค่า Gini Index หรือ Entropy ที่เกิดจากการแบ่งข้อมูลในแต่ละโหนดของต้นไม้หลายต้น ซึ่งเป็นประโยชน์อย่างยิ่งในการวิเคราะห์ปัจจัยที่มีผลต่อผลลัพธ์ และนำไปใช้ในกระบวนการเลือกตัวแปร (Feature Selection) เพื่อปรับปรุงประสิทธิภาพของโมเดล

3) Random Forest มีความทนทานต่อข้อมูลที่มี Missing Values และ Outliers ได้ดี โดยเนื่องจากลักษณะการรวมผลจากหลายต้นไม้ช่วยลดผลกระทบจากค่าผิดปกติในข้อมูลบางชุดย่อย และไม่จำเป็นต้องมีการเติมค่าข้อมูล (Imputation) ก่อนการฝึกเสมอไป ทำให้เหมาะสมกับการประยุกต์ในสถานการณ์ที่ข้อมูลไม่สมบูรณ์ (Breiman, 2001; Liaw & Wiener, 2002)

แม้ Random Forest จะมีจุดแข็งหลายประการ แต่ก็มีข้อจำกัดที่ควรพิจารณา โดยเฉพาะในด้านการตีความผลลัพธ์ที่ซับซ้อนและไม่สามารถอธิบายความสัมพันธ์เชิงลึกระหว่างตัวแปรได้อย่างชัดเจน จึงมักถูกจัดให้อยู่ในกลุ่มของ “โมเดลกล่องดำ” (Black-Box Models) นอกจากนี้ การฝึกและทดสอบโมเดล Random Forest ต้องใช้ทรัพยากรคำนวณสูง ทั้งในแง่ของหน่วยประมวลผลและหน่วยความจำ โดยเฉพาะเมื่อมีการฝึกต้นไม้จำนวนมากหรือใช้กับชุดข้อมูลขนาดใหญ่ ซึ่งอาจเป็นอุปสรรคในแง่ของเวลาและประสิทธิภาพของการใช้งาน (Breiman, 2001; Liaw & Wiener, 2002)

2.6.3 XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) เป็นโมเดลที่พัฒนาขึ้นจากแนวคิด Gradient Boosting โดยเน้นการเพิ่มประสิทธิภาพทั้งด้านความเร็ว ความแม่นยำ และความยืดหยุ่นของโมเดลด้วยคุณสมบัติที่โดดเด่น XGBoost จึงได้รับความนิยมอย่างแพร่หลายในวงการวิเคราะห์ข้อมูลเชิงแข่งขัน เช่น Kaggle รวมถึงการใช้งานจริงในภาคธุรกิจ

หัวใจสำคัญของ XGBoost คือการฝึกต้นไม้แบบลำดับ (Stage-Wise Additive) โดยในแต่ละรอบการฝึก ต้นไม้ใหม่จะถูกสร้างขึ้นเพื่อเรียนรู้จากข้อผิดพลาด (Residual) ของโมเดลก่อนหน้า จากนั้นจะใช้ Gradient Descent ในการปรับพารามิเตอร์ให้ลดค่าฟังก์ชันสูญเสีย (Loss Function) ลงเรื่อยๆ พร้อมทั้งใช้ฟังก์ชัน Regularization ควบคุมความซับซ้อนของโครงสร้างต้นไม้ เพื่อลดความเสี่ยงของการเกิด Overfitting โดยสมการการทำนายโดยรวมของ XGBoost มีลักษณะดังนี้

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{2.5}$$

โดยที่

f_k คือฟังก์ชันการทำนายของต้นไม้ต้นที่ k

ฟังก์ชันวัตถุประสงค์ (Objective Function) ที่ใช้ใน XGBoost ประกอบด้วยสองส่วน ได้แก่ ค่าความสูญเสีย (Loss Function) และพารามิเตอร์ควบคุมความซับซ้อน (Regularization Term)

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.6}$$

โดยที่

l คือฟังก์ชัน Loss เช่น Logistic Loss หรือ Mean Squared Error

Ω(*fk*) คือ Regularization term ซึ่งอาจรวมถึงจำนวนโหนดในต้นไม้ ความลึก และค่าพารามิเตอร์ของโหนดใบ เพื่อควบคุมไม่ให้โมเดลซับซ้อนเกินไป

กระบวนการฝึกโมเดลใน XGBoost ประกอบด้วย

1) ฝึกต้นไม้ทีละต้นตามแนวคิด Stage---Wise Additive Learning
2) แต่ละต้นจะพยายามลดข้อผิดพลาดจากต้นก่อนหน้า โดยเรียนรู้จาก Gradient ของ Loss Function
3) ใช้เทคนิคเสริมเพื่อป้องกัน Overfitting ได้แก่ 1) Shrinkage (Learning Rate) เป็นการปรับขนาดของการอัปเดตในแต่ละรอบ 2) Column Subsampling เลือกใช้เพียงบางตัวแปรในแต่ละต้นไม้ เพื่อเพิ่มความหลากหลายของโมเดล และ 3) Early Stopping เป็นการหยุดการฝึกเมื่อประสิทธิภาพบนชุด Validation ไม่ดีขึ้น

XGBoost ได้รับการยอมรับอย่างกว้างขวางว่าเป็นหนึ่งในโมเดลที่มีประสิทธิภาพสูงในงานด้านการจำแนก (Classification) และการถดถอย (Regression) โดยเฉพาะอย่างยิ่งกับข้อมูลเชิงตาราง (Structured/Tabular Data) ซึ่งมีความสำคัญอย่างยิ่งในการวิเคราะห์ทางการเงิน โมเดลนี้ได้รับคะแนนสูงในแพลตฟอร์มการแข่งขันด้าน Machine Learning อย่าง Kaggle เนื่องจากมีประสิทธิภาพสูงในการจัดการข้อมูลโครงสร้างตาราง อีกทั้งยังสามารถรองรับข้อมูลที่มีค่าขาดหาย (Missing Values) ได้โดยไม่จำเป็นต้องเติมค่าก่อนการฝึก ซึ่งช่วยลดภาระในการเตรียมข้อมูล นอกจากนี้ยังมีความยืดหยุ่นสูงในการปรับแต่งพารามิเตอร์ต่างๆ เช่น จำนวนต้นไม้ ความลึก หรือค่าความซับซ้อนของโมเดล จึงสามารถปรับให้เหมาะสมกับลักษณะของข้อมูลได้อย่างมีประสิทธิภาพ

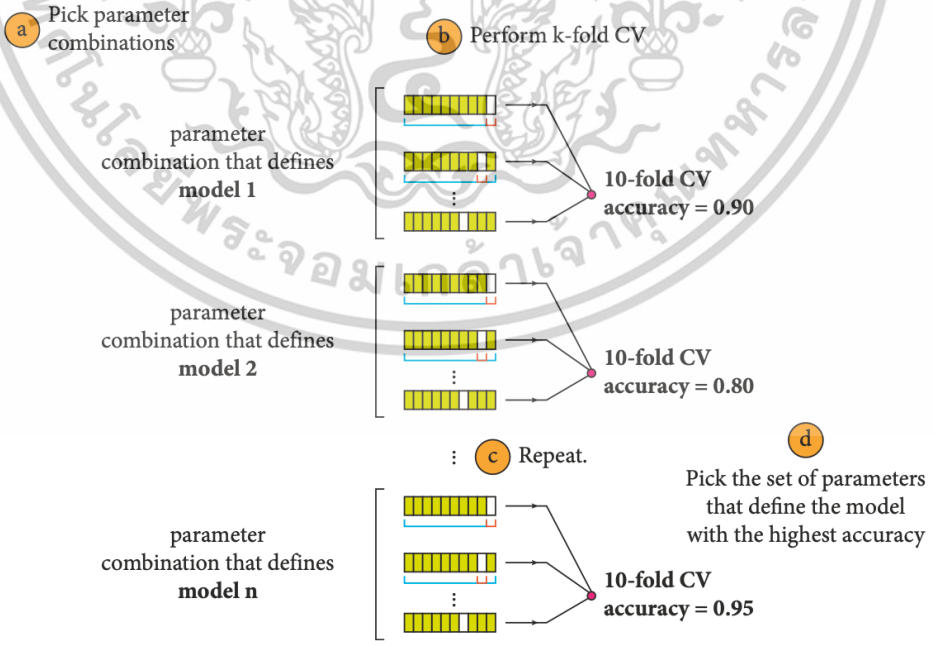
อย่างไรก็ตาม แม้ XGBoost จะมีจุดแข็งหลายประการ แต่ก็มีข้อจำกัดที่ควรพิจารณาในการใช้งานจริง โดยเฉพาะในการประยุกต์ใช้ในระบบขององค์กร เนื่องจากต้องอาศัยการปรับจูน

พารามิเตอร์จำนวนมากเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ซึ่งอาจใช้เวลาและทรัพยากรสูงในการทดลองหาค่าที่เหมาะสมที่สุด อีกทั้งการตีความผลลัพธ์ของโมเดลในเชิงลึกก็เป็นเรื่องที่ซับซ้อน ทำให้ XGBoost จัดอยู่ในประเภท “กล่องดำ” (Black-Box Model) ซึ่งอาจไม่เหมาะสมในบริบทที่ต้องการความโปร่งใสในการตัดสินใจ (Chen & Guestrin, 2016; Friedman, 2001)

2.7 แนวคิดและทฤษฎีเกี่ยวข้องกับการเลือกไฮเปอร์พารามิเตอร์แบบกริด (Grid Search Cross-Validation)

การเลือกไฮเปอร์พารามิเตอร์ (Hyperparameter Tuning) เป็นขั้นตอนสำคัญในกระบวนการพัฒนา Machine Learning ที่มีผลต่อประสิทธิภาพของโมเดลอย่างมีนัยสำคัญ ไฮเปอร์พารามิเตอร์คือพารามิเตอร์ที่ไม่ถูกปรับค่าผ่านกระบวนการเรียนรู้โดยตรง แต่ต้องกำหนดไว้ล่วงหน้า เช่น ค่าความลึกของต้นไม้ (Max_Depth) ใน Decision Tree ตัวเลือกที่เหมาะสมของไฮเปอร์พาพารามิเตอร์จะช่วยเพิ่มความแม่นยำ ลดความซับซ้อนเกินจำเป็น (Overfitting) และทำให้โมเดลทั่วไปกับข้อมูลใหม่ได้ดีขึ้น (Bergstra & Bengio, 2012)

Grid Search Cross-Validation เป็นวิธีมาตรฐานและได้รับความนิยมสูงในการเลือกไฮเปอร์พารามิเตอร์ โดยแนวคิดหลักคือการกำหนดชุดค่าของไฮเปอร์พารามิเตอร์ที่ต้องการทดลอง จากนั้นสร้าง "กริด" แล้วนำไปประเมินโมเดลแต่ละชุดด้วยเทคนิค Cross-Validation ซึ่งหมายถึงการแบ่งข้อมูลออกเป็นหลายส่วน (Folds) เพื่อฝึกและทดสอบโมเดลซ้ำๆ ในแต่ละชุดค่าพารามิเตอร์ (Kohavi, 1995) วิธีนี้ช่วยให้ได้ผลการประเมินที่มีความเสถียรและลดอคติที่อาจเกิดจากการแบ่งข้อมูลครั้งเดียว



รูปที่ 2.6 ขั้นตอนการทำงานของ Grid Search Cross-Validation

ที่มา : Fathi et al. (2021)

จากรูปที่ 2.6 เป็นกระบวนการ Grid Search Cross-Validation จะทำงานโดยอัตโนมัติผ่านขั้นตอนต่อเนื่อง ได้แก่

1) Pick Parameter Combinations: เริ่มจากการกำหนดชุดของค่าพารามิเตอร์ที่ต้องการทดลอง เช่น Max_Depth หรือ Min_Samples_Split
2) Perform K-Fold CV: สำหรับแต่ละชุดค่าพารามิเตอร์ ทำการแบ่งข้อมูลเป็น K ส่วน (เช่น 5 หรือ 10 ส่วน) เพื่อทำการ Cross-Validation โดยในแต่ละรอบจะนำข้อมูลบางส่วนมาใช้ฝึกสอนโมเดล และใช้ส่วนที่เหลือในการทดสอบประสิทธิภาพของโมเดลเพื่อวัดค่าต่างๆ เช่น Accuracy, Recall หรือ F1-Score สำหรับชุดพารามิเตอร์นั้นๆ แล้วคำนวณค่าเฉลี่ยประสิทธิภาพจากทุก Fold (เช่น Accuracy เฉลี่ย 0.95, 0.80)
3) Repeat: ขั้นตอนนี้จะทำซ้ำกับทุกชุดพารามิเตอร์ที่กำหนดไว้
4) Select Best Parameter: เลือกชุดค่าพารามิเตอร์ที่ให้ประสิทธิภาพดีที่สุดเพื่อนำไปใช้กับโมเดล

ข้อดีของ Grid Search Cross-Validation คือความครอบคลุมและความมั่นใจว่าไม่มีการมองข้ามค่าพารามิเตอร์ที่เป็นไปได้ทั้งหมด (Exhaustive Search) ซึ่งช่วยป้องกันปัญหาการเลือกค่าที่ไม่เหมาะสมโดยอาศัยประสบการณ์หรือความรู้สึก อย่างไรก็ตามข้อจำกัดสำคัญคืออาจใช้ทรัพยากรคำนวณสูง เมื่อจำนวนไฮเปอร์พารามิเตอร์หรือขอบเขตของค่าแต่ละตัวมีมาก (Curse of Dimensionality) ทำให้ในกรณีที่มีพารามิเตอร์จำนวนมากหรือชุดข้อมูลขนาดใหญ่ นักวิจัยอาจเลือกใช้วิธีสุ่ม (Random Search) หรือ Bayesian Optimization แทน เพื่อเพิ่มประสิทธิภาพและลดเวลาคำนวณ (Bergstra & Bengio, 2012) Grid Search Cross-Validation จึงเป็นเครื่องมือสำคัญที่ช่วยให้กระบวนการเลือกไฮเปอร์พารามิเตอร์มีความเป็นระบบ โดยเฉพาะเมื่อนำไปใช้กับงานที่ต้องการผลลัพธ์ที่มั่นใจและโปร่งใส (Pedregosa et al., 2011)

2.8 แนวคิดและทฤษฎีเกี่ยวข้องกับการประเมินประสิทธิภาพในการทำนายของโมเดล

การประเมินประสิทธิภาพของโมเดลเป็นขั้นตอนสำคัญในการวิเคราะห์โมเดล Machine Learning เพื่อให้ทราบถึงความสามารถในการทำนายและข้อจำกัดของโมเดล โดยเฉพาะในปัญหาการจำแนกประเภท (Classification) ซึ่งนิยมใช้เมทริกซ์ต่างๆ เช่น Confusion Matrix และ Classification Report สำหรับการวิเคราะห์เชิงลึก

2.8.1 เมทริกซ์ความสับสน (Confusion Matrix)

Confusion Matrix คือเครื่องมือที่ช่วยประเมินและวิเคราะห์ผลการทำนายของโมเดลจำแนกประเภท โดยจะแสดงจำนวนของแต่ละกรณีที่โมเดลทำนายถูกหรือผิดเมื่อเปรียบเทียบกับค่าจริง (Fawcett, 2006; Powers, 2011) Confusion Matrix ประกอบด้วย 4 องค์ประกอบหลักดังนี้

1) True Positive (TP) คือ จำนวนตัวอย่างที่โมเดลสามารถทำนายได้ถูกต้องว่าอยู่ในกลุ่มบวก (Positive Class)
2) True Negative (TN) คือ จำนวนตัวอย่างที่โมเดลทำนายได้ถูกต้องว่าอยู่ในกลุ่มลบ (Negative Class)
3) False Positive (FP) คือ จำนวนตัวอย่างที่โมเดลทำนายผิดโดยระบุว่าเป็นกลุ่มบวก ทั้งที่จริงแล้วเป็นกลุ่มลบ (Type I Error)
4) False Negative (FN) คือ จำนวนตัวอย่างที่โมเดลทำนายผิดโดยระบุว่าเป็นกลุ่มลบ ทั้งที่จริงแล้วเป็นกลุ่มบวก (Type II Error)

เมทริกซ์ความสับสน (Confusion Matrix) นี้ทำให้ผู้ใช้งานเข้าใจประสิทธิภาพของโมเดลได้อย่างลึกซึ้ง ทั้งในด้านการวัดความถูกต้องและการวิเคราะห์ข้อผิดพลาดจากการทำนาย Confusion Matrix สามารถนำเสนอได้ 2 รูปแบบ คือ

1) ค่าที่ทำนายกับค่าจริง (Predict–Actual)

Predicted \ Actual	Default	Non default
Default	True Positive (TP)	False Positive (FP)
Non default	False Negative (FN)	True Negative (TN)

รูปที่ 2.7 เมทริกซ์ความสับสน (Confusion Matrix) แบบค่าที่ทำนายกับค่าจริง (Predict-Actual)

2) ค่าจริงกับค่าที่ทำนาย (Actual-Predict)

Actual \ Predicted	Default	Non default
Default	True Positive (TP)	False Negative (FN)
Non default	False Positive (FP)	True Negative (TN)

รูปที่ 2.8 เมทริกซ์ความสับสน (Confusion Matrix) แบบค่าจริงกับค่าที่ทำนาย (Actual-Predict)

Confusion Matrix ทำให้ผู้ใช้งานเข้าใจลักษณะข้อผิดพลาดของโมเดลได้อย่างลึกซึ้ง ทั้งด้านการวัดความถูกต้องและการวิเคราะห์ข้อผิดพลาดในแต่ละกลุ่มเป้าหมาย (Géron, 2019)

2.8.2 การจำแนกประเภท (Classification Report)

Classification Report เป็นการแสดงผลวิเคราะห์ประสิทธิภาพของโมเดลจำแนกประเภทในระบบ Machine Learning โดยจะรายงานตัวชี้วัดที่สำคัญ เช่น ความถูกต้อง (Accuracy) ความระลึก (Recall) ความแม่นยำ (Precision) และค่าประสิทธิภาพโดยรวม (F1-Score) ซึ่งแต่ละค่ามีความสำคัญและช่วยสะท้อนประสิทธิภาพของโมเดลในแง่มุมต่างๆ (Scikit-Learn Developers, n.d.) ตัวชี้วัดที่สำคัญในการจำแนกประเภท (Classification Report) ประกอบด้วย

1) ค่าความเที่ยงตรงหรือความถูกต้อง (Accuracy) คือ อัตราส่วนของจำนวนข้อมูลที่ทำนายได้ถูกต้องทั้งหมดต่อจำนวนข้อมูลทั้งหมด คำนวณโดยใช้สูตร

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (2.7)$$

2) ค่าความครบถ้วนหรือค่าความระลึก (Recall คือ อัตราส่วนของจำนวนข้อมูลที่ทำนายได้ถูกต้องในคลาสบวกต่อจำนวนข้อมูลที่ทำนายได้จริงของคลาสบวก คำนวณโดยใช้สูตร

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2.8)$$

3) ค่าความแม่นยำ (Precision) คือ อัตราส่วนของจำนวนข้อมูลที่ทำนายได้ถูกต้องในคลาสบวกต่อข้อมูลที่ถูกต้องในคลาสบวกและจำนวนข้อมูลที่ทำนายผิดพลาดในคลาส คำนวณโดยใช้สูตร

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2.9)$$

4) ค่าประสิทธิภาพโดยรวม (F1-Score หรือ F-Measure) เป็นค่าเฉลี่ยแบบ Harmonic ของค่าความระลึก (Recall) และค่าความแม่นยำ (Precision) ให้ค่าระหว่าง 0 ถึง 1 ซึ่งค่าที่สูงขึ้นหมายความว่าประสิทธิภาพของโมเดลดีขึ้น คำนวณโดยใช้สูตร

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.10)$$

โดยค่า F1-Score สามารถอธิบายได้ดังนี้

1) ถ้าค่า F1-Score มีค่าสูง หมายความว่า Precision และ Recall มีค่าดีทั้งคู่
2) ถ้าค่า F1-Score ต่ำ แสดงว่ามีอย่างน้อยหนึ่งค่าที่ต่ำ หรือทั้งสองค่าต่ำ (Géron, 2019)

Classification Report ยังมักแสดงค่า “Support” คือ จำนวนตัวอย่างจริงในแต่ละกลุ่ม เพื่อประกอบการวิเคราะห์เชิงปริมาณและเปรียบเทียบประสิทธิภาพของโมเดลอย่างเป็นระบบ โดยเฉพาะในปัญหาที่ข้อมูลแต่ละกลุ่มมีขนาดไม่เท่ากัน (Imbalanced Data) (Géron, 2019).

2.8.3 ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

ROC-AUC เป็นเครื่องมือที่ใช้วัดประสิทธิภาพของโมเดลการจำแนกประเภท (Classification Model) โดยเฉพาะอย่างยิ่งในบริบทของการประเมินความสามารถในการแยกแยะกลุ่มเป้าหมาย เช่น กลุ่มลูกหนี้ที่เสี่ยงผิดนัดและไม่เสี่ยง ซึ่งมีการนำไปใช้อย่างแพร่หลายในอุตสาหกรรมธนาคารและการเงิน เช่น การให้คะแนนเครดิต (Credit Scoring) และการสร้างโมเดลความเสี่ยง (Risk Modeling) (Siddiqi, 2006; Anderson, 2007) ROC (Receiver Operating Characteristic) เป็นกราฟที่แสดงความสัมพันธ์ระหว่างค่า True Positive Rate (TPR) และ False Positive Rate (FPR) ที่ได้จากการทำนายภายใต้ Threshold หลายค่า โดย

$$\text{TPR} = \text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{2.11}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.12}$$

เส้น ROC จะแสดงผลลัพธ์ของโมเดลในมิติของความสามารถในการแยกแยะตัวแปรเป้าหมาย โดยไม่อิงค่า threshold ใดๆ เป็นการเฉพาะ

AUC (Area Under the Curve) คือค่าพื้นที่ใต้กราฟ ROC ซึ่งใช้แทนระดับประสิทธิภาพของโมเดล โดยมีค่าระหว่าง 0 ถึง 1 โดย

AUC = 1.0 หมายถึง โมเดลสามารถแยกแยะกลุ่มเป้าหมายได้อย่างสมบูรณ์แบบ

AUC = 0.5 หมายถึง โมเดลไม่มีความสามารถในการจำแนกเลย และเทียบเท่าการเดาสุ่ม

AUC < 0.5 อาจหมายถึงโมเดลมีพฤติกรรมตรงกันข้ามกับที่ต้องการ

จุดเด่นของ ROC-AUC คือความสามารถในการประเมินคุณภาพของโมเดลโดยไม่ขึ้นอยู่กับ threshold ที่ใช้ในการจำแนก ทำให้เหมาะสมกับกรณีที่ต้องการเปรียบเทียบโมเดลหลายแบบ หรือใช้ในกรณีที่ Class Imbalance มีผลต่อ Precision หรือ Recall โดยตรง ด้วยเหตุนี้ ROC-AUC จึงได้รับ

ความนิยมในงานวิเคราะห์สินเชื่อและระบบให้คะแนนความเสี่ยงที่ต้องการความแม่นยำสูงในการจำแนกกลุ่มเป้าหมาย เช่น ลูกหนี้ดีและลูกหนี้เสีย

2.8.4 KS (Kolmogorov-Smirnov Statistic)

เป็นสถิติที่ใช้วัดความแตกต่างระหว่างการกระจายของกลุ่ม “Good” และ “Bad” (เช่น กลุ่มไม่ผิดนัดชำระหนี้กับกลุ่มผิดนัดชำระหนี้) จากผลลัพธ์ที่โมเดลทำนาย หลักการและวิธีคิดคือ คำนวณ Cumulative Distribution Function (CDF) ของ “Good” และ “Bad” ตามค่าคะแนนที่โมเดลทำนาย KS คือ “ค่าสูงสุดของความแตกต่างระหว่าง CDF ของ Good และ Bad” ตลอดช่วงคะแนน

$$KS = \max_{x} \left| F_{\text{good}}(x) - F_{\text{bad}}(x) \right| \tag{2.13}$$

โดยที่

$F_{good}(x)$ คือ สัดส่วนสะสมของกลุ่ม Good ที่ได้คะแนนไม่เกิน x

$F_{bad}(x)$ คือ สัดส่วนสะสมของกลุ่ม Bad ที่ได้คะแนนไม่เกิน x

การตีความค่า KS (Kolmogorov-Smirnov Statistic) มักยึดตามเกณฑ์มาตรฐานที่ใช้ในอุตสาหกรรมการเงินและงานด้านการจำแนกประเภท โดยทั่วไป ค่า KS ที่ดีควรมีค่ามากกว่า 0.30 หรือ 30% ซึ่งบ่งชี้ว่าโมเดลมีพลังในการแยกแยะกลุ่มเป้าหมายได้ในระดับที่ยอมรับได้ หากโมเดลมีค่า KS อยู่ในช่วงระหว่าง 0.40 ถึง 0.50 แสดงว่าโมเดลมีความสามารถในการจำแนกกลุ่ม Good (เช่น ลูกหนี้ที่ไม่ผิดนัด) และ Bad (เช่น ลูกหนี้ที่ผิดนัด) ได้อย่างชัดเจนและมีประสิทธิภาพสูง ในทางกลับกัน หากค่า KS ต่ำกว่า 0.20 อาจเป็นสัญญาณว่าโมเดลมีความสามารถในการจำแนกที่ต่ำ และอาจไม่เหมาะสมต่อการนำไปใช้งานในบริบทจริง โดยเฉพาะอย่างยิ่งในระบบการตัดสินใจที่ต้องการความแม่นยำ เช่น ระบบให้สินเชื่อหรือระบบประเมินความเสี่ยง

ตัวอย่างการตีความ เช่น หากโมเดลมีค่า KS เท่ากับ 0.45 หรือ 45% หมายความว่า ณ จุดที่โมเดลสามารถแยกแยะได้ดีที่สุด ความแตกต่างระหว่างสัดส่วนสะสมของกลุ่ม Good และ Bad มีมากถึง 45% ซึ่งถือว่าอยู่ในระดับที่ดีและแสดงถึงศักยภาพของโมเดลในการวิเคราะห์ข้อมูลและสนับสนุนการตัดสินใจได้อย่างมีประสิทธิภาพ

ข้อดีของดัชนี KS (Kolmogorov-Smirnov Statistic) คือความเรียบง่ายในการใช้งานและการตีความที่ชัดเจน โดย KS สามารถบ่งชี้ได้อย่างตรงไปตรงมาว่าโมเดลสามารถแยกแยะกลุ่มเป้าหมาย เช่น กลุ่มลูกหนี้ดีและกลุ่มเสี่ยงผิดนัด ได้ดีเพียงใด โดยไม่จำเป็นต้องกำหนดค่า threshold เฉพาะเจาะจงเหมือนกับดัชนี Accuracy หรือ Recall ที่ต้องอิงกับจุดตัดในการจำแนก ซึ่งอาจทำให้เกิดอคติในกรณีที่ข้อมูลมีความไม่สมดุล นอกจากนี้ KS ยังเป็นเครื่องมือที่ได้รับความนิยมในการใช้งานควบคู่กับ ROC-AUC เพื่อประเมินประสิทธิภาพของโมเดลจำแนกประเภท โดยเฉพาะอย่างยิ่งในบริบทของการประเมินความเสี่ยงด้านสินเชื่อและการวิเคราะห์ลูกหนี้ในภาคการเงิน

เนื่องจากสามารถให้ข้อมูลเชิงปฏิบัติที่เป็นประโยชน์ต่อการตัดสินใจในระดับธุรกิจได้อย่างเป็นรูปธรรม

2.8.5 GINI (Gini Coefficient)

ค่า GINI หรือ Gini Coefficient เป็นตัวชี้วัดที่ใช้ประเมินความสามารถในการจำแนก (Discriminatory Power) ของโมเดลจำแนกประเภท โดยเฉพาะในบริบทของการให้คะแนนเครดิต (Credit Scoring) และการวิเคราะห์ความเสี่ยงทางการเงิน (Risk Modeling) ค่า GINI ถูกใช้กันอย่างแพร่หลายในอุตสาหกรรมการเงินเพื่อวัดประสิทธิภาพของโมเดลในการแยกแยะกลุ่มลูกหนี้ที่มีความเสี่ยงผิดนัด (Bad) ออกจากกลุ่มลูกหนี้ปกติ (Good) ค่า GINI มีความสัมพันธ์โดยตรงกับค่า AUC (Area Under the Receiver Operating Characteristic Curve: ROC-AUC) ซึ่งเป็นค่าพื้นที่ใต้กราฟ ROC โดยสามารถคำนวณค่า GINI ได้จากสูตรต่อไปนี้:

$$GINI = 2 \times AUC - 1 \tag{2.14}$$

โดยที่ AUC เป็นตัวสะท้อนถึงความสามารถในการแยกแยะกลุ่มเป้าหมายของโมเดล โดยค่า GINI จะมีค่าระหว่าง 0 ถึง 1 โดย

ค่า GINI ใกล้ 1 แสดงว่าโมเดลสามารถจำแนกกลุ่มได้อย่างสมบูรณ์

ค่า GINI = 0 หมายถึงโมเดลไม่มีความสามารถในการจำแนก (เทียบเท่าการสุ่ม)

การตีความค่าดังกล่าวสามารถอ้างอิงจากเกณฑ์มาตรฐานที่ใช้ในวงการวิเคราะห์ความเสี่ยงได้ดังตารางต่อไปนี้

ตารางที่ 2.1 การตีความค่า GINI

ค่า GINI	**ความสามารถในการจำแนก**
< 0.3	อ่อน (Weak)
0.3 - 0.5	ปานกลาง (Medium)
0.5 - 0.7	ดี (Strong)
> 0.7	ดีมาก (Very Strong)

การแปลผลในทางธุรกิจจะพิจารณาว่าค่า GINI ที่สูงมีนัยสำคัญว่าตัวแบบสามารถแยกแยะกลุ่มลูกค้าได้อย่างมีประสิทธิภาพ ซึ่งช่วยเพิ่มความแม่นยำในการให้สินเชื่อ ลดความเสี่ยง และเพิ่มประสิทธิภาพของกระบวนการตัดสินใจด้านเครดิตได้อย่างมีนัยสำคัญ (Siddiqi, 2006; Anderson, 2007) ในทางตรงกันข้าม หากค่า GINI อยู่ในช่วงต่ำกว่า 0.3 จะถือว่าโมเดลมีพลังจำแนกต่ำเกินไปและไม่ควรนำไปใช้ในสภาพแวดล้อมจริงที่ต้องการความแม่นยำสูง

2.8.6 การทดสอบสมมติฐานด้วย McNemar's Test

McNemar's Test เป็นสถิติที่ใช้สำหรับทดสอบความแตกต่างของสัดส่วนระหว่างตัวแปรที่จัดอยู่ในรูปแบบ "จับคู่" (Paired Nominal Data) โดยเฉพาะกรณีที่ต้องการเปรียบเทียบผลลัพธ์ของโมเดลสองโมเดลหรือสองวิธีการที่นำไปใช้กับตัวอย่างเดียวกัน เช่น การเปรียบเทียบความแม่นยำของโมเดล Machine Learning สองแบบ โดยพิจารณาผลลัพธ์ที่โมเดลทำนายผิด -ถูกในชุดข้อมูลเดียวกัน

McNemar's Test จะทำงานกับข้อมูลประเภท 2x2 Contingency Table ที่ประกอบด้วยผลลัพธ์ของแต่ละวิธีการ/โมเดลในรูปแบบ binary (เช่น ถูก/ผิด, ใช่/ไม่ใช่) และทดสอบว่าความน่าจะเป็นของการเปลี่ยนแปลงระหว่างสองวิธีการมีความแตกต่างกันหรือไม่ ดังตารางต่อไปนี้

ตารางที่ 2.2 ตัวอย่างการทำนายที่แตกต่างกันระหว่างโมเดล A และโมเดล B สำหรับการทดสอบ McNemar's Test

กลุ่มย่อย	โมเดล B ถูก	โมเดล B ผิด
โมเดล A ถูก	a	b
โมเดล A ผิด	c	d

McNemar's Test จะพิจารณาเฉพาะค่าที่ไม่ตรงกัน (Off-Diagonal) คือ b กับ c

สมมติฐาน

สมมติฐานศูนย์ (H_0): ความน่าจะเป็นที่โมเดล A และโมเดล B ทำนายแตกต่างกันในทิศทางหนึ่ง เท่ากับอีกทิศทางหนึ่ง (b = c)

สมมติฐานทางเลือก (H_1): ความน่าจะเป็นที่โมเดล A และโมเดล B ทำนายแตกต่างกันในทิศทางหนึ่ง ไม่เท่ากับอีกทิศทางหนึ่ง (b ≠ c)

สูตรการคำนวณ

$$\chi^2 = \frac{(b - c)^2}{b + c} \tag{2.15}$$

โดยที่

b คือ จำนวนตัวอย่างที่โมเดล A ผิด แต่โมเดล B ถูก

c คือ จำนวนตัวอย่างที่โมเดล A ถูก แต่โมเดล B ผิด

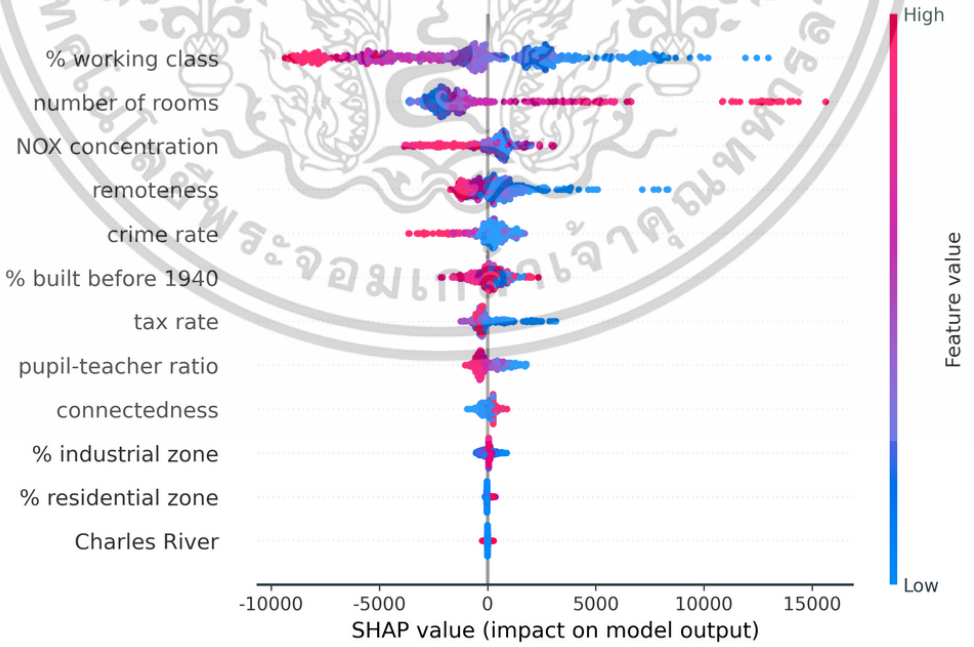
โดยค่าที่ได้จะนำไปเปรียบเทียบกับตาราง **Chi-Square Distribution** ที่ระดับอิสระ 1 (df = 1) เพื่อพิจารณาค่า p-value

2.9 แนวคิดและทฤษฎีเกี่ยวกับ Feature Importance

Feature Importance คือ แนวคิดเกี่ยวกับการวัด “ความสำคัญเชิงสถิติหรือเชิงโมเดล” ของตัวแปรแต่ละตัว (Feature หรือ Predictor) ว่ามีผลต่อผลลัพธ์ของโมเดลมากน้อยเพียงใด (Molnar, 2022) Explainability หรือ “การอธิบายโมเดล” คือ กระบวนการทำให้โมเดล Machine Learning โดยเฉพาะแบบ Black-Box สามารถอธิบายได้ว่าทำไมจึงให้ผลทำนายเช่นนั้น (Doshi-Velez & Kim, 2017) เทคนิคการวัด Feature Importance มีหลายเทคนิค โดยในกาศึกษานี้จะใช้ SHAP (SHapley Additive exPlanations) เป็นแนวทางในการอธิบายการทำงานของโมเดล

SHAP (SHapley Additive exPlanations) เป็นอีกหนึ่งเทคนิคสำคัญที่ถูกพัฒนาขึ้นเพื่อใช้ตีความโมเดลโดยเฉพาะในระดับรายบุคคล (Instance-Level Explanation) โดยอ้างอิงจาก Shapley Value ซึ่งมีต้นกำเนิดมาจากทฤษฎีเกม (Cooperative Game Theory) โดยนักคณิตศาสตร์ Lloyd Shapley (1953) หลักการของ SHAP คือการคำนวณค่าผลกระทบของแต่ละฟีเจอร์โดยเปรียบเสมือนว่าแต่ละฟีเจอร์เป็นผู้เล่นในเกม และผลลัพธ์ของโมเดลเป็นผลรวมของคะแนนที่แต่ละผู้เล่นมีส่วนร่วม (Lundberg & Lee, 2017) โดยการรวมค่าผลกระทบของฟีเจอร์ทั้งหมดจะเท่ากับค่าทำนายของโมเดลในกรณีนั้นๆ เสมอ

SHAP สามารถอธิบายได้ทั้งในระดับ Global และ Local กล่าวคือสามารถบอกได้ว่าโดยรวมแล้วฟีเจอร์ใดสำคัญที่สุด (Global Importance) และในแต่ละกรณีเฉพาะบุคคล ฟีเจอร์ใดทำให้โมเดลตัดสินใจเช่นนั้น (Local Explanation) ซึ่งเหมาะสมอย่างยิ่งสำหรับการวิเคราะห์ความเสี่ยงรายบุคคล เช่น ลูกค้ารายใดมีแนวโน้มผิดนัด และเพราะเหตุใด



รูปที่ 2.9 ตัวอย่าง Beeswarm Plot ของค่า SHAP

ที่มา: Cooper (2021)

Beeswarm Plot เป็นเครื่องมือที่สำคัญของ SHAP (SHapley Additive exPlanations) สำหรับการอธิบายความสำคัญของฟีเจอร์และการกระจายอิทธิพลของแต่ละตัวแปรที่มีต่อผลลัพธ์ของ Machine Learning โดยแผนภาพนี้สามารถให้ข้อมูลเชิงลึกได้ทั้งในเชิงปริมาณและเชิงคุณภาพ ดังนี้

1) โครงสร้างของ Beeswarm Plot

แกน Y แสดงรายชื่อของตัวแปรอินพุต (input variables) ที่เรียงลำดับจากบนลงล่างตามค่าเฉลี่ยสัมบูรณ์ของ SHAP (Mean Absolute SHAP values) ซึ่งบ่งบอกถึงความสำคัญของแต่ละฟีเจอร์ต่อการทำนายของโมเดลในภาพรวม

แกน X แสดงค่า SHAP (SHAP Value) ของแต่ละตัวอย่างในชุดข้อมูล ซึ่งสะท้อนถึงผลกระทบของค่าฟีเจอร์นั้นๆ ต่อการทำนายของโมเดล (ค่าเป็นบวกช่วยเพิ่มค่าทำนาย, ค่าเป็นลบช่วยลดค่าทำนาย)

จุดแต่ละจุดแทนตัวอย่างข้อมูลหนึ่งแถว จุดเหล่านี้จะกระจายตามแกน X ตามค่า SHAP ของแต่ละกรณี หากบริเวณใดมีความหนาแน่นสูง จุดจะเรียงซ้อนกันในแนวตั้งคล้ายฝูงผึ้ง

2) ความหมายของสี

สีของแต่ละจุดแสดงถึงค่าดิบของตัวแปรในแต่ละกรณี (ไม่ใช่ค่า SHAP) โดยใช้การไล่ระดับสีตั้งแต่สีน้ำเงิน (ค่าต่ำ) ถึงสีชมพู/แดง (ค่าสูง)

การพิจารณาการกระจายของสีในแต่ละแถว ช่วยให้เข้าใจทิศทางและความสัมพันธ์ระหว่างค่าตัวแปรกับผลกระทบต่อการทำนาย เช่น หากจุดสีแดงกระจายไปทางด้านขวา แสดงว่าค่าสูงของฟีเจอร์นั้นมีแนวโน้มเพิ่มค่าทำนายของโมเดล

3) การตีความ Beeswarm Plot

Beeswarm plot ไม่ได้แสดงเฉพาะลำดับความสำคัญของฟีเจอร์ (เหมือน Bar Plot) แต่ยังช่วยให้เห็นความสัมพันธ์เชิงลึกระหว่างค่าฟีเจอร์กับการทำนาย เช่น สามารถสังเกตได้ว่าค่าฟีเจอร์ที่สูงหรือต่ำมีแนวโน้มทำให้ผลทำนายเปลี่ยนไปในทิศทางใด

การกระจายตัวของค่า SHAP ตามแกน X สะท้อนถึงระดับอิทธิพลของแต่ละฟีเจอร์ ตัวแปรที่มีการกระจายกว้างแสดงว่ามีผลกระทบต่อผลลัพธ์สูง ส่วนตัวแปรที่มีการกระจายแคบมีผลกระทบจำกัด

Beeswarm Plot ให้ข้อมูลทั้งเชิงปริมาณและเชิงคุณภาพในแผนภาพเดียว ช่วยให้เข้าใจการตัดสินใจของโมเดลได้ชัดเจนมากขึ้น เหมาะกับการสื่อสารให้ทั้งผู้เชี่ยวชาญและผู้มีส่วนได้ส่วนเสียที่ไม่ใช่สายเทคนิค

2.10 แนวคิดและทฤษฎีเกี่ยวกับ Credit Scoring

Credit Scoring (เครดิตสกอริ่ง) คือ กระบวนการให้คะแนนความน่าเชื่อถือทางการเงินของลูกค้าหรือผู้ขอกู้ โดยใช้วิธีการทางสถิติหรือเทคนิค Machine Learning ในการประเมินความเสี่ยงที่

ลูกค้าจะผิดนัดชำระหนี้ (Default) ภายในระยะเวลาที่กำหนด ซึ่งคะแนนเครดิตนี้จะถูกนำไปใช้ประกอบการตัดสินใจอนุมัติสินเชื่อ กำหนดวงเงิน หรือปรับอัตราดอกเบี้ยให้เหมาะสมกับความเสี่ยงของแต่ละบุคคล (Hand et al., 1997; Thomas, 2000)

หลักการสำคัญของ Credit Scoring ประกอบด้วยขั้นตอน ดังนี้

- การรวบรวมข้อมูล (Data Collection): รวบรวมข้อมูลคุณลักษณะของผู้กู้ เช่น อายุ รายได้ ประวัติสินเชื่อ พฤติกรรมการชำระหนี้ เป็นต้น
- การสร้างโมเดล (Model Development): ใช้เทคนิคทาง หรือ Machine Learning ในการวิเคราะห์ข้อมูลเพื่อตรวจสอบปัจจัยที่มีผลต่อความเสี่ยงผิดนัด
- การคำนวณคะแนน (Score Calculation): แปลงผลลัพธ์ความน่าจะเป็นของการผิดนัดชำระหนี้ (Probability of Default) ให้เป็นคะแนน (Credit Score) ด้วยสูตร Probability-to-Score Transformation เช่น

$$\text{Score} = A - B \cdot \ln\left(\frac{\text{Probability}}{1 - \text{Probability}}\right) \tag{2.16}$$

- การกำหนดช่วงคะแนนและการนำไปใช้ (Score Interpretation & Policy): แบ่งช่วงคะแนนออกเป็นกลุ่มความเสี่ยง (Risk Groups) เช่น Good, Moderate, High Risk, Very High ตามมาตรฐานสากลหรือเงื่อนไขของแต่ละธนาคาร (Thomas et al., 2002)

Credit Scoring มีบทบาทสำคัญในระบบวิเคราะห์รายบุคคลของสถาบันการเงิน เนื่องจากช่วยลดต้นทุน เวลา และอคติจากการตัดสินใจโดยบุคคลได้อย่างชัดเจน Credit Scoring ช่วยเร่งกระบวนการอนุมัติสินเชื่อจากที่เคยใช้เวลานานหลายวันหรือหลายสัปดาห์ พัฒนาเป็นใช้เวลาเพียงไม่กี่ชั่วโมงหรือหลายวันเท่านั้น ซึ่งส่งผลให้ลูกค้าสะดวกและช่วยลดค่าใช้จ่ายให้กับเมื่อนำมาใช้จริง นอกจากนี้ Credit Scoring ยังส่งเสริมความเป็นธรรมในการพิจารณาสินเชื่อ เนื่องจากระบบจะใช้มาตรฐานการประเมินเดียวกันกับทุกผู้สมัคร ไม่คำนึงถึงเชื้อชาติ เพศ หรือกลุ่มที่ได้รับการคุ้มครอง ทำให้ลดความเสี่ยงจากการตีความอคติหรือการปฏิบัติที่ไม่เท่าเทียม ดังนั้น Credit Scoring จึงช่วยเปิดโอกาสให้ผู้มีประวัติดีได้รับการเข้าถึงสินเชื่อมากขึ้น โดยเฉพาะผู้ที่อาจไม่ผ่านการพิจารณาด้วยวิธีดั้งเดิม ส่งผลให้ผู้มีคุณสมบัติเหมาะสมได้รับผลประโยชน์มากขึ้น

2.11 งานวิจัยที่เกี่ยวข้อง

Hand & Henley (2021) เป็นการวิเคราะห์ Credit Scoring ทั่วโลก โดยศึกษาตัวแปรสำคัญที่มีผลต่อความเสี่ยงผิดนัดชำระหนี้ ทั้งในบริบทธนาคารและผู้ปล่อยสินเชื่อประเภทต่างๆ โดยเก็บ

ข้อมูล Credit History, Employment Length, Income, Debt to Income Ratio, Loan to Value Ratio และนำมาวิเคราะห์หาความสัมพันธ์กับอัตราการผิดนัด งานวิจัยนี้ใช้โมเดล Logistic Regression, Tree-based Models และเทคนิคการถ่วงน้ำหนัก (Weighting) เพื่อจัดการกับข้อมูลไม่สมดุล และเน้นประเมินผลด้วย ROC-AUC, Recall และ KS ข้อค้นพบของงานวิจัยนี้ตอกย้ำถึงความสำคัญของการเลือกฟีเจอร์และตัวชี้วัด ROC-AUC สำหรับโมเดล Credit Scoring โดยเฉพาะในกลุ่มข้อมูลที่ไม่สมดุลสูง อันเป็นแนวทางในการเลือกใช้ metric ที่เหมาะสมและการออกแบบโมเดลที่เน้นความสามารถในการอธิบายหรือแสดงเหตุผลเบื้องหลังการตัดสินใจ

Chen et al. (2021) ได้ดำเนินการทดลองเปรียบเทียบโมเดล XGBoost, Random Forest, Logistic Regression กับเทคนิค OverSampling (SMOTE, ADASYN) ในปัญหาข้อมูลสินเชื่อที่มี Class Imbalance สูง งานวิจัยนี้ใช้ชุดข้อมูลสินเชื่อที่มีสัดส่วน Default ต่ำกว่า 10% และเปรียบเทียบ performance ของโมเดลต่างๆ ผลการศึกษาพบว่า XGBoost ร่วมกับ SMOTE ให้ค่า ROC-AUC สูงสุดและมีความเสถียร แม้จะทดสอบซ้ำหลายครั้งบนชุดข้อมูลที่ต่างกัน ในขณะที่ Logistic Regression แม้จะตีความได้ดีแต่มีประสิทธิภาพต่ำกว่าในด้านการแยกกลุ่ม Default

Siddiqi (2017) และ Moody's Analytics (2017) ได้รวบรวมและยืนยันหลักปฏิบัติของการพัฒนา Credit Scoring Model ที่มีคุณภาพสูงสำหรับข้อมูลที่ไม่สมดุล พบว่าค่า ROC-AUC เป็นตัวชี้วัดมาตรฐานสากล เพราะสามารถประเมินศักยภาพของโมเดลในการแยกกลุ่ม ลูกหนี้ดี กับ กลุ่มเสี่ยงผิดนัด ได้อย่างแม่นยำทั้งในระดับ Development และ Production ข้อค้นพบสำคัญ ผู้วิจัยควรเลือก ROC-AUC เป็นตัวชี้วัดหลักในการคัดเลือกโมเดลที่เหมาะสมที่สุดสำหรับการนำไปใช้ Credit Scoring ในระบบงานจริง

Hand และ Henley (1997) เป็นการทบทวนองค์ความรู้เกี่ยวกับวิธีการจำแนกประเภทในตลาดสินเชื่อผู้บริโภค โดยเน้นที่เทคนิคสถิติต่างๆ สำหรับการสร้างโมเดล Credit Scoring งานนี้นำเสนอข้อดีข้อเสียของโมเดลพื้นฐาน ได้แก่ Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbors, Decision Trees, Neural Networks และเปรียบเทียบการประเมินโมเดลผ่านตัวชี้วัด ROC-AUC, Gini Coefficient, และ KS Statistic ผู้วิจัยเน้นว่าข้อมูลที่ใช้ในการสร้างโมเดล Credit Scoring มักประกอบด้วยตัวแปรพื้นฐาน เช่น อายุ เพศ รายได้ สถานภาพสมรส อัตราส่วนทางการเงิน และประวัติการผิดนัดชำระหนี้ พร้อมทั้งแสดงให้เห็นว่าการใช้ ROC Curve และ AUC ช่วยให้สามารถประเมินประสิทธิภาพโมเดลได้ดีกว่า Accuracy โดยเฉพาะในกรณีที่ข้อมูลไม่สมดุล (กลุ่มลูกหนี้ดีมีมากกว่ากลุ่มเสีย) งานนี้ยังเน้นย้ำว่า Gini Coefficient และ KS Statistic เป็นมาตรฐานสากลที่นิยมใช้ในธนาคารและสถาบันการเงินทั่วโลก ผลสรุปของงานนี้พบว่า Logistic Regression เป็นโมเดลที่ได้รับความนิยมสูงสุดในเชิงปฏิบัติ เพราะตีความง่ายและมีประสิทธิภาพใกล้เคียงกับเทคนิคซับซ้อนอื่นๆ ในบริบท Credit Scoring

Siddiqi (2017) นับเป็นคู่มือมาตรฐานในการพัฒนา Credit Risk Scorecard โดยครอบคลุมตั้งแต่การเลือกตัวแปร การเตรียมข้อมูล เทคนิค Binning และ WOE (Weight of Evidence) ไป

จนถึงการสร้างและประเมินโมเดล ตัวอย่างตัวแปรที่นำมาใช้ ได้แก่ ข้อมูลประชากร (อายุ เพศ สถานภาพสมรส) รายได้ วงเงินกู้ ประสบการณ์ทางธุรกิจ อัตราส่วนทางการเงิน พฤติกรรมการชำระหนี้ และประวัติผิดนัด Siddiqi แนะนำให้ใช้ Logistic Regression เป็นโมเดลหลักเนื่องจากความสามารถในการตีความและการรองรับข้อมูล imbalanced อย่างเหมาะสม นอกจากนี้ ยังเน้นการใช้ ROC-AUC, Gini, KS เป็นตัวชี้วัดประสิทธิภาพหลัก พร้อมยกตัวอย่างการประเมินโมเดลในธนาคารจริง ซึ่งค่าที่ธนาคารยอมรับได้ควรมี ROC-AUC มากกว่า 0.70 และ KS มากกว่า 0.30 สำหรับโมเดลที่ผ่านเกณฑ์ ทั้งนี้ยังกล่าวถึง PSI (Population Stability Index) ในการตรวจสอบเสถียรภาพโมเดลเมื่อใช้งานจริงกับข้อมูลใหม่

Brown และ Mues (2012) เป็นการทดลองเชิงเปรียบเทียบประสิทธิภาพของโมเดลหลายแบบสำหรับ Credit Scoring ในชุดข้อมูลที่ไม่สมดุล (เช่น อัตราผิดนัดต่ำ) โดยเปรียบเทียบทั้ง Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Machines, และ Neural Networks นอกจากนี้ยังเปรียบเทียบวิธี Sampling เช่น Random Undersampling, SMOTE, และ Cluster-Based Sampling รวมถึงการทำ Feature Selection และ Preprocessing (เช่น WOE, Scaling) ในการทดลอง ผู้วิจัยใช้ตัวแปรเชิงประชากร (อายุ เพศ เป็นต้น) รายได้ อัตราส่วนทางการเงิน และประวัติการผิดนัด พร้อมใช้ ROC-AUC, Gini, KS, Recall, Precision, F1-Score เป็นตัวชี้วัดประสิทธิภาพ ผลการศึกษาพบว่า Logistic Regression และ Tree-Based Models (โดยเฉพาะ Random Forest และ Gradient Boosting) สามารถจัดการกับปัญหาข้อมูลไม่สมดุลได้ดี เมื่อผสมกับเทคนิค Sampling ที่เหมาะสม สำหรับค่า ROC-AUC ในงานวิจัยนี้ โมเดลที่ดีที่สุดได้ ROC-AUC อยู่ที่ 0.78-0.82 และ KS อยู่ที่ 0.35-0.41 ทั้งนี้ ROC-AUC และ Gini เป็นตัวชี้วัดที่สะท้อน Performance ที่แท้จริงได้ดีที่สุด

Altman และ Sabato (2007) มุ่งเน้นไปที่การสร้างและประเมินโมเดล Credit Risk สำหรับธุรกิจ SMEs ในสหรัฐฯ โดยเปรียบเทียบ Logistic Regression และ Linear Discriminant Analysis ผู้วิจัยใช้ข้อมูลจากบริษัท SMEs จำนวนมาก โดยเลือกตัวแปรทั้งด้านประชากร ประเภทธุรกิจ ประสบการณ์ธุรกิจ รายได้ วงเงิน หนี้สิน ค้ำประกัน และอัตราส่วนทางการเงินเป็นตัวแปรอิสระ ผลการศึกษาเปรียบเทียบพบว่า Logistic Regression ให้ผลลัพธ์ดีกว่า Discriminant Analysis เล็กน้อย โดยได้ค่า ROC-AUC ในช่วง 0.73–0.77 และ Gini Coefficient ในช่วง 0.46–0.54 งานนี้ชี้ให้เห็นว่า ตัวแปรทางพฤติกรรมและการเงินของบริษัทมีอิทธิพลสูงต่อโอกาสเกิด NPL และสามารถใช้พัฒนาระบบ Scoring ที่มีความแม่นยำสำหรับธุรกิจขนาดเล็ก

Agarwal และคณะ (2020) ศึกษาการทำนาย First Payment Default (FPD) ในสินเชื่อจำนองของสหรัฐ โดยนำข้อมูลลูกหนี้ด้านรายได้ เงื่อนไขสัญญา พฤติกรรม และประวัติเครดิต มาสร้างโมเดลเปรียบเทียบหลายแบบ ได้แก่ Logistic Regression, Random Forest, Gradient Boosting, และ Neural Networks พร้อมทั้งประเมินด้วย ROC-AUC, KS ผลการทดลองพบว่า Logistic Regression ยังคงเป็น Baseline ที่ตีความง่ายและได้ประสิทธิภาพดี แต่ Random Forest

และ Gradient Boosting ให้ค่า ROC-AUC สูงกว่าเล็กน้อยในบางเซ็ตข้อมูล สำหรับการทำนาย FPD พบว่า ROC-AUC ของโมเดลที่ดีที่สุดอยู่ที่ 0.76–0.80 และค่า KS ในช่วง 0.32–0.40 ผลวิจัยนี้ยืนยันว่า FPD เป็นปัจจัยสำคัญที่ทำนายโอกาสผิดนัดระยะยาวและควรให้ความสำคัญในระบบประเมินความเสี่ยงของสถาบันการเงิน

จากการทบทวนงานวิจัยจึงได้รวบรวมและกำหนดตัวแปรเบื้องต้นที่มีอิทธิพลกับผิดนัดชำระหนี้งวดแรกหรือ First Payment Default (FPD) สรุปดังตารางต่อไปนี้

ตารางที่ 2.3 สรุปผลศึกษางานวิจัยที่เกี่ยวข้อง

งานวิจัย	ตัวแปรที่งานวิจัยใช้	โมเดล	เทคนิคจัดการข้อมูลไม่สมดุล	ตัวชี้วัด
Hand & Henley (2021)	Credit History, Employment Length, Income, DTI, LTV	Logistic Regression, Tree-based	Weighting	ROC-AUC, KS, Recall
Chen et al. (2021)	เพศ, อายุ, รายได้, วงเงิน, พฤติกรรม, ประวัติผิดนัด	XGBoost	SMOTE	ROC-AUC, Accuracy
Siddiqi (2017), Moody's Analytics (2017)	อายุ, เพศ, รายได้, วงเงิน, พฤติกรรม, อัตราส่วนทางการเงิน, ประวัติผิดนัด	Logistic Regression (Scorecard)	WOE binning/Balance	ROC-AUC, Gini, KS, PSI
Hand & Henley (1997)	อายุ, เพศ, รายได้, สถานภาพสมรส, อัตราส่วนการเงิน, ประวัติผิดนัด	Logistic Regression		ROC-AUC, Gini, KS
Brown & Mues (2012)	อายุ, เพศ, รายได้, พฤติกรรม, อัตราส่วนการเงิน	Random Forest, GBM	SMOTE, Under Sampling	ROC-AUC, Gini, KS, Recall, F1
Altman & Sabato (2007)	อายุ, รายได้, ประสบการณ์, วงเงิน, หนี้, ค้ำประกัน	Logistic Regression		ROC-AUC, Gini
Agarwal et al. (2020)	รายได้, เงื่อนไขสัญญา พฤติกรรมการเงิน,ประวัติเครดิต	Random Forest, GBM		ROC-AUC, KS, Lift

จากตารางที่ 2.3 สรุปการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง ทำให้ผู้วิจัยสามารถกำหนดตัวแปรอิสระที่มีอิทธิพลต่อการผิดนัดชำระงวดแรกมีอยู่ดังนี้ ข้อมูลประชากรศาสตร์ เช่น เพศ อายุ ระดับการศึกษา สถานภาพสมรส ข้อมูลการเงิน เช่น รายได้ต่อเดือน รายได้สุทธิ รายจ่าย รายได้

ครัวเรือน ประวัติการกู้ยืม เช่น จำนวนวงเงินที่ขอ วงเงินอนุมัติ ประวัติการกู้ยืมในอดีต ประวัติผิดนัดพฤติกรรมธุรกรรม เช่นประเภทอาชีพ ลักษณะธุรกิจ ข้อมูลการใช้จ่าย อัตราส่วนทางการเงิน เช่น อัตราส่วนยอดขายต่อยอดหนี้ อัตราส่วนหนี้ต่อรายได้ ระยะเวลากู้ และ ข้อมูลอื่นๆ เช่น ระยะเวลาทำงาน ประสบการณ์ทางธุรกิจ คะแนนเครดิต ส่วนในด้านเทคนิคการจัดการข้อมูลและโมเดล Machine Learning ที่จะนำมาศึกษาและเปรียบเทียบ ผู้วิจัยได้เลือกเทคนิคจัดการข้อมูลไม่สมดุล ได้แก่ SMOTE, SMOTEENN และ Random Under Sampling ใช้สำหรับจัดการข้อมูลให้สมดุลกับกลุ่มปกติ และในส่วนของโมเดล Machine Learning ผู้วิจัยได้เลือกโมเดล XGBoost รวมถึง Random Forest ที่ให้ผลดีกับข้อมูลที่ไม่สมดุล โดยเฉพาะเมื่อใช้ร่วมกับ SMOTE หรือ SMOTEENN และความยืดหยุ่นในการแยกกลุ่มลูกค้าที่เสี่ยง และโมเดลที่สามารถอธิบายความสัมพันธ์ระหว่างตัวแปรได้ดี เช่น Logistic Regression ก็ยังคงเป็นที่นิยมอยู่ ในส่วนของตัวชี้วัดที่ควรใช้ในการเลือกโมเดล (Evaluation Metrics) พบว่า ROC-AUC เป็นตัวชี้วัดมาตรฐานสากล สำหรับข้อมูลไม่สมดุล เพราะแสดงศักยภาพการแยกกลุ่มได้ดีและ KS (Kolmogorov-Smirnov) วัดระยะห่างสูงสุดระหว่าง Distribution ของกลุ่มดีกับกลุ่มเสี่ยง เหมาะสำหรับ Credit Scoring ค่า KS ที่สูงกว่า 0.30 ถือว่าผ่านเกณฑ์อุตสาหกรรมสำหรับโมเดล Credit Scoring และการวิเคราะห์ด้วยเทคนิค SHAP (SHapley Additive exPlanations) เพื่อวิเคราะห์ปัจจัยที่มีอิทธิพลต่อการผิดนัดชำระหนี้งวดแรก

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงประยุกต์ โดยมุ่งเน้นพัฒนาโมเดลทำนายความเสี่ยงการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) สำหรับกลุ่มลูกค้าสินเชื่อธุรกิจขนาดเล็ก (Micro Business Loans) ของสถาบันการเงินแห่งหนึ่งในประเทศไทย การดำเนินงานตลอดกระบวนการอ้างอิงแนวคิดมาตรฐานอุตสาหกรรมคือ CRISP-DM (Cross-Industry Standard Process for Data Mining) ซึ่งได้รับความนิยมอย่างแพร่หลายในแวดวงวิทยาศาสตร์ข้อมูล (Data Science) มีขั้นตอนดังนี้

1) การทำความเข้าใจธุรกิจ (Business Understanding)
2) ทำความเข้าใจข้อมูล (Data Understanding)
3) การเตรียมข้อมูล (Data Preparation)
4) สร้างโมเดล (Modeling)
5) ประเมินผล (Evaluation)
6) นำไปใช้งาน (Deployment)

3.1 การทำความเข้าใจธุรกิจ (Business Understanding)

การเข้าใจปัญหาเชิงธุรกิจถือเป็นจุดเริ่มต้นสำคัญของการพัฒนาเครื่องมือวิเคราะห์ความเสี่ยงทางสินเชื่อ โดยเฉพาะ การผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ซึ่งเป็นตัวชี้วัดสำคัญที่สะท้อนถึงคุณภาพของพอร์ตสินเชื่อ และมีผลโดยตรงต่อระดับหนี้ที่ไม่ก่อให้เกิดรายได้ (Non-Performing Loan: NPL) ของธนาคาร การป้องกัน FPD ตั้งแต่กระบวนการอนุมัติสินเชื่อ จะช่วยลดความเสี่ยงการเกิดหนี้เสียในระยะยาว ส่งเสริมให้กระบวนการคัดกรองลูกค้ามีประสิทธิภาพมากขึ้น และช่วยให้ธนาคารสามารถวางแนวทางอนุมัติที่เหมาะสมกับระดับความเสี่ยงของลูกค้าแต่ละกลุ่ม เพื่อให้สอดคล้องกับเป้าหมายดังกล่าว งานวิจัยนี้จึงนำเสนอการพัฒนาคะแนนความเสี่ยง (FPD Score) จากโมเดลทำนายโอกาสผิดนัดชำระหนี้ โดยนำมาใช้เป็นเครื่องมือในการกรองและจัดกลุ่มลูกค้าในกระบวนการอนุมัติสินเชื่อ พร้อมแมปกับนโยบายการอนุมัติที่ชัดเจนและเหมาะสมกับระดับความเสี่ยง ดังรายละเอียดในตารางต่อไปนี้

ตารางที่ 3.1 การจำแนกกลุ่มความเสี่ยงของลูกค้าโดยใช้ช่วงคะแนน FICO

ช่วงคะแนน (FICO)	กลุ่มความเสี่ยง (Risk Group)	ความหมาย/แนวโน้ม
800–850	Excellent (ยอดเยี่ยม)	อนุมัติทันที
740–799	Good (ดี)	อนุมัติ พร้อมตรวจสอบเบื้องต้น
670–739	Moderate (ปานกลาง)	ขอเอกสารเพิ่ม/วงเงินจำกัด
580–669	High Risk (เสี่ยงสูง)	พิจารณาเข้มงวด
300–579	Very High (เสี่ยงมาก)	ปฏิเสธ/อนุมัติแบบพิเศษ

จากตารางที่ 3.1 (FICO, 2024) เป็นแนวทางการประยุกต์ใช้ FPD Score เพื่อระบบ Early Warning ในการบริหาร NPL หนึ่งในประเด็นสำคัญของการบริหาร NPL คือการมี ระบบเตือนภัยล่วงหน้า (Early Warning System) ที่ช่วยให้ธนาคารหรือสถาบันการเงินสามารถ ตรวจพบความเสี่ยงของลูกค้าก่อนเกิดปัญหาหนี้เสีย ได้อย่างรวดเร็วและมีประสิทธิภาพ

1) ตรวจจับกลุ่มเสี่ยงได้ก่อนเกิดปัญหา FPD Score ที่สร้างจากโมเดล Machine Learning สามารถใช้เป็นดัชนี Early Warning ได้โดยตรง เพราะคะแนนที่ออกมาเป็นการสะท้อน “ความน่าจะเป็นในการผิดนัดชำระหนี้ตั้งแต่งวดแรก” หากลูกค้าได้รับคะแนนในช่วง “High Risk” หรือ “Very High” ธนาคารจะสามารถติดตามลูกค้ากลุ่มนี้อย่างใกล้ชิดตั้งแต่ก่อนอนุมัติ หรือวางแผนกลยุทธ์การติดตามหนี้ (Pre-delinquency Action) ได้อย่างแม่นยำยิ่งขึ้น

2) การออกแบบกระบวนการเตือนภัย ธนาคารสามารถตั้งเกณฑ์หรือ Threshold ของ FPD Score ให้สอดคล้องกับนโยบายความเสี่ยง เช่น หากลูกค้ากลุ่ม Moderate หรือ High Risk มีสัดส่วนสูงขึ้นในพอร์ต อาจต้องทบทวน/ปรับเงื่อนไขอนุมัติหรือมาตรการควบคุม แจ้งเตือนให้ฝ่ายบริหารสินเชื่อ (Credit Officer) พิจารณารายงานวิเคราะห์ความเสี่ยงก่อนการอนุมัติวงเงินใหม่หรือเพิ่มวงเงิน

3) เชื่อมโยงกับกระบวนการ Collection/ติดตามหนี้ เมื่อระบบ Early Warning ตรวจจับลูกค้าเสี่ยงสูงได้ล่วงหน้า ฝ่ายติดตามหนี้สามารถใช้ข้อมูลนี้วางแผน ติดต่อลูกค้าก่อนครบกำหนดชำระ ให้คำแนะนำด้านการเงิน หรือปรับโครงสร้างหนี้ล่วงหน้า ลดความเสี่ยงที่จะเข้าสู่สถานะ NPL

4) ประโยชน์ต่อการบริหารพอร์ตสินเชื่อ ลดอัตราการเกิด NPL ในระยะยาว สร้างมาตรการป้องกันเชิงรุก ไม่ใช่แค่ “แก้ปัญหา” เมื่อหนี้เสียเกิดขึ้นแล้ว ทำให้กระบวนการบริหารความเสี่ยงและการอนุมัติสินเชื่อโปร่งใส ตัดสินใจได้ด้วยข้อมูลและหลักฐานเชิงประจักษ์

ตัวอย่าง: ในกรณีที่พบว่ากลุ่มลูกค้าใหม่ที่เข้าสู่กระบวนการอนุมัติ มีคะแนน FPD อยู่ในกลุ่ม High Risk หรือ Very High เป็นจำนวนมาก ระบบ Early Warning สามารถแจ้งเตือนให้ฝ่ายบริหารสินเชื่อและทีม Collections วางมาตรการป้องกัน อาทิ การขอหลักประกันเพิ่ม การจำกัดวงเงิน หรือการติดตามหลังอนุมัติอย่างใกล้ชิด เพื่อป้องกันการเกิด NPL ตั้งแต่ต้น

3.2 ทำความเข้าใจข้อมูล (Data Understanding)

วิเคราะห์แหล่งที่มาของข้อมูล ประเภทของตัวแปร (ชนิดข้อมูล) ตรวจสอบความสมบูรณ์ ค่าขาดหาย (Missing Value), Outlier ความเบ้ (Skewness) ความสัมพันธ์เบื้องต้น (Correlation) และความสมดุลของกลุ่มเป้าหมาย (Class Imbalance) โดยเป้าหมายของหัวข้อนี้คือ เข้าใจข้อมูลอย่างรอบด้านเพื่อวางแผนการเตรียมข้อมูล (Data Preparation)

3.2.1 เครื่องมือและทรัพยากรที่ใช้

แพลตฟอร์มการทำงาน (Development Platform) Jupyter Notebook บนแพลตฟอร์ม Anaconda Navigator และ SAS Enterprise Guide ใช้ภาษา SQL (Structured Query Language) และ SAS Data Step สำหรับดึงข้อมูลและจัดการข้อมูลเบื้องต้น รวมถึงภาษาไพธอน (Python Programming Language) สำหรับการประมวลผลข้อมูล วิเคราะห์ข้อมูล และสร้างโมเดล

ตารางที่ 3.2 ไลบรารีบน Python ที่ใช้ในงานวิจัย

ไลบรารี (Library)	**วัตถุประสงค์การใช้**
pandas	ประมวลผลข้อมูลแบบตาราง
numpy	คำนวณเชิงตัวเลขและอาเรย์
matplotlib	สร้างกราฟและ Visualization
seaborn	Visualization และกราฟสถิติ
scipy	จัดการกับอาร์เรย์เชิงตัวเลข (Numerical Array)
imblearn	เทคนิคจัดการข้อมูลที่ไม่สมดุล
scikit-learn	สร้างโมเดลการเรียนรู้ของเครื่อง
xgboost	สร้างโมเดล Xgboost
shap	อธิบายฟีเจอร์มีผลต่อการพยากรณ์ของโมเดล

3.2.2 ประชากรและกลุ่มตัวอย่าง

ประชากร (Population) คือ ลูกค้าสินเชื่อธุรกิจขนาดเล็กที่สมัครและได้รับการพิจารณาสินเชื่อกับสถาบันการเงินแห่งหนึ่งในประเทศไทย

กลุ่มตัวอย่าง (Sample) คือ ข้อมูลการสมัครสินเชื่อธุรกิจขนาดเล็กที่บันทึกผลสถานะผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในช่วงระยะเวลาตั้งแต่วันที่ 1 กันยายน พ.ศ. 2562 ถึง 1 มีนาคม พ.ศ. 2567 กลุ่มตัวอย่างประกอบด้วยตัวแปรต้น (Independent Variables) ที่เกี่ยวข้องกับคุณสมบัติของลูกค้าและรายละเอียดสินเชื่อ และตัวแปรตาม (Dependent Variable) คือ FPD

3.2.3 ตัวแปรในการวิจัย

ขั้นตอนการอธิบายข้อมูล (Data Description) ถือเป็นกระบวนการสำคัญที่ช่วยให้ผู้วิจัยสามารถทำความเข้าใจข้อมูลในภาพรวมก่อนดำเนินการวิเคราะห์เชิงลึกในขั้นตอนถัดไป โดยมุ่งเน้นการสำรวจลักษณะของข้อมูลทั้งในเชิงโครงสร้าง (Structure) เชิงปริมาณ (Quantitative) และเชิงคุณภาพ (Qualitative) เพื่อให้สามารถระบุปัญหาที่อาจเกิดขึ้น เช่น ค่าที่ขาดหาย (Missing Values), ค่าผิดปกติ (Outliers), ความไม่สมดุลของกลุ่มข้อมูล (Imbalanced Data), และความสัมพันธ์ระหว่างตัวแปรต่างๆ (Correlation) ซึ่งล้วนเป็นประเด็นที่ส่งผลต่อความแม่นยำและประสิทธิภาพของโมเดลทำนายที่พัฒนาขึ้นในภายหลัง งานวิจัยนี้รวบรวมตัวแปรที่คาดว่าเป็นปัจจัยที่ส่งผลต่อการผิดนัดชำระหนี้ จำนวน 30 ตัวแปร ประกอบด้วยตัวแปรอิสระ 29 ตัวแปรและตัวแปรตาม 1 ตัวแปร โดยแบ่งเป็นข้อมูลเชิงกลุ่มจำนวน 5 ตัวแปร และเป็นอัตราส่วน (Ratio) จำนวน 24 ตัวแปร ดังตารางที่ 3.3

ตารางที่ 3.3 คุณลักษณะของตัวแปรที่ใช้ในงานวิจัย

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ประเภทข้อมูลตามระดับมาตรวัด
1	Gender	เพศ	นามบัญญัติ (Nominal)
2	CustAge	อายุ (ปี)	อัตราส่วน (Ratio)
3	MaritalStatus	สถานภาพการสมรส	นามบัญญัติ (Nominal)
4	Region	ภูมิภาคที่อยู่อาศัย	นามบัญญัติ (Nominal)
5	BusinessType	ประเภทธุรกิจ	นามบัญญัติ (Nominal)
6	AppliedAmount	ยอดสินเชื่อที่สมัคร (บาท)	อัตราส่วน (Ratio)
7	TotalCreditAmount	วงเงินรวม (บาท)	อัตราส่วน (Ratio)
8	Installment	ค่างวดต่อเดือน (บาท)	อัตราส่วน (Ratio)
9	LTV	อัตราส่วนการให้สินเชื่อเทียบกับมูลค่าหลักประกัน (ร้อยละ)	อัตราส่วน (Ratio)
10	Tenor	ระยะเวลากู้ (เดือน)	อัตราส่วน (Ratio)
11	AppraisalAmount	ราคาประเมินหลักทรัพย์ (บาท)	อัตราส่วน (Ratio)

ตารางที่ 3.3 (ต่อ) คุณลักษณะของตัวแปรที่ใช้ในงานวิจัย

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ประเภทข้อมูลตามระดับมาตรวัด
12	BusinessExpYear	อายุประสบการณ์ธุรกิจ (ปี)	อัตราส่วน (Ratio)
13	SalesAmt	ยอดขายต่อเดือน (บาท)	อัตราส่วน (Ratio)
14	TotalIncome	รายได้รวมต่อเดือน (บาท)	อัตราส่วน (Ratio)
15	TotalDebt	หนี้สินรวมทั้งหมด (บาท)	อัตราส่วน (Ratio)
16	DSCR	สัดส่วนรายได้ที่สามารถใช้สำหรับการชำระหนี้กับจำนวนหนี้ที่ต้องชำระ (ร้อยละ)	อัตราส่วน (Ratio)
17	TTRec	จำนวนบัญชีทั้งหมดในระบบ (บัญชี)	อัตราส่วน (Ratio)
18	RiskLevel	ระดับความเสี่ยง (ต่ำ/กลาง/สูง)	นามบัญญัติ (Nominal)
19	ActRecNo	จำนวนบัญชีที่ Active ทั้งหมดในระบบ (บัญชี)	อัตราส่วน (Ratio)
20	ActRecAmt	จำนวนเงินที่ Active ทั้งหมดในระบบ (บาท)	อัตราส่วน (Ratio)
21	Last6MonReqNo	จำนวนครั้งที่ขอกู้รอบ 6 เดือน (ครั้ง)	อัตราส่วน (Ratio)
22	SumLimitAmt	ยอดรวมวงเงินทั้งหมดในระบบ (บาท)	อัตราส่วน (Ratio)
23	Sale_per_TotalDebt	อัตราส่วนยอดขายต่อยอดหนี้ทั้งหมด (ร้อยละ)	อัตราส่วน (Ratio)
24	CreditAmt_per_Installment	อัตราส่วนยอดค่างวดต่อยอดเงินกู้ (ร้อยละ)	อัตราส่วน (Ratio)
25	DebtToIncomeRatio	ภาระหนี้ทั้งหมดเทียบกับรายได้ทั้งหมด (ร้อยละ)	อัตราส่วน (Ratio)
26	InstallmentToIncomeRatio	ภาระค่างวดต่อเดือนเทียบกับรายได้ (ร้อยละ)	อัตราส่วน (Ratio)

ตารางที่ 3.3 (ต่อ) คุณลักษณะของตัวแปรที่ใช้ในงานวิจัย

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ประเภทข้อมูลตามระดับมาตรวัด
27	CreditUtilization	อัตราส่วนยอดขายต่อยอดหนี้ทั้งหมด (ร้อยละ)	อัตราส่วน (Ratio)
28	Tenor_per_Income	ระยะเวลากู้เทียบกับรายได้ (ร้อยละ)	อัตราส่วน (Ratio)
29	HasGuarantor	การมีผู้ค้ำประกัน (มี/ไม่มี)	นามบัญญัติ (Nominal)
30	FPD	สถานะผิดนัดชำระหนี้งวดแรก (ผิดนัด/ไม่ผิดนัด)	นามบัญญัติ (Nominal)

จากตารางที่ 3.3 ตัวแปรตามได้แก่สถานะผิดนัดชำระหนี้งวดแรก (FPD) จะมีค่าที่เป็นไปได้โดย 0 คือ ไม่ผิดนัดชำระหนี้งวดแรก (Non-FPD) และ 1 คือ ผิดนัดชำระหนี้งวดแรก (FPD: ผิดนัดชำระหนี้งวดแรกเกิน 30 วัน)

3.3.4 สำรวจข้อมูลเบื้องต้น

ทำการสำรวจข้อมูลเบื้องต้นเพื่อวางแผนการเตรียมข้อมูล (Data Preparation) ขั้นแรกสำรวจจำนวนทั้งหมด สัดส่วนตัวแปรตามและข้อมูลขาดหาย สำรวจข้อมูลทั้งชุดพบว่ามีจำนวน 3,013 บัญชีและมี 30 ตัวแปรทั้งหมด ไม่พบข้อมูลขาดหาย สัดส่วนของตัวแปรตาม โดย 0 คือไม่ผิดนัดชำระหนี้งวดแรกมีสัดส่วนประมาณร้อยละ 95.6 และ 1 คือผิดนัดชำระหนี้งวดแรกมีสัดส่วนประมาณร้อยละ 4.4 ดังนั้นจากการสำรวจข้อมูลเบื้องต้นปัญหานี้เป็นปัญหาข้อมูลไม่สมดุลค่อนข้างมาก (Imbalance Problem) มีตัวแปรอิสระชนิด อัตราส่วน (Ratio) จำนวน 24 ตัวแปร และตัวแปรอิสระชนิดนามบัญญัติ (Nominal) จำนวน 5 ตัวแปร

3.3 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลเป็นกระบวนการสำคัญโดยมีวัตถุประสงค์เพื่อปรับปรุงคุณภาพของข้อมูลให้มีความพร้อมในการป้อนเข้าสู่กระบวนการสร้างโมเดลโดยมีรายละเอียดดังนี้

3.3.1 การทำความสะอาดข้อมูล (Data Cleansing)

ในการเตรียมข้อมูลเพื่อการวิเคราะห์และการสร้างโมเดล Machine Learning จำเป็นต้องดำเนินการ "ทำความสะอาดข้อมูล" (Data Cleansing) อย่างเป็นระบบ เพื่อให้ได้ชุดข้อมูลที่มีคุณภาพสูง มีความสมบูรณ์ ถูกต้อง และเหมาะสมต่อการเรียนรู้ของโมเดล ในงานวิจัยนี้ผู้วิจัยได้ดำเนินการทำความสะอาดข้อมูลตามขั้นตอนดังนี้

ขั้นตอนที่ 1 ตรวจสอบข้อมูลซ้ำซ้อน (Duplicate Records) โดยพิจารณาความซ้ำกันของข้อมูลในระดับแถว หากพบว่าแถวใดมีค่าทุกคอลัมน์เหมือนกันทั้งหมดจะถูกลบทิ้ง อย่างไรก็ตาม ผลการประมวลผลพบว่าไม่มีข้อมูลใดที่ซ้ำกันทุกค่าในระดับแถว (Row-Level Duplication) จึงไม่มีการลบแถวหรือคอลัมน์ในขั้นตอนนี้

ขั้นตอนที่ 2 ตรวจสอบคอลัมน์ที่มีข้อมูลสูญหายจำนวนมาก โดยกำหนดเกณฑ์ว่าหากคอลัมน์ใดมีค่าข้อมูลสูญหายเกินร้อยละ 30 ของจำนวนแถวทั้งหมด จะถือว่าไม่สามารถใช้ในการวิเคราะห์ได้อย่างมีประสิทธิภาพ จึงควรถูกตัดออกจากชุดข้อมูล อย่างไรก็ตาม จากการประเมินไม่พบคอลัมน์ใดที่มีค่าข้อมูลสูญหายเกินเกณฑ์ดังกล่าว จึงไม่มีการลบคอลัมน์ในขั้นตอนนี้

ขั้นตอนที่ 3 ตรวจสอบแถวที่มีข้อมูลสูญหาย โดยผู้วิจัยได้กำหนดเงื่อนไขว่าหากแถวใดมีข้อมูลสูญหายตั้งแต่ 3 คอลัมน์ขึ้นไป จะพิจารณาว่าเป็นข้อมูลที่ไม่สมบูรณ์และควรถูกลบออกจากชุดข้อมูล จากการประเมินไม่พบแถวใดที่มีค่าข้อมูลสูญหายเกินเกณฑ์ดังกล่าว จึงไม่มีการลบแถวในขั้นตอนนี้

ขั้นตอนที่ 4 ตรวจสอบคอลัมน์ที่ไม่มีความหลากหลายของข้อมูล (Zero Variance Features) โดยพิจารณาว่าหากคอลัมน์ใดมีค่าซ้ำกันทั้งหมด หรือมีค่าหนึ่งค่าปรากฏมากกว่าร้อยละ 99 ของข้อมูลทั้งคอลัมน์ จะถือว่าไม่มีประโยชน์ในการวิเคราะห์และควรถูกลบทิ้ง อย่างไรก็ตามจากการตรวจสอบพบว่าไม่มีคอลัมน์ใดเข้าข่ายตามเกณฑ์ จึงไม่มีการลบคอลัมน์ในขั้นตอนนี้

ขั้นตอนที่ 5 ตรวจสอบคอลัมน์ที่มีความหลากหลายน้อย (Low Variance Features) โดยวิเคราะห์ค่าความแปรปรวน (Variance) และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) หากคอลัมน์ใดมีค่าเบี่ยงเบนมาตรฐานต่ำกว่า 0.01 หรือมีค่าที่พบมากที่สุดครอบคลุมข้อมูลในสัดส่วนสูงเกินไป จะพิจารณาลบทิ้ง ผลการวิเคราะห์พบว่าคอลัมน์ทั้งหมดมีค่าความแปรปรวนสูงกว่าเกณฑ์ที่กำหนด จึงไม่มีการลบคอลัมน์ใดเพิ่มเติมในขั้นตอนนี้

โดยสรุปจากกระบวนการทำความสะอาดข้อมูล พบว่าข้อมูลผ่านการตรวจสอบโดยกระบวนการทั้ง 5 ขั้นตอนไม่พบขั้นตอนใดที่ต้องปรับปรุงส่งผลให้จำนวนแถวและคอลัมน์ยังคงเดิม ซึ่งการดำเนินการดังกล่าวเป็นกระบวนการที่จำเป็นที่จะต้องตรวจสอบทุกครั้งเพื่อทำให้ได้ชุดข้อมูลมีความสมบูรณ์ โปร่งใส และเหมาะสมต่อการนำไปใช้ในการวิเคราะห์และการสร้างโมเดลอย่างมีประสิทธิภาพในขั้นตอนต่อไป

3.3.2 การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูลเป็นขั้นตอนสำคัญที่ช่วยให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์หรือสร้างโมเดลทางสถิติและแมชชีนเลิร์นนิง โดยมีรายละเอียดดังนี้

1) ข้อมูลเชิงตัวเลข (Numerical Data) ข้อมูลประเภทนี้ประกอบด้วยตัวแปรที่มีลักษณะเป็นตัวเลข หรือมีค่าต่อเนื่อง เช่น อายุ รายได้รวม เป็นต้น สำหรับข้อมูลเชิงตัวเลขไม่จำเป็นต้องแปลงค่าก่อนนำเข้าโมเดลแต่ในบางกรณีเพื่อให้ข้อมูลอยู่ในช่วงที่เหมาะสมต่อการเรียนรู้ของโมเดล จะมีการ

ปรับขนาดข้อมูล เช่น การทำ Normalization หรือ Standardization และตรวจสอบความสัมพันธ์ระหว่างกันเองของตัวแปรอิสระเชิงตัวเลข

```
# สรุปคู่ที่ correlation > 0.9 (พิจารณาตัด)
corr_pairs = corr_matrix.abs().unstack().sort_values(ascending=False)
corr_pairs = corr_pairs[corr_pairs < 1] # ตัด diagonal
print("Top correlated pairs:")
display(corr_pairs[corr_pairs > 0.9])
```

```
Top correlated pairs:
Installment        TotalCreditAmount    0.998767
TotalCreditAmount  Installment          0.998767
DebtToIncomeRatio  DSCR                 0.963620
DSCR               DebtToIncomeRatio    0.963620
TTRec              ActRecNo             0.910065
ActRecNo           TTRec                0.910065
dtype: float64
```

รูปที่ 3.1 สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ของตัวแปรอิสระ

จากรูปที่ 3.1 ผู้วิจัยได้ตรวจสอบค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ของตัวแปรอิสระ และตัดตัวแปรที่มีค่าสัมประสิทธิ์ที่มากกว่า 0.9 คือ วงเงินรวม (TotalCreditAmount), สัดส่วนรายได้ที่สามารถใช้สำหรับการชำระหนี้กับจำนวนหนี้ที่ต้องชำระ (DSCR) และจำนวนบัญชีที่ Active ทั้งหมดในระบบ (ActRecNo) เพื่อลดปัญหา Multicollinearity ซึ่งช่วยให้โมเดลมีเสถียรภาพ สามารถตีความผลลัพธ์ได้ง่ายขึ้น และลดโอกาสเกิด Overfitting ในการทดสอบกับข้อมูลใหม่

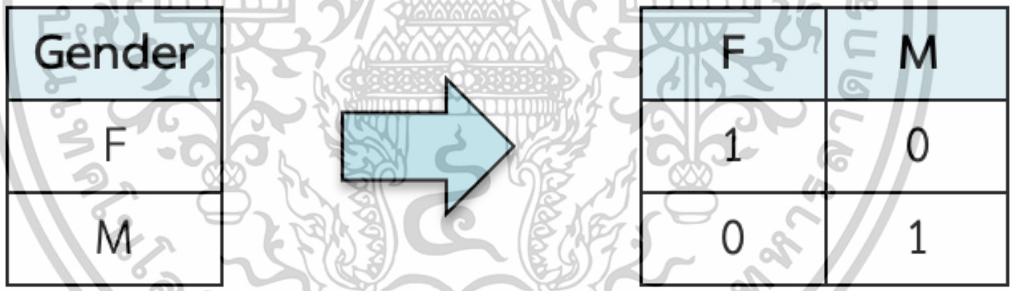
2) ข้อมูลเชิงกลุ่ม (Categorical Data) ข้อมูลประเภทนี้ประกอบด้วยตัวแปรที่แสดงถึงกลุ่มหรือประเภท ซึ่งไม่สามารถนำไปประมวลผลด้วยโมเดลเชิงตัวเลขได้โดยตรง จำเป็นต้องแปลงข้อมูลให้อยู่ในรูปแบบที่โมเดลสามารถนำไปใช้งานได้ โดยข้อมูลเชิงกลุ่มสามารถแบ่งย่อยได้อีก 2 ประเภท ได้แก่

3) ข้อมูลนามบัญญัติ (Nominal Data) เป็นข้อมูลที่แสดงถึงประเภทหรือกลุ่มซึ่งไม่มีลำดับความสำคัญ ตัวอย่างเช่น เพศ (ชาย/หญิง), ภูมิภาค (เหนือ/กลาง/ตะวันออกเฉียงเหนือ/ใต้) เป็นต้น ตัวแปรประเภทนี้จะถูกแปลงโดยใช้เทคนิค One-Hot Encoding ตามที่แสดงในตารางที่ 3.4

ตารางที่ 3.4 คุณลักษณะของตัวแปรต้นที่เป็นข้อมูลนามบัญญัติ

ชื่อตัวแปร	**ค่าที่เป็นไปได้**
Gender	ชาย, หญิง
MaritalStatus	โสด, สมรส, สมรสไม่จดทะเบียน, หย่าร้าง, หม้าย
Region	ภาคกลาง, ภาคเหนือ, ภาคตะวันออกเฉียงเหนือ, ภาคตะวันออก, ภาคใต้, กรุงเทพมหานคร
BusinessType	เกษตรกรรม, ก่อสร้าง, การเงิน, อุตสาหกรรมการผลิต, อสังหาริมทรัพย์, บริการ, การค้า, ขนส่ง
HasGuarantor	มีผู้ค้ำประกัน, ไม่มีผู้ค้ำประกัน

การเข้ารหัสแบบวัน-ฮอต (One Hot Encoding) วิธีนี้นิยมใช้กับข้อมูลเชิงกลุ่มที่ไม่มีลำดับ (Nominal Data) โดยจะเปลี่ยนค่าของแต่ละกลุ่มให้เป็นคอลัมน์ใหม่ และแทนค่าด้วย 0 หรือ 1 ตามการปรากฏของข้อมูลในแต่ละประเภท ในตัวอย่างรูปที่ 3.2 ตัวแปร “Gender ที่ประกอบด้วยกลุ่ม “F” และ “M” จะถูกแยกออกเป็นคอลัมน์ย่อยตามแต่ละกลุ่ม และระบุค่า 1 ในตำแหน่งของกลุ่มที่ตรงกัน ส่วนค่าที่ไม่ตรงกันจะเป็น 0 ทำให้แต่ละกลุ่มสามารถแสดงอยู่บนมิติข้อมูลใหม่ที่ไม่ทับซ้อนกัน



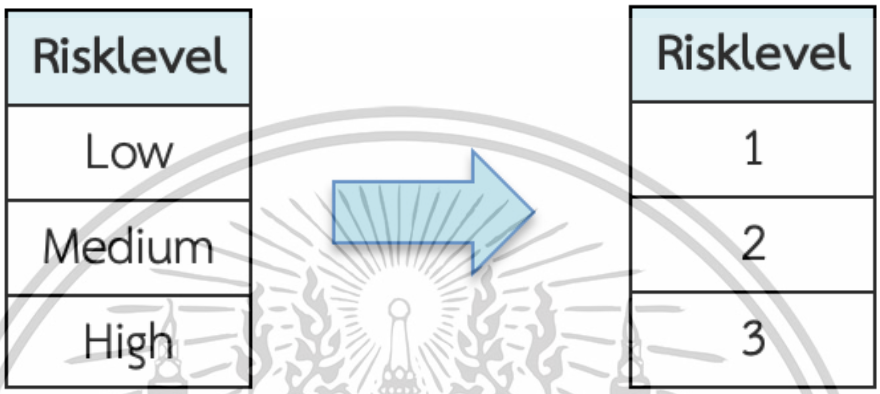
รูปที่ 3.2 การแปลงข้อมูลเชิงกลุ่มแบบไม่มีลำดับด้วยวิธี One Hot Encoding

3) ข้อมูลเชิงอันดับ (Ordinal Data) เป็นข้อมูลที่แสดงถึงลำดับขั้นหรือระดับ ซึ่งลำดับของข้อมูลมีความหมาย ตัวอย่างเช่น ระดับความเสี่ยง (ต่ำ/ปานกลาง/สูง) เป็นต้น การแปลงข้อมูลประเภทนี้นิยมใช้เทคนิค Label Encoding หรือการแทนค่าด้วยตัวเลขที่สะท้อนลำดับชั้นตามที่แสดงในตารางที่ 3.5

ตารางที่ 3.5 คุณลักษณะของตัวแปรต้นที่เป็นข้อมูลเชิงอันดับ

ชื่อตัวแปร	**ค่าที่เป็นไปได้**
Risklevel	ความเสี่ยงต่ำ, ความเสี่ยงปานกลาง, ความเสี่ยงสูง

การเข้ารหัสแบบ **Label Encoding** วิธีนี้นิยมใช้กับข้อมูลเชิงกลุ่มที่มีลำดับ (Ordinal Data) โดยจะเปลี่ยนค่าของแต่ละกลุ่มให้เป็นค่าตัวเลขแทน และแทนค่าด้วย 0, 1, 2 ตามลำดับความหมายของข้อมูล ในตัวอย่าง รูปที่ 3.3 ตัวแปร “RiskLevel” ที่ประกอบด้วยกลุ่ม “ต่ำ”, “ปานกลาง” และ “สูง” จะถูกแปลงโดยกำหนดลำดับของแต่ละกลุ่ม เช่น “ต่ำ” = 0, “ปานกลาง” = 1 และ “สูง” = 2 โดยค่าที่มากขึ้นหมายถึงความเสี่ยงที่สูงขึ้น



รูปที่ 3.3 การแปลงข้อมูลเชิงกลุ่มแบบมีลำดับด้วยวิธี Label Encoding

หลังจากดำเนินการเตรียมข้อมูลเบื้องต้นและแปลงค่าตัวแปรให้เหมาะสมกับการนำไปวิเคราะห์แล้ว ขั้นตอนถัดไปคือการคัดเลือกตัวแปรต้น (Features) ที่มีความเกี่ยวข้องและเหมาะสมต่อการสร้างโมเดลเพื่อทำนายการผิดนัดชำระหนี้ โดยพิจารณาจากความสัมพันธ์เชิงทฤษฎี ความพร้อมของข้อมูล และความสามารถในการแปลผล ตัวแปรเหล่านี้ประกอบด้วยทั้งข้อมูลเชิงปริมาณและข้อมูลเชิงกลุ่ม ซึ่งผ่านกระบวนการแปลงค่าด้วยวิธีต่างๆ ให้สามารถนำเข้าโมเดลทางสถิติและ Machine Learning ได้อย่างเหมาะสม ดังตารางที่ 3.6 ที่แสดงรายชื่อตัวแปรที่ผ่านการคัดเลือกพร้อมระบุเทคนิคที่ใช้ในการแปลงข้อมูลแต่ละประเภทก่อนนำเข้าสู่กระบวนการสร้างโมเดล

ตารางที่ 3.6 สรุปวิธีการแปลงข้อมูลของแต่ละตัวแปร

ลำดับ	ชื่อตัวแปร	วิธีการแปลง
1	Gender	One Hot Encoding
2	CustAge	Standardization
3	MaritalStatus	One Hot Encoding
4	Region	One Hot Encoding
5	BusinessType	One Hot Encoding
6	AppliedAmount	Standardization
7	Installment	Standardization

ตารางที่ 3.6 (ต่อ) สรุปวิธีการแปลงข้อมูลของแต่ละตัวแปร

ลำดับ	ชื่อตัวแปร	วิธีการแปลง
8	LTV	Standardization
9	Tenor	Standardization
10	AppraisalAmount	Standardization
11	BusinessExpYear	Standardization
12	SalesAmt	Standardization
13	TotalIncome	Standardization
14	TotalDebt	Standardization
15	TTRec	Standardization
16	RiskLevel	Label Encoding
17	ActRecAmt	Standardization
18	Last6MonReqNo	Standardization
19	SumLimitAmt	Standardization
20	Sale_per_TotalDebt	Standardization
21	CreditAmt_per_Installment	Standardization
22	DebtToIncomeRatio	Standardization
23	InstallmentToIncomeRatio	Standardization
24	CreditUtilization	Standardization
25	Tenor_per_Income	Standardization
26	HasGuarantor	One Hot Encoding

3.4 การสร้างโมเดล (Modeling)

เลือกใช้โมเดลที่ได้เลือกจากการทบทวนงานวิจัยที่เกี่ยวข้องได้แก่ Logistic Regression (การถดถอยโลจิสติก), Random Forest (ป่าไม้สุ่ม), XGBoost (เอ็กซ์จีบูสต์) ปรับจูนค่าพารามิเตอร์ (Hyperparameter Tuning) เช่น learning_rate, max_depth, n_estimators เป็นต้น โดยใช้ GridSearchCV และใช้ Cross-validation (K-Fold) เพื่อประเมินความน่าเชื่อถือของโมเดล เป้าหมายในหัวข้อนี้คือ ได้โมเดลที่มีประสิทธิภาพสูงสุดตามตัวชี้วัดที่กำหนดและเลือกโมเดลที่ตอบโจทย์ทั้งเชิงธุรกิจและเชิงเทคนิค มีขั้นตอนดังต่อไปนี้

1) แบ่งข้อมูลที่ได้ทำการแปลงค่าและปรับแต่งจากขั้นตอนก่อนหน้าเรียบร้อยแล้วเป็นชุดข้อมูลฝึกฝนและข้อมูลทดสอบ

```python
from sklearn.model_selection import train_test_split

# แบ่งข้อมูล (stratify ให้รักษาสัดส่วน class)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)

print(f"Shape X_train: {X_train.shape}, y_train: {y_train.shape}")
print(f"Shape X_test: {X_test.shape}, y_test: {y_test.shape}")

# โชว์สัดส่วน (percent) class ของ train/test
print("\n== Class distribution ==")
print("Train:")
display(y_train.value_counts(normalize=True).rename("proportion") * 100)
print("Test:")
display(y_test.value_counts(normalize=True).rename("proportion") * 100)
```

```
Shape X_train: (2410, 29), y_train: (2410,)
Shape X_test: (603, 29), y_test: (603,)

== Class distribution ==
Train:
fpd
0    95.560166
1     4.439834
Name: proportion, dtype: float64
Test:
fpd
0    95.522388
1     4.477612
Name: proportion, dtype: float64
```

รูปที่ 3.4 การแบ่งข้อมูลฝึกฝนและข้อมูลทดสอบ

จากรูปที่ 3.4 แบ่งข้อมูลออกเป็น 2 ส่วน คือ ชุดฝึกฝน (Training Set) 80% สำหรับฝึกโมเดล และชุดทดสอบ (Testing Set) 20% สำหรับประเมินประสิทธิภาพโมเดล โดยใช้ Stratified Sampling เพื่อรักษาสัดส่วนของกลุ่มเป้าหมาย (FPD) ให้เท่าเทียมในทั้งสองชุดข้อมูล

2) การจัดการข้อมูลไม่สมดุล (Imbalanced Data Handling)

จากการสำรวจข้อมูล FPD มีสัดส่วนกลุ่มผิดนัด (Minority Class) น้อยกว่ากลุ่มปกติ (Majority Class) มาก หากไม่แก้ไข โมเดลจะเรียนรู้แยกแยะกลุ่มผิดนัดได้ไม่ดี จึงใช้เทคนิค SMOTE (Synthetic Minority Over-Sampling Technique) SMOTEENN และ Random Undersampling ในการจัดการปัญหาข้อมูลไม่สมดุล

```python
samplers = {
    "SMOTE": SMOTE(random_state=42),
    "SMOTEENN": SMOTEENN(random_state=42),
    "RandomUnderSampler": RandomUnderSampler(random_state=42)
}

# ใช้ X_train_num, y_train กับ samplers
for name, sampler in samplers.items():
    X_res, y_res = sampler.fit_resample(X_train_num, y_train)
    print(f"\n{name}:")
    print(f"X_res shape: {X_res.shape}, y_res shape: {y_res.shape}")
    class_dist = pd.Series(y_res).value_counts(normalize=True).rename("proportion") * 100
    display(class_dist)
```

```
SMOTE:
X_res shape: (4606, 47), y_res shape: (4606,)
fpd
0    50.0
1    50.0
Name: proportion, dtype: float64
SMOTEENN:
X_res shape: (3731, 47), y_res shape: (3731,)
fpd
1    58.402573
0    41.597427
Name: proportion, dtype: float64
RandomUnderSampler:
X_res shape: (214, 47), y_res shape: (214,)
fpd
0    50.0
1    50.0
Name: proportion, dtype: float64
```

รูปที่ 3.5 การจัดการข้อมูลไม่สมดุล

ข้อควรระวังการทำ Resampling เฉพาะบนชุดข้อมูลฝึกฝน (Training Set) เท่านั้น และตรวจสอบผล Sampling ว่าข้อมูลใหม่ที่สร้างขึ้นไม่ผิดธรรมชาติของข้อมูลเดิม

3) การตั้งค่าพารามิเตอร์แบบกริดสำหรับแต่ละโมเดล

ตามช่วงที่เหมาะสมกับข้อมูลที่ไม่สมดุลและทรัพยากรสำหรับการสร้างโมเดล เนื่องจากการตั้งค่ามีผลกับทรัพยากรที่ใช้ในการสร้าง หากมีเพียงพอสามารถกำหนดค่าให้ละเอียดขึ้นได้ การปรับพารามิเตอร์แบบกริดแสดงดังรูปที่ 3.6

```
grids = {
    "LogisticRegression": {
        'clf__C': [0.01, 0.1, 1, 10],
        'clf__solver': ['liblinear'],
        'clf__penalty': ['l1', 'l2'],
        'clf__class_weight': ['balanced', None]
    },
    "RandomForest": {
        'clf__n_estimators': [100, 200],
        'clf__max_depth': [4, 6, 10],
        'clf__min_samples_split': [2, 5],
        'clf__min_samples_leaf': [1, 2],
        'clf__class_weight': ['balanced', None]
    },
    "XGBoost": {
        'clf__learning_rate': [0.01, 0.05, 0.1],
        'clf__max_depth': [3, 5],
        'clf__n_estimators': [100, 150],
        'clf__scale_pos_weight': [scale_pos_weight],
        'clf__subsample': [0.8, 1.0],
        'clf__colsample_bytree': [0.8, 1.0]
    }
```

รูปที่ 3.6 กำหนดค่าพารามิเตอร์

หลังจากตั้งค่าพารามิเตอร์แบบกริดสำหรับแต่ละโมเดลแล้วจึงดำเนินการค้นหาค่าพารามิเตอร์ที่เหมาะสมที่สุด (Best Parameter) สำหรับแต่ละโมเดลที่ใช้เทคนิคการจัดการข้อมูลไม่สมดุลต่างๆ ผลลัพธ์ที่ได้จากกระบวนการปรับแต่งโมเดลแต่ละชุดจะแสดงไว้ในตารางที่ 3.7

ตารางที่ 3.7 พารามิเตอร์ที่ดีที่สุดของแต่ละโมเดล

เทคนิคปรับสมดุลข้อมูล	โมเดล	พารามิเตอร์ที่เหมาะสม	พื้นที่ใต้กราฟ ROC
SMOTE	LR	{'clf__C': 10, 'clf__class_weight': 'balanced' 'clf__penalty': 'l2', 'clf__solver': 'liblinear'}	0.6828
SMOTE	RF	{'clf__class_weight': 'balanced', 'clf__max_depth': 10, ‘clf__min_samples_leaf’: 2, 'clf__min_samples_split': 2, 'clf__n_estimators': 200}	0.7239

ตารางที่ 3.7 (ต่อ) พารามิเตอร์ที่ดีที่สุดของแต่ละโมเดล

เทคนิคปรับสมดุลข้อมูล	โมเดล	พารามิเตอร์ที่เหมาะสม	พื้นที่ใต้กราฟ ROC
SMOTE	XGBoost	{'clf__max_depth': 3, 'clf__n_estimators': 150, 'clf__colsample_bytree': 0.08, 'clf__learning_rate': 0.05, 'clf__scale_pos_weight': 21.523364485981308}, 'clf__subsample': 1	0.6974
SMOTEENN	LR	{'clf__C': 10, 'clf__class_weight': 'balanced', 'clf__penalty': 'l1', 'clf__solver': 'liblinear'}	0.7016
SMOTEENN	RF	{'clf__class_weight': 'balanced', 'clf__max_depth': 10, 'clf__min_samples_leaf': 2, 'clf__min_samples_split': 5, 'clf__n_estimators': 200}	0.7144
SMOTEENN	XGBoost	{'clf__max_depth': 5, 'clf__n_estimators': 150, 'clf__colsample_bytree': 1, 'clf__learning_rate': 0.1, 'clf__scale_pos_weight': 21.523364485981308}, 'clf__subsample': 0.8	0.7098
RUS	LR	{'clf__C': 10, 'clf__class_weight': 'balanced', 'clf__penalty': 'l2', 'clf__solver': 'liblinear'}	0.7008

ตารางที่ 3.7 (ต่อ) พารามิเตอร์ที่ดีที่สุดของแต่ละโมเดล

เทคนิคปรับสมดุลข้อมูล	โมเดล	พารามิเตอร์ที่เหมาะสม	พื้นที่ใต้กราฟ ROC
RUS	RF	{'clf__class_weight': 'balanced', 'clf__max_depth': 10, 'clf__min_samples_leaf': 2, 'clf__min_samples_split': 5, 'clf__n_estimators': 200}	0.7329
RUS	XGBoost	{'clf__max_depth': 5, 'clf__n_estimators': 150, 'clf__colsample_bytree': 1, 'clf__learning_rate': 0.05, 'clf__scale_pos_weight': 21.523364485981308}, 'clf__subsample': 0.8	0.7163

จากตารางที่ 3.7 ผลการปรับพารามิเตอร์กริดสำหรับโมเดล Logistic Regression พบว่าพารามิเตอร์ที่เหมาะสม ได้แก่ C=10, class_weight='balanced', และ penalty='l1' หรือ 'l2' ขึ้นกับเทคนิค Sampling ที่ใช้ โดยค่า ROC-AUC อยู่ในช่วงร้อยละ 68-70 ซึ่งแม้จะต่ำกว่ากลุ่ม Tree-Based Model แต่ยังคงให้ผลการจำแนกที่น่าพอใจในกรณีที่ต้องการโมเดลที่อธิบายผลได้ง่าย ด้าน Random Forest พบว่าใช้ max_depth=10, min_samples_leaf=2, class_weight='balanced' และจำนวนต้นไม้ที่เหมาะสมที่ 200 ต้น ช่วยเพิ่ม Performance ของโมเดล โดยเฉพาะเมื่อจับคู่กับ RUS จะได้ค่า ROC-AUC สูงสุดที่ร้อยละ 73.29 ซึ่งสูงกว่าการใช้ SMOTE หรือ SMOTEENN เล็กน้อย แสดงให้เห็นถึงประโยชน์ของการลดข้อมูลในกลุ่มที่มีมากเกินไป ในขณะที่ XGBoost ซึ่งเป็นโมเดลที่ได้รับความนิยมในงานด้านการจัดการข้อมูลไม่สมดุล ให้ผลลัพธ์ที่ดีเช่นกัน โดยการปรับ scale_pos_weight ตามอัตราส่วนข้อมูลจริง (21.52 ในชุดข้อมูลนี้) และใช้ max_depth=3-5, n_estimators=150, learning_rate=0.05-0.1 จะได้ค่า ROC-AUC ในช่วงร้อยละ 69-72 แสดงถึงศักยภาพของ XGBoost ในการจัดการกับข้อมูลที่มีความไม่สมดุลสูงได้อย่างมีประสิทธิภาพ โดยสรุปผลการปรับพารามิเตอร์แสดงให้เห็นว่าการเลือกใช้เทคนิค Sampling ร่วมกับการ Tune พารามิเตอร์หลักของแต่ละโมเดลสามารถช่วยเพิ่มประสิทธิภาพในการจำแนกกลุ่มลูกค้าที่มีแนวโน้มผิดนัดชำระหนี้ได้ดียิ่งขึ้น ทั้งนี้ Random Forest และ XGBoost ถือเป็นตัวเลือกที่เหมาะสมที่สุดสำหรับปัญหานี้ในข้อมูลชุดนี้ เมื่อพิจารณาจากค่า ROC-AUC และโครงสร้างโมเดลที่รองรับความสัมพันธ์แบบไม่เชิงเส้นระหว่างตัวแปรได้ดีกว่า Logistic Regression นอกจากค่าพารามิเตอร์ที่เหมาะสมจะช่วยเพิ่มประสิทธิภาพในการทำนายของโมเดลแล้ว การปรับค่าพารามิเตอร์ยังมีผลโดยตรงต่อการใช้

ทรัพยากรคอมพิวเตอร์ทั้งในแง่ของเวลาในการประมวลผลและหน่วยความจำที่ต้องใช้ ด้วยข้อจำกัดของทรัพยากรคอมพิวเตอร์ในงานวิจัยนี้ เช่น เวลาในการประมวลผลที่มีจำกัดและหน่วยความจำที่ไม่สูงมาก การปรับค่าพารามิเตอร์จึงต้องคำนึงถึง สมดุลระหว่างประสิทธิภาพของโมเดลกับการใช้ทรัพยากร การเลือกใช้ค่าพารามิเตอร์ที่เหมาะสมและไม่สูงเกินความจำเป็นจะช่วยให้โมเดลมีความแม่นยำในระดับดี ในขณะที่ยังสามารถประมวลผลได้ภายใต้ข้อจำกัดที่มีอยู่

4) ขั้นตอนการสร้างโมเดล

ในขั้นตอนการสร้างโมเดล ผู้วิจัยได้ดำเนินการตามกระบวนการที่มีความเป็นระบบ โดยเริ่มจากการแบ่งข้อมูลออกเป็น 2 ชุดหลัก ได้แก่ ชุดข้อมูลฝึกฝน (Training Data) และชุดข้อมูลทดสอบ (Testing Data) โดยได้เลือกใช้วิธีการแบ่งข้อมูลแบบไขว้ (Hold-Out) สัดส่วนชุดข้อมูลฝึกฝนร้อยละ 80 และชุดข้อมูลทดสอบร้อยละ20 โดยรักษาสัดส่วนคลาสน้อย และผ่านกระบวนการปรับจูนไฮเปอร์พารามิเตอร์โดยใช้วิธี Grid Search โดยมีขั้นตอนดำเนินการที่สามารถสรุปไว้ในตาราง 3.8 ดังนี้

ตารางที่ 3.8 ขั้นตอนการสร้างโมเดล

ขั้นตอน	Logistic Regression	Random Forest	XGBoost
การเลือกตัวแปร	ใช้ตัวแปรทั้งหมดที่ผ่านการเตรียมข้อมูล		
ตั้งค่าพารามิเตอร์	clf__C, clf__class_weight, clf__penalty, clf__solver	clf_class_weight, clf_max_depthclf_min, samples_leaf, clf_min_samples_split, clf_n_estimators	clf__max_depth, clf__n_estimators, clf__colsample_bytree, clf__learning_rate, clf__scale_pos_weight, clf__subsample
การแบ่งข้อมูล	Hold-Out รักษาสัดส่วนคลาสน้อย		
การฝึกฝนและทดสอบ	ชุดข้อมูลฝึกฝน 80% และชุดข้อมูลทดสอบ 20%		
เทคนิคการจัดการข้อมูล	SMOTE, SMOTEENN, Random Undersampling (ข้อมูลฝึกฝน)		
ตรวจสอบประสิทธิภาพ	K-Fold Cross Validation (k=5) บนชุดข้อมูลฝึกฝน, ชุดข้อมูลทดสอบ		
การบันทึกผลลัพธ์	ROC-AUC, Accuracy, Precision, Recall, F1-Score, KS, Confusion Matrix		
เปรียบเทียบประสิทธิภาพ	ใช้ตัวชี้วัด ROC-AUC และ KS		
การสรุปผล	คัดเลือกโมเดลที่ดีที่สุดสำหรับนำไปประยุกต์ใช้		

จากตาราง 3.8 ขั้นตอนการสร้างโมเดลในการทำนายการผิดนัดชำระหนี้งวดแรกในงานวิจัยนี้ ประกอบด้วยการเลือกตัวแปรทั้งหมดที่ผ่านการเตรียมข้อมูลล่วงหน้า จากนั้นดำเนินการตั้งค่าพารามิเตอร์สำหรับแต่ละโมเดล ได้แก่ Logistic Regression, Random Forest และ XGBoost โดยเลือกพารามิเตอร์สำคัญ เช่น clf__C, clf__class_weight, clf__penalty และ clf__solver สำหรับ Logistic Regression, clf__class_weight, clf__max_depth, clf__min_samples_leaf, clf__min_samples_split และ clf__n_estimators สำหรับ Random Forest และพารามิเตอร์กลุ่ม clf__max_depth, clf__n_estimators, clf__colsample_bytree, clf__learning_rate, clf__scale_pos_weight และ clf__subsample สำหรับ XGBoost ในการแบ่งชุดข้อมูล ใช้วิธี Hold-Out โดยรักษาสัดส่วนของกลุ่มเป้าหมาย (Class น้อย) แยกเป็นชุดข้อมูลฝึกฝน 80% และชุดข้อมูลทดสอบ 20% สำหรับการจัดการปัญหาข้อมูลไม่สมดุล ใช้เทคนิค SMOTE, SMOTEENN และ Random Undersampling บนชุดข้อมูลฝึกฝนเท่านั้น เพื่อให้ชุดข้อมูลทดสอบคงสภาพปัญหาจริง กระบวนการฝึกฝนและทดสอบโมเดลใช้หลักการ K-Fold Cross Validation (k=5) บนชุดข้อมูลฝึกฝน เพื่อประเมินประสิทธิภาพโมเดลในแต่ละเทคนิคการจัดการข้อมูลไม่สมดุลและโมเดลอย่างรอบด้าน ก่อนนำไปทดสอบกับชุดข้อมูลทดสอบอีกครั้งหนึ่ง ทั้งนี้ ได้บันทึกผลลัพธ์โดยใช้ตัวชี้วัดสำคัญ ได้แก่ ROC-AUC, Accuracy, Precision, Recall, F1-Score, KS และ Confusion Matrix ในส่วนการเปรียบเทียบประสิทธิภาพของโมเดล ใช้ ROC-AUC และ KS เป็นตัวชี้วัดหลักสำหรับการเลือกโมเดลที่ดีที่สุด ซึ่งโมเดลที่ผ่านการคัดเลือกจะถูกนำไปประยุกต์ใช้ในการทำนายการผิดนัดชำระหนี้งวดแรกบนข้อมูลใหม่หรือในระบบงานจริงต่อไป

3.5 การประเมินผล (Evaluation)

ในการประเมินประสิทธิภาพของโมเดลทำนายการผิดนัดชำระหนี้ (First Payment Default) งานวิจัยนี้ได้ดำเนินการตามขั้นตอนที่เป็นมาตรฐานโดยแบ่งเป็น 3 ระดับหลัก ได้แก่

1) การประเมินด้วย Cross-Validation (CV) ในชุดข้อมูลฝึกฝน (Training Set) ถูกแบ่งออกเป็น 5 ส่วนเท่าๆ กัน (K-Fold Cross Validation) เพื่อให้โมเดลได้รับการฝึกฝนและทดสอบบนชุดข้อมูลที่แตกต่างกันในแต่ละรอบ ช่วยลดปัญหา Overfitting และเพิ่มความน่าเชื่อถือของผลลัพธ์ ตัวชี้วัดหลักที่ใช้ประเมินแต่ละ Fold ได้แก่ ค่า ROC-AUC และค่า KS (Kolmogorov–Smirnov Statistic) รวมถึงการสรุปผลโดยเฉลี่ย (Mean) ของแต่ละตัวชี้วัดตลอดทุก Fold ตัวชี้วัดนี้สอดคล้องกับปัญหาตามที่ได้ทบทวนวรรณกรรม และเป็นตัวชี้วัดหลักในการคัดเลือกโมเดลที่เหมาะสมที่สุด

2) การทดสอบบนชุดข้อมูลทดสอบ (Test Set) หลังจากคัดเลือกพารามิเตอร์ที่ดีที่สุดจากการ Cross-Validation โมเดลแต่ละแบบจะถูกนำไปฝึกฝนใหม่บนชุดข้อมูลฝึกฝนทั้งหมด และทดสอบบนชุดข้อมูลทดสอบที่กันไว้ (Hold-Out) ซึ่งไม่เคยใช้ในการฝึกฝนหรือปรับพารามิเตอร์มาก่อน ผลลัพธ์จะถูกประเมินด้วยตัวชี้วัดเดียวกันกับที่ใช้ใน Cross-Validation เพื่อเปรียบเทียบและยืนยันประสิทธิภาพของโมเดลบนข้อมูลจริง

3) การใช้ Confusion Matrix และการวิเคราะห์เชิงลึก เพื่อทำความเข้าใจผลลัพธ์ของโมเดลอย่างรอบด้าน งานวิจัยนี้ได้นำเสนอ Confusion Matrix ทั้งในระดับ Cross-Validation และข้อมูลทดสอบ (Test Set) ช่วยให้เห็นสัดส่วนของกลุ่มที่ถูกและผิดแยกได้ชัดเจน เพื่อประกอบการตัดสินใจเลือกโมเดลที่เหมาะสมที่สุด

การประเมินดังกล่าวทำให้มั่นใจได้ว่าโมเดลที่เลือกมีประสิทธิภาพในการแยกแยะกลุ่มลูกค้าที่มีความเสี่ยงผิดนัดได้อย่างถูกต้องแม่นยำ และผลของโมเดลที่ดีที่สุดยังสามารถนำไปใช้งานจริงในการบริหารความเสี่ยงสินเชื่อ เช่น Credit Scoring หรือ FPD Score ต่อไป

3.6 การนำโมเดลไปใช้ (Deployment)

ผลของโมเดลที่ผ่านการประเมินสามารถนำไปใช้งานจริงในการบริหารความเสี่ยงสินเชื่อ เช่น Credit Scoring หรือ FPD Score ยังสามารถนำไปพัฒนาเป็นระบบ Credit Scoring System หรือ API (Application Programming Interface) เป้าหมายในหัวข้อนี้คือเปลี่ยนผลลัพธ์การศึกษาให้ใช้งานได้จริงในกระบวนการธุรกิจและช่วยลดความเสี่ยงและยกระดับคุณภาพสินเชื่อได้อย่างมีประสิทธิภาพ และยังสามารถเชื่อมต่อ (Integrate) Pipeline ที่พัฒนาไว้กับ Application ผ่าน API หรือระบบภายในองค์กร (เช่น ระบบอนุมัติสินเชื่อ Mobile App, Web Portal, Core Banking) ได้โดยตรงในหลายรูปแบบ ขึ้นกับโครงสร้างและความพร้อมของ IT Infrastructure

```python
proba = xgb_best.predict_proba(X_test)[:, 1]  # ได้ Probability FPD = 1

# สูตรคำนวณ Score (เทียบเคียงแบบแบงก์นิยม)
def probability_to_score(prob, A=600, B=174.6):
    odds = np.clip(prob/(1-prob), 1e-6, 1e6)
    score = A - B * np.log(odds)
    # Optional: Limit score in range
    score = np.clip(score, 300, 850)
    return score

# คำนวณ Credit Score
scores = probability_to_score(proba)
df_score = X_test.copy()
df_score['FPD_Prob'] = proba
df_score['CreditScore'] = scores

# === 4.5.2 การกำหนดช่วงคะแนนและนโยบายการอนุมัติ ===

# กำหนดช่วงคะแนนและกลุ่มความเสี่ยง (Cutoff สามารถปรับได้ตามโจทย์ business)
score_bins = [0, 599, 649, 699, 749, 850]
risk_labels = ['Very High', 'High Risk', 'Moderate', 'Good', 'Excellent']
df_score['RiskGroup'] = pd.cut(df_score['CreditScore'], bins=score_bins, labels=risk_labels, right=True)

# Mapping นโยบายตัวอย่าง
approval_policy = {
    'Excellent': 'อนุมัติทันที',
    'Good': 'อนุมัติ พร้อมตรวจสอบเบื้องต้น',
    'Moderate': 'ขอเอกสารเพิ่ม/วงเงินจำกัด',
    'High Risk': 'พิจารณาเข้มงวด',
    'Very High': 'ปฏิเสธ/อนุมัติแบบพิเศษ'
}
df_score['Policy'] = df_score['RiskGroup'].map(approval_policy)
```

รูปที่ 3.7 สร้างคะแนนความเสี่ยงจาก Best Model

จากรูปที่ 3.7 การแปลงค่าความน่าจะเป็น (Prop หรือ Probability) จากโมเดล Machine Learning เป็น คะแนนมาตรฐาน (Credit Score) ที่ธนาคารใช้กันทั่วไป สำหรับนำไปปรับใช้ตามนโยบายของธนาคาร

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในบทนี้ผู้วิจัยจะกล่าวถึงผลการวิเคราะห์การทำนายการผิดนัดชำระหนี้งวดแรกในสินเชื่อธุรกิจขนาดเล็กโดยใช้เทคนิคต่างๆจากการทบทวนวรรณกรรมจึงได้เลือกเทคนิคดังนี้ เทคนิคการจัดการข้อมูลที่ไม่สมดุล 3 เทคนิค ได้แก่เทคนิค SMOTE (Synthetic Minority Over-Sampling) และSMOTEENN (SMOTE Edited Nearest Neighbors) ร่วมกับเทคนิคการเรียนรู้ของเครื่อง 3 วิธี ประกอบไปด้วยอัลกอรึทึม การถดถอยโลจิสติก (Logistic Regression), การสุ่มป่าไม้ (Random Forest), XGBoost (Extreme Gradient Boosting)

ข้อมูลที่ใช้พัฒนาโมเดลเป็นข้อมูลรายการยื่นคำขอกู้สินเชื่อตั้งแต่วันที่ 1 กันยายน พ.ศ. 2562 ถึงวันที่ 1 มีนาคม พ.ศ. 2567 โดยมีข้อมูลทั้งหมด 3,013 บัญชี ที่ผ่านการแปลงตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณ มีการแบ่งข้อมูลเป็นชุดข้อมูลฝึกฝน (Training Set) ร้อยละ80 ของข้อมูลทั้งหมดและอีกร้อยละ 20 จะเป็นชุดข้อมูลสำหรับทดสอบ (Testing Set) มีการดำเนินตรวจสอบความถูกต้อง (Model Validation) ด้วยวิธี K-Fold Cross-Validation (k=5) และการประเมินประสิทธิภาพโมเดลจะพิจารณาจากค่าพื้นที่ใต้โค้ง ROC-AUC (Area Under the Receiver Operating Characteristic Curve) ที่มีค่าที่สูงสุด และพิจารณาร่วมกับค่า KS (Kolmogorov-Smirnov) วัดระยะห่างสูงสุดระหว่าง Distribution ของกลุ่มดีกับกลุ่มเสี่ยงตามเกณฑ์มาตรฐานสำหรับโมเดล Credit Scoring นอกจากนี้งานวิจัยนี้ได้ดำเนินการวิเคราะห์ความสำคัญของตัวแปรจากโมเดลที่มีการประเมินว่ามีประสิทธิภาพที่ดีที่สุดจากการวัดผลจากตัวชี้วัดที่ระบุไว้ เพื่อระบุปัจจัยที่สำคัญที่ส่งผลต่อการผัดนัดชำระหนี้งวดแรก และแปลผลจากการทดลองเพื่อเป็นแนวทางนำไปใช้จริงกับธุรกิจ โดยในบทนี้จะประกอบไปด้วยหัวข้อหลักดังนี้

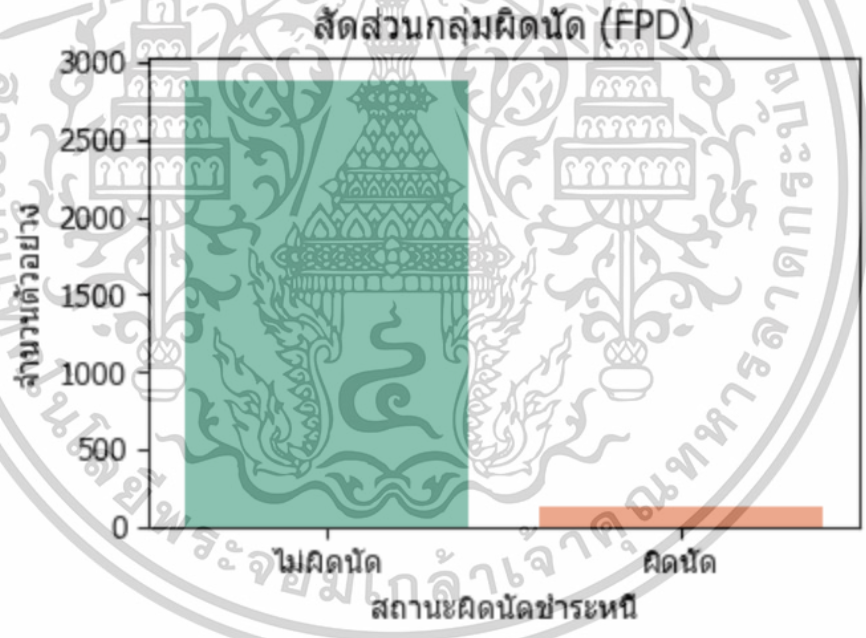
1. สถิติเชิงพรรณนา (Descriptive Statistics)
2. ผลเปรียบเทียบประสิทธิภาพโมเดลแต่ละเทคนิคจัดการข้อมูลไม่สมดุล (Sampling Technique)
3. ผลการประเมินโมเดล (Model Evaluation)
4. การวิเคราะห์ความสำคัญของตัวแปร (Feature Importance)
5. การประยุกต์ใช้โมเดลเพื่อการสร้างคะแนนความเสี่ยง (FPD Score)
6. อภิปรายผล

4.1 สถิติเชิงพรรณนา (Descriptive Statistics)

การศึกษานี้ดำเนินการโดยใช้ชุดข้อมูลจากสถาบันการเงินแห่งหนึ่งในประเทศไทย ซึ่งได้ทำการคัดเลือกมาจากฐานข้อมูล โดยเป็นข้อมูลรายการยื่นคำขอกู้สินเชื่อตั้งแต่วันที่ 1 กันยายน พ.ศ. 2562 ถึงวันที่ 1 มีนาคม พ.ศ. 2567 โดยมีข้อมูลทั้งหมด 3,013 บัญชี โดยมีรายละเอียดตามหัวข้อย่อยดังต่อไปนี้

4.1.1 การสำรวจตัวแปรตาม (Target Variable)

ในงานวิจัยที่มุ่งเน้นการพัฒนาโมเดลเพื่อทำนายโอกาสผิดนัดชำระหนี้งวดแรกของลูกค้าสินเชื่อธุรกิจขนาดเล็ก ในการศึกษานี้ ตัวแปรตามที่ใช้ คือ “FPD” (First Payment Default) ซึ่งแบ่งออกเป็น 2 สถานะ ได้แก่ 0 หมายถึง “ไม่ผิดนัดชำระ” และ 1 หมายถึง “ผิดนัดชำระในงวดแรกที่ถึงกำหนดเกิน 30 วัน” เพื่อให้เห็นภาพรวมของสถานะการผิดนัดชำระหนี้งวดแรกในกลุ่มตัวอย่าง ผู้วิจัยได้สรุปข้อมูลการกระจายตัวของตัวแปรตาม “FPD” (First Payment Default) ซึ่งแสดงผลในรูปที่ 4.1 ดังนี้



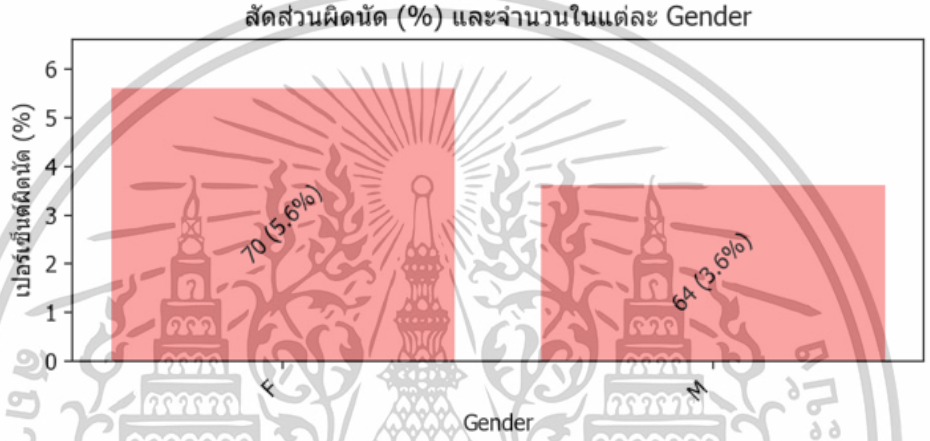
รูปที่ 4.1 สัดส่วนกลุ่มผิดนัดชำระหนี้งวดแรก (FPD) และไม่ผิดนัด

จากกราฟจะเห็นได้อย่างชัดเจนว่า กลุ่มที่ไม่ผิดนัดชำระหนี้งวดแรก (Non-FPD) มีจำนวนมากถึง 2,879 ราย ขณะที่ กลุ่มที่ผิดนัดชำระหนี้งวดแรก (FPD) มีเพียง 134 ราย คิดเป็นสัดส่วนประมาณ 95.55% ต่อ 4.45% ตามลำดับ สะท้อนถึง ลักษณะปัญหาข้อมูลไม่สมดุล (Imbalanced Data) ซึ่งเป็นปัญหาสำคัญในงานวิจัยนี้ เนื่องจากโมเดล Machine Learning ทั่วไปมักจะมีแนวโน้มทำนาย “กลุ่มใหญ่” (Non-FPD) ได้ดี แต่ไม่สามารถตรวจจับ “กลุ่มผิดนัด” (FPD) ซึ่งเป็นกลุ่มเป้าหมายสำคัญได้อย่างแม่นยำ ดังนั้น จึงมีความจำเป็นต้องใช้เทคนิคการจัดการข้อมูลไม่สมดุล

เช่น SMOTE, Random Undersampling เป็นต้น ในขั้นตอนต่อไปของการศึกษา เพื่อช่วยเพิ่มความแม่นยำในการทำนายกลุ่มที่มีความเสี่ยงผิดนัดชำระหนี้งวดแรก

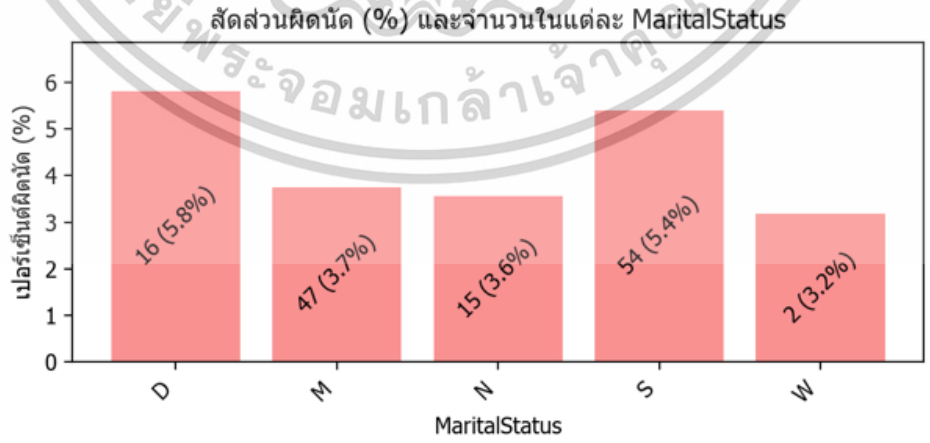
4.1.2 การสำรวจตัวแปรอิสระเชิงคุณภาพ

การสำรวจกลุ่มตัวแปรเชิงคุณภาพกับสถานะการผิดนัดชำระหนี้งวดแรก ในการวิเคราะห์ปัจจัยที่อาจมีผลต่อการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ของสินเชื่อธุรกิจขนาดเล็ก ได้ทำการสำรวจและเปรียบเทียบสัดส่วนและจำนวนลูกค้าที่ผิดนัดในแต่ละกลุ่มของตัวแปรเชิงคุณภาพหลักดังต่อไปนี้



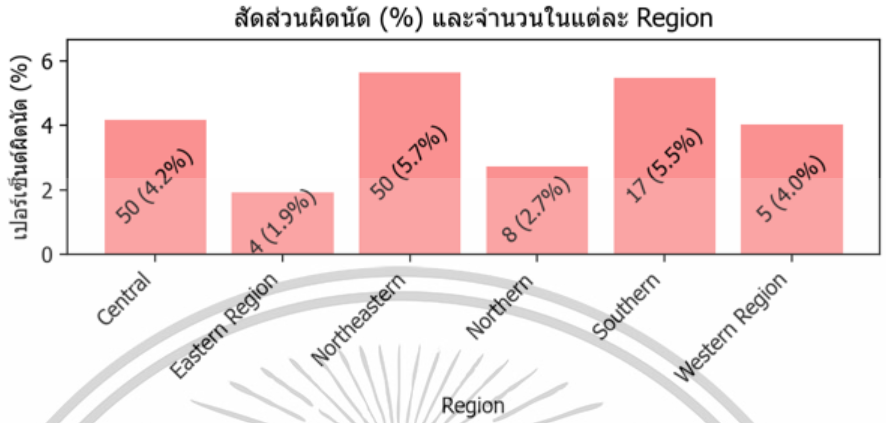
รูปที่ 4.2 สัดส่วนและจำนวนผู้ผิดนัดชำระหนี้งวดแรกจำแนกตามเพศ

จากรูปที่ 4.2 พบว่า เพศ (Gender) พบว่ากลุ่มลูกค้าหญิงมีสัดส่วนการผิดนัดสูงกว่ากลุ่มลูกค้าชาย โดยเพศหญิงมีอัตราผิดนัดร้อยละ 5.6 ในขณะที่เพศชายอยู่ที่ร้อยละ 3.6 สะท้อนให้เห็นถึงความแตกต่างของความเสี่ยงในแต่ละเพศ



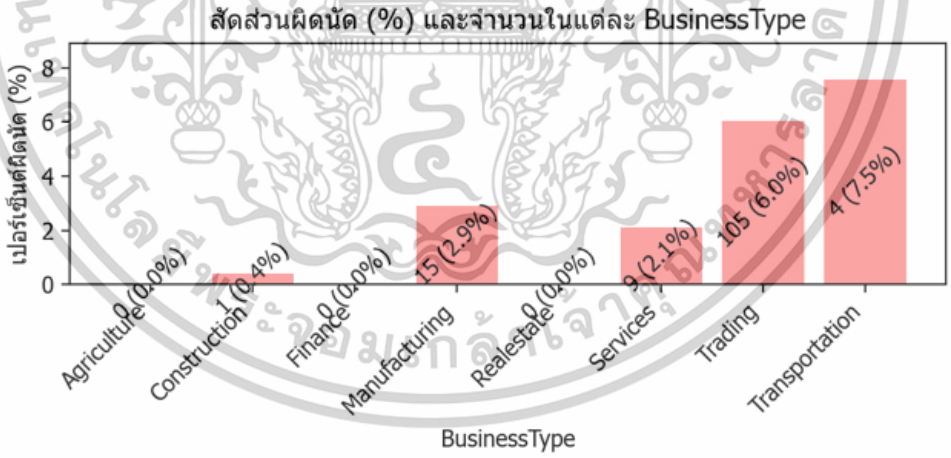
รูปที่ 4.3 สัดส่วนและจำนวนผู้ผิดนัดชำระหนี้งวดแรกจำแนกสถานภาพ

จากรูปที่ 4.3 พบว่า สถานภาพ (MaritalStatus) กลุ่มลูกค้าที่หย่าร้าง (D) และโสด (S) มีสัดส่วนการผิดนัดสูงกว่ากลุ่มสมรส (M) และกลุ่มอื่นๆ โดยกลุ่มหย่าร้างมีอัตราผิดนัดร้อยละ 5.8 และกลุ่มโสดร้อยละ 5.4 ในขณะที่กลุ่มสมรสมีสัดส่วนต่ำกว่า (ร้อยละ 3.7)



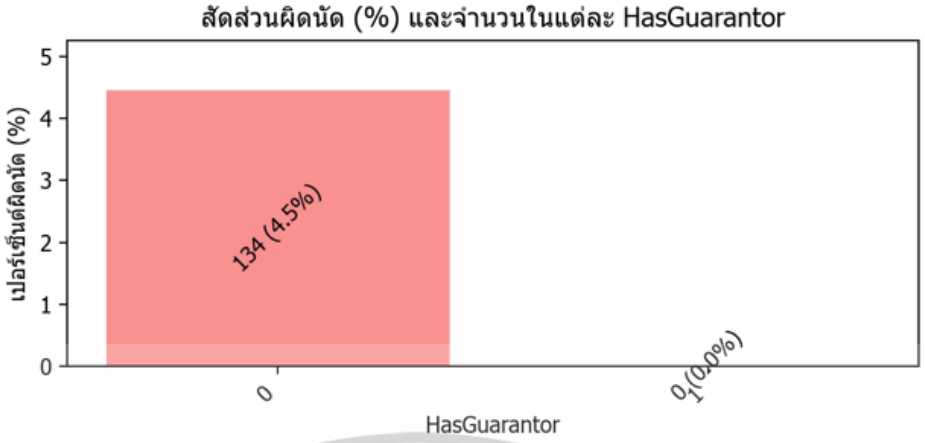
รูปที่ 4.4 สัดส่วนและจำนวนผู้ผิดนัดชำระหนี้งวดแรกจำแนกตามภูมิภาค

จากรูปที่ 4.4 วิเคราะห์แต่ละตัวแปรดังนี้ ภูมิภาค (Region) ภูมิภาคที่มีอัตราผิดนัดสูง ได้แก่ ภาคตะวันออกเฉียงเหนือ (Northeastern) และภาคใต้ (Southern) ที่มีสัดส่วนผิดนัดที่ร้อยละ 5.7 และ 5.5 ตามลำดับ ขณะที่ภาคกลางและภูมิภาคอื่นมีอัตราผิดนัดต่ำกว่า



รูปที่ 4.5 สัดส่วนและจำนวนผู้ผิดนัดชำระหนี้งวดแรกจำแนกตามประเภทธุรกิจ

จากรูปที่ 4.5 ในส่วนตัวแปรประเภทธุรกิจ (BusinessType) ลูกค้าที่ประกอบธุรกิจประเภทขนส่ง (Transportation) และค้าขาย (Trading) มีอัตราการผิดนัดสูงสุดที่ร้อยละ (7.5 และ 6.0 ตามลำดับ) เมื่อเทียบกับธุรกิจประเภทอื่น เช่น ผลิต การเกษตร บริการ หรืออสังหาริมทรัพย์ ซึ่งมีอัตราผิดนัดต่ำมากหรือเป็นศูนย์ในบางกลุ่ม



รูปที่ 4.6 สัดส่วนและจำนวนผู้ผิดนัดชำระหนี้งวดแรกจำแนกตามการมีผู้ค้ำประกัน

จากรูปที่ 4.6 พบว่า กลุ่มที่ไม่มีผู้ค้ำประกัน (0) มีสัดส่วนการผิดนัดสูงที่ร้อยละ 4.5 ขณะที่กลุ่มที่มีผู้ค้ำประกัน (1) แทบไม่พบการผิดนัดเลย สะท้อนถึงบทบาทของผู้ค้ำประกันในการลดความเสี่ยงผิดนัด

4.1.3 การสำรวจตัวแปรอิสระเชิงปริมาณ

ในการศึกษานี้มีการพิจารณาตัวแปรเชิงปริมาณหลากหลายด้าน ทั้งข้อมูลประชากรศาสตร์ ข้อมูลทางการเงิน ข้อมูลธุรกิจ และพฤติกรรมสินเชื่อ โดยสถิติพรรณนาที่นำเสนอประกอบด้วย ค่าเฉลี่ย ค่าต่ำสุด และค่าสูงสุด เพื่อสะท้อนถึงช่วงการกระจายของแต่ละตัวแปร ดังตาราง 4.1

ตารางที่ 4.1 สถิติพรรณนาของตัวแปรเชิงปริมาณ

ตัวแปร	ค่าเฉลี่ย	ค่าต่ำสุด	ค่าสูงสุด
CustAge	44.1900	23.0000	80.0000
AppliedAmount	8,508,000.7600	100,000.0000	30,750,000.0000
TotalCreditAmount	6,982,783.3100	100,000.0000	31,614,000.0000
LTV	63.1200	0.0000	300.0000
Tenor	122.8800	36.0000	240.0000
AppraisalAmount	2,586,169.0700	0.0000	82,730,000.0000
BusinessExpYear	7.9900	0.0000	39.0000
SalesAmt	8,436,785.4400	24,000.0000	373,625,833.0000
TotalIncome	713,142.0200	11,220.0000	27,189,874.8500
Installment	102,770.7600	1,400.0000	462,800.0000
TotalDebt	317,731.8300	7,100.0000	2,838,285.3200
DSCR	2.3400	1.0100	50.8900
TTRec	17.1100	0.0000	239.0000
ActRecNo	11.9300	0.0000	91.0000

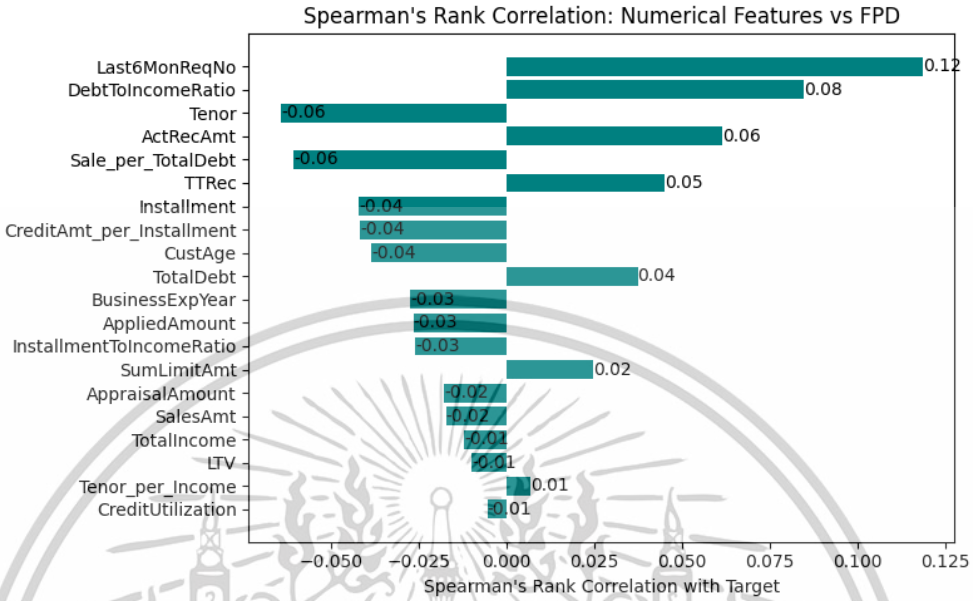
ตารางที่ 4.1 (ต่อ) สถิติพรรณนาของตัวแปรเชิงปริมาณ

ตัวแปร	ค่าเฉลี่ย	ค่าต่ำสุด	ค่าสูงสุด
ActRecAmt	11,253,141.5000	0.0000	312,424,512.0000
Last6MonReqNo	3.3300	0.0000	62.0000
SumLimitAmt	41,649,951.1200	0.0000	2,906,742,934.0000
Sale_per_TotalDebt	26.1400	1.1100	957.3200
CreditAmt_per_Installment	67.7200	29.8600	109.3900
DebtToIncomeRatio	0.5900	0.0200	0.9900
InstallmentToIncomeRatio	0.2100	0.0000	0.8900
CreditUtilization	0.6800	0.0000	37.0400
Tenor_per_Income	0.0000	0.0000	0.0200

จากตารางที่ 4.1 จากตารางแสดงค่าสถิติเชิงพรรณนาเบื้องต้นของข้อมูลลูกค้าสินเชื่อธุรกิจขนาดเล็ก พบว่า อายุ (CustAge) ในกลุ่มตัวอย่างมีอายุเฉลี่ย 44 ปี (ต่ำสุด 23 ปี สูงสุด 80 ปี) สะท้อนว่าผู้กู้ส่วนใหญ่เป็นผู้ใหญ่หรือวัยทำงาน ยอดเงินที่ยื่นกู้ (AppliedAmount) มีค่าเฉลี่ยประมาณ 8.5 ล้านบาท โดยกลุ่มสูงสุดขอยื่นถึง 30.75 ล้านบาท แสดงถึงความหลากหลายของขนาดกิจการที่เข้าถึงสินเชื่อ ระยะเวลากู้ (Tenor)ส่วนใหญ่มีระยะเวลากู้เฉลี่ย 122 เดือน หรือราว 10 ปี สะท้อนถึงความต้องการเงินทุนระยะยาวของธุรกิจ LTV (อัตราส่วนสินเชื่อต่อมูลค่าหลักประกัน) มีค่าเฉลี่ยอยู่ที่ 63% บ่งชี้ว่าส่วนใหญ่มีหลักประกันเพียงพอต่อวงเงินกู้ รายได้รวม (TotalIncome) เฉลี่ย 713,142 บาทต่อปี ขณะที่รายได้ต่ำสุดเพียง 11,220 บาท แสดงถึงความแตกต่างทางศักยภาพการเงินของผู้กู้แต่ละราย ภาระหนี้ต่อรายได้ (DebtToIncomeRatio) มีค่าเฉลี่ย 0.59 หรือประมาณ 59% บ่งชี้ถึงภาระหนี้ที่ค่อนข้างสูงในบางกลุ่ม ประสบการณ์ธุรกิจ (BusinessExpYear) ค่าเฉลี่ย 7.99 ปี สะท้อนว่าผู้ขอสินเชื่อส่วนใหญ่มีประสบการณ์พอสมควร แต่ยังมีบางรายเพิ่งเริ่มต้น (ต่ำสุด 0 ปี) และหนี้สินรวม (TotalDebt) เฉลี่ย 317,731 บาท ขณะที่สูงสุดถึง 2.83 ล้านบาท พบ Insight สำคัญ บางตัวแปรเช่น SalesAmt, SumLimitAmt, AppraisalAmount มีค่าต่ำสุดเป็นศูนย์ อาจบ่งชี้ข้อมูลที่ขาดหายไปหรือค่าผิดปกติหรือกิจการขนาดเล็กมาก จำเป็นต้องพิจารณาแยกกลุ่มวิเคราะห์เพิ่มเติม จากการสำรวจข้อมูล พบว่าผู้กู้มีลักษณะทางประชากรและการเงินที่หลากหลาย ตั้งแต่ผู้ที่มีรายได้และยอดเงินกู้ต่ำ ไปจนถึงผู้ที่มีกำลังทรัพย์สูง ทั้งนี้ภาระหนี้ต่อรายได้ที่อยู่ในระดับสูงและการขาดหลักประกัน อาจเป็นปัจจัยเสี่ยงที่ควรให้ความสำคัญในการวิเคราะห์หาสาเหตุของ FPD ต่อไป

เพื่อประเมินความสัมพันธ์ระหว่างตัวแปรอิสระเชิงตัวเลขแต่ละตัวกับโอกาสผิดนัดชำระหนี้ (First Payment Default: FPD) ผู้วิจัยได้ใช้สถิติ Spearman’s Rank Correlation ซึ่งเป็นสถิติที่เหมาะสมสำหรับการวัดความสัมพันธ์เชิงอันดับ (Rank) ระหว่างตัวแปร โดยไม่จำกัดอยู่แค่ความสัมพันธ์เชิงเส้น (linear relationship) เท่านั้น ทั้งนี้ Spearman’s Rank Correlation ยัง

สามารถประเมินความสัมพันธ์ระหว่างตัวแปรในกรณีที่ข้อมูลมีลักษณะเบ้ (skewed) หรือมี outlier ได้ดี ดังรูปที่ 4.7



รูปที่ 4.7 Spearman’s Rank Correlation ระหว่างตัวแปรอิสระเชิงตัวเลขกับตัวแปรตาม

จากรูปที่ 4.7 การประเมินความสัมพันธ์ระหว่างตัวแปรอิสระเชิงตัวเลขแต่ละตัวกับการผิดนัดชำระหนี้งวดแรก (FPD) ด้วยสถิติ Spearman’s Rank Correlation พบว่า ตัวแปรทั้งหมดในชุดข้อมูลมีค่าความสัมพันธ์กับ FPD อยู่ในระดับต่ำมาก โดยไม่มีตัวแปรใดที่มีค่าสัมประสิทธิ์ความสัมพันธ์เกิน 0.12 ทั้งในเชิงบวกและเชิงลบ ตัวแปรที่มีความสัมพันธ์กับ FPD ในทิศทางบวกมากที่สุด ได้แก่ Last6MonReqNo (จำนวนครั้งที่มีการขอสินเชื่อใน 6 เดือนล่าสุด) มีค่าสัมประสิทธิ์เท่ากับ 0.12 DebtToIncomeRatio (อัตราหนี้สินต่อรายได้) มีค่าสัมประสิทธิ์เท่ากับ 0.08 Tenor, ActRecAmt และ Sale_per_TotalDebt มีค่าสัมประสิทธิ์ประมาณ 0.06 ในส่วนตัวแปรอื่นๆ เช่น Installment, CreditAmt_per_Installment และ CustAge มีค่าสัมประสิทธิ์ความสัมพันธ์ใกล้ศูนย์มาก ทั้งนี้ ค่าสัมประสิทธิ์ที่ได้บ่งชี้ว่าความสัมพันธ์เชิงอันดับระหว่างตัวแปรเหล่านี้กับการผิดนัดชำระหนี้นั้นอยู่ในระดับต่ำ สะท้อนให้เห็นว่าการใช้ตัวแปรอิสระเชิงตัวเลขแต่ละตัวแปรเพียงลำพังอาจไม่สามารถอธิบายหรือพยากรณ์การผิดนัดชำระหนี้ได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม การใช้ตัวแปรหลายตัวร่วมกันในโมเดลอาจช่วยเพิ่มประสิทธิภาพในการทำนายได้ดียิ่งขึ้น

4.2 ผลเปรียบเทียบประสิทธิภาพโมเดลแต่ละเทคนิคจัดการข้อมูลไม่สมดุล (Sampling Technique)

เพื่อเปรียบเทียบความสามารถของโมเดลในการทำนายการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ได้ดำเนินการเปรียบเทียบประสิทธิภาพของโมเดล Machine Learning ที่

ได้รับความนิยมในงานวิจัยด้านข้อมูลไม่สมดุล ได้แก่ XGBoost Random Forest และ Logistic Regression โดยนำแต่ละโมเดลมาทดสอบร่วมกับเทคนิคจัดการข้อมูลไม่สมดุล 3 วิธี ได้แก่ SMOTE, SMOTEENN และ Random Undersampling การเปรียบเทียบดำเนินการบนชุดข้อมูลที่ผ่านการเตรียมตัวแปรอย่างเหมาะสมตามกระบวนการที่กำหนด จากนั้นแบ่งชุดข้อมูลออกเป็น2 ชุด คือข้อมูลฝึกฝนร้อยละ 80 และข้อมูลทดสอบร้อยละ20 พร้อมทั้งตรวจสอบความถูกต้องด้วย Cross-Validation 5 fold ผ่านการปรับจูนพารามิเตอร์ด้วยกริด และเลือกพารามิเตอร์ที่ดีที่สุดของทุกโมเดล ทั้งนี้ผลการเปรียบเทียบใช้ตัวชี้วัดหลัก ได้แก่ ค่าพื้นที่ใต้กราฟ ROC (ROC-AUC) และค่า KS ร่วมกับค่าความไว (Recall) ค่า Precision และ F1-Score ที่ได้จาก Cross-Validation (CV) บน ข้อมูลชุดฝึกฝน (Train Set) และข้อมูลชุดทดสอบ (Test Set) รายละเอียดการเปรียบเทียบผลลัพธ์ของแต่ละเทคนิค นำเสนอในหัวข้อย่อยถัดไป

4.2.1 ผลของเทคนิค SMOTE ร่วมกับ Logistic Regression, Random Forest และ XGBoost

ตารางที่แสดงด้านล่างนี้สรุปผลการประเมินประสิทธิภาพของโมเดล Machine Learning แต่ละเทคนิคได้แก่ Logistic Regression, Random Forest และ XGBoost ที่เทรนโดยใช้เทคนิคจัดการข้อมูลไม่สมดุล SMOTE บนข้อมูลสินเชื่อธุรกิจขนาดเล็ก เพื่อทำนายการผิดนัดชำระหนี้งวดแรก (FPD) แสดงผลในตาราง Confusion Matrix และเปรียบเทียบค่า Cross-Validation (CV) บนข้อมูลชุดฝึกฝน (Train Set) และข้อมูลชุดทดสอบ (Test Set) แสดงผลตามตัวชี้วัด ได้แก่ ROC-AUC, KS, Recall, Precision, และ F1-Score

ตารางที่ 4.2 เมทริกซ์ความสับสน (Confusion Matrix) ของโมเดลภายใต้เทคนิค SMOTE

Sampling	Model	Set	TN (%)	FP (%)	FN (%)	TP (%)
SMOTE	LR	Train	69.9925	30.0074	37.4458	62.5541
SMOTE	LR	Test	68.0555	31.9444	25.9259	74.0740
SMOTE	RF	Train	88.7969	11.2030	72.9870	27.0129
SMOTE	RF	Test	90.2777	9.7222	55.5555	44.4444
SMOTE	XGBoost	Train	94.6595	5.3404	74.7619	25.2381
SMOTE	XGBoost	Test	95.4861	4.5138	70.3703	29.6296

จากตารางที่ 4.2 ซึ่งแสดง Confusion Matrix ของโมเดล Logistic Regression (LR), Random Forest (RF) และ XGBoost หลังการปรับสมดุลข้อมูลด้วยเทคนิค SMOTE เมื่อเปรียบเทียบระหว่างข้อมูลชุดฝึกฝน (Train Set) และข้อมูลชุดทดสอบ (Test Set) สามารถสรุปได้ดังนี้

Logistic Regression (LR): ค่าความถูกต้องของการจำแนกกลุ่มปกติ (TN และ FP) บนชุดทดสอบ (Test Set) ใกล้เคียงกับชุดฝึกฝน (Train Set) ขณะที่ค่า TP บนชุดทดสอบสูงกว่าชุดฝึกฝนอย่างชัดเจน (ร้อยละ 74.07 เทียบกับร้อยละ 62.55) และค่า FN ลดลง แสดงว่าไม่มีสัญญาณของการเรียนรู้เกิน (Overfitting) อย่างชัดเจน ความแตกต่างนี้อาจเกิดจากความแปรปรวนของชุดข้อมูลหรือชุดทดสอบมีลักษณะเอื้อต่อการทำนายของโมเดล

Random Forest (RF): ค่า TN และ FP บนชุดทดสอบ (Test Set) สูงกว่าชุดฝึกฝน (Train Set) เล็กน้อย ขณะที่ค่า TP เพิ่มขึ้นอย่างมีนัยสำคัญ (ร้อยละ 44.44 เทียบกับร้อยละ 27.01) แสดงว่าโมเดลมีความสามารถในการจำแนกกลุ่มผิดนัดได้ดีขึ้นในชุดทดสอบ โดยไม่พบ Overfitting อย่างชัดเจน ซึ่งอาจบ่งชี้ถึงศักยภาพในการจำแนกที่ดีขึ้นของโมเดล หรือความง่ายของข้อมูลชุดทดสอบในการแยกกลุ่มผิดนัด

XGBoost: ผลการทำนายของโมเดลบนชุดทดสอบ (Test Set) และชุดฝึกฝน (Train Set) มีความใกล้เคียงกันในทุกค่าของ Confusion Matrix (TN, FP, FN, TP) แสดงถึงความเสถียรของโมเดลและไม่พบสัญญาณ Overfitting สะท้อนว่าโมเดล XGBoost สามารถเรียนรู้และทำนายได้อย่างสม่ำเสมอในข้อมูลทั้งสองชุด

ตารางที่ 4.3 เปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression, Random Forest และ XGBoost ภายใต้เทคนิค SMOTE

Sampling	Model	Set	ROC-AUC	KS	Recall	Precision	F1-Score
SMOTE	LR	Train	0.7154	0.4055	0.6255	0.0888	0.1555
SMOTE	LR	Test	0.7544	0.4485	0.7407	0.0980	0.1732
SMOTE	RF	Train	0.7235	0.4035	0.2701	0.1010	0.1461
SMOTE	RF	Test	0.7694	0.4097	0.4444	0.1765	0.2526
SMOTE	XGBoost	Train	0.7193	0.4076	0.2523	0.1775	0.2067
SMOTE	XGBoost	Test	**0.8142**	0.5012	0.2962	0.2353	0.2623

จากตารางที่ 4.3 แสดงการเปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression, Random Forest และ XGBoost โดยใช้เทคนิค SMOTE ในการจัดสมดุลข้อมูล และประเมินผลทั้งบนชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) ด้วยตัวชี้วัดหลัก คือ ROC-AUC และ KS พบว่าแต่ละโมเดลมีจุดเด่นและข้อจำกัดที่แตกต่างกัน โมเดล Logistic Regression ให้ค่า Recall สูงที่สุดที่ร้อยละ 74.07 บนชุดทดสอบ (Test Set) สะท้อนถึงความสามารถในการตรวจจับกลุ่มผู้ผิดนัดชำระหนี้ (FPD) ได้ดีอย่างไรก็ตาม ค่า ROC-AUC ร้อยละ 75.44 และ KS ร้อยละ 44.85 อยู่ในระดับรองลงมาจาก XGBoost ส่วนค่า Precision และ F1-Score ค่อนข้างต่ำ แสดงถึงอัตราการทำนายผิดว่าเป็น FPD ในกลุ่มลูกค้าปกติยังมีอยู่มาก ขณะที่ Random Forest มีค่า ROC-AUC ที่ร้อยละ 76.94

และ KS ร้อยละ 40.97 ใกล้เคียงกับ Logistic Regression แต่ค่า Recall และ F1-Score ต่ำกว่า ข้อดีของโมเดลนี้คือค่า Precision และ F1-Score สูงกว่า Logistic Regression เล็กน้อย แต่โดยรวมแล้วประสิทธิภาพการแยกกลุ่มผิดนัดยังไม่เด่นชัด และโมเดล XGBoost ให้ค่า ROC-AUC สูงที่สุดที่ร้อยละ 81.42 และ KS สูงที่สุดร้อยละ 50.12 ในกลุ่มโมเดลที่เปรียบเทียบกัน สะท้อนถึงศักยภาพในการแยกแยะกลุ่มผู้ผิดนัดและกลุ่มปกติได้ดีที่สุด แม้ว่าค่า Recall จะต่ำกว่าทั้งสองโมเดลก่อนหน้าที่ร้อยละ 29.62 แต่ Precision และ F1-Score กลับสูงสุดในกลุ่ม หมายความว่า โมเดลนี้หากทำนายว่าลูกค้าจะผิดนัด โอกาสที่จะทำนายถูกมีสูงกว่าโมเดลอื่น

สรุปการเลือกโมเดลที่ดีที่สุดในเทคนิคการจัดการข้อมูลที่ไม่สมดุล **SMOTE** เมื่อพิจารณาจากตัวชี้วัดหลัก ROC-AUC และ KS โมเดล **XGBoost** คือโมเดลที่มีประสิทธิภาพดีที่สุด ในการแยกกลุ่มระหว่างกลุ่มผู้กู้ที่ผิดนัดและไม่ผิดนัด

4.2.2 ผลของเทคนิค SMOTEENN ร่วมกับ Logistic Regression, Random Forest และ XGBoost

ตารางที่แสดงด้านล่างนี้สรุปผลการประเมินประสิทธิภาพของโมเดล Machine Learning แต่ละเทคนิคได้แก่ Logistic Regression, Random Forest และ XGBoost ที่เทรนโดยใช้เทคนิคจัดการข้อมูลไม่สมดุล SMOTEENN บนข้อมูลสินเชื่อธุรกิจขนาดเล็ก เพื่อทำนายการผิดนัดชำระหนี้งวดแรก (FPD) แสดงผลในตาราง Confusion Matrix และเปรียบเทียบค่า Cross-Validation (CV) บนชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) แสดงผลตามตัวชี้วัด ได้แก่ ROC-AUC, KS, Recall, Precision, และ F1-Score

ตารางที่ 4.4 เมทริกซ์ความสับสน (Confusion Matrix) ของโมเดลภายใต้เทคนิค SMOTEENN

Sampling	Model	Set	TN (%)	FP (%)	FN (%)	TP (%)
SMOTE	LR	Train	61.4384	38.5615	27.1428	72.8571
SMOTE	LR	Test	58.6805	41.3194	18.5185	81.4814
SMOTE	RF	Train	81.1533	18.8466	56.4069	43.5930
SMOTE	RF	Test	82.1180	17.8819	48.1481	51.8518
SMOTE	XGBoost	Train	91.3142	8.6857	74.0692	25.9307
SMOTE	XGBoost	Test	92.5347	7.4652	62.9629	37.0370

จากตารางที่ 4.4 ซึ่งแสดง Confusion Matrix ของโมเดล Logistic Regression (LR), Random Forest (RF) และ XGBoost หลังการปรับสมดุลข้อมูลด้วยเทคนิค SMOTEENN เมื่อเปรียบเทียบระหว่างชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) สามารถสรุปผลได้ดังนี้

Logistic Regression (LR): ชุดฝึกฝน (Train Set) ให้ค่า TP เท่ากับร้อยละ 72.86 และ FP เท่ากับร้อยละ 38.56 ขณะที่ชุดทดสอบ (Test Set) ให้ค่า TP เท่ากับร้อยละ 81.48 และ FP เท่ากับร้อยละ 41.32 สะท้อนว่าโมเดลมีแนวโน้มการทำนายในทิศทางเดียวกันทั้งสองชุด แม้ค่า FP จะเพิ่มขึ้นเล็กน้อยในชุดทดสอบ แต่ไม่พบสัญญาณ Overfitting อย่างชัดเจน ความผันผวนดังกล่าวอาจเกิดจากสัดส่วนกลุ่มผิดนัดที่แตกต่างกันเล็กน้อยระหว่างชุดข้อมูล

Random Forest (RF): ชุดฝึกฝน (Train Set) มีค่า TN เท่ากับร้อยละ 81.15 และ TP เท่ากับร้อยละ 43.59 ส่วนชุดทดสอบ (Test Set) มีค่า TN เท่ากับร้อยละ 82.12 และ TP เท่ากับร้อยละ 51.85 ผลลัพธ์แสดงให้เห็นว่าโมเดลมีความสามารถในการจำแนกได้ดีขึ้นบนชุดทดสอบ โดยไม่มีสัญญาณ Overfitting ค่าทั้งสองชุดมีแนวโน้มใกล้เคียงกัน แสดงถึงศักยภาพในการ Generalize ไปยังข้อมูลใหม่ได้อย่างเหมาะสม

XGBoost: ชุดฝึกฝน (Train Set) ให้ค่า TN เท่ากับร้อยละ 91.31 และ TP เท่ากับร้อยละ 25.93 ส่วนชุดทดสอบ (Test Set) มีค่า TN เท่ากับร้อยละ 92.53 และ TP เท่ากับร้อยละ 37.04 ค่า TP บนชุดทดสอบเพิ่มขึ้นเล็กน้อย ขณะที่ TN คงที่อยู่ในระดับสูง สะท้อนถึงความเสถียรของโมเดลในการทำนาย และไม่พบสัญญาณ Overfitting ที่ชัดเจน

สรุปโมเดลทั้งสามแสดงผลลัพธ์บนชุดฝึกฝนและชุดทดสอบที่มีแนวโน้มสอดคล้องกัน ไม่พบสัญญาณ Overfitting อย่างเด่นชัดในโมเดลใดๆ สะท้อนถึงความสามารถในการเรียนรู้ที่เหมาะสมและการประยุกต์ใช้เทคนิค SMOTEENN ที่มีประสิทธิภาพในการปรับสมดุลข้อมูล

ตารางที่ 4.5 เปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression, Random Forest และ XGBoost ภายใต้เทคนิค SMOTEENN

Sampling	Model	Set	ROC-AUC	KS	Recall	Precision	F1-Score
SMOTEENN	LR	Train	0.7129	0.3959	0.7285	0.0808	0.1455
SMOTEENN	LR	Test	0.7491	0.4589	0.8148	0.0846	0.1533
SMOTEENN	RF	Train	0.7355	0.4233	0.4359	0.0951	0.1556
SMOTEENN	RF	Test	**0.7556**	0.3796	0.5185	0.1196	0.1944
SMOTEENN	XGBoost	Train	0.7258	0.4362	0.2593	0.1158	0.1579
SMOTEENN	XGBoost	Test	0.7516	0.3813	0.3703	0.1886	0.2500

จากตารางที่ 4.5 จากผลการประเมินประสิทธิภาพโมเดลภายใต้เทคนิค SMOTEENN พบว่า Random Forest ให้ค่า ROC-AUC บนชุดทดสอบสูงที่สุดที่ ร้อยละ 75.56 รองลงมาคือ XGBoost ร้อยละ 75.16 และ Logistic Regression ร้อยละ 74.91 ซึ่งสะท้อนถึงความสามารถในการจำแนกกลุ่มลูกค้าที่ผิดนัดชำระหนี้กับกลุ่มปกติได้ดีในภาพรวม อย่างไรก็ตาม เมื่อพิจารณาตัวชี้วัด KS ซึ่งเป็นอีกหนึ่งตัวชี้วัดสำคัญ พบว่า Logistic Regression มีค่า KS สูงที่สุดที่ร้อยละ 45.89 สะท้อนศักยภาพ

ในการแยกแยะกลุ่มตัวอย่างได้ดีที่สุดในกลุ่มนี้ ขณะที่ Random Forest และ XGBoost มีค่า KS ใกล้เคียงกันที่ประมาณ ร้อยละ 38 ในประเด็น Recall ซึ่งแสดงถึงความสามารถในการจับกลุ่มลูกค้าที่ผิดนัด (Sensitivity) พบว่า Logistic Regression ยังคงโดดเด่นด้วยค่า Recall สูงถึงร้อยละ 81.48 ขณะที่ Random Forest และ XGBoost มี Recall ต่ำกว่าคือร้อยละ 51.85 และร้อยละ37.03 ตามลำดับ แต่ XGBoost ให้ค่า Precision และ F1-Score สูงสุดในกลุ่มนี้ สะท้อนความสมดุลของ Precision และ Recall ได้ดีกว่า เมื่อพิจารณาค่าชี้วัดบนชุดข้อมูลฝึกฝนและทดสอบ พบว่าทุกโมเดลไม่มีสัญญาณ Overfitting ค่าชี้วัดสำคัญบนชุดทดสอบและฝึกฝนอยู่ในระดับใกล้เคียงกัน แสดงให้เห็นถึงความสามารถในการ Generalize ของโมเดลบนข้อมูลจริง โดยสรุปจากการใช้ตัวชี้วัด ROC-AUC เป็นหลัก **Random Forest** ถือเป็นโมเดลที่มีประสิทธิภาพดีที่สุดภายใต้เทคนิค **SMOTEENN**

4.2.3 ผลของเทคนิค Random Undersampling ร่วมกับ Logistic Regression, Random Forest และ XGBoost

ตารางที่แสดงด้านล่างนี้สรุปผลการประเมินประสิทธิภาพของโมเดล Machine Learning แต่ละเทคนิคได้แก่ Logistic Regression, Random Forest และ XGBoost ที่เทรนโดยใช้เทคนิคจัดการข้อมูลไม่สมดุล Random Undersampling บนข้อมูลสินเชื่อธุรกิจขนาดเล็ก เพื่อทำนายการผิดนัดชำระหนี้งวดแรก (FPD) แสดงผลในตาราง Confusion Matrix และเปรียบเทียบค่า Cross-Validation (CV) บนชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) แสดงผลตามตัวชี้วัด ได้แก่ ROC-AUC, KS, Recall, Precision, และ F1-Score

ตารางที่ 4.6 เมทริกซ์ความสับสน (Confusion Matrix) ของโมเดลภายใต้เทคนิค Random Undersampling

Sampling	Model	Set	TN (%)	FP (%)	FN (%)	TP (%)
SMOTE	LR	Train	60.6579	39.3420	25.3679	74.6320
SMOTE	LR	Test	60.2430	39.7569	22.2222	77.7777
SMOTE	RF	Train	57.6605	42.3394	24.5021	75.4978
SMOTE	RF	Test	59.7222	40.2777	29.6296	70.3703
SMOTE	XGBoost	Train	59.8754	40.1245	32.8138	67.1861
SMOTE	XGBoost	Test	61.9791	38.0208	29.6296	70.3703

จากตารางที่ 4.6 ซึ่งแสดง Confusion Matrix ของโมเดล Logistic Regression (LR), Random Forest (RF) และ XGBoost หลังการปรับสมดุลข้อมูลด้วยเทคนิค Random Undersampling เมื่อเปรียบเทียบระหว่างชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) สามารถสรุปได้ดังนี้

Logistic Regression (LR): ผลลัพธ์ของโมเดลในชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) มีความใกล้เคียงกัน โดยค่าความแตกต่างของ TN และ FP ไม่เกินร้อยละ 1 ส่วนความแตกต่างของ FN และ TP อยู่ในช่วงประมาณร้อยละ 3 ค่า TP บนชุดทดสอบอยู่ที่ร้อยละ 77.78 สะท้อนถึงความสามารถในการทำนายกลุ่มผิดนัดได้ดี แม้ว่า FP จะยังอยู่ในระดับค่อนข้างสูง (ร้อยละ 39.76) อย่างไรก็ตาม ไม่พบสัญญาณ Overfitting อย่างชัดเจน

Random Forest (RF): ผลการทำนายของชุดฝึกฝนและชุดทดสอบใกล้เคียงกันในทุกค่าของ Confusion Matrix โดย TP บนชุดทดสอบอยู่ที่ร้อยละ 70.37 เทียบกับร้อยละ 75.50 บนชุดฝึกฝน สะท้อนถึงการลดลงเพียงเล็กน้อย ไม่แสดงอาการ Overfitting ชัดเจน โมเดล RF มีความสมดุลในการจำแนกทั้งกลุ่มปกติและกลุ่มผิดนัด แม้ค่า FP ยังคงสูงอยู่

XGBoost: ผลลัพธ์ระหว่างชุดฝึกฝนและชุดทดสอบมีแนวโน้มสอดคล้องกัน โดย TP บนชุดทดสอบสูงกว่าชุดฝึกฝนเล็กน้อย แสดงถึงศักยภาพของโมเดลในการ Generalize ไปยังข้อมูลใหม่ โดยเฉพาะในด้านการทำนายกลุ่มปกติ (TN) ที่มีความแม่นยำสูงกว่าโมเดลอื่นในกลุ่มเดียวกัน และไม่พบสัญญาณ Overfitting

สรุปโมเดลทั้งสามแสดงแนวโน้มผลการทำนายที่ใกล้เคียงกันระหว่างชุดฝึกฝน (Train Set) และชุดทดสอบ (Test Set) โดยไม่พบสัญญาณ Overfitting อย่างชัดเจน แสดงให้เห็นถึงความเสถียรของโมเดลภายใต้เทคนิคการสุ่มลดข้อมูล (Random Under Sampling)

ตารางที่ 4.7 เปรียบเทียบประสิทธิภาพของโมเดล Logistic Regression, Random Forest และ XGBoost ภายใต้เทคนิค Random Undersampling

Sampling	Model	Set	ROC-AUC	KS	Recall	Precision	F1-Score
RUS	LR	Train	0.6917	0.3697	0.7463	0.0807	0.1457
RUS	LR	Test	0.7116	0.4160	0.7777	0.0840	0.1516
RUS	RF	Train	0.7327	0.3904	0.7549	0.0766	0.1390
RUS	RF	Test	**0.7218**	0.3703	0.7037	0.0757	0.1366
RUS	XGBoost	Train	0.7061	0.3560	0.6718	0.0728	0.1313
RUS	XGBoost	Test	0.7083	0.4149	0.7037	0.0798	0.1433

จากตารางที่ 4.7 จากผลการประเมินประสิทธิภาพโมเดลภายใต้เทคนิค Random Undersampling (RUS) พบว่า Random Forest เป็นโมเดลที่ให้ค่า ROC-AUC สูงที่สุดในกลุ่ม ร้อยละ 72.18 แสดงถึงความสามารถในการจำแนกกลุ่มลูกค้าที่ผิดนัดและไม่ผิดนัดได้ดีในภาพรวม ขณะที่ Logistic Regression แม้จะมี ROC-AUC รองลงมา คือร้อยละ 71.16 แต่ให้ค่า KS สูงสุด (0.4160) สะท้อนศักยภาพในการแยกแยะกลุ่มตัวอย่างได้ดีที่สุดในกลุ่มนี้ ทั้งนี้ XGBoost ให้ค่า ROC-AUC และ KS ใกล้เคียงกับ Logistic Regression แต่ยังคงเป็นรองเล็กน้อย ในส่วนของ Recall พบว่า

Logistic Regression ให้ค่า Recall สูงสุดที่ร้อยละ 77.77 รองลงมาคือ Random Forest และ XGBoost ที่มีค่าเท่ากันที่ ร้อยละ 70.37 ขณะที่ Precision และ F1-Score ของทั้งสามโมเดลยังอยู่ในระดับต่ำ สะท้อนถึงความท้าทายในการจำแนกกลุ่มผิดนัดชำระหนี้จากข้อมูลที่ไม่สมดุล เมื่อเปรียบเทียบผลลัพธ์ระหว่างชุดฝึกฝนและชุดทดสอบ พบว่าค่าตัวชี้วัดหลักทั้ง ROC-AUC, KS, Recall และ Precision มีความใกล้เคียงกันในแต่ละโมเดล แสดงว่าไม่มีอาการ Overfitting อย่างชัดเจน โมเดลมีความสามารถในการ generalize ไปยังข้อมูลใหม่ได้ดี โดยสรุป จากตัวชี้วัด ROC-AUC เป็นหลัก **Random Forest** คือโมเดลที่เหมาะสมที่สุดภายใต้เทคนิค **Random Undersampling (RUS)** ในขณะที่ Logistic Regression เด่นด้าน KS และ Recall

4.3 ผลการประเมินโมเดล (Model Evaluation)

เพื่อประเมินประสิทธิภาพของโมเดลที่สร้างขึ้น ผู้วิจัยได้นำชุดข้อมูลทดสอบ (Test Set) มาใช้ในการทดสอบผลการทำนายการผิดนัดชำระหนี้งวดแรก (FPD) จากโมเดล Machine Learning ที่ถูกประเมินผลโมเดลว่าดีที่สุดภายใต้แต่ละเทคนิคการจัดการข้อมูลไม่สมดุล (Sampling Technique) ได้แก่ SMOTE, SMOTEENN และ Random Undersampling จากหัวข้อก่อนหน้า ผลการวัดประสิทธิภาพโมเดลนำเสนอผ่านตัวชี้วัดหลัก ได้แก่ ค่าพื้นที่ใต้กราฟ ROC (ROC-AUC) และค่า KS (Kolmogorov-Smirnov Statistic) คือสถิติที่ใช้วัดพลังในการแยกแยะของโมเดล Classification โดยเฉพาะในงานด้าน Credit Scoring และ Risk Model โดยพิจารณาจากจุดที่ค่า KS สูงสุด ซึ่งเป็นมาตรฐานที่นิยมมากในงาน Credit Scoring ทั่วโลก และพิจารณาค่า GINI ช่วยดูประสิทธิภาพโดยรวมให้เป็นไปตามเกณฑ์มาตรฐาน โดยที่ขึ้นไป KS มากกว่า 0.40 และ Gini มากกว่า 0.50 คือเกณฑ์ขั้นต่ำที่แนะนำ ในอุตสาหกรรมธนาคารพาณิชย์ไทยและเป็นแนวปฏิบัติในการจัดทำ Credit Scoring สำหรับนำไปประยุกต์ใช้จริง (ธนาคารแห่งประเทศไทย, 2566)

4.3.1 การประเมินประสิทธิภาพของโมเดลโดยใช้ตัวชี้วัด ROC-AUC, KS และ Gini

จากการประเมินโมเดลภายใต้แต่ละเทคนิคการจัดการข้อมูลไม่สมดุล ได้แก่ SMOTE, SMOTEENN และ Random Undersampling ผู้วิจัยได้คัดเลือกโมเดลที่มีผลการเรียนรู้ดีที่สุดจากแต่ละเทคนิคมาทดสอบกับชุดข้อมูลทดสอบ (Test Set) โดยใช้ตัวชี้วัดหลักในการประเมิน ได้แก่ ค่า ROC-AUC, ค่า KS และค่า Gini เพื่อสะท้อนความสามารถของโมเดลในการจำแนกกลุ่มลูกค้าที่มีความเสี่ยงต่อการผิดนัดชำระหนี้ได้อย่างมีประสิทธิภาพผลการประเมินดังกล่าวสรุปไว้ในตารางที่ 4.8

ตารางที่ 4.8 สรุปผลการประเมินประสิทธิภาพโมเดล Machine Learning ในการทำนายการผิดนัดชำระหนี้งวดแรก (FPD) บนชุดข้อมูลทดสอบ (Test Set)

Sampling	Model	ROC-AUC	KS	GINI
SMOTE	XGBoost	0.8142	0.5012	0.6283
SMOTEENN	Random Forest	0.7556	0.3796	0.5113
Random Undersampling	Random Forest	0.7218	0.3703	0.4436

จากตาราง 4.8 จากการเปรียบเทียบประสิทธิภาพของโมเดลทั้งสามจากการประเมินผลโมเดลที่ดีที่สุดใน ภายใต้เทคนิคจัดการข้อมูลไม่สมดุล SMOTE, SMOTEENN, Random Undersampling) โดยใช้ตัวชี้วัดหลักคือ ROC-AUC และ KS บนชุดข้อมูลทดสอบ (Test Set) พบว่าเทคนิค SMOTE ร่วมกับ XGBoost ให้ค่า ROC-AUC สูงสุดที่ร้อยละ 81.42 และ KS สูงสุดที่ร้อยละ 50.12 แสดงให้เห็นถึงความสามารถในการจำแนกกลุ่มลูกค้าที่เสี่ยงผิดนัดได้ดีที่สุดในกลุ่มนี้เทคนิค SMOTEENN Random Forest ให้ค่า ROC-AUC สูงสุดที่ร้อยละ 75.56 เทคนิค Random Undersampling (RUS) Random Forest ให้ค่า ROC-AUC สูงสุดที่ร้อยละ 72.18 ผลลัพธ์ของโมเดลทั้งสามมีค่าตัวชี้วัดไม่แตกต่างกันมาก ข้อสรุปพิจารณาเฉพาะ ค่าตัวชี้วัดหลักบน Test set **XGBoost ร่วมกับเทคนิค SMOTE** คือ โมเดลที่ดีที่สุด สำหรับข้อมูลนี้ ด้วยค่า ROC-AUC ที่สูงที่สุดที่ร้อยละ 81.42 และ KS ที่ร้อยละ 50.12

4.3.2 การทดสอบสมมติฐานด้วย McNemar's Test

ภายหลังจากการประเมินและเปรียบเทียบประสิทธิภาพของโมเดลทั้งสาม ได้แก่ Logistic Regression, Random Forest และ XGBoost ภายใต้เทคนิคจัดการข้อมูลไม่สมดุล (SMOTE, SMOTEENN, Random Undersampling) โดยใช้ตัวชี้วัดหลักคือ ROC-AUC และ KS บนชุดข้อมูลทดสอบ (Test set) ผลการทดลองพบว่า โมเดล SMOTE ร่วมกับ XGBoost มีค่าประสิทธิภาพสูงสุด อย่างไรก็ตาม เพื่อเสริมความน่าเชื่อถือของข้อค้นพบในเชิงสถิติ ผู้วิจัยจึงดำเนินการทดสอบสมมติฐานโดยใช้ McNemar's Test เพื่อวิเคราะห์ว่า โมเดล SMOTE ร่วมกับ XGBoost มีความแตกต่างจากโมเดลอื่นอย่างมีนัยสำคัญหรือไม่ การทดสอบ McNemar's Test เป็นเทคนิคที่เหมาะสำหรับการเปรียบเทียบผลการจำแนกประเภทของโมเดลสองแบบที่ทดสอบกับกลุ่มตัวอย่างเดียวกัน (Paired Comparison) โดยให้ความสำคัญกับกรณีที่โมเดลหนึ่งทำนายถูก ในขณะที่อีกโมเดลหนึ่งทำนายผิด ซึ่งจะวิเคราะห์จากตัวแปร b และ c ในตาราง 2×2 ตามนิยามและสมการ (2.15) ที่นำเสนอในบทที่ 2

การทดสอบกำหนดสมมติฐานดังนี้

H_0**:** ผลการทำนายของโมเดลทั้งสองไม่มีความแตกต่างกัน

H_1**:** ผลการทำนายของโมเดลทั้งสองมีความแตกต่างกัน

ตารางที่ 4.9 ผลการเปรียบเทียบระหว่าง SMOTE ร่วมกับ XGBoost กับโมเดลอื่นๆ

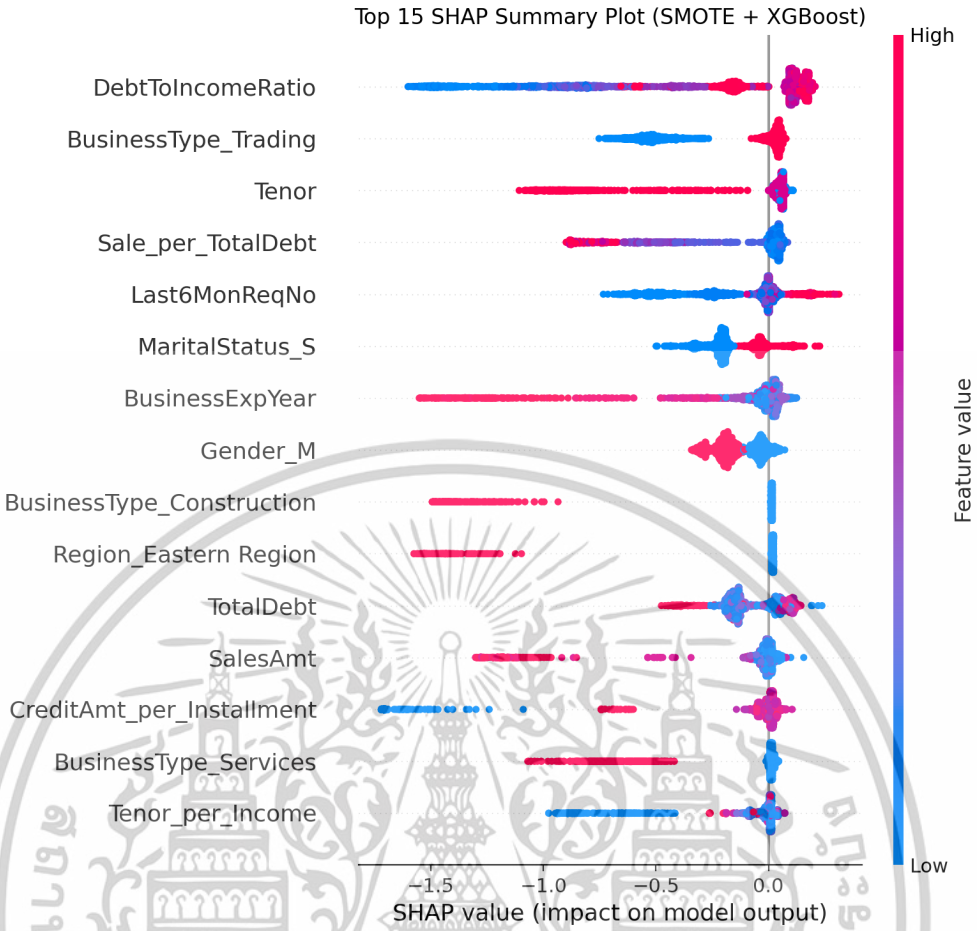
Model Pair	โมเดล A ผิด - B ถูก	โมเดล A ถูก - B ผิด	χ^2	p-value
SMOTE ร่วมกับ XGBoost vs SMOTEENN Random Forest	14	271	14.0	6.56×10^{-63}
SMOTE ร่วมกับ XGBoost vs RUS Random Forest	41	151	41.0	5.35×10^{-16}

จากตารางที่ 4.9 พบว่าทุกคู่เปรียบเทียบมีค่า p-value ต่ำกว่าระดับนัยสำคัญ 0.05 อย่างชัดเจน **จึงปฏิเสธ H_0** กล่าวคือ ผลการทำนายของโมเดล SMOTE ร่วมกับ XGBoost มีความแตกต่างจากโมเดลอื่นๆ อย่างมีนัยสำคัญทางสถิติ

จากการทดสอบ McNemar's Test ในหัวข้อนี้ช่วยส่งเสริมหลักฐานทางสถิติที่ชัดเจนว่า SMOTE ร่วมกับ XGBoost ให้ผลลัพธ์ที่แตกต่างและเหนือกว่าโมเดลอื่นๆ เมื่อใช้ข้อมูลเดียวกัน

4.4 ผลการวิเคราะห์ความสำคัญของตัวแปร (Feature Importance)

การวิเคราะห์ความสำคัญของตัวแปร (Feature Importance Analysis) เพื่อทำความเข้าใจและตีความปัจจัยที่ส่งผลต่อการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มผู้กู้สินเชื่อธุรกิจขนาดเล็ก งานวิจัยนี้ได้ทำการวิเคราะห์ความสำคัญของตัวแปร โดยอาศัยเทคนิค SHAP (SHapley Additive exPlanations) ซึ่งเป็นเทคนิคที่ได้รับความนิยมและมีความแม่นยำสูงในการอธิบายการตัดสินใจของโมเดล Machine Learning การวิเคราะห์ดังกล่าวช่วยให้สามารถระบุได้ว่าตัวแปรใดมีอิทธิพลต่อผลการทำนายของโมเดลมากที่สุด และแต่ละตัวแปรส่งผลต่อการเพิ่มหรือลดความเสี่ยงในการผิดนัดชำระหนี้อย่างไร โดยเนื้อหาต่อไปนี้จะนำเสนอผลการวิเคราะห์ความสำคัญของตัวแปรหลัก (Top Features) จากโมเดล XGBoost ที่ได้ประสิทธิภาพสูงสุดในงานวิจัยนี้ ซึ่งผ่านการปรับสมดุลข้อมูลและการปรับแต่งพารามิเตอร์อย่างเหมาะสม ดังแสดงในรูป



รูปที่ 4.7 ค่าอิทธิพลของตัวแปรด้วยวิธี SHAP ของโมเดล XGBoost ด้วยเทคนิค SMOTE

จากรูป 4.7 จะเห็นได้ว่า ตัวแปร DebtToIncomeRatio เมื่อมีค่ามาก (จุดสีแดง) จะส่งผลผลักดันต่อโมเดลให้ทำนายความเสี่ยงผิดนัดชำระหนี้สูงขึ้น (กระจายทางขวา)

สรุปภาพรวมจากการวิเคราะห์ค่า SHAP (SHapley Additive exPlanations) เพื่อประเมินอิทธิพลของแต่ละปัจจัยต่อการทำนายความเสี่ยงการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ของโมเดล SMOTE ร่วมกับ XGBoost พบว่า มีปัจจัยหลักที่ส่งผลต่อผลลัพธ์ของโมเดลดังนี้

DebtToIncomeRatio (อัตราส่วนหนี้สินต่อรายได้) ปัจจัยนี้สะท้อนถึงภาระหนี้ที่ผู้กู้ต้องชำระเมื่อเทียบกับรายได้ที่ได้รับ โดยค่าที่สูงบ่งชี้ถึงความเสี่ยงในการผิดนัดชำระหนี้มากขึ้น ซึ่งสอดคล้องกับแนวนโยบายการบริหารความเสี่ยงของสถาบันการเงินที่มุ่งจำกัดการปล่อยสินเชื่อให้กับลูกค้าที่มีความสามารถในการชำระหนี้เพียงพอ

BusinessType_Trading (ประเภทธุรกิจ: การค้า) ผู้กู้ที่ประกอบธุรกิจการค้ามีแนวโน้มที่จะมีความเสี่ยง FPD สูงกว่าธุรกิจประเภทอื่น ธนาคารสามารถใช้ข้อมูลนี้ในการพิจารณาออกแบบผลิตภัณฑ์สินเชื่อหรือกำหนดหลักเกณฑ์อนุมัติสินเชื่อให้เหมาะสมกับลักษณะความเสี่ยงของแต่ละประเภทธุรกิจ

Tenor (ระยะเวลากู้) ระยะเวลากู้ที่ยาวขึ้นมีความสัมพันธ์เชิงบวกกับโอกาสผิดนัดชำระหนี้ การกำหนดระยะเวลากู้ที่เหมาะสม หรือมีการทบทวนการปล่อยกู้ในกลุ่มที่ต้องการระยะเวลานาน อาจช่วยลดความเสี่ยง FPD ได้

Sale_per_TotalDebt (อัตราส่วนยอดขายต่อยอดหนี้) สะท้อนถึงประสิทธิภาพในการหมุนเวียนรายได้เพื่อชำระหนี้ ยิ่งอัตราส่วนนี้ต่ำ ความเสี่ยงที่จะผิดนัดชำระหนี้ยิ่งสูง ซึ่งสอดคล้องกับหลักการประเมินความสามารถในการชำระหนี้ของผู้กู้ในภาคธนาคาร

Last6MonReqNo (จำนวนการขอสินเชื่อใน 6 เดือนที่ผ่านมา) จำนวนการขอสินเชื่อที่สูงในช่วงเวลาดังกล่าว บ่งชี้ถึงความต้องการเงินทุนหรือสภาพคล่องที่อาจไม่มั่นคง อาจเป็นสัญญาณเตือนในการประเมินลูกค้า

MaritalStatus_S (สถานภาพสมรส: โสด) พบว่าผู้กู้ที่ยังไม่มีคู่สมรสมีความเสี่ยง FPD สูงกว่ากลุ่มที่สมรสแล้ว ซึ่งอาจสะท้อนถึงความมั่นคงทางรายได้หรือภาระผูกพันที่แตกต่างกัน

BusinessExpYear (ประสบการณ์ในการดำเนินธุรกิจ) ผู้กู้ที่มีประสบการณ์ดำเนินธุรกิจน้อยมีแนวโน้มเกิด FPD สูง สถาบันการเงินควรพิจารณาข้อมูลด้านประสบการณ์เป็นส่วนหนึ่งของเกณฑ์อนุมัติ

Gender_M (เพศชาย) ผลการวิเคราะห์ชี้ให้เห็นว่าเพศมีผลในระดับหนึ่งต่อความเสี่ยงผิดนัด

Region_Eastern Region (ภูมิภาคตะวันออก) ปัจจัยภูมิภาคสะท้อนถึงบริบทเศรษฐกิจท้องถิ่นที่อาจมีความเสี่ยงเฉพาะตัว

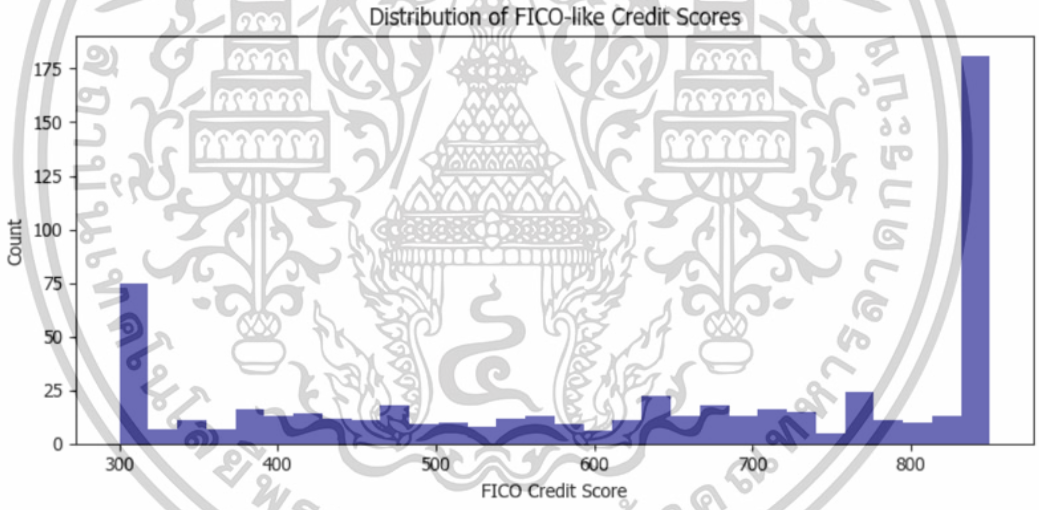
ปัจจัยอื่นๆ เช่น TotalDebt, SalesAmt, Tenor_per_Income, BusinessType Services, CreditAmt_per_Installment ปัจจัยเหล่านี้สะท้อนถึงโครงสร้างรายได้ ภาระหนี้ และรูปแบบธุรกิจ ซึ่งสามารถนำไปประกอบการวิเคราะห์ความเสี่ยงในระดับรายลูกค้า

สรุปโดยรวมการใช้ SHAP ร่วมกับ XGBoost ภายใต้เทคนิค SMOTE ช่วยให้อธิบายการตัดสินใจของโมเดลได้อย่างโปร่งใสและสามารถนำไปใช้จริงในการบริหารความเสี่ยงสินเชื่อของสถาบันการเงิน โดยการวิเคราะห์ปัจจัยสำคัญเชิงลึกสามารถตอบโจทย์การพัฒนากลยุทธ์สินเชื่อและสนับสนุนการตัดสินใจทางธุรกิจได้อย่างมีประสิทธิภาพ

4.6 การประยุกต์ใช้โมเดลเพื่อการสร้างคะแนนความเสี่ยง (FPD Score)

หลังจากที่ได้ดำเนินการพัฒนาและประเมินประสิทธิภาพของโมเดลทำนายการผิดนัดชำระหนี้งวดแรก (FPD) โดยใช้ข้อมูลและเทคนิคต่างๆ ที่เหมาะสม ขั้นตอนถัดไปคือการนำโมเดลที่ดีที่สุดมาประยุกต์ใช้ในเชิงธุรกิจ เพื่อสร้าง คะแนนความเสี่ยงผิดนัดชำระหนี้งวดแรก (FPD Score) สำหรับลูกค้ารายใหม่ พร้อมทั้งแนวทางกำหนดนโยบายจากคะแนนดังกล่าว เพื่อให้การตัดสินใจมีมาตรฐานและโปร่งใสมากยิ่งขึ้น

ในหัวข้อนี้ ได้นำโมเดลที่ผ่านการคัดเลือกและมีประสิทธิภาพสูงสุด (Best Model) มาสร้างคะแนนความเสี่ยง (FPD Score) ให้กับลูกค้าใหม่แต่ละราย โดยคะแนนดังกล่าวได้จากการแปลงค่าความน่าจะเป็น (Probability) ที่โมเดลคาดการณ์ได้ให้อยู่ในรูปแบบของ Credit Score ตามสูตรมาตรฐาน (Probability-to-Score Transformation) ซึ่งทำให้ธนาคารสามารถนำไปใช้เทียบเคียงหรือปรับใช้ร่วมกับเกณฑ์คะแนนที่ใช้อยู่เดิมได้อย่างสะดวก เนื่องจากการแปลผลคะแนนเพียงอย่างเดียวอาจยังไม่ตอบโจทย์การตัดสินใจเชิงธุรกิจ จึงได้กำหนดช่วงคะแนน (Score Bands) และกลุ่มความเสี่ยง (Risk Group) เพื่อแบ่งกลุ่มลูกค้าตามระดับความเสี่ยงที่แตกต่างกัน เช่น Excellent, Good, Moderate, High Risk, และ Very High พร้อมทั้งกำหนดนโยบายการอนุมัติสำหรับแต่ละกลุ่มอย่างชัดเจน (เช่น อนุมัติทันที, ตรวจสอบเพิ่ม, ขอเอกสารเพิ่ม, พิจารณาเข้มงวด, หรือปฏิเสธ) เพื่อช่วยให้การอนุมัติมีมาตรฐานและลดความเสี่ยงที่อาจเกิดขึ้น หัวข้อนี้นำเสนอการวิเคราะห์การกระจายตัวของ Credit Score และสัดส่วนของลูกค้าในแต่ละกลุ่มความเสี่ยง ซึ่งช่วยให้เห็นภาพรวมของคุณภาพลูกค้า ซึ่งสามารถใช้เป็นข้อมูลและแนวทางประกอบการปรับนโยบายหรือวางแผนกลยุทธ์ให้เหมาะกับแนวทางของแต่ละธุรกิจ



รูปที่ 4.8 การกระจายคะแนนเครดิตบนชุดข้อมูลทดสอบ

จากรูป 4.11 พบกลุ่มคะแนนกระจุกตัวที่ขอบทั้งสองด้าน (Bimodal Distribution) คะแนนส่วนหนึ่งมีค่าอยู่ใกล้ 300 (ต่ำสุด) และอีกกลุ่มมีค่าอยู่ใกล้ 850 (สูงสุด) อย่างชัดเจน กลุ่มที่มีคะแนน 850 มากที่สุด แสดงถึงกลุ่มลูกค้าที่โมเดลประเมินว่ามีความเสี่ยงต่ำมาก ในขณะที่กลุ่มที่มีคะแนนกระจุกตัวที่ 300 คือกลุ่มที่โมเดลประเมินว่ามี ความเสี่ยงสูงมาก (มีโอกาสผิดนัดสูง) คะแนนในช่วงกลางกระจายค่อนข้างบาง สะท้อนว่าข้อมูลของกลุ่มลูกค้าส่วนใหญ่ถูกโมเดลแบ่งความเสี่ยงอย่างชัดเจน (ชัดเจนว่าควรอนุมัติหรือไม่อนุมัติ)

ตารางที่ 4.10 การจัดกลุ่มคะแนนเครดิต (Credit Score) และการกำหนดนโยบายอนุมัติสินเชื่อจากข้อมูลชุดทดสอบ (Test Set)

ช่วงคะแนน (FICO Score)	กลุ่มความเสี่ยง (Risk Group)	นโยบาย	จำนวนผู้กู้	สัดส่วน (%)
800 – 850	Excellent	อนุมัติทันที	220	36.5%
740 – 799	Very Good	อนุมัติแบบเร่งรัด/ตรวจสอบน้อย	45	7.3%
670 – 739	Good	อนุมัติ พร้อมตรวจสอบเบื้องต้น	60	9.9%
580 – 669	Fair	ขอเอกสารเพิ่ม/จำกัดวงเงิน	55	9.1%
300 – 579	Poor	ปฏิเสธ/อนุมัติแบบพิเศษ	223	37.0%
รวม			**603**	**100%**

จากตารางที่ 4.10 ผลการ mapping คะแนน FICO-Like จากโมเดลกับกลุ่มความเสี่ยง พบว่า กลุ่มที่มีคะแนน “Poor” และ “Excellent” มีจำนวนสูงสุดในชุดข้อมูลทดสอบ ตามตารางข้างต้น ทั้งนี้ การแบ่งกลุ่มนี้สอดคล้องกับมาตรฐาน FICO และสามารถนำไปประยุกต์ใช้ในการกำหนดนโยบายการอนุมัติสินเชื่อเพื่อบริหารความเสี่ยง NPL การแจ้งเตือนต่างๆ ดังนั้นการนำโมเดลไปใช้งานจริง ธนาคารสามารถเลือกกำหนดนโยบายอนุมัติสินเชื่อได้ตามแต่ละกลุ่มความเสี่ยงให้เหมาะกับนโยบายของแต่ละหน่วยงานที่เกี่ยวข้อง

ข้อเสนอแนะสำหรับธุรกิจ ธนาคารควรนำคะแนนนี้ไปทดสอบกับฐานลูกค้าจริง และอาจพิจารณาปรับช่วงคะแนน Cut-Off ให้เหมาะสมกับสัดส่วนของลูกค้าในแต่ละกลุ่มความเสี่ยง สำหรับลูกค้าที่มีคะแนน 300 หรือ 850 อาจเป็นกลุ่มที่สามารถตัดสินใจอนุมัติหรือปฏิเสธได้อย่างมั่นใจ สำหรับกลุ่มที่มีคะแนนอยู่ช่วงกลาง (400-700) อาจพิจารณานโยบายเสริม เช่น ขอเอกสารเพิ่มเติม หรือกำหนดมาตรการบริหารความเสี่ยงพิเศษ

4.7 อภิปรายผล

จากผลการศึกษานี้ พบว่าโมเดลที่ผสานเทคนิค SMOTE กับ XGBoost ให้ประสิทธิภาพโดยรวมสูงที่สุดในการจำแนกความเสี่ยงการผิดนัดชำระหนี้งวดแรก (FPD) ในกลุ่มสินเชื่อธุรกิจขนาดเล็ก เมื่อพิจารณาจากตัวชี้วัดหลัก ได้แก่ ค่าพื้นที่ใต้โค้ง ROC (ROC-AUC) และค่า KS Statistic ซึ่งเป็นมาตรฐานที่ใช้กันอย่างแพร่หลายในการประเมินโมเดลด้านความเสี่ยงในสถาบันการเงิน (Hand & Henley, 1997; Brown & Mues, 2012; Moody’s Analytics, 2017) โมเดลดังกล่าวมีศักยภาพในการจำแนกกลุ่มลูกค้าที่มีความเสี่ยงสูงได้อย่างแม่นยำมากกว่าเทคนิคและโมเดลอื่น ทั้งในชุดข้อมูลฝึกฝนและทดสอบ และค่าประสิทธิภาพที่ได้ยังผ่านเกณฑ์มาตรฐาน ROC-AUC มากกว่าเท่ากับ 0.7

และ KS มากกว่าเท่ากับ 0.3 ซึ่งเป็นเกณฑ์ที่ธนาคารพาณิชย์และองค์กรกำกับดูแลในประเทศไทยใช้อ้างอิง (ธนาคารแห่งประเทศไทย, 2561; Moody's Analytics, 2017)

ในเชิงเปรียบเทียบ แม้โมเดล Random Forest ภายใต้เทคนิค SMOTEENN และ Random Undersampling จะให้ค่า Recall ที่สูงกว่าในบางกรณี แต่กลับมีค่า Precision ต่ำกว่า ขณะที่ค่าตัวชี้วัด ROC-AUC และ KS ยังคงต่ำกว่าโมเดล XGBoost อย่างมีนัยสำคัญ ทั้งนี้สอดคล้องกับผลการศึกษา Chen et al. (2021) ซึ่งระบุว่า SMOTE ร่วมกับ XGBoost สามารถเพิ่มความแม่นยำในการพยากรณ์ความเสี่ยงผิดนัดชำระหนี้ในข้อมูลที่มีความไม่สมดุลสูงได้ดีกว่าเทคนิคดั้งเดิม

เมื่อวิเคราะห์ความสำคัญของตัวแปรผ่านการใช้ SHAP พบว่า อัตราส่วนหนี้ต่อรายได้ (Debt-to-Income Ratio) เป็นปัจจัยที่มีอิทธิพลสูงสุดต่อการทำนายความเสี่ยง FPD รองลงมาคือประเภทธุรกิจ ระยะเวลากู้ จำนวนครั้งขอสินเชื่อย้อนหลัง สถานภาพสมรส และประสบการณ์ทางธุรกิจ ซึ่งปัจจัยเหล่านี้สอดคล้องกับทฤษฎีการให้สินเชื่อและงานวิจัยที่เกี่ยวข้อง (Berger & Udell, 2006) ที่ระบุถึงความสำคัญของข้อมูลทางเศรษฐกิจ สถานะทางสังคม และพฤติกรรมทางการเงินในการพยากรณ์ความเสี่ยงผิดนัด

การนำโมเดลที่ผ่านการคัดเลือกไปพัฒนาเป็นระบบ Credit Scoring โดยแปลงค่าความน่าจะเป็นเป็นคะแนน FICO-Like พบว่าสามารถแบ่งกลุ่มความเสี่ยงได้อย่างชัดเจนและนำไปสู่การกำหนดนโยบายอนุมัติสินเชื่อที่แตกต่างกันในแต่ละกลุ่ม ทั้งนี้ส่งผลให้สถาบันการเงินสามารถบริหารความเสี่ยง NPL ได้อย่างเหมาะสมและสอดคล้องกับกลยุทธ์องค์กร

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์หลักสองประการ ได้แก่ (1) ศึกษาปัจจัยสำคัญที่มีผลต่อการเกิดการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มลูกค้าสินเชื่อธุรกิจขนาดเล็ก (2) สร้างโมเดลที่เหมาะสมสำหรับลูกค้าที่มีโอกาสการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มสินเชื่อธุรกิจขนาดเล็กภายใต้สภาวะข้อมูลที่ไม่สมดุล

1) ปัจจัยสำคัญที่มีผลต่อการเกิดการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD)

การวิเคราะห์ความสำคัญของตัวแปรด้วยเทคนิค SHAP พบว่า “อัตราส่วนหนี้ต่อรายได้” (Debt-to-Income Ratio) เป็นปัจจัยที่มีอิทธิพลสูงสุดในการทำนายความเสี่ยงต่อการผิดชำระหนี้งวดแรก รองลงมาคือ ประเภทธุรกิจ ระยะเวลากู้ จำนวนครั้งขอสินเชื่อย้อนหลัง สถานภาพสมรส และประสบการณ์ทางธุรกิจ ผลดังกล่าวสอดคล้องกับทฤษฎีและงานวิจัยที่เกี่ยวข้องซึ่งชี้ให้เห็นว่าข้อมูลทางการเงิน ข้อมูลเชิงประชากร และพฤติกรรมการกู้ยืมมีบทบาทสำคัญในการพยากรณ์ความเสี่ยงผิดนัดชำระหนี้

2) สร้างโมเดลที่เหมาะสมสำหรับลูกค้าที่มีโอกาสการผิดนัดชำระหนี้งวดแรก (First Payment Default: FPD) ในกลุ่มสินเชื่อธุรกิจขนาดเล็กภายใต้สภาวะข้อมูลที่ไม่สมดุล

จากการวิเคราะห์ผลการทดสอบโมเดลพบว่า การจัดการข้อมูลไม่สมดุลด้วย **SMOTE** สามารถเพิ่มประสิทธิภาพของโมเดล **XGBoost** ได้อย่างมีนัยสำคัญ โดยสามารถแยกแยะกลุ่มลูกค้าที่มีความเสี่ยงสูงในการผิดนัดชำระหนี้งวดแรกได้แม่นยำมากกว่าเทคนิคอื่น ทั้งนี้ ผลการประเมินด้วยตัวชี้วัด ROC-AUC, KS และ GINI อยู่ในระดับที่สูงกว่ามาตรฐานขั้นต่ำที่สถาบันการเงินไทยยอมรับ (ROC-AUC มากกว่าร้อยละ 70, KS มากกว่าร้อยละ 40, GINI มากกว่าร้อยละ 40) (ธนาคารแห่งประเทศไทย, 2566; Moody’s Analytics, 2017)

การเลือกใช้โมเดล XGBoost ร่วมกับการปรับสมดุลข้อมูลด้วย SMOTE มีความเหมาะสมในเชิงปฏิบัติสำหรับการบริหารความเสี่ยงสินเชื่อธุรกิจขนาดเล็ก เนื่องจากสามารถนำไปต่อยอดเป็นระบบ Credit Scoring ที่ปรับระดับคะแนนความเสี่ยง (Risk Banding) ตามช่วงคะแนนแบบ FICO ซึ่งช่วยให้สถาบันการเงินสามารถบริหารความเสี่ยงได้อย่างมีประสิทธิภาพยิ่งขึ้น ในขณะเดียวกัน โมเดล Random Forest แม้จะให้ค่า Recall สูงในบางกรณี แต่มีค่า Precision ต่ำและมีประสิทธิภาพโดยรวม (ROC-AUC, KS) ต่ำกว่า XGBoost ซึ่งสอดคล้องกับข้อค้นพบของงานวิจัย

สากล อาทิ Chen et al. (2021) ที่ระบุว่าโมเดล XGBoost ผสานกับการจัดการข้อมูลไม่สมดุลมีศักยภาพสูงในการพยากรณ์ความเสี่ยงด้านสินเชื่อ

5.2 ข้อเสนอแนะ

จากผลการวิจัยที่ได้ดำเนินการเปรียบเทียบประสิทธิภาพของโมเดลการทำนายการผิดนัดชำระหนี้งวดแรกในสินเชื่อธุรกิจขนาดเล็ก ผู้วิจัยขอเสนอข้อเสนอแนะสำหรับการวิจัยในอนาคตและข้อเสนอแนะเชิงนโยบายและเชิงปฏิบัติ ดังนี้

1) ข้อเสนอแนะสำหรับการวิจัยในอนาคต

ผลการศึกษานี้มีข้อจำกัดในด้านความครอบคลุมของข้อมูลที่นำมาศึกษา ซึ่งเป็นข้อมูลจากธนาคารเดียวและช่วงเวลาที่จำกัด ดังนั้นจึงควรมีการทดสอบโมเดลกับข้อมูลจากแหล่งอื่นหรือช่วงเวลาอื่นเพิ่มเติมก่อนนำไปใช้งานจริง เพื่อยืนยันความสามารถในการ Generalize ของโมเดล อีกทั้งในอนาคตอาจพัฒนาโมเดลโดยเพิ่มเทคนิคขั้นสูงหรือรวมข้อมูลเชิงพฤติกรรม (Behavioral Data) และข้อมูลทางเลือก (Alternative Data) เพื่อยกระดับความแม่นยำในการทำนายความเสี่ยงการผิดนัดชำระหนี้งวดแรก และพิจารณาขยายขอบเขตไปยังการทำนายความเสี่ยง NPL ระยะยาว หรือปรับใช้กับกลุ่มลูกค้ารายย่อยในภาคธุรกิจอื่นๆ

2) ข้อเสนอแนะเชิงนโยบายและเชิงปฏิบัติ

2.1) การนำโมเดลที่ได้ไปใช้ในกระบวนการพิจารณาสินเชื่อ สถาบันการเงินสามารถนำโมเดล SMOTE + XGBoost ที่ผ่านการทดสอบและปรับแต่งแล้วไปประยุกต์ใช้จริงในขั้นตอนการอนุมัติสินเชื่อ เพื่อเพิ่มความแม่นยำในการประเมินความเสี่ยง และลดอัตราการเกิด NPL (Non-Performing Loan) ในระยะยาว

2.2) การนำ Credit Scoring ไปต่อยอดในระบบบริหารจัดการความเสี่ยง การสร้าง Credit Score ที่ได้จากโมเดลนี้ สามารถนำไปแบ่งกลุ่มลูกค้าตามช่วงคะแนนและกำหนดนโยบายอนุมัติที่เหมาะสมกับระดับความเสี่ยงได้ อาทิ กลุ่ม "Excellent" สามารถอนุมัติอัตโนมัติ ขณะที่กลุ่ม "Fair" หรือ "Poor" ควรมีขั้นตอนพิจารณาเพิ่มเติมหรือขอเอกสารประกอบ

2.3) การนำปัจจัยสำคัญไปใช้ปรับปรุงการบริหารจัดการข้อมูลลูกค้า สถาบันการเงินควรให้ความสำคัญกับการเก็บรวบรวมข้อมูลที่เกี่ยวข้องกับปัจจัยสำคัญ เช่น รายละเอียดรายได้ ประเภทธุรกิจ ประวัติการขอสินเชื่อย้อนหลัง เพื่อให้โมเดลมีความแม่นยำและสามารถปรับปรุงได้อย่างต่อเนื่อง

เอกสารอ้างอิง

ธนาคารกสิกรไทย จำกัด (มหาชน). (2566). รายงานประจำปี 2566. สืบค้นจาก https:// www.kasikornbank.com/en/IR/FinancialReports/AnnualReports/Pages/AnnualReports.aspx

ธนาคารแห่งประเทศไทย. (2563). คู่มือแนวทางการประเมินประสิทธิภาพโมเดลทางการเงิน. สืบค้นจาก https://www.bot.or.th/Thai/FinancialInstitutions/Regulation/DocLib_Regulation/Model_Validation_Guideline_2020.pdf

ธนาคารแห่งประเทศไทย. (2563). แนวปฏิบัติที่ดีในการบริหารความเสี่ยงด้านเครดิตสำหรับสินเชื่อรายย่อย. สืบค้นจาก https://www.bot.or.th/th/research-paper

ธนาคารแห่งประเทศไทย. (2566). รายงานสถานการณ์สินเชื่อธุรกิจขนาดเล็กของไทย. สืบค้นจากจาก https://www.bot.or.th/th/news-and-media/news/news-2023/news-2030316b.html

ธนาคารแห่งประเทศไทย. (2566). *รายงานเสถียรภาพระบบการเงินไทย ปี 2566*. กรุงเทพฯ:ธนาคารแห่งประเทศไทย.

สำนักงานเศรษฐกิจการคลัง. (2566). *แนวทางประเมินโมเดลสินเชื่อธุรกิจขนาดเล็ก*. กรุงเทพฯ: กระทรวงการคลัง.

Agarwal, S., Ambrose, B., Chomsisengphet, S., & Sanders, A. (2020). Predicting first payment defaults in the subprime mortgage market. *Journal of Real Estate Finance and Economics, 61*(1), 1–29. https://doi.org/10.1007/s11146-019-09722-7

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus, 43*(3), 332–357. https://doi.org/10.1111/j.1467-6281.2007.00234.x

Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Bank for International Settlements. (2023). *Credit risk management: Principles for the management of credit risk*. https://www.bis.org/publ/bcbs75.htm

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDDExplorations Newsletter, 6(1), 20–29.

เอกสารอ้างอิง (ต่อ)

Basel Committee on Banking Supervision. (2000, September). Principles for the management of credit risk (BCBS Paper No. 75). Bank for International Settlements.

Berger, A. N., & Udell,G.F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking & Finance, 30*(11), 2945–2966.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.034

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM Consortium.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, C., Li, X., & Wang, J. (2021). Credit scoring with imbalanced data: A comparative study of machine learning models. *Expert Systems with Applications, 183*, 115311. https://doi.org/10.1016/j.eswa.2021.115311

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, J., Li, X., & Wang, S. (2021). Credit scoring using machine learning techniques: A systematic literature review. *Financial Innovation, 7*, 1–23.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEESymposium Series on Computational Intelligence* (pp. 159–166).

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan), 1–30.

เอกสารอ้างอิง (ต่อ)

Fathi, A., Rezaei, S., & Mohammadi, M. R. (2021). Hyperparameter tuning and optimal design of machine learning models: A systematic literature review. *Applied Soft Computing, 107*, 107443.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133-3181.

FICO. (2024). Understanding first payment default and early risk indicators. Retrieved June 24, 2025, from https://www.fico.com

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Funding Societies. (2024). *SME financing: Designed to support small businesses*. Retrieved June 24, 2025, from https://fundingsocieties.com/

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Hu, W., & Li, X. (2013). Visual detection in imbalanced datasets using SMOTE. In Proceedings of the International Conference on Machine Learning.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, 1137–1143.

Koç, U., & Sevgili, T. (2020). Consumer loans' first payment default detection: A predictive model. *Turkish Journal of Electrical Engineering & Computer Sciences, 28*(1), 167–181. https://doi.org/10.3906/elk-1809-190

เอกสารอ้างอิง (ต่อ)

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1–5. https://imbalanced-learn.org

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Liu, Y., Wang, H., & Lin, C. (2020). Early default prediction using machine learning: A case study in China. *Journal of Risk and Financial Management, 13*(4), 65. https://doi.org/10.3390/jrfm13040065

Liu, Y., Wang, K., & Lin, M. (2020). Predicting first payment default in microfinance: Machine learning evidence from emerging markets. *Journal of Risk and Financial Management, 13*(12), 315.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

Mitchell, T. M. (1997). *Machine Learning* (1st ed.). McGraw-Hill. ISBN 978-0-07-0428072.

Moody's Analytics. (2017). *ROC curve, AUC and Gini explained*. Retrieved June 24, 2025, from https://www.moodysanalytics.com/risk-perspectives-magazine/credit-risk/roc-curve-auc-and-gini-explained

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rodríguez, R. A. (2024). A required debt service coverage ratio related to the economic value of the asset involved. *Journal of Financial Risk Management, 13*, 618–642. https://doi.org/10.4236/jfrm.2024.134029

Sammut, C., & Webb, G. I. (2017). *Encyclopedia of Machine Learning and Data Mining*. Springer.

Siddiqi, N. (2017). *Credit risk scorecards: Developing and implementing intelligent credit scoring* (2nd ed.). Wiley.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.

เอกสารอ้างอิง (ต่อ)

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101.

Thomas, L. C. (2009). *Consumer credit models: Pricing, profit and portfolios*. Oxford University Press.scirp.org+12scirp.org+12ideas.repec.org+12

Train in Data. (2022). Random under-sampling: dealing with imbalanced datasets in machine learning. Retrieved from https://www.trainindata.com

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.

Zhou, H., Liu, Y., & Lin, M. (2022). Ensemble machine learning for credit risk prediction: A comparison study. *Financial Innovation, 8*(1), 27.

ประวัติผู้เขียน

ชื่อ	นางสาวภรภัทร ชินคำวงค์
วัน เดือน ปีเกิด	21 พฤศจิกายน 2532
ที่อยู่ปัจจุบัน	111/356 หมู่4 ตำบลบางแก้ว อำเภอบางพลี จังหวัดสมุทรปราการ 10540
ประวัติการศึกษา	2555 วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา