การเรียนรู้ของเครื่องเพื่อการทำนายการผิดนัดชำระของลูกหนี้บัตรเครดิต

สกุลกาญจน์ ทองคำ

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

การเรียนรู้ของเครื่องเพื่อการทำนายการผิดนัดชำระของลูกหนี้บัตรเครดิต

สกุลกาญจน์ ทองคำ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566

MACHINE LEARNING MODELS
FOR CREDIT CARD DEFAULT PREDICTION

SAKULKRAN THONGKHAM

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

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2023

สารนิพนธ์

เรื่อง

การเรียนรู้ของเครื่องเพื่อการทำนายการผิดนัดชำระของลูกหนี้บัตรเครดิต

ของ

สกุลกาญจน์ ทองคำ

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... กรรมการ

(อาจารย์ ดร.ศุภร คนธภักดี)

ชื่อเรื่อง	การเรียนรู้ของเครื่องเพื่อการทำนายการผิดนัดชำระของลูกหนี้บัตรเครดิต
ผู้วิจัย	สกุลกาญจน์ ทองคำ
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2566
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. นุวีย์ วิวัฒนวัฒนา

งานวิจัยนี้มุ่งศึกษาการทำนายลูกหนี้บัตรเครดิตที่มีโอกาสผิดนัดชำระ โดยใช้การเรียนรู้ของเครื่อง (Machine Learning) เป็นเครื่องมือสร้างแบบจำลองการจำแนกประเภทลูกหนี้แบบมีผู้สอน (Supervised Learning) ประเภท Classification ด้วยการทดสอบกับข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต โดยมีข้อมูลรายการธุรกรรม จำนวน 1,048,575 แถว และข้อมูลลูกหนี้ จำนวน 438,557 แถว จากเว็บไซต์ Kaggle.com ผู้วิจัยสร้างแบบจำลองเพื่อจำแนกประเภทลูกหนี้ที่มีความสามารถในการชำระหนี้บัตรเครดิต เป็น 2 ประเภท ได้แก่ ลูกหนี้ปกติและลูกหนี้ผิดนัดชำระ ประกอบด้วยอัลกอริทึม 3 วิธี ได้แก่ 1.) Logistic Regression 2.) XGBoost และ 3.) CatBoost เพื่อหาแบบจำลองที่มีประสิทธิภาพมากที่สุดในการจำแนกประเภทลูกหนี้ ผลการศึกษาพบว่า วิธีทำนายแบบ XGBoost ให้ค่าความถูกต้อง 98 เปอร์เซ็นต์ ที่จำนวนต้นไม้ 15 ต้น กับอัตราการเรียนรู้ที่ 0.1 วิธีทำนายแบบ CatBoost ให้ค่าความถูกต้อง 97 เปอร์เซ็นต์ ที่จำนวนต้นไม้ 7 ต้น กับอัตราการเรียนรู้ที่ 0.1 และวิธีทำนายแบบ Logistic Regression ให้ค่าความถูกต้อง 62 เปอร์เซ็นต์ เมื่อเปรียบเทียบค่า Confusion Matrix พบว่าแบบจำลอง Random Forest และ Catboost ให้ผลลัพธ์สูงสุดใกล้เคียงกัน

คำสำคัญ : การอนุมัติสินเชื่อ, เทคนิคการเรียนรู้ของเครื่อง, การทำนายลูกค้า, จำแนกประเภทลูกหนี้

Title	MACHINE LEARNING MODELS FOR CREDIT CARD DEFAULT PREDICTION
Author	SAKULKRAN THONGKHAM
Degree	MASTER OF SCIENCE
Academic Year	2023
Thesis Advisor	Assistant Professor Dr. Nuwee Wiwatwattana

This thesis aimed to study predictive analysis among credit card holders who could create a non-performing loan using machine learning to set up supervised learning in the classification character. The learning machine tested credit card loan transaction data with 1,048,575 rows of transaction lists and 438,557 rows of credit card customer data selected from Kaggle.com. The process functioned by designing the model to divide credit card customers into two groups: normal customers and non-performing loan customers with the aid of machine learning and classification supervised learning. This machine learning had three algorithms: (1) Logistic Regression; (2) XGBoost; and (3) CatBoost, to explore the most effective model to analyze credit card customers. The study depicted that the XGBoost algorithm provided 98% accuracy at 15 Depth with 0.1 degree of learning rate, the Catboost algorithm provided 97% accuracy with 7 Depth and 0.1 degree of learning rate, and the logistic regression algorithm provided 62% of accuracy. The output from the confusion matrix table pointed that the XGBoost algorithm and CatBoost algorithm maintained the most effective outcome in close proximity.

Keyword : Credit card approval, Prediction algorithm, Machine learning, Non-Performing loan

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร. นุวีย์ วิวัฒนวัฒนา อาจารย์ที่ปรึกษา ที่ให้คำปรึกษาตั้งแต่เริ่มต้นจนเสร็จสมบูรณ์และช่วยตรวจสอบความถูกต้องในด้านข้อมูลทางวิชาการ รวมถึงตรวจสอบความเรียบร้อย ความสวยงามของการใช้คำในสารนิพนธ์ในทุกขั้นตอน

ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์และอาจารย์ทุกท่าน ที่ให้ความรู้และคำแนะนำที่เป็นประโยชน์ในการปรับปรุงสารนิพนธ์ให้ดียิ่งขึ้น

ขอกราบขอบพระคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ สำหรับทุนสนับสนุนการนำเสนอผลงานวิจัยของนิสิตบัณฑิตศึกษาในงานประชุมวิชาการ ทำให้ได้รับประสบการณ์ที่ดีในการเผยแพร่และแลกเปลี่ยนความรู้กับผู้นำเสนอท่านอื่น

สุดท้ายนี้ขอขอบพระคุณครอบครัวของผู้วิจัยที่ให้โอกาสในการศึกษาและเป็นกำลังใจให้จนสำเร็จการศึกษา รวมถึงขอบคุณรุ่นพี่ในสาขาวิชาที่คอยให้ความช่วยเหลือและคอยให้คำแนะนำทั้งในช่วงเวลาเรียนและช่วงเวลาในการทำรูปเล่มสารนิพนธ์

สกุลกาญจน์ ทองคำ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1. ภูมิหลัง

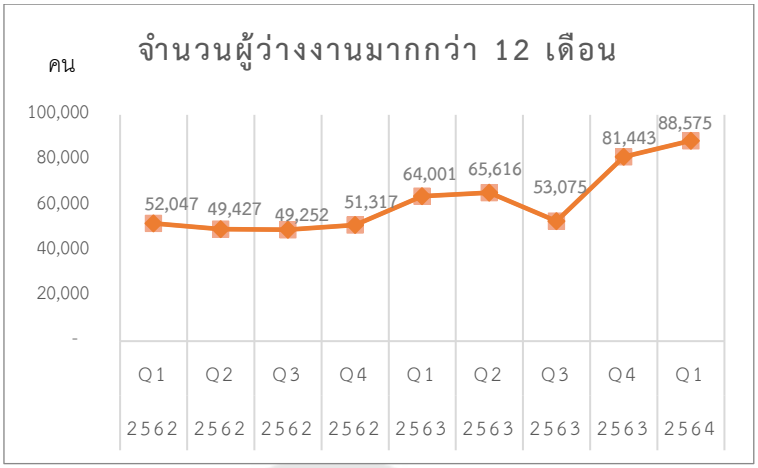
ธุรกิจบัตรเครดิตเป็นการให้บริการของสถาบันการเงินต่างๆ ซึ่งได้รับความนิยมกันอย่างแพร่หลายในปัจจุบัน เนื่องจากมีความสะดวกสำหรับใช้จ่ายอีกทั้งมีความปลอดภัยในการพกพาเงินสดเป็นจำนวนมาก โดยสามารถใช้ในการชำระค่าสินค้า และค่าบริการหรือหนี้แทนการชำระด้วยเงินสด รวมทั้งผู้ใช้บัตรเครดิตสามารถเบิกถอนเงินสดได้โดยไม่จำเป็นต้องมีเงินสด การเปลี่ยนแปลงด้านพฤติกรรมการใช้จ่ายนี้สะท้อนให้เห็นว่าบัตรเครดิตเข้ามามีบทบาทตามยุคสมัยสังคมไร้เงินสด

จากการขยายตัวของการใช้จ่ายผ่านบัตรเครดิต แม้จะมีส่วนในการกระตุ้นและช่วยให้เงินลงทุนหมุนเวียนในระบบเศรษฐกิจเพิ่มมากขึ้น แต่อีกแง่มุมหนึ่งก็เป็นต้นตอของปัญหาต่างๆ ตามมา ทั้งปัญหาการเปลี่ยนแปลงคุณภาพลูกหนี้ โดยสถาบันการเงินมีเกณฑ์การจัดชั้นลูกหนี้เชิงปริมาณพิจารณาจากระยะเวลาค้างชำระ สามารถจำแนกประเภทลูกหนี้ได้ 3 กลุ่มหลักๆ ได้แก่ 1.) ลูกหนี้ปกติ (Performing Loan – PL) ลูกหนี้ที่มีระยะเวลาค้างชำระน้อยกว่า 1 เดือน 2.) ลูกหนี้กล่าวถึงเป็นพิเศษ (Special Mention – SM) ลูกหนี้ที่มีระยะเวลาค้างชำระมากกว่า 1 เดือนแต่ไม่เกิน 3 เดือน และ 3.) ลูกหนี้ไม่ก่อให้เกิดรายได้ (Non-Performing Loans – NPLs) ลูกหนี้ที่มีระยะเวลาค้างชำระมากกว่า 3 เดือนดังภาพประกอบที่ 1 ซึ่งหากเกิดลูกหนี้ไม่ก่อให้เกิดรายได้ อันเนื่องมาจากการใช้จ่ายฟุ่มเฟือยจนเกินความสามารถชำระหนี้คืนในอนาคต หรือปัญหาสภาพเศรษฐกิจที่อาจเป็นผลกระทบต่อลูกหนี้ อันเป็นเหตุทำให้ไม่สามารถจ่ายหนี้ได้ตามกำหนด อีกทั้งสินเชื่อบัตรเครดิตเป็นสินเชื่อไม่มีหลักประกัน (Clean Loan) หากเกิดปัญหาลูกหนี้ไม่สามารถชำระหนี้ได้ตามกำหนดความสูญเสียที่เกิดขึ้นจะส่งผลกระทบต่อการขาดทุนไปจนถึงการล้มละลายของสถาบันการเงินก็เป็นได้ ดังนั้นหากมองในภาพความเสี่ยงด้านเครดิต สินเชื่อบัตรเครดิตถือได้ว่าเป็นสินเชื่อที่มีความเสี่ยงสูงเนื่องจากเป็นสินเชื่อไม่มีหลักประกัน และผลกระทบทำให้ลูกหนี้ไม่มีความสามารถในการชำระหนี้ จากปัจจัยภายในอันเนื่องมาจากพฤติกรรมของลูกหนี้และปัจจัยภายนอกจากสภาพเศรษฐกิจที่ส่งผลให้ลูกหนี้บางกลุ่มขาดรายได้และไม่สามารถชำระหนี้ได้ตามกำหนด

ภาพประกอบ 1 แสดงภาพปัจจัยเชิงปริมาณที่ใช้ในการพิจารณาจัดชั้นสินทรัพย์

ที่มา (ธนาคารแห่งประเทศไทย, 2016)

สืบเนื่องจากสถานะการแพร่ระบาดโควิด-19 ในปัจจุบันส่งผลกระทบต่อสภาพเศรษฐกิจและมีแนวโน้มรุนแรงเพิ่มมากขึ้น ผลกระทบดังกล่าวเป็นสาเหตุหลักทำให้อัตราผู้ว่างงานมีแนวโน้มเพิ่มมากขึ้น และมีแนวโน้มเป็นผู้ว่างงานระยะยาวมากขึ้น (ว่างงานมากกว่า 12 เดือน) ดังภาพประกอบที่ 2 ซึ่งผลกระทบดังกล่าวทำให้กลุ่มแรงงานในวิสาหกิจขนาดกลางและขนาดย่อม (MSME) อาจถูกลดชั่วโมงการทำงาน เพราะเมื่อเศรษฐกิจไม่ดีนายจ้างจึงพยายามรักษาลูกจ้างไว้โดยลดชั่วโมงการทำงานก่อน เมื่อมีความจำเป็นจึงค่อยเลิกจ้าง หรือกลุ่มแรงงานในภาคการท่องเที่ยวอาจถูกเลิกจ้างมากขึ้น และต้องหาอาชีพใหม่ ตำแหน่งงานอาจไม่เพียงพอรองรับเด็กจบใหม่ หากจัดกลุ่มด้านความเสี่ยงตามสภาพเศรษฐกิจในปัจจุบันจะพบว่ากลุ่มอาชีพที่ได้รับผลกระทบต่อการแพร่ระบาดโควิด-19 นั้นอาจมีปัญหาด้านการเงินและเป็นสาเหตุที่ทำให้ลูกหนี้กลุ่มอาชีพที่ได้รับผลกระทบไม่สามารถชำระหนี้ได้ตามกำหนดจนกลายเป็นลูกหนี้ไม่ก่อให้เกิดรายได้



ภาพประกอบ 2 แสดงภาพจำนวนผู้ว่างงานมากกว่า 12 เดือน ระหว่าง ไตรมาส 1/2562 ถึง ไตรมาส 1/2564

ที่มา (สำนักงานสภาพัฒนาเศรษฐกิจและสังคมแห่งชาติ, 2021)

ดังนั้นหากสถาบันการเงินสามารถทำการแบ่งกลุ่มลูกหนี้ (Customer Segmentation) ที่มีคุณลักษณะและพฤติกรรมที่เหมือนกันหรือใกล้เคียงกันจะถูกจัดกลุ่มไว้ด้วยกัน ส่วนลูกหนี้ที่มีคุณลักษณะและพฤติกรรมที่แตกต่างกันหรือไม่เหมือนกันจะถูกจัดกลุ่มที่แยกจากกันก็จะสามารถแยกกลุ่มลูกหนี้ที่มีความสามารถในการชำระหนี้เพื่ออนุมัติสินเชื่อ และไม่อนุมัติสินเชื่อต่อกลุ่มลูกหนี้ที่มีความเสี่ยงในการชำระหนี้ ซึ่งอาจเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ในอนาคต การจำแนกประเภทลูกหนี้ดังกล่าวเพื่อให้สถาบันการเงินสามารถลดความเสี่ยงต่อผลการขาดทุนด้านเครดิตของสถาบันการเงิน ซึ่งในอดีตสถาบันการเงินมีกระบวนการทำงานโดยจะกระจายอำนาจการอนุมัติสินเชื่อตามพื้นที่ โดยใช้บุคลากรเป็นผู้วิเคราะห์สินเชื่อ เนื่องจากจำนวนปริมาณการใช้งานและข้อมูลที่มีจำนวนน้อยจึงสามารถจัดการได้ครอบคลุม

แต่เนื่องด้วยปัจจุบันจำนวนข้อมูลที่มีปริมาณมหาศาลอีกทั้งความนิยมในการใช้บัตรเครดิตมีจำนวนมากจึงทำให้ยากแก่การจำแนกลูกหนี้ได้ในระยะเวลาอันสั้น และอาจเกิดความเสี่ยงต่อสินทรัพย์ได้หากผู้อนุมัติสินเชื่อไม่มีความชำนาญเพียงพอ แต่ด้วยเทคนิคการเรียนรู้ของเครื่องสามารถรองรับข้อมูลจำนวนมากและสามารถแยกประเภทลูกหนี้ได้ทันที เพื่อให้สถาบันการเงินสามารถแก้ไขปัญหาดังกล่าวอีกทั้งยังเป็นการลดระยะเวลาในการทำงานและเพิ่มความพึงพอใจต่อลูกหนี้ รวมถึงลดต้นทุนด้านทรัพยากรบุคคลในการวิเคราะห์สินเชื่อ โดยงานวิจัยมีจุดประสงค์เพื่อศึกษาข้อมูลรายการธุรกรรมบัตรเครดิตและข้อมูลลูกหนี้ เพื่อจำแนกประเภทข้อมูลลูกหนี้โดยอาศัยความคล้ายคลึงกันของลักษณะและพฤติกรรมของผู้กู้ โดยการนำเทคนิคการ

เรียนรู้ของเครื่องจักรประเภท การจำแนกประเภทของข้อมูลเข้ามาประยุกต์ใช้ (Classification) จะทำให้สามารถจำแนกประเภทลูกหนี้ที่มีความสามารถในการชำระหนี้ และยังสามารถนำมาประยุกต์ใช้กับการทำงานในหน่วยงานต่างๆได้ ก็จะสามารถเสริมประสิทธิภาพการทำงานให้ดียิ่งขึ้น งานวิจัยนี้จึงศึกษาวิธีการจำแนกประเภทลูกหนี้จากข้อมูลรายการธุรกรรมและข้อมูลลูกหนี้ โดยใช้เทคนิคการเรียนรู้เครื่อง จำแนกประเภทลูกหนี้ลูกหนี้ เพื่อผลลัพธ์ที่ได้จะสามาถนำมาวิเคราะห์พฤติกรรมลูกหนี้และกำหนดกลยุทธ์ทางการตลาดต่อไป

งานวิจัยนี้เน้นการศึกษาการทำนายลูกหนี้ที่มีโอกาสผิดนัดชำระ โดยใช้การเรียนรู้ของเครื่อง (Machine Learning) เป็นเครื่องมือสำหรับสร้างแบบจำลองในการจำแนกประเภทลูกหนี้แบบมีผู้สอน (Supervised Learning) ประเภท Classification โดยทดลองกับข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต ประกอบด้วย 2 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 1,048,575 แถวและข้อมูลลูกหนี้ ประกอบด้วย 17 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 438,557 แถว

2. จุดประสงค์ของงานวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อสร้างแบบจำลองแบ่งกลุ่มลูกหนี้ที่มีความสามารถในการชำระหนี้โดยจำแนกประเภทลูกหนี้เป็น 2 ประเภทคือ 1 ลูกหนี้ปกติ 2 ลูกหนี้ผิดนัดชำระโดยใช้เทคนิค 1. Logistic Regression 2. Extreme Gradient Boosting (XGBoost) 3.CatBoost ประเภท Classification

2. เพื่อศึกษาว่าการสร้างแบบจำลองวิธีใด มีประสิทธิภาพมากที่สุดในการทำนายลูกหนี้กับข้อมูลที่ใช้ในการทดสอบ

3. ขอบเขตของการวิจัย

งานวิจัยนี้ศึกษาการจำแนกประเภทลูกหนี้ที่มีคุณลักษณะและพฤติกรรมที่ใกล้เคียงกันหรือเหมือนกันออกเป็น 1. ลูกหนี้ปกติ 2. ลูกหนี้ผิดนัดชำระโดยใช้เทคนิค Machine Learning เป็นเครื่องมือสำหรับสร้างแบบจำลองในการจำแนกประเภทลูกหนี้ โดยข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต ประกอบด้วย 2 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 1,048,575 แถวและข้อมูลลูกหนี้ ประกอบด้วย 17 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 438,557 แถว จากแหล่งข้อมูลสาธารณะ Kaggle.com โดยข้อมูลถูกเก็บในรูปแบบตาราง จากนั้นจึงใช้ Machine learning มาเป็นเครื่องมือสำหรับ สร้างแบบจำลองเพื่อจำแนกประเภทลูกหนี้ ถูกพัฒนาด้วยภาษา Python โดยใช้วิศวกรรมคุณลักษณะ (Feature Engineering) และใช้เทคนิค 1. Logistic Regression 2. Extreme Gradient Boosting (XGBoost) 3. CatBoost ในการทำ

Classification เพื่อหาแบบจำลองที่มีประสิทธิภาพในการจำแนกประเภทลูกหนี้ที่ดี โดยประกอบด้วย 5 วิธีการวัดประสิทธิภาพคือ Accuracy, Precision, Recall, AUC และ ROC ผู้วิจัยยังสนใจเปรียบเทียบผลของการจัดการข้อมูลโดยใช้เทคนิค ก่อนและหลังใช้เทคนิค Synthetic Minority Oversampling Technique

ตัวแปรข้อมูลลูกหนี้

- เพศ
- อาชีพ
- ประเภทรายได้
- ระดับการศึกษา
- สถานภาพการสมรส
- จำนวนบุตร
- จำนวนสมาชิกครอบครัว
- ประเภทที่อยู่อาศัย
- รายได้ต่อปี
- วันเกิดสะสม (วัน)
- วันที่เริ่มทำงาน
- Flag ระบุรายละเอียดลูกหนี้มีโทรศัพท์มือถือหรือไม่
- Flag ระบุรายละเอียดลูกหนี้มีเบอร์โทรที่ทำงานหรือไม่
- Flag ระบุรายละเอียดลูกหนี้มีเบอร์โทรศัพท์หรือไม่
- Flag ระบุรายละเอียดลูกหนี้มี E-mail หรือไม่
- Flag ระบุรายละเอียดลูกหนี้มีรถหรือไม่
- Flag ระบุรายละเอียดลูกหนี้มีอสังหาริมทรัพย์หรือไม่

ตัวแปรข้อมูลรายการธุรกรรม

- สถานะบัญชี
- จำนวนเดือนค้างชำระ

4. ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถนำแบบจำลองที่ได้ ไปประยุกต์ใช้กับแอปพลิเคชันเพื่อประเมินเบื้องต้นว่าลูกหนี้สามารถกู้สินเชื่อได้ หรือไม่

2. สถาบันการเงินสามารถลดต้นทุนในกระบวนการวิเคราะห์สินเชื่อในด้านทรัพยากรบุคคล ปริมาณเอกสาร เป็นต้น

3. สถาบันการเงินสามารถลดระยะเวลาในการทำงานและปรับปรุงความเร็วในการให้บริการแก่ลูกค้าเพื่อความพึงพอใจของลูกค้าที่เพิ่มขึ้น

บทที่ 2
วรรณกรรม และงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาและงานวิจัยที่เกี่ยวข้องและนำเสนอตามหัวข้อดังต่อไปนี้

1. ทฤษฎีเกี่ยวกับอัลกอริทึมในการจำแนกประเภท (Classification Algorithms)
2. วิศวกรรมคุณลักษณะ (Feature Engineering)
3. ทฤษฎีเกี่ยวกับการวัดประสิทธิภาพแบบจำลองการจำแนกประเภท
4. การจัดการกับข้อมูลที่ไม่สมดุล
5. งานวิจัยที่เกี่ยวข้อง

1. **ทฤษฎีเกี่ยวกับอัลกอริทึมในการจำแนกประเภท** (Classification Algorithms)

เป็นเทคนิคการเรียนรู้แบบ Supervised โดยเทคนิค Classification Algorithm มีหลายรูปแบบได้แก่

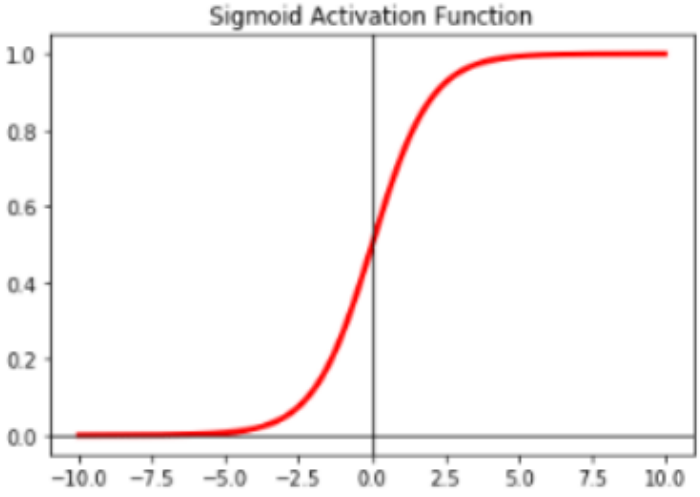
1.1 Logistic Regression

การวิเคราะห์การถดถอยแบบโลจิสติกเป็นวิธีทั่วไปของ Machine Learning เพื่อใช้ในการแก้ปัญหาการจัดกลุ่มซึ่งการวิเคราะห์การถดถอยโลจิสติกแบ่งออกเป็น 2 ประเภทคือ (1) การวิเคราะห์การถดถอยโลจิสติกทวิ (Binary Logistic Regression) (2) การวิเคราะห์การถดถอยโลจิสติกพหุกลุ่ม (Multinomial Logistic Regression การวิเคราะห์การถดถอยโลจิสติกทั้ง 2 ประเภท แตกต่างกันด้านตัวแปร โดยการวิเคราะห์การถดถอยโลจิสติกทวิใช้ตัวแปรตามที่แบ่งออกเป็น 2 กลุ่มย่อย (Dichotomous Variable) มี 2 ค่า คือมีค่าเป็น 0 กับ 1 เช่น กลุ่มลูกหนี้ปกติกับกลุ่มลูกหนี้ผิดนัดชำระ ส่วนการวิเคราะห์การถดถอยโลจิสติกพหุกลุ่มใช้กับตัวแปรตามที่มีค่ามากกว่า 2 กลุ่ม (Polytomous Variable) เช่นธนาคารมีการให้บริการสูง ปานกลาง และต่ำการวิเคราะห์การถดถอยโลจิสติกมีเป้าหมายเพื่อทำนายโอกาสที่เกิดเหตุการณ์ที่สนใจ โดยการวิเคราะห์การถดถอยแบบโลจิสติกเป็นอัลกอริทึมเชิงเส้น ที่ถูกเปลี่ยนสมการเส้นตรงโดยใช้ฟังก์ชัน Sigmoid กับการถดถอยเชิงเส้น ดังภาพประกอบที่ 3 ซึ่งค่าที่ได้จะมีค่าความน่าจะเป็นที่จะเกิดเหตุการณ์ ดังสมการที่ (1)

$$E(Y|x) = \frac{e^{(X\beta)}}{1+e^{(X\beta)}} \quad (1)$$

โดยที่

$E(Y\|x)$	คือ ความน่าจะเป็นที่ x อยู่ในคลาส Y
X	คือ จุดข้อมูลที่ต้องการทำนาย
β	คือ ค่าสัมประสิทธิ์



ภาพประกอบ 3 Sigmoid Activation Function

ที่มา (Kandel & Castelli, 2020)

ข้อดีของ Logistic Regression คือ ใช้ทรัพยากรและเวลาในการประมวลแบบจำลองน้อยเนื่องจากโมเดลเป็นแบบเรียบง่าย ซึ่งหมายความว่าประสิทธิภาพการคำนวณสูง

ข้อเสียของ Logistic Regression คือ แบบจำลองโอกาสเกิดปัญหาจำลองมีความถูกต้องในการทำนายเป้าหมายต่ำไป (Under Fitting) และปัญหาแบบจำลองทำนายได้ไม่แม่นยำหากข้อมูลมีความไม่สมบูรณ์

1.2 Extreme Gradient Boosting (XGBoost)

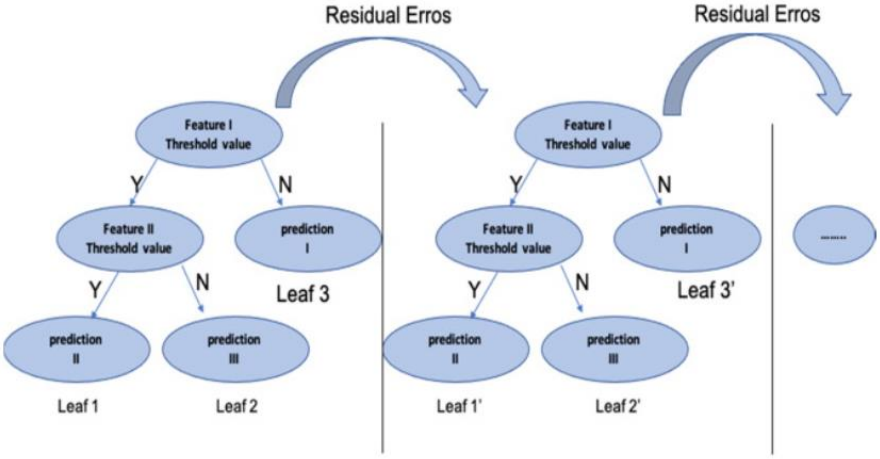
XGBoost เป็นอัลกอริทึมที่ถูกพัฒนาขึ้นมาจาก Gradient Tree Boosting สำหรับจัดการกับงานที่มีข้อมูลขนาดใหญ่อย่างมีประสิทธิภาพและใช้เวลาในการพัฒนาแบบจำลองที่เหมาะสม ซึ่งมีการประยุกต์ใช้งานในด้านงานวิจัยต่างๆ ตั้งแต่งานด้านประเมินความเสี่ยงสินเชื่อไปจนถึงการวินิจฉัยโรคมะเร็ง โดยถูกพัฒนาโดยภาษา Python สำหรับพัฒนาแบบจำลอง ปัจจุบันจึงเป็นที่นิยมสำหรับใช้พัฒนาแบบจำลองที่มีข้อมูลขนาดใหญ่ ซึ่งมีหลายงานวิจัยอภิปรายผลจากจากการนำอัลกอริทึม XGBoost นั้นสามารถใช้ในการจัดการกับปัญหาความไม่สมดุลกันของชุดข้อมูลได้ดี ซึ่งสามารถทำงานได้อย่างมีประสิทธิภาพเหนือกว่าวิธีอื่น

โดยมีแนวคิดคือการปรับปรุงแบบจำลองโดยค่อยๆ พัฒนาแบบจำลองที่มีประสิทธิภาพต่ำหลายๆ แบบจำลองตามลำดับเพื่อสร้างแบบจำลองที่มีประสิทธิภาพในท้ายที่สุด ซึ่งจะนำผลลัพธ์จากการทำนายของแบบจำลองไปใช้งานเป็นข้อมูลขาเข้าของแบบจำลองถัดไป โดยมีวัตถุประสงค์เพื่อพยายามลดค่าของความผิดพลาด ดังภาพประกอบที่ 4 กระบวนการทำงานแบบนี้เพื่อให้การทำนายมีความแม่นยำเพราะการเรียนรู้ของต้นไม้จะต่อเนื่องและมีความลึกมากพอ แบบจำลองจะหยุดการเรียนรู้โดยที่ไม่เหลือรูปแบบของการทำนายที่ผิดพลาดจากการสร้างต้นไม้ก่อนหน้าที่เรียนไปแล้ว ดังสมการที่ (2)

$$Obj^{(t)} = -\frac{1}{2}\frac{\left(\Sigma_{i\epsilon I_j} g_i\right)^2}{\Sigma_{i\epsilon I_j} h_i + \lambda} + \gamma \qquad (2)$$

โดยที่

$Obj^{(t)}$	คือ เป้าหมายของการทำนาย
I_j	คือ จุดข้อมูลที่ถูกกำหนดด้วยจำนวนใบไม้
g	คือ Loss Function ลำดับที่ 1
h	คือ Loss Function ลำดับที่ 2
λ	คือ Regularization Parameter
γ	คือ Gamma



ภาพประกอบ 4 แสดงแผนผังของต้นไม้ XGBoost

ที่มา (Ibrahem Ahmed Osman, Najah Ahmed, Chow, Feng Huang, & El-Shafie, 2021)

หลักการสำหรับสร้างต้นไม้

โหนดภายใน (Internal Node)

แต่ละโหนดจะถูกแยกออกตามแต่ละตัวแปร

เงื่อนไขของกิ่งก้าน (Edge) ระบุได้ว่าข้อมูลใดสามารถไหลผ่านไปได้

ใบไม้ (Leaves)

จุดของข้อมูลจนถึงใบไม้จะถูกกำหนดด้วยน้ำหนัก

ค่าถ่วงน้ำหนักจะใช้ในการทำนาย

ข้อดีของ XGBoost คือ แก้ปัญหาข้อมูลที่มีความแปรปรวนจากการข้อมูลที่มีค่าว่าง และสามารถลดปัญหา Overfitting

ข้อเสียของ XGBoost คือ ใช้การทรัพยากรและเวลาในการประมวลผลนานหากมีจำนวนตัวแปรและข้อมูลที่มาก

1.3 CatBoost

อัลกอริทึม CatBoost เป็นวิธี Gradient Boosting ชนิดหนึ่ง ซึ่งถูกพัฒนาเพื่อจัดการกับข้อมูลจำพวกตัวแปรประเภทและเพิ่มประสิทธิภาพการทำงานในแบบจำลองเพื่อให้สามารถทำนายได้รวดเร็ว โดยหลักการของอัลกอริทึมสมมุติว่ามีชุดข้อมูล (Data Set) D={ (Xi,Yi) }, …, n.

และ $Y_i \in R$ คือชุดของเป้าหมายของการทำนาย (Label Set) ขั้นแรก CatBoost จะสุ่มเรียงลำดับข้อมูลทั้งหมด สำหรับค่าบางค่าในหมวดหมู่ของแต่ละตัวอย่าง เมื่อแปลงเป็นค่าตัวเลข ค่าเฉลี่ยจะถูกใช้ตามค่าเป้าหมายของการทำนาย ตัวอย่างก่อนหน้านี้ และน้ำหนักลำดับความสำคัญจะถูกเพิ่ม

ข้อดีของ CatBoost คือสามารถประมวลผลตัวแปรประเภทหมวดหมู่จากชุดข้อมูลที่โดยผ่านกระบวนการแปลงตัวแปรประเภทหมวดหมู่ให้เป็นค่าตัวเลข (One-Hot Encoding) แต่หากมีจำนวนของตัวแปรหมวดหมู่อาจพบปัญหาความลึกของต้นไม้ (Tree Depth) ที่มากเกินไป ดังนั้น CatBoost จะใช้วิธีการทางสถิติในการประมวลผลตัวแปรประเภทหมวดหมู่ ซึ่งทำให้โมเดลนี้มีความแม่นยำในการจำแนกประเภทได้ดี นอกจากนี้ CatBoost ยังสามารถช่วยลดปัญหาการปรับแต่งตัวแปร (Hyperparameter Tuning) , ลดปัญหาข้อมูลมีความถูกต้องในการทำนายเป้าหมายสูงเกินไป (Overfitting) และมีความรวดเร็วในการทำนาย

ข้อเสียของ CatBoost คือ ใช้การทรัพยากรและเวลาในการประมวลผลนานหากจำนวนตัวแปรที่มีจำนวนเยอะซึ่งอาจส่งผลต่อการทดสอบแบบจำลอง และการตั้งค่าตัวเลขสุ่มที่ต่างกันอาจส่งผลกระทบต่อการทำนายแบบจำลอง

2. วิศวกรรมคุณลักษณะ (Feature Engineering)

วิศวกรรมคุณลักษณะ คือ กระบวนการหนึ่งในการเตรียมข้อมูล (Data Preparation) โดยวัตถุประสงค์ของการทำ Feature Engineering เพื่อเป็นการทำให้แบบจำลองนั้นมีประสิทธิภาพ รวมถึงสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม หรืออาจจะนำคุณลักษณะเดิมผ่านกระบวนการคำนวนเพื่อให้ได้ซึ่งคุณลักษณะใหม่ โดยมีกระบวนการที่เกี่ยวข้องดังต่อไปนี้ การเติมข้อมูลที่ขาดหาย (Imputation) การจัดการข้อมูลผิดปกติ (Handling Outliers) การแปลงข้อมูล (Log Transform) การแปลงข้อมูลให้อยู่ในรูปแบบตัวเลข (One-Hot Encoding / Embedding) การจัดกลุ่มข้อมูล (Grouping Operations) และการสเกลหรือปรับช่วงของข้อมูล (Normalize, Standardize) เพื่อเป็นประโยชน์ในการจำแนกประเภทลูกหนี้

3. ทฤษฎีเกี่ยวกับการวัดประสิทธิภาพแบบจำลองการจำแนกประเภท

กระบวนการวัดประสิทธิภาพแบบจำลอง คือ เป็นเทคนิคเพื่อใช้วัดประสิทธิภาพแบบจำลอง โดยมีวิธีการต่างๆ เพื่อที่จะนำเสนอประสิทธิภาพแบบจำลองเช่น Confusion Matrix , Accuracy , Recall , Precision , F1 Score เป็นต้น

Confusion Matrix คือตารางวัดประสิทธิภาพของ การเรียนรู้ของเครื่อง (Machine Learning) ประเภท Classification Algorithm โดยลักษณะของ Confusion Matrix จะเป็นผลรูปแบบตารางตามภาพประกอบที่ 5

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

ภาพประกอบ 5 แสดงภาพตาราง Confusion Matrix

จากภาพแสดงผลตัวแปรผลของการทำนายอัลกอริทึมเพื่อใช้ในการวัดผลประสิทธิโดยมีรายละเอียดดังนี้

True Positive (TP) คือ สิ่งที่แบบจำลองทำนายว่า “จริง” และ มีค่าเป็น “ จริง ”

True Negative (TN) True Negative (TN) คือ สิ่งที่โแบบจำลองทำนายว่า “ไม่จริง” และมีค่า “ ไม่จริง ”

False Positive (FP) False Positive (FP) คือ สิ่งที่แบบจำลองทำนายว่า “จริง” แต่ มีค่าเป็น “ไม่จริง”

False Negative (FN) False Negative (FN) คือ สิ่งที่แบบจำลองทำนายว่า “ไม่จริง” แต่มีค่าเป็น “จริง” ดังภาพประกอบที่ 5

โดยแต่ละค่าจะเก็บผลของการทำนายเพื่อนำไปคำนวณหาค่าประสิทธิของการทำนายซึ่งปัจจุบันนิยม 3 ค่าได้แก่

ค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้องของแบบจำลอง โดยพิจารณารวมทุกคลาส หากข้อมูลคลาสมีความสมดุลค่าความถูกต้องจะมีประสิทธิภาพ แต่หากข้อมูลคลาสไม่มีความสมดุลค่าความถูกต้องจะมีประสิทธิภาพต่ำ แสดงได้ดังสมการที่ (3)

$$\frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{3}$$

ค่าเรียกคืน (Recall) เป็นการวัดความถูกต้องของแบบจำลอง โดยพิจารณาแยกทีละคลาส โดยใช้เปอร์เซ็น ของ ความถูกต้อง (True Positive) แสดงได้ดังสมการที่ (4)

$$\frac{TP}{TP+FN} \tag{4}$$

ค่าความแม่นยำ (Precision) เป็นการวัดความแม่นยำของข้อมูล โดยพิจารณาแยกทีละคลาส แสดงได้ดังสมการที่ (5)

$$\frac{TP}{TP+FP} \tag{5}$$

Receiver Operating Characteristic Curve (ROC)

ROC คือกราฟที่แสดงประสิทธิภาพของแบบจำลองการจัดประเภทที่เกณฑ์การจัดประเภททั้งหมด กราฟนี้แสดงพารามิเตอร์ 2 รายการดังนี้

1. True Positive Rate (TPR) ค่าที่บอกว่าแบบจำลองทำนายได้ว่าจริง เป็นอัตราส่วนเท่าไรของจริงทั้งหมด แสดงได้ดังสมการที่ (6)

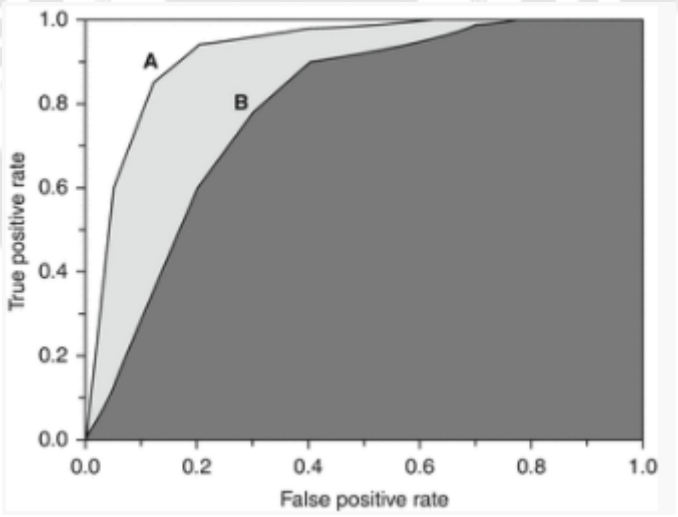
$$TPR = \frac{TP}{TP+FN} \tag{6}$$

2. False Positive Rate (FPR) ค่าที่บอกว่าแบบจำลองทำนายได้ว่าผิดจริง เป็นอัตราส่วนเท่าไรของที่ผิดจริงทั้งหมด แสดงได้ดังสมการที่ (7)

$$FPR = \frac{FP}{FP+TN} \tag{7}$$

Area Under the ROC Curve (AUC)

AUC คือพื้นที่สองมิติทั้งหมดภายใต้เส้นโค้ง ROC ทั้งหมด ดังภาพประกอบที่ 6

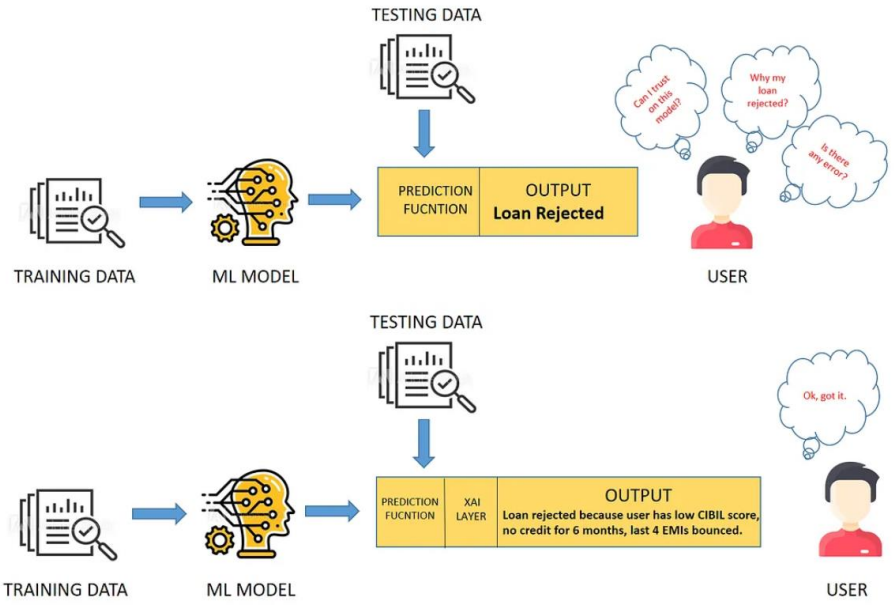


ภาพประกอบ 6 แสดงภาพ ROC และ AUC

ที่มา (Melo, 2013)

Local Interpretable Model-agnostic Explanations (LIME)

อัลกอริทึมเพื่อใช้ในการ อธิบายผลการทำนายแบบจำลองในรูปที่น่าเชื่อถือและเข้าใจได้ โดยแสดงออกมาในรูปแบบรูปภาพความสัมพันธ์ระหว่างฟีเจอร์กับข้อมูล ซึ่งสามารถอธิบายได้ทีละ 1 ข้อมูล



ภาพประกอบ 7 แสดงภาพการทำงานของ LIME

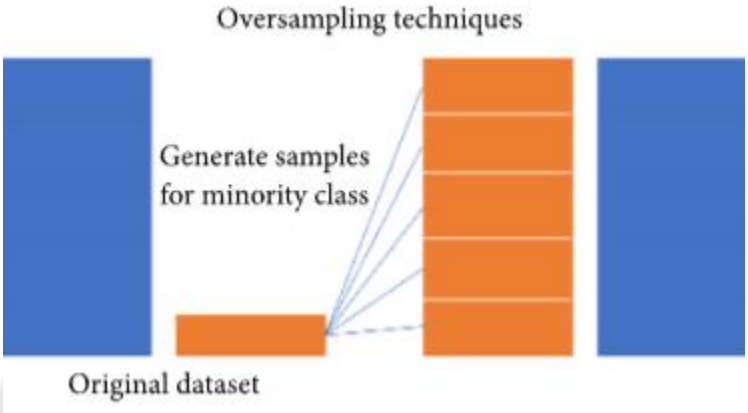
ที่มา (Solanki, 2020)

จากภาพประกอบที่ 7 แสดงการเปลี่ยนเทียบผลการทำนายระหว่างประโยชน์เมื่อนำ LIME มาใช้จะทำให้ผู้ใช้งาน (USER) สามารถเข้าใจเหตุผลที่โดนปฏิเสธการกู้สินเชื่อ เนื่องจากลูกหนี้มีคะแนน CIBIL (Credit Information Bureau) ต่ำ, ไม่มีข้อมูลเครดิตในช่วงเวลา 6 เดือน และค้างชำระ 4 เดือน ซึ่งข้อมูลดังกล่าวจะทำให้ผู้พัฒนาแบบจำลองเข้าใจการทำงานของแบบจำลองและพนักงานอนุมัติสินเชื่อสามารถตัดสินใจได้ว่าสามารถเชื่อถือแบบจำลองได้หรือไม่

4. **การจัดการกับข้อมูลที่ไม่สมดุลแบบ** Synthetic Minority Oversampling Technique (SMOTE)

เทคนิคการปรับข้อมูลกรณีข้อมูลมีความไม่สมดุลสาเหตุเนื่องจากข้อมูลระหว่างคลาสมีความแตกต่างกันอย่างมาก ถือได้ว่าเป็นเทคนิคหนึ่งสำหรับวิธีการสร้างตัวอย่างโดยการสังเคราะห์ข้อมูลขึ้นมาจากการสุ่มข้อมูลเกินขนาด (Oversampling) ดังภาพประกอบที่ 8 โดยส่วนมากจะพบใน

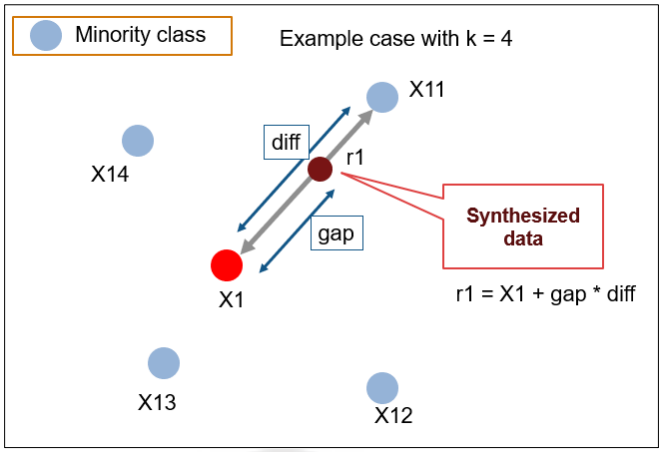
ชุดข้อมูลธุรกิจธนาคาร หากเมื่อต้องการจำแนกประเภทลูกหนี้ ระหว่างลูกหนี้ปกติกับลูกหนี้ผิดนัดชำระ เพื่อนำมาพัฒนาโมเดลจะพบปัญหาว่าชุดข้อมูลลูกหนี้ผิดนัดชำระมีปริมาณข้อมูลน้อยมาก สาเหตุเพราะหากธนาคารมีลูกหนี้ผิดนัดชำระเยอะก็จะไม่สามารถดำเนินธุรกิจได้ ดังนั้นเพื่อให้ข้อมูลมีความสมดุลจึงต้องทำการปรับข้อมูลเพื่อให้ วัตถุประสงค์เพื่อทำให้ข้อมูลมีความสมบูรณ์มากขึ้น



ภาพประกอบ 8 แสดงภาพเทคนิคก่อน-หลังการปรับข้อมูลโดยเทคนิค SMOTE

ที่มา (Le, Vo, Vo, Lee, & Baik, 2019)

เทคนิค SMOTE นิยมใช้สำหรับการทำ Oversampling โดยการเพิ่มข้อมูลในคลาสที่น้อย จำนวน 1 ค่าหลังจากนั้นพิจารณาค่าข้อมูลใกล้เคียงอีกนำนวน K ค่า (K-Nearest Neighbor) แล้วคำนวณค่าระยะทาง (Euclidean Distance) ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียง จากนั้นจึงสร้างข้อมูลเทียมระหว่างค่าของข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงตัวที่ให้ค่าระยะทางน้อยที่สุด ดังภาพประกอบที่ 9



ภาพประกอบ 9 แสดงภาพ Synthetic Minority Over-Sampling Technique

ที่มา : (SATPATHY, 2020)

5. งานวิจัยที่เกี่ยวข้อง

1. บทความวิจัยเรื่อง Predictive Analytics for Default of Credit Card Clients (Bacová & Babic , 2021)

งานวิจัยนี้ได้กล่าวถึงการทำนายลูกหนี้ที่มีโอกาสเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ ซึ่งงานวิจัยนี้ใช้ข้อมูลของผู้ใช้บัตรเครดิตจากไต้หวัน ช่วงเดือนเมษายนถึงกันยายน 2005 โดยนำข้อมูลทั้งหมดมาวิเคราะห์ข้อมูลซึ่งพิจารณาจาก Customer Value หลังนำข้อมูลมาวิเคราะห์พบว่าข้อมูลที่ใช้ระบุกลุ่มเป้าหมายลูกหนี้ (Target) มีความไม่สมดุลของข้อมูลจึงได้นำเทคนิค SMOTE Imbalance มาใช้ปรับข้อมูลให้มีความเหมาะสมก่อนนำข้อมูลมาทดสอบแบบจำลอง หลังจากนั้นจึงได้นำเทคนิค Classification Algorithm มาใช้ทั้งหมดห้าวิธีได้แก่ Random forest , Bagging , AdaBoost , XGBoost , Gradient Boosting สำหรับการวัดผลประสิทธิภาพแบบจำลองได้ใช้วิธีวัดผลทั้งหมดห้าค่าได้แก่ Accuracy , Precision , Recall , ROC , AUC ซึ่งการวัดผลแบ่งเป็นสองกลุ่มชุดข้อมูลได้แก่ ข้อมูลก่อนทำการปรับความสมดุลของข้อมูล (Original Dataset) และ ข้อมูลที่ทำการปรับความสมดุลของข้อมูลด้วยเทคนิค SMOTE (Processed Data) จากผลการทดสอบวัดประสิทธิภาพแบบจำลองด้วยวิธีก่อนและหลังปรับความสมดุลของข้อมูล ซึ่งได้ผลดังนี้เมื่อวัดผลแบบจำลองจากข้อมูล Original Dataset ได้ผลดังนี้ AdaBoost : 0.7762 และ Gradient Boosting : 0.7825 และวัดผล Processed Data ได้ผลดังนี้ AdaBoost : 0.7751 และ Gradient Boosting : 0.7828 สรุปพบว่าผลการวัดประสิทธิภาพจากค่า ROC ระหว่างข้อมูล Original Dataset และ Processed Data มีผลใกล้เคียงกัน

2. **บทความวิจัยเรื่อง** The Application of Machine Learning Algorithms Credit Card Default Prediction (Yue Yu , 2020)

งานวิจัยนี้ได้กล่าวถึงการทำนายลูกหนี้ที่มีโอกาสเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ ซึ่งงานวิจัยนี้ใช้ข้อมูลของผู้ใช้บัตรเครดิตจากไต้หวัน ช่วงเดือนเมษายนถึงกันยายน 2005 โดยนำข้อมูลทั้งหมดมาวิเคราะห์ข้อมูลซึ่งพิจารณาจาก Customer Value หลังนำข้อมูลมาวิเคราะห์พบว่าข้อมูลที่ใช้ระบุกลุ่มเป้าหมายลูกหนี้ (Target) มีความไม่สมดุลของข้อมูลจึงได้นำเทคนิค Weighted Model มาใช้ปรับข้อมูลให้มีความเหมาะสมก่อนนำข้อมูลมาทดสอบแบบจำลอง หลังจากนั้นจึงได้นำเทคนิค Classification Algorithm มาใช้ทั้งหมดสี่วิธีได้แก่ Logistic Regression , Decision Tree, AdaBoost , Random Forest สำหรับการวัดผลประสิทธิภาพแบบจำลองได้ใช้วิธีวัดผลทั้งหมดสิงค่าได้แก่ Accuracy , Precision ซึ่งการวัดผลแบ่งเป็นสองกลุ่มชุดข้อมูลได้แก่ ข้อมูลก่อนทำการปรับความสมดุลของข้อมูล และ ข้อมูลที่ทำการปรับความสมดุลของข้อมูลด้วยเทคนิค Weighted Model จากผลการทดสอบวัดประสิทธิภาพแบบจำลองด้วยวิธีก่อนและหลังปรับความสมดุลของข้อมูล Random Forest สามารถทำนายได้ดี Accuracy : 99.27% ทั้ง 2 แบบไม่ว่าจะเป็นข้อมูลก่อนทำการปรับความสมดุลของข้อมูล (Original Dataset) หรือข้อมูลที่ทำการปรับความสมดุลของข้อมูลด้วยเทคนิค Weighted Model

3. **บทความวิจัยเรื่อง** Loan Repayment Behavior Prediction of Provident Fund Users Using a Stacking-Based Model (Liling Ke **และทีม** , 2021)

งานวิจัยนี้ได้กล่าวถึงการศึกษาพฤติกรรมการชำระเงินกู้ โดยผู้วิจัยได้ใช้ข้อมูลรายการชำระเงินจากสินเชื่อที่อยู่อาศัยของประเทศจีน ซึ่งชุดข้อมูลดังกล่าวเป็นข้อมูลเป็นที่เข้าชั้นความลับจึงไม่สามารถเปิดเผยที่มาของข้อมูลได้โดยงานวิจัยนี้พบปัญหาลูกหนี้ไม่ชำระเงินกู้ตามระยะเวลาที่กำหนดทำให้เกิดปัญหาต่อการดำเนินธุรกิจ 2 เรื่องหลักคือ

1. ลูกหนี้ชำระเงินกู้ก่อนกำหนดตามสัญญาเงินกู้ ซึ่งส่งผลทำให้ธนาคารขาดรายได้ดอกเบี้ยที่คาดว่าจะได้รับ
2. กระแสเงินสดที่ไม่สามารถคาดการณ์ได้ส่งผลกระทบต่อการจัดการสภาพคล่องของสินทรัพย์ในการดำเนินธุรกิจ

โดยวัตถุประสงค์ของงานวิจัยเพื่อแยกประเภทของลูกหนี้ออกเป็น 3 แบบได้แก่

1. ลูกหนี้ชำระเงินกู้ตามระยะเวลาที่กำหนด (Class : 0)
2. ลูกหนี้ชำระเงินกู้ก่อนระยะเวลาที่กำหนด (Class : 1)
3. ลูกหนี้ชำระเงินกู้เกินระยะเวลาที่กำหนด (Class : 2)

ทั้งนี้การแยกกลุ่มลูกหนี้ออกเป็น 3 กลุ่มเพื่อให้ธุรกิจสามารถเตรียมแผนในการจัดการสินทรัพย์ของธุรกิจ และเตรียมความพร้อมจัดการความเสี่ยงด้านสภาพคล่องเพื่อให้ธุรกิจสามารถดำเนินงานต่อไปได้อย่างต่อเนื่อง

ปัญหาของชุดข้อมูลที่พบว่าข้อมูลที่ใช้ในการพัฒนาแบบจำลองนั้นมีความไม่สมดุลโดยข้อมูล Class : 1 มีจำนวนมากเกินไปซึ่งแตกต่างจาก Class : 2 ที่มีจำนวนน้อยมาก ดังนั้นผู้วิจัยจึงได้ใช้เทคนิคปรับความไม่สมดุลของข้อมูลดังต่อไปนี้

Class 0 และ Class 1 มีความไม่สมดุลของข้อมูลผู้วิจัยจึงได้ใช้เทคนิค Undersampling เนื่องจากข้อมูลมากเกินไป โดยการปรับขนาดของข้อมูลให้เป็น 20,000 ชุดข้อมูล

จัดการ Class 2 โดยใช้เทคนิค SMOTE ด้วยการใช้ Overampling เนื่องจากข้อมูลน้อยเกินไป

หลังจากทำการปรับข้อมูลเรียบร้อยแล้วผู้วิจัยได้นำเทคนิค Classification Algorithm มาใช้ทั้งหมดเจ็ดวิธีได้แก่ Logistic Regression , Random Forest , AdaBoost , XGBoost , CatBoost , LightGBM , Stacking Model สำหรับการวัดผลประสิทธิภาพแบบจำลองนั้น ผู้วิจัยวัดผลการวัดประสิทธิภาพแบบจำลองทั้งหมดห้าค่าได้แก่ Accuracy Score , Recall , F1 , AUC , Kappa Score ซึ่งได้ผลคือ Stacking Model มีประสิทธิภาพที่ดีเมื่อเปรียบเทียบกับอีก 6 แบบจำลอง โดยเมื่อวัดประสิทธิภาพแบบจำลองพบว่าค่า AUC มีค่าใกล้เคียงถึง 0.95

4. บทความวิจัยเรื่อง Machine Learning Models for Mortgage Default Prediction in Pakistan (2021 , Kamran Meer)

งานวิจัยนี้ได้กล่าวถึงการทำนายลูกหนี้ที่มีโอกาสเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ ซึ่งงานวิจัยนี้ใช้ข้อมูลของลูกหนี้สินเชื่อที่อยู่อาศัยของธนาคารกลางของรัฐปากีสถาน ช่วงเดือนมกราคม 2017 ถึงมิถุนายน 2020 โดยนำข้อมูลทั้งหมด 5,960 Records มาวิเคราะห์ข้อมูลซึ่งพิจารณาจากตัวแปรทั้งหมด 13 ตัวแปรได้แก่ Loan defaulted or repaid , Amount of loan approved , Amount due on the existing mortgage , Current value of the property เป็นต้น จากการสำรวจข้อมูลผู้วิจัยพบว่าตัวแปร Loan defaulted or repaid สามารถนำมาใช้เป็นเป้าหมายในการทำนายได้ (Target) ซึ่งสามารถแบ่งได้ 2 ประเภทดังนี้

Repaid (Negative Class): 4,771 Records (80.05%)

Loan Defaulted (Positive Class) : 1,189 Records (19.95%)

โดยผู้วิจัยได้นำเทคนิค Classification Algorithm มาใช้ทั้งหมดสี่วิธีได้แก่ โดย แบบจำลอง Logistic Regression จะแบ่งออกเป็น 2 รูปแบบ ได้แก่ Logistic Regression (L1 Regularization) กำหนดขนาดข้อมูลแบบดั้งเดิม และ Logistic Regression (L2 Regularization) ปรับขนาดข้อมูล โดยทั้ง 2 รูปแบบได้ปรับพารามิเตอร์กำหนดค่า C = 0.05 (จากค่าตั้งต้น C= 1) เพื่อใช้ในการทดสอบแบบจำลอง , แบบจำลอง Random Forest ปรับจำนวน Tree = 300 เพื่อใช้ในการทดสอบแบบจำลอง , แบบจำลอง Gradient boosting ปรับจำนวน Tree = 300 และปรับระดับความลึกของต้นไม้ (Max Depth) = 5 เพื่อใช้ในการทดสอบแบบจำลอง ซึ่งข้อมูลถูกแบ่งอัตราส่วน 70:30 โดยแบ่งข้อมูลเพื่อใช้ในการทดสอบแบบจำลอง (Training Set : 70) และข้อมูลสำหรับวัดประสิทธิภาพแบบจำลอง (Test Set : 30) การวัดผลประสิทธิภาพแบบจำลองได้ใช้วิธีวัดผลทั้งหมดสี่ค่าได้แก่ True Positive Rate , False Positive Rate , AUROC ซึ่งผลการวัดประสิทธิภาพแบบจำลองพบว่า Random Forest และ Gradient Boosting มีประสิทธิภาพในการทำนาย (AUROC : 0.96)

5 . บทความวิจัยเรื่อง Predictive Analytics for Loan Default in Banking Sector Using Machine Learning Techniques (2018 , Salma Khaled Shaheen & Essam ElFakharany)

งานวิจัยนี้ได้ล่าวถึงการทำนายลูกหนี้ที่มีโอกาสเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ ซึ่งงานวิจัยนี้ใช้ข้อมูลลูกหนี้ธนาคารอียิปต์ ช่วงเดือนพฤษภาคม 2005 จนถึง ธันวาคม 2017 โดยมีข้อมูลทั้งหมด 2,954,168 แถว ซึ่งชุดข้อมูลดังกล่าวเป็นข้อมูลชั้นความลับจึงไม่สามารถเปิดเผยข้อมูลได้ โดยเป้าหมายของงานวิจัยนี้เพื่อจำแนกลูกหนี้ออกเป็น 2 ประเภทโดยผู้วิจัยได้นำแบบจำลองเพื่อใช้ในการทำนายผลมาใช้ทั้งหมดสี่วิธีได้แก่ K-NN , Logistic Regression , Random Forest , Gradient Boosting อัตราส่วนใช้ทดสอบแบบจำลอง 70:30 ซึ่งผลการวัดประสิทธิภาพแบบจำลองพบว่า Random Forest มีค่า Accuracy : 91.7% และ Precision : 95.83% และ Gradient Boosted มีค่า Accuracy : 91.7% และ Precision : 95.83%

6 . บทความวิจัยเรื่อง Loan Default Prediction with Machine Learning Techniques (2020 , LiLi Lai)

งานวิจัยนี้ได้กล่าวถึงการพัฒนาแบบจำลองโดยใช้เทคนิคการเรียนรู้ของเครื่อง วัตถุประสงค์ของงานวิจัยเพื่อให้แบบจำลอง Adaboost สามารถทำนายได้แม่นยำที่สุด ซึ่งงานวิจัยนี้ใช้ข้อมูลลูกหนี้ของ Xiamen International Bank องค์ประกอบสำหรับของข้อมูลเพื่อนำมาวิจัยมีทั้งหมด 3 ตาราง ได้แก่ 1. ข้อมูลลูกหนี้ (User Attributes) 2. ข้อมูลการกู้ยืม (Lending Related Information) 3. ข้อมูลเครดิตลูกหนี้ (Information Related to User Credit Reporting) โดยข้อมูลเครดิตลูกหนี้ทางผู้วิจัยไม่สามารถนำมาใช้ในการพัฒนาแบบจำลองได้เนื่องจากเป็นข้อมูลส่วนบุคคลที่มีความอ่อนไหว ผู้วิจัยได้นำเทคนิค Classification Algorithm มาใช้ทั้งหมดห้าวิธีได้แก่ XGBoost, Random Forest (RF), AdaBoost, K Nearest Neighbors (KNN), Multilayer Perceptrons (MLP) และทุกๆแบบจำลองผู้วิจัยได้ใช้ Hyper Parameter Search เพื่อหาค่าที่เหมาะสมที่สามารถทำให้แบบจำลองสามารถทำนายได้แม่นยำที่สุดซึ่งผลการวัดประสิทธิภาพแบบจำลองพบว่าผู้วิจัยสามารถปรับ ค่า Parameter เพื่อทำให้แบบจำลองมี ค่า Accuracy = 100% โดยพบว่าแบบจำลอง Adaboost ที่กำหนด base_estimator_max_depth : 20 และ n_estimators : 100 มีความแม่นยำ 100%

7 . **บทความวิจัยเรื่อง** Prediction of the Borrowers' Payback to the Loan with Lending Club Data (2020 , Xiaoqi Sun)

งานวิจัยนี้ได้กล่าวถึงการทำนายว่าผู้กู้จะชำระคืนเงินกู้หรือไม่ โดยเป้าหมายของแบบจำลองเพื่อให้ได้ผลลัพธ์ไบนารีคลาส ซึ่งงานวิจัยนี้ใช้ข้อมูลจากการให้กู้ยืมแบบ peer-to-peer (P2P) ของ Lending Club Data จาก Kaggle.com มีข้อมูลทั้งหมด 890,000 แถว และ 145 ตัวแปรระยะเวลาของข้อมูลช่วงเดือนมกราคม 2007 ถึงมิถุนายน 2015 โดยงานวิจัยได้ทำกระบวนการทำความสะอาดเพื่อหลีกเลี่ยงข้อมูลที่ไม่สมบูรณ์ ซึ่งได้เลือกข้อมูลที่นำมาทดสอบแบบจำลอง 1,500 รายการจากข้อมูลทั้งหมด ข้อมูลตัวแปรประเภทตัวเลข 10 ตัวแปร และ ข้อมูลตัวแปรประเภทหมวดหมู่ 8 ตัวแปร และเพื่อนำตัวแปรมาเป็นเครดิตให้กับลูกหนี้ได้แก่ ข้อมูลตัวแปร CIBIL Score (Credit History) , มูลค่าธุรกิจ (Business Value) , ทรัพย์สินลูกหนี้ (Assets of Customer) ฯลฯ หากพบข้อมูลที่มีค่าว่าง จะทดแทนด้วย ค่ากลาง ค่ามัธยฐาน และค่าฐานนียม เนื่องจากด้วยจำนวนข้อมูลที่น้อยจึงเป็นสาเหตุที่ทางผู้วิจัยไม่สามารถตัดข้อมูลทิ้งได้ การทดสอบแบบจำลองจะแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ 80:20 หรือ 70:30 โดยส่วนหลักจะใช้ในการทดสอบแบบจำลองและส่วนรองจะใช้ในการวัดประสิทธิภาพแบบจำลองผู้วิจัยได้แบบจำลอง Logistic Regression, SVM, KNN เพื่อนำมาใช้ในการทำนายและ วัดผลการวัดประสิทธิภาพแบบจำลองโดยห้าใช้วิธี Confusion Metrics , Accuracy , Precision , Recall , F1 Score โดยผลการทดสอบแบบจำลองพบว่าค่าความแม่นยำ (Accuracy) = 81% ซึ่งแบบจำลองสามารถทำนายการอนุมัติและการปฏิเสธการอนุมัติสินเชื่อได้อย่างเหมาะสมตามหลักการธุรกิจธนาคารอีกด้วย เช่น แบบจำลองจะอนุมัติสินเชื่อแก่ผู้ที่มีรายได้มากและผู้ที่ขออนุมัติสินเชื่อที่มียอดวงเงินต่ำจะมีโอกาสกู้สูง

ตาราง 1 สรุปงานวิจัยที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัยและผู้วิจัย	วัตถุประสงค์	กลุ่มตัวอย่าง	การเลือกตัวแปร	แบบจำลอง	การวัดประสิทธิภาพ	ผลของการวัด
1	ผู้วิจัย Bacová & Babic ชื่องานวิจัย Predictive Analytics for Default of Credit Card Clients	ทำนายผู้กู้ออกเป็น 2 กลุ่ม	ข้อมูลทั้งหมด : 30,000 แถว	ตัวแปรทั้งหมด : 23 ตัวแปร	จัดเตรียมข้อมูล 2 กลุ่ม ได้แก่ 1. Original Dataset : ข้อมูลปกติ 2. Processed Data : ข้อมูลผ่านกระบวนการเตรียมข้อมูล เช่น การทำ Imbalance SMOTE Classification Random forest , Bagging , AdaBoost , XGBoost , Gradient Boosting	Accuracy Precision Recall ROC AUC	Original Dataset AdaBoost : 0.7762 Gradient Boosting : 0.7825 Processed Data AdaBoost : 0.7751 Gradient Boosting : 0.7828

ตาราง 1 (ต่อ)

ลำดับ	ชื่องานวิจัยและผู้วิจัย	วัตถุประสงค์	กลุ่มตัวอย่าง	การเลือกตัวแปร	แบบจำลอง	การวัดประสิทธิภาพ	ผลของการวัด
2	ผู้วิจัย Yue Yu ชื่องานวิจัย The Application of Machine Learning Algorithms in Credit Card Default Prediction	แยกประเภทของลูกหนี้ออกเป็น 2 กลุ่ม	ข้อมูลทั้งหมด : 30,000 แถว	ตัวแปรทั้งหมด 23 ตัวแปร	Imbalance Weighted Model (1 : 3.5) Classification Logistic Regression , Decision Tree , AdaBoost , Random Forest	Accuracy Precision	Random Forest สามารถทำนายได้ดี Accuracy : 99.27% ทั้ง 2 แบบไม่ว่าจะเป็นข้อมูลดังเดิมหรือผ่านกระบวนการ Weigted

ตาราง 1 (ต่อ)

ลำดับ	ชื่องานวิจัยและผู้วิจัย	วัตถุประสงค์	กลุ่มตัวอย่าง	การเลือกตัวแปร	แบบจำลอง	การวัดประสิทธิภาพ	ผลของการวัด
3	ผู้วิจัย Liling Ke และทีม ชื่องานวิจัย Loan Repayment Behavior Prediction of Provident Fund Users Using a Stacking-Based Model	แยกประเภทของลูกหนี้ออกเป็น 3 แบบได้แก่ 1. ลูกหนี้ชำระเงินกู้ตามระยะเวลาที่กำหนด (Class : 0) 2. ลูกหนี้ชำระเงินกู้ก่อนระยะเวลาที่กำหนด (Class : 1) 3. ลูกหนี้ชำระเงินกู้เกินระยะเวลาที่กำหนด (Class : 2)	ไม่เปิดเผย	ตัวแปรทั้งหมด 31 ตัวแปร	Imbalance Random Undersampling Algorithm : class 0 , class 1 Oversampling (SMOTE) : class 2 Classification Logistic Regression , Random Forest , AdaBoost , XGBoost , CatBoost , LightGBM , Stacking Model	Accuracy Score Recall F1 AUC Kappa Score	Stacking Model ใกล้เคียง 0.95

ตาราง 1 (ต่อ)

ลำดับ	ชื่องานวิจัยและผู้วิจัย	วัตถุประสงค์	กลุ่มตัวอย่าง	การเลือกตัวแปร	แบบจำลอง	การวัดประสิทธิภาพ	ผลของการวัด
4	ผู้วิจัย Kamran Meer ชื่องานวิจัย Machine Learning Models for Mortgage Default Prediction in Pakistan	แยกประเภทของลูกหนี้ออกเป็น 2 กลุ่ม	ข้อมูลทั้หมด 5,960 แถว	ตัวแปรทั้งหมด 13 ตัวแปร	Classification Logistic Regression , Random Forest , Gradient Boosting	True Positive Rate False Positive Rate AUROC	Random Forest และ Gradient Boosting มีประสิทธิภาพในการทำนาย (AUROC : 0.96)
5	ผู้วิจัย Salma Khaled Shaheen & Essam ElFakharany ชื่องานวิจัย Predictive Analytics for Loan Default in Banking Sector Using Machine Learning Techniques	- แยกประเภทของลูกหนี้ออกเป็น 2 กลุ่ม - เพื่อให้ AdaBoosts ทำนายได้แม่นยำที่สุดด้วยการปรับ Hyperparameter	ข้อมูลทั้หมด 132,029 แถว	ตัวแปรทั้งหมด 10 ตัวแปร	Classification XGBoost, Random Forest (RF), AdaBoost , K Nearest Neighbors (KNN), Multilayer Perceptrons(MLP)	Confusion Matrix 6 ค่า ได้แก่ Accuracy Precision Sensitivity (TP Rate) Specificity (TP Rate) Fall Out (FP Rate) Miss Rate (FN Rate)	Random Forest : 91.7% Accuracy and 95.83% Precision Gradient Boosted : 91.9% Accuracy and 79.68%

ตาราง 1 (ต่อ)

ลำดับ	ชื่องานวิจัยและผู้วิจัย	วัตถุประสงค์	กลุ่มตัวอย่าง	การเลือกตัวแปร	แบบจำลอง	การวัดประสิทธิภาพ	ผลของการวัด
6	ผู้วิจัย Lili Lai ชื่องานวิจัย Loan Default Prediction with Machine Learning Techniques	- แยกประเภทของลูกหนี้ออกเป็น 2 - เพื่อให้ AdaBoosts ทำนายได้แม่นยำที่สุดด้วยการปรับ Hyperparameter	ข้อมูลทั้หมด : 132,029 แถว	ตัวแปรทั้งหมด :	Classification XGBoost, Random Forest (RF), AdaBoost, K Nearest Neighbors (KNN), Multilayer Perceptrons(MLP)	ROC AUC Accuracy	Adaboost ที่กำหนด base_estimator_max_depth : 20 และ n_estimators : 100 มีความแม่นยำ 100%
7	ผู้วิจัย Xiaoqi Sun ชื่องานวิจัย Prediction of the Borrowers' Payback to the Loan with Lending Club Data	- แยกประเภทของลูกหนี้ออกเป็น 2	ข้อมูลทั้หมด : 1,500 แถว	ตัวแปรทั้งหมด : 12 ตัวแปร	Feature Engineering Imputation Handling Outliers Binning Log Transform One Hot Encoding Classification Logistic Regression , Random Forest , Decision Tree Model	Confusion Metrics Accuracy Precision Recall F1 Score	ผลการทดสอบแบบจำลองพบว่าค่าความแม่นยำ (Accuracy) = 81%

บทที่ 3
การดำเนินงานวิจัย

ในงานวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

3.1 กระบวนการทำงานของแบบจำลอง

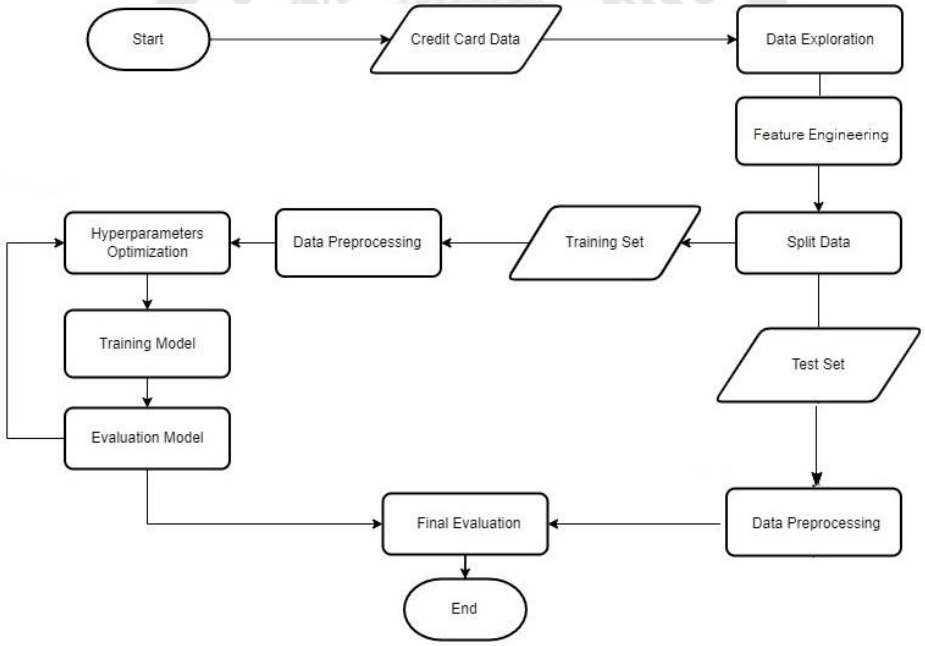
3.2 การเก็บรวบรวมข้อมูล (Data Collection)

3.3 การสำรวจข้อมูล (Exploratory Data Analysis: EDA)

3.4 การเตรียมข้อมูล (Data Preparation)

3.5 การสร้างแบบจำลองพยากรณ์

1. กระบวนการทำงานของแบบจำลอง



ภาพประกอบ 10 แสดงกระบวนการทำงานของแบบจำลอง

จากภาพประกอบที่ 10 ได้อธิบายถึงกระบวนการสร้างแบบจำลองการจำแนกประเภทและการวัดผลลัพธ์จากการวิเคราะห์ข้อมูล โดยเริ่มจากขั้นตอนการนำเข้าข้อมูล (Credit Card Data) การตรวจสอบข้อมูล การพิจารณานำข้อมูลมาใช้ในการวิเคราะห์ (Data Exploration) เพื่อทำความเข้าใจข้อมูลที่นำมาใช้มาพัฒนาแบบจำลอง จากนั้นจะทำการวิศวกรรมคุณลักษณะ (Feature Engineering) โดยการจัดการกับข้อมูลที่มีค่าว่าง จัดกลุ่มข้อมูลให้มีความเหมาะสม และตัดฟีเจอร์ที่

มีความซ้ำซ้อน จากนั้นแบ่งข้อมูล (Split Data) ออกเป็นข้อมูลฝึกสอน (Train Data) และข้อมูลทดสอบ (Test Data) และทำการแปลงข้อมูล (Pre-Processing) ที่ได้ทำการเก็บรวบรวมมาให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ได้ ที่ข้อมูลฝึกสอนเท่านั้นเพื่อทำให้ข้อมูลอยู่ในรูปแบบที่ง่ายต่อการทำงานและประสิทธิภาพของแบบจำลอง กระบวนการที่ใช้คือการเปลี่ยนข้อมูลให้อยู่ในรูปแบบของตัวเลข การจัดการกับข้อมูลที่ไม่สมดุลด้วยการสังเคราะห์ข้อมูลในกลุ่มที่น้อยให้เพิ่มขึ้นและการปรับให้สเกลของข้อมูลอยู่ในช่วงใกล้เคียงกัน จากนั้นนำไปสร้างและทดสอบประสิทธิภาพแบบจำลอง

ขั้นตอนการสร้างแบบจำลองในการทำนาย ได้สร้างแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) สองแบบเพื่อเปรียบเทียบประสิทธิภาพกัน คือ แบบจำลอง Logistic Regression เปรียบเทียบกับแบบจำลอง XGBoost และ CatBoost เพื่อจำแนกประเภทลูกหนี้และทำการปรับจูนพารามิเตอร์ (Hyper parameter Optimization) เพื่อเพิ่มประสิทธิภาพของแบบจำลอง จากนั้นทำการประเมินผลประสิทธิภาพของแบบจำลอง (Evaluation Model) และขั้นตอนสุดท้ายขั้นตอนการนำเสนอผลการวัดประสิทธิภาพแบบจำลองในการทำนายลูกหนี้ไม่ก่อให้เกิดรายได้ ผ่านข้อมูลทดสอบ (Test Data) โดยใช้ค่า Confusion Matrix, ROC, AUC, Accuracy, Precision, Recall, F1-Score

2. การเก็บรวบรวมข้อมูล (Data Collection)

ในงานวิจัยนี้ได้ใช้การทำธุรกรรมของสินเชื่อบัตรเครดิต จากเว็บ Kaggle.com โดยมีทั้งหมด 2 ตารางประกอบด้วย ตาราง ข้อมูลลูกหนี้ มีจำนวนตัวแปรทั้งหมด 17 ตัวแปร และมีข้อมูลจำนวนทั้งหมด 438,557 แถว ดังตาราง 2 และตาราง รายการธุรกรรม มีจำนวนตัวแปรทั้งหมด 2 ตัวแปร และมีข้อมูลจำนวนทั้งหมด 1,048,575 แถว ดังตาราง 3

ตาราง 2 แสดงตัวแปรของข้อมูลลูกหนี้ที่ใช้สำหรับพัฒนาแบบจำลอง

Field Name	Data Type	คำอธิบายข้อมูล
ID	Int64	รหัสลูกหนี้
CODE_GENDER	Object	เพศ (M = ชาย, F = หญิง)
FLAG_OWN_CAR	Object	Flag ระบุรายละเอียดลูกหนี้มีรถหรือไม่ (Y = มี , N = ไม่มี)
FLAG_OWN_REALTY	Object	Flag ระบุรายละเอียดลูกหนี้มีอสังหาริมทรัพย์หรือไม่ (Y = มี , N = ไม่มี)
CNT_CHILDREN	Int64	จำนวนบุตร
AMT_INCOME_TOTAL	Float64	รายได้ต่อปี
NAME_INCOME_TYPE	Object	ประเภทรายได้
NAME_EDUCATION_TYPE	Object	ระดับการศึกษา
NAME_FAMILY_STATUS	Object	สถานภาพการสมรส
NAME_HOUSING_TYPE	Object	ประเภทที่อยู่อาศัย
DAYS_BIRTH	Int64	วันเกิด (Date)
DAYS_EMPLOYED	Int64	วันที่เริ่มทำงาน (Date)
FLAG_MOBIL	Object	Flag ระบุรายละเอียดลูกหนี้มีโทรศัพท์มือถือหรือไม่ (Y = มี , N = ไม่มี)
FLAG_WORK_PHONE	Object	Flag ระบุรายละเอียดลูกหนี้มีเบอร์โทรที่ทำงานหรือไม่ (Y = มี , N = ไม่มี)
FLAG_PHONE	Object	Flag ระบุรายละเอียดลูกหนี้มีเบอร์โทรศัพท์หรือไม่ (Y = มี , N = ไม่มี)
FLAG_EMAIL	Object	Flag ระบุรายละเอียดลูกหนี้มี E-mail หรือไม่ (Y = มี , N = ไม่มี)
OCCUPATION_TYPE	Object	อาชีพ
CNT_FAM_MEMBERS	Float64	จำนวนสมาชิกครอบครัว

ตาราง 3 แสดงตัวแปรของข้อมูลรายกาธุรกรรมที่ใช้สำหรับพัฒนาแบบจำลอง

Field Name	Data Type	**คำอธิบายข้อมูล**
ID	Int64	รหัสลูกหนี้
MONTHS_BALANCE	Int64	จำนวนเดือนสะสม
STATUS	Object	สถานะ

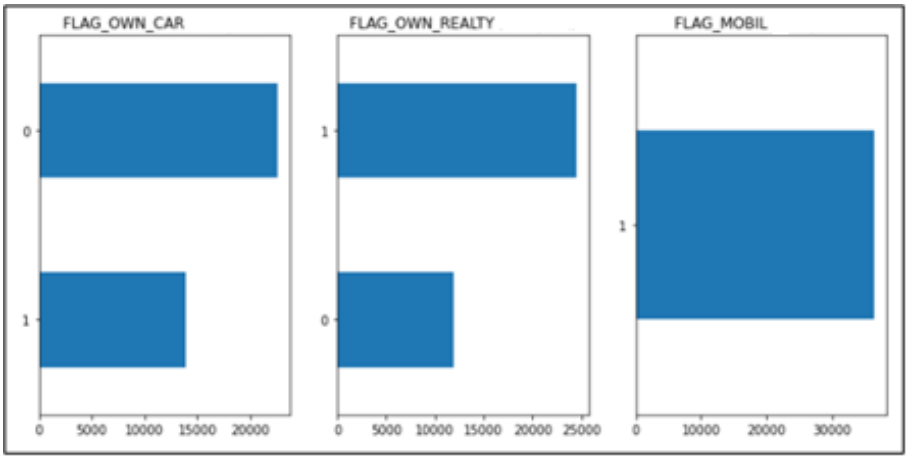
3. การสำรวจข้อมูล (Exploratory Data Analysis: EDA)

งานวิจัยนี้ใช้ภาษาไพทอน (Python) ในการวิเคราะห์ข้อมูลและทำ Machine Learning เริ่มต้นด้วยการนำเข้าโมดูลสำคัญสำหรับการสร้างแบบจำลองต่อมานำเข้าข้อมูลและข้อมูลที่ใช้สำหรับสร้างแบบจำลอง ดังภาพประกอบที่ 11

ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS
5008804	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage
5008805	M	Y	Y	0	427500.0	Working	Higher education	Civil marriage
5008806	M	Y	Y	0	112500.0	Working	Secondary / secondary special	Married
5008808	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married
5008809	F	N	Y	0	270000.0	Commercial associate	Secondary / secondary special	Single / not married

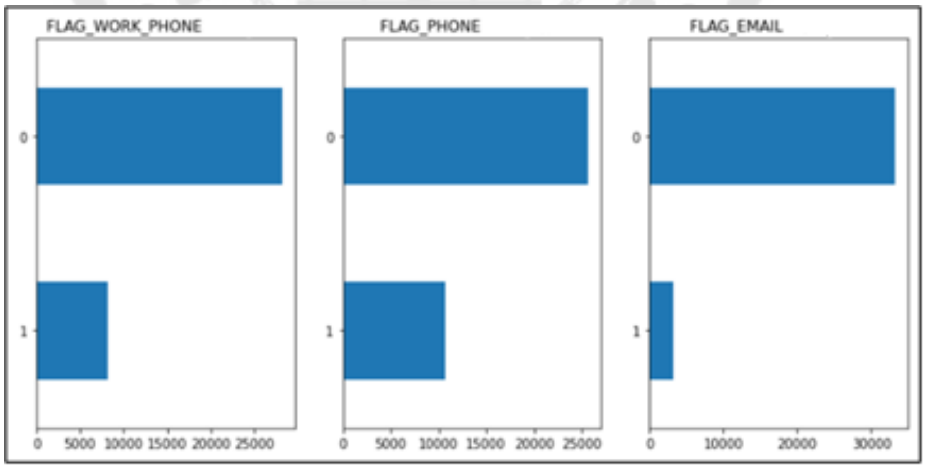
ภาพประกอบ 11 ตัวอย่างตารางข้อมูลเพื่อใช้ในการพัฒนาแบบจำลอง

เริ่มกระบวนการ EDA เพื่อหาข้อมูลเชิงลึกจากข้อมูลตาราง application_record (ตารางข้อมูลลูกหนี้) และ credit_record (ตารางข้อรายการธุรกรรม)



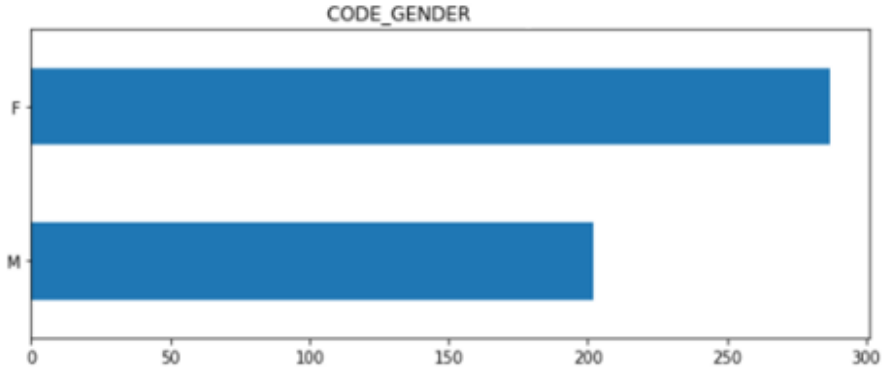
ภาพประกอบ 12 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์

Flag ระบุรายละเอียดลูกหนี้มีรถ , Flag ระบุรายละเอียดลูกหนี้มีอสังหาริมทรัพย์ , Flag ระบุรายละเอียดลูกหนี้มีโทรศัพท์มือถือ ตามลำดับ จากภาพประกอบที่ 12 พบว่า FLAG_MOBIL ไม่สามารถแบ่งประเภทได้จึงไม่สามารถนำ Feature ดังกล่าวมาใช้ในการพัฒนาแบบจำลอง ดังนั้นผู้วิจัยจึงได้ทำการตัดตัวแปรดังกล่าวออกจากทดสอบแบบจำลอง



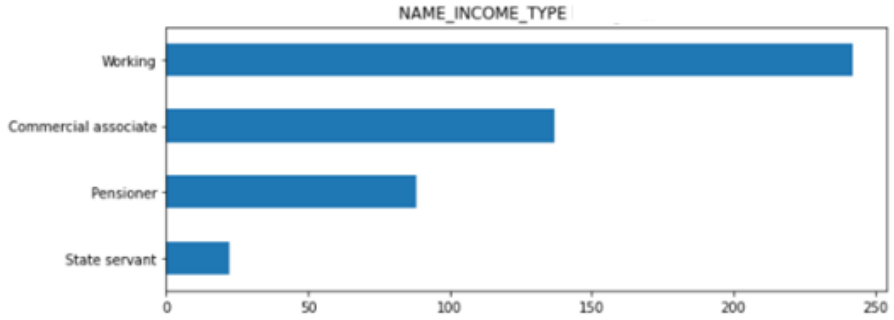
ภาพประกอบ 13 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์

Flag ระบุรายละเอียดลูกหนี้มีเบอร์โทรที่ทำงาน , Flag ระบุรายละเอียดลูกหนี้มีเบอร์โทรศัพท์, Flag ระบุรายละเอียดลูกหนี้มี E-mail ตามลำดับจากภาพประกอบที่ 13 โดยส่วนมากลูกหนี้ไม่มีโทรศัพท์ที่ทำงาน ที่บ้าน และ Email



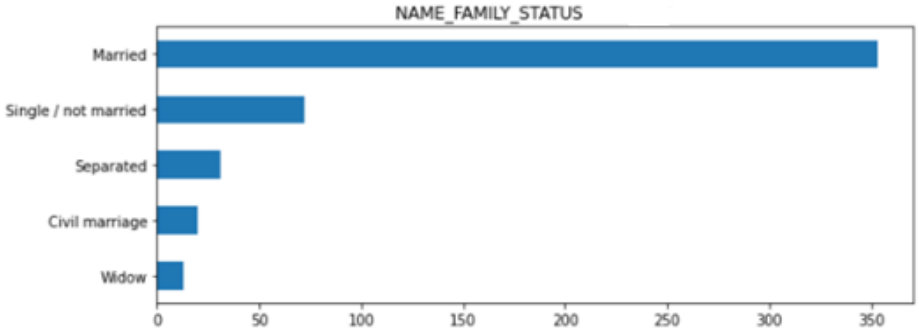
ภาพประกอบ 14 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์เพศ

จากภาพประกอบที่ 14 พบว่าลูกหนี้สินเชื่อบัตรเครดิตมีจำนวนเพศหญิงมากกว่าเพศชาย



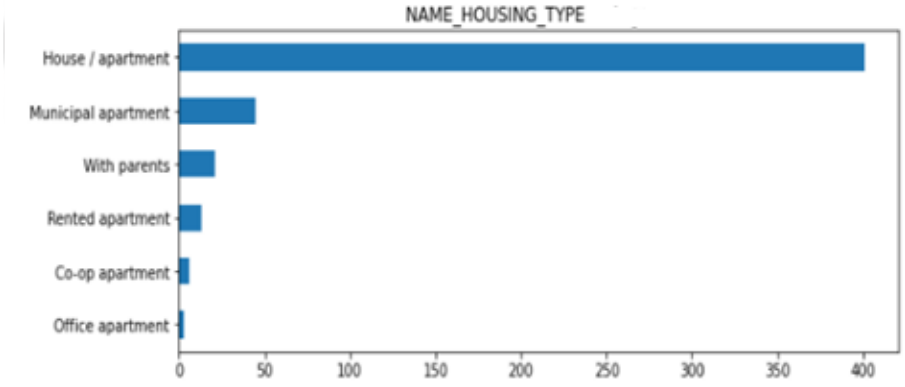
ภาพประกอบ 15 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์ประเภทรายได้

จากภาพประกอบที่ 15 พบว่าลูกหนี้โดยส่วนมากเป็นทำงานซึ่งแตกกันกันแต่ละอาชีพ



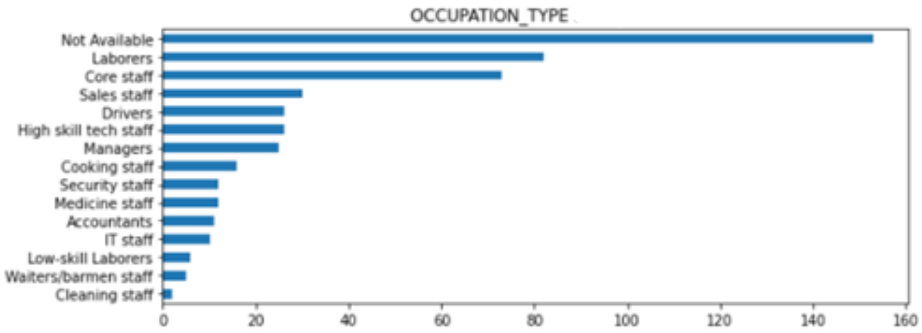
ภาพประกอบ 16 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์สถานภาพการสมรส

จากภาพประกอบที่16 พบว่าลูกหนี้โดยส่วนมากแต่งงานแล้วมีเพียงส่วนน้อยเท่านั้นที่สถานะโสด หรือเป็นหม้าย



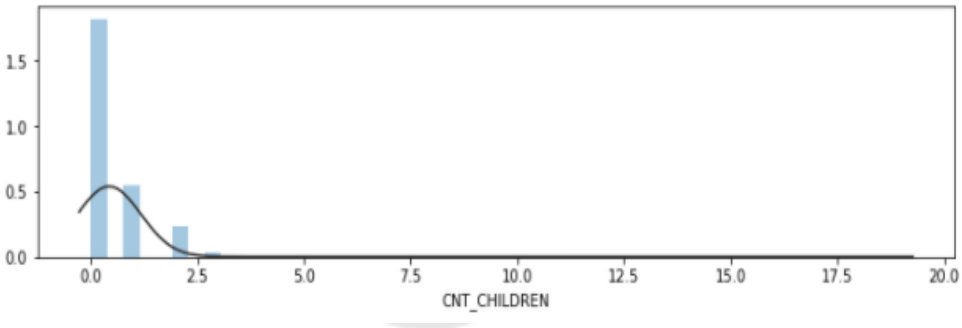
ภาพประกอบ 17 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์ประเภทที่อยู่อาศัย

จากภาพประกอบที่ 17 แสดงข้อมูลจำนวนประเภทที่อยู่อาศัยโดยส่วนมากลูกหนี้อาศัยอยู่บ้านหรือพาร์ตเมนต์



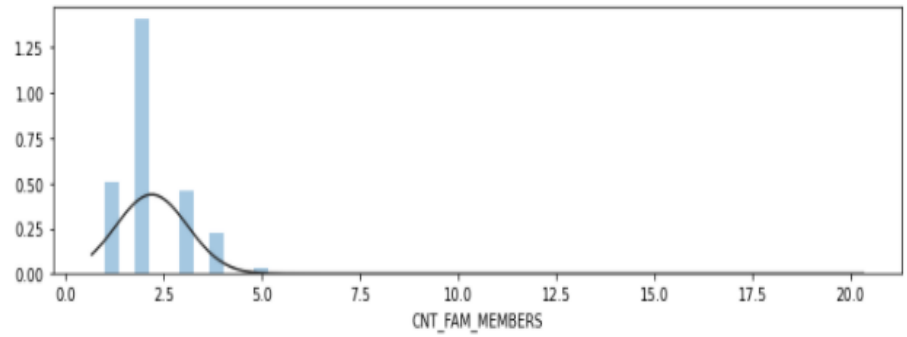
ภาพประกอบ 18 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์อาชีพ

จากภาพประกอบที่ 18 ลูกหนี้ที่มีอาชีพที่ไม่สามารถระบุได้ (Not Available) มีจำนวนมากที่สุด แต่หากเฉลี่ยรายได้ของกลุ่มอาชีพนี้แล้วพบว่า กลุ่มลูกหนี้ดังกล่าวมีรายได้ที่ไม่สูงมากนัก ดังรูปภาพที่ 28 กราฟระหว่างค่าเฉลี่ยรายได้กับอาชีพโดยเรียงข้อมูลค่าเฉลี่ยรายได้จากมากไปน้อย ที่พบว่ารายได้ที่มากที่สุดคือ Managers



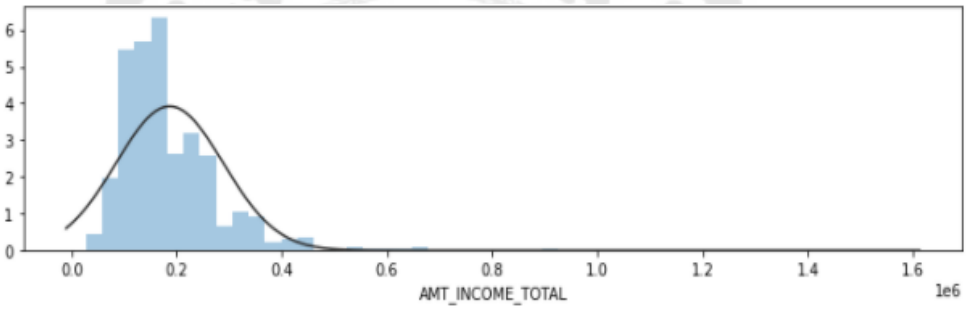
ภาพประกอบ 19 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์จำนวนบุตร

จากภาพประกอบที่ 19 จำนวนบุตรของลูกหนี้อยู่ช่วง 0 – 3 คน



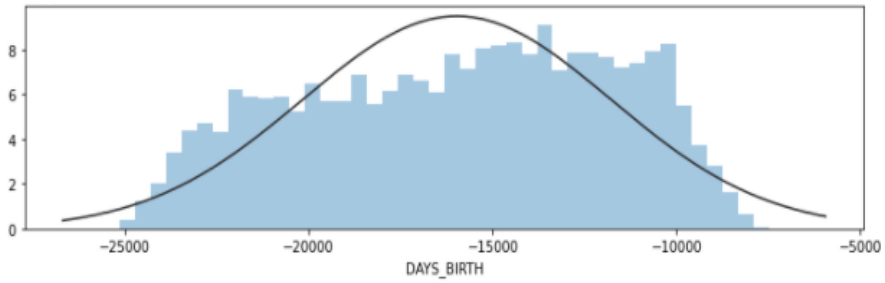
ภาพประกอบ 20 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์จำนวนสมาชิกครอบครัว

จากภาพประกอบที่ 20 จำนวนสมาชิกครอบครัวอยู่ช่วง 1 – 5 คน



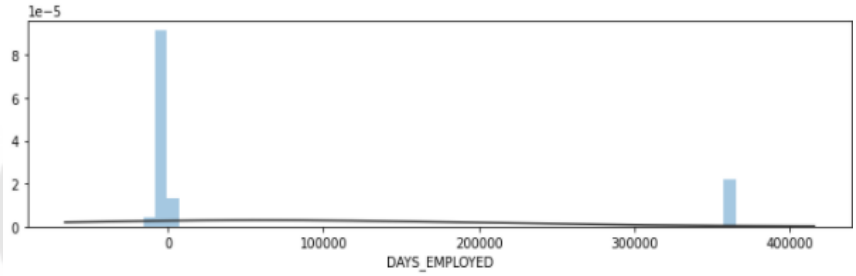
ภาพประกอบ 21 รายได้รวมต่อปีของลูกหนี้

จากภาพประกอบที่ 21 ส่วนมากแล้วนั้นรายได้ต่อปีของลูกหนี้ที่กู้สินเชื่อค่อนข้างต่ำ

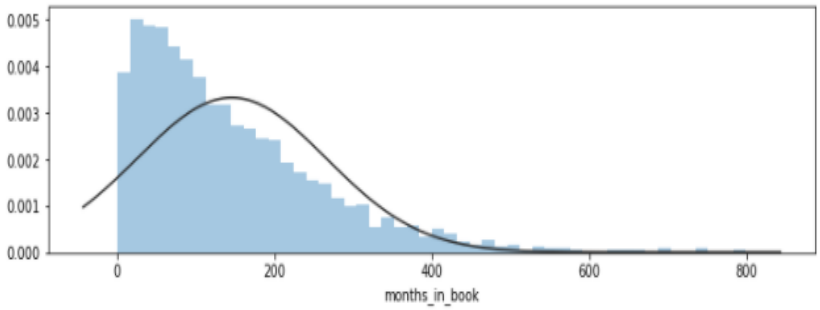


ภาพประกอบ 22 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์จำนวนวันเกิด

จากภาพประกอบที่ 22 DAY_BIRTH ช่วงอายุของลูกหนี้ที่กู้สินเชื่อบัตรเครดิตอยู่ระหว่าง 10,000/365 วัน (27 ปี) ถึง 15,000/365 (41 ปี)

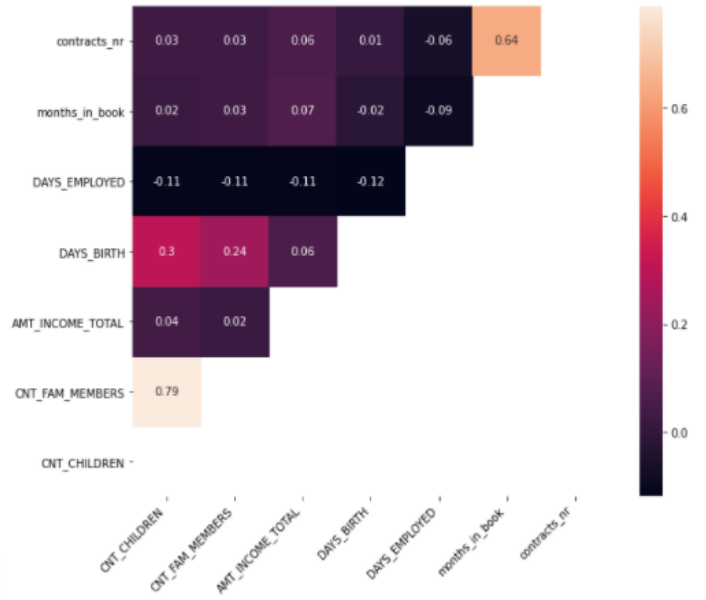


ภาพประกอบ 23 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์จำนวนวันที่เริ่มทำงาน



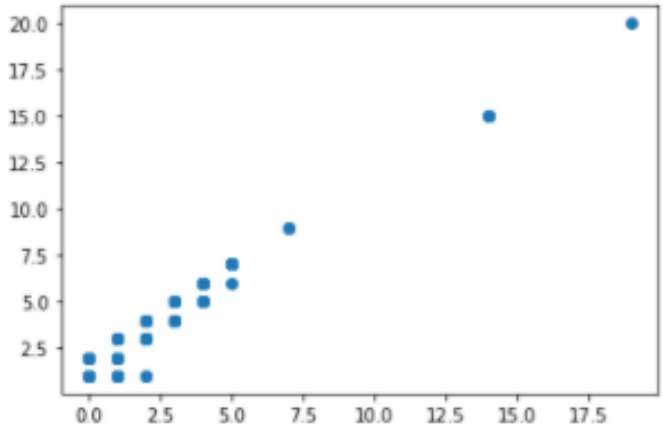
ภาพประกอบ 24 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของคอลัมน์จำนวนเดือนที่ลูกหนี้กู้

จากภาพประกอบที่ 24 แสดงการกระจายตัวของข้อมูลจำนวนเดือนกูทั้งหมดซึ่งโดยส่วนใหญ่แล้วอยู่ในช่วง 10 – 120 เดือน



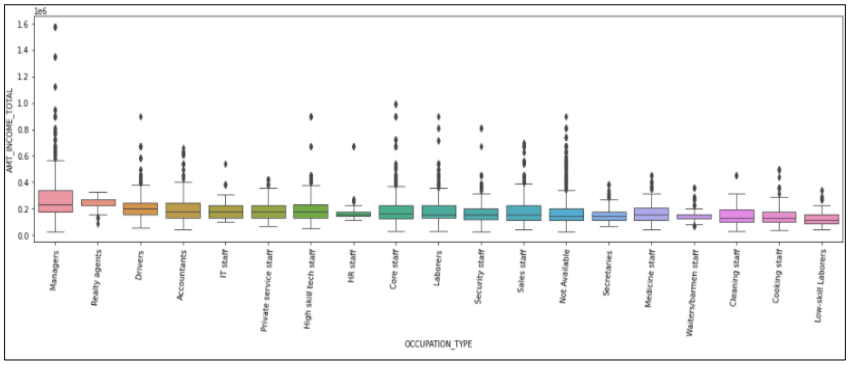
ภาพประกอบ 25 แสดงการวิเคราะห์ความสัมพันธ์ของแต่ละคอลัมน์

จากภาพประกอบ 25 แสดงความสัมพันธ์ของแต่ละตัวแปรซึ่งจากกราฟแสดงให้เห็นว่าความสัมพันธ์ระหว่างตัวแปร CNT_FAM_MEMBERS กับ CNT_CHILDREN มีค่าความสัมพันธ์ไปทางบวกมากถึง 0.79

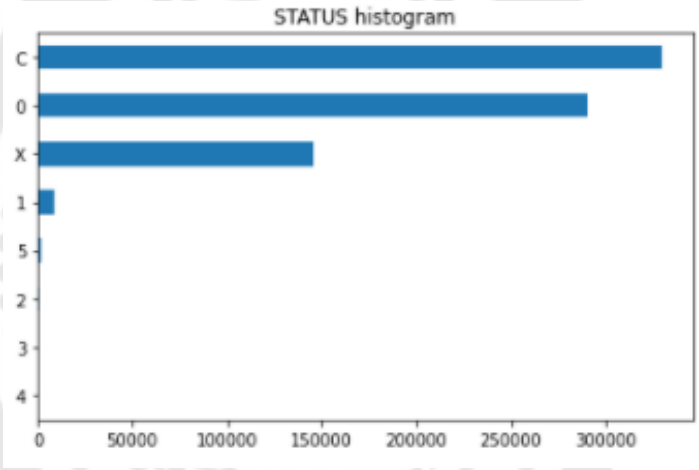


ภาพประกอบ 26 แสดง Correlation ระหว่าง CNT_FAM_MEMBERS และ CNT_CHILDREN

จากภาพประกอบที่ 26 ความสัมพันธ์ระหว่าง CNT_FAM_MEMBERS และ CNT_CHILDREN นั้นเป็นไปในทิศทางเดียวกันเนื่องจากจำนวนสมาชิกครอบครัวที่มากจะทำให้จำนวนบุตรมากตามไปด้วย ดังนั้นจึงไม่นำ CNT_CHILDREN (จำนวนบุตร) มาใช้ในการพัฒนาแบบจำลอง และจะเลือกใช้ CNT_FAM_MEMBERS (จำนวนสมาชิกครอบครัว) แทน



ภาพประกอบ 27 แสดงกราฟระหว่างค่าเฉลี่ยรายได้กับอาชีพโดยเรียงข้อมูลค่าเฉลี่ยรายได้จากมากไปน้อย



ภาพประกอบ 28 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของตัวแปรสถานะลูกหนี้ (STATUS)

จากภาพประกอบที่ 28 ตัวแปรที่นำมาใช้เป็น Target คือ STATUS (สถานะลูกหนี้) ลักษณะกระจายตัวของข้อมูลพบว่าอัตรส่วนลูกหนี้ที่มีความสามารถในการชำระหนี้ (C , 0 , X) มากกว่าลูกหนี้ที่ไม่สามารถชำระหนี้ (1 , 2 , 3 , 4 , 5) อย่างมาก ซึ่งในทางธุกิจธนาคารถือได้ว่าเป็นเรื่องปกติเนื่องจาก ธนาคารจำเป็นต้องบริหารพอร์ตให้ลูกหนี้มาชำระหนี้ได้ตรงตามเงื่อนไขที่ทางธนาคารกำหนด แต่ในทางทดสอบแบบจำลองอาจทำให้ประสิทธิภาพแบบจำลองต่ำได้เนื่องจากข้อมูลมีความไม่สมดุล ดังนั้นผู้วิจัยจะนำเทคนิค SMOTH มาปรับข้อมูลที่มีความไม่สมดุล

วิศวกรรมคุณลักษณะ (Feature Engineering)

ตาราง 4 แสดงข้อมูลจำนวนแถว NAME_INCOME_TYPE

NAME_INCOME_TYPE	Rows
Working	18,819
Commercial associate	8,490
Pensioner	6,152
State servant	2,985
Student	11

จากตาราง 4 แสดงจำนวนข้อมูลแถวของตาราง NAME_INCOME_TYPE และได้ทำการจัดกลุ่มใหม่เพื่อใช้ในการพัฒนาแบบจำลองโดยมีการจัดกลุ่มดังตาราง 5

ตาราง 5 แสดงตารางการจัดกลุ่ม NAME_INCOME_TYPE

NAME_INCOME_TYPE	New NAME_INCOME_TYPE
Working	Working
Commercial associate	Working
Pensioner	Pensioner
State servant	Working
Student	Student

หลังจากจัดกลุ่มเรียบร้อยจะได้ผลลัพธ์ดังตาราง 6 โดยวิเคราะห์ว่ากลุ่ม Commercial associate และ State servant ให้เป็น Working เพราะประเภทรายได้ถือว่าเป็นกลุ่มเดียวกับ Working

ตาราง 6 แสดงข้อมูลหลังทำการจัดกลุ่ม NAME_INCOME_TYPE

NAME_INCOME_TYPE	Rows
Working	30,294
Pensioner	6,152
Student	11

จากตาราง 6 แสดงข้อมูลหลังทำการจัดกลุ่มใหม่ NAME_INCOME_TYPE เพื่อใช้ในการพัฒนาแบบจำลอง

ตาราง 7 แสดงข้อมูลจำนวนแถว NAME_EDUCATION_TYPE

NAME_EDUCATION_TYPE	Rows
Secondary / secondary special	24,777
Higher education	9,864
Incomplete higher	1,410
Lower secondary	374
Academic degree	32

จากตาราง 7 แสดงจำนวนข้อมูลแถวของตาราง NAME_INCOME_TYPE และได้ทำการจัดกลุ่มใหม่เพื่อใช้ในการพัฒนาแบบจำลองโดยมีการจัดกลุ่มดังตาราง 8

ตาราง 8 แสดงตารางการจัดกลุ่ม NAME_EDUCATION_TYPE

NAME_EDUCATION_TYPE	New NAME_EDUCATION_TYPE
Secondary / secondary special	Secondary / secondary special
Higher education	Higher education
Incomplete higher	Secondary / secondary special
Lower secondary	Basic
Academic degree	Higher education

หลังจากจัดกลุ่มเรียบร้อยจะได้ผลลัพธ์ดังตาราง 9 โดยวิเคราะห์ว่ากลุ่ม Incomplete higher และ Lower secondary ให้เป็น Basic เพราะมีระดับการศึกษาที่คล้ายคลึงกัน

ตาราง 9 แสดงข้อมูลจำนวนหลังทำการจัดกลุ่ม NAME_EDUCATION_TYPE

NAME_EDUCATION_TYPE	Rows
Secondary / secondary special	26,187
Higher education	9,896
Basic	374

จากตาราง 9 แสดงข้อมูลหลังทำการจัดกลุ่มใหม่ NAME_EDUCATION_TYPE เพื่อใช้ในการพัฒนาแบบจำลอง

ตาราง 10 แสดงข้อมูลจำนวนแถว NAME_FAMILY_STATUS

NAME_FAMILY_STATUS	Rows
Married	25,048
Single / not married	4,829
Civil marriage	2,945
Separated	2,103
Widow	1,532

จากตาราง 10 แสดงจำนวนข้อมูลแถวของตาราง NAME_INCOME_TYPE และได้ทำการจัดกลุ่มใหม่เพื่อใช้ในการพัฒนาแบบจำลองโดยมีการจัดกลุ่มดังตาราง 11

ตาราง 11 แสดงตารางการจัดกลุ่ม NAME_FAMILY_STATUS

NAME_FAMILY_STATUS	New NAME_FAMILY_STATUS
Married	Married
Single / not married	Single / not married
Civil marriage	Married
Separated	Separated
Widow	Widow

หลังจากจัดกลุ่มเรียบร้อยจะได้ผลลัพธ์ดังตาราง 12 โดยวิเคราะห์ว่ากลุ่ม Married และ Civil marriage จัดให้เป็น Married เพราะเปรียบเสมือนสมรสแล้ว

ตาราง 12 แสดงข้อมูลจำนวนหลังทำการจัดกลุ่ม NAME_FAMILY_STATUS

NAME_FAMILY_STATUS	Rows
Married	27,993
Single / not married	4,829
Separated	2,103
Widow	1,532

จากตาราง 12 แสดงข้อมูลหลังทำการจัดกลุ่มใหม่ NAME_FAMILY_STATUS เพื่อใช้ในการพัฒนาแบบจำลอง

ตาราง 13 แสดงข้อมูลจำนวนแถว NAME_HOUSING_TYPE

NAME_HOUSING_TYPE	Rows
House / apartment	32,548
With parents	1,776
Municipal apartment	1,128
Rented apartment	575
Office apartment	262
Co-op apartment	1168

จากตาราง 13 แสดงจำนวนข้อมูลแถวของตาราง NAME_HOUSING_TYPE และได้ทำการจัดกลุ่มใหม่เพื่อใช้ในการพัฒนาแบบจำลองโดยมีการจัดกลุ่มดังตาราง 14

ตาราง 14 แสดงตารางการจัดกลุ่ม NAME_HOUSING_TYPE

NAME_HOUSING_TYPE	New NAME_HOUSING_TYPE
House / apartment	Rented apartment
With parents	With parents
Municipal apartment	Municipal apartment
Rented apartment	Rented apartment
Office apartment	Municipal apartment
Co-op apartment	Rented apartment

หลังจากจัดกลุ่มเรียบร้อยจะได้ผลลัพธ์ดังตาราง 15 โดยวิเคราะห์ว่ากลุ่ม House / apartment และ Co-op apartment ให้เป็น Rented apartment

ตาราง 15 แสดงข้อมูลจำนวนหลังทำการจัดกลุ่ม NAME_HOUSING_TYPE

NAME_HOUSING_TYPE	Rows
Rented apartment	33,291
With parents	1,776
Municipal or Office apartment	1,390

จากตาราง 15 แสดงข้อมูลหลังทำการจัดกลุ่มใหม่ NAME_HOUSING_TYPE เพื่อใช้ในการพัฒนาแบบจำลอง

ตาราง 16 แสดงข้อมูลจำนวนแถว OCCUPATION_TYPE

OCCUPATION_TYPE	Rows
Not Available	11,323
Laborers	6,211
Core staff	3,591
Sales staff	3,485
Managers	3,012
Drivers	2,138
High skill tech staff	1,383
Accountants	1,241
Medicine staff	1,207
Cooking staff	655
Security staff	592
Cleaning staff	551
Private service staff	344
Low-skill Laborers	175
Waiters/barmen staff	174
Secretaries	151
HR staff	85
Realty agents	79
IT staff	60

จากตาราง 16 แสดงจำนวนข้อมูลแถวของตาราง OCCUPATION_TYPE และได้ทำการจัดกลุ่มใหม่เพื่อใช้ในการพัฒนาแบบจำลองโดยมีการจัดกลุ่มดังตาราง 17

ตาราง 17 แสดงตารางการจัดกลุ่ม OCCUPATION_TYPE

OCCUPATION_TYPE	New OCCUPATION_TYPE
Not Available	Group 3
Laborers	Group 3
Core staff	Group 2
Sales staff	Group 3
Managers	Group 1
Drivers	Group 1
High skill tech staff	Group 2
Accountants	Group 1
Medicine staff	Group 4
Cooking staff	Group 4
Security staff	Group 3
Cleaning staff	Group 4
Private service staff	Group 2
Low-skill Laborers	Group 4
Waiters/barmen staff	Group 4
Secretaries	Group 3
HR staff	Group 2
Realty agents	Group 1
IT staff	Group 2

โดยจัดกลุ่มตามรายได้เงินเดือน โดยมีรายละเอียดดังนี้

Group 1 : กลุ่มผู้มีรายได้สูง

Group 2 : กลุ่มผู้มีรายได้ดี

Group 3 : กลุ่มผู้มีรายได้พอใช้

Group 4 : กลุ่มผู้มีรายได้ต่ำ

วัตถุประสงค์ของการจัดกลุ่มเพื่อใช้ในการแบ่งกลุ่มความเสี่ยง เช่น กลุ่มผู้มีรายได้สูงมีความเสี่ยงที่ต่ำ หรือ กลุ่มผู้มีรายได้ต่ำมีความเสี่ยงที่สูง

ตาราง 18 แสดงข้อมูลจำนวนหลังทำการจัดกลุ่ม OCCUPATION_TYPE

OCCUPATION_TYPE	Rows
Group 1	6,470
Group 2	5,463
Group 3	21,762
Group 4	2,762

จากตาราง 18 แสดงข้อมูลหลังทำการจัดกลุ่มใหม่ OCCUPATION_TYPE เพื่อใช้ในการพัฒนาแบบจำลอง

4. การเตรียมข้อมูล (Data Preprocessing)

ผู้วิจัยทำการแบ่งข้อมูลด้วย Train_Test_Split ที่สัดส่วน 80% สำหรับข้อมูลในการเรียนรู้โดยข้อมูลทั้งหมด 28,774 แถวและ 20% สำหรับข้อมูลในการทดสอบข้อมูลทั้งหมด 7,292 แถว โดยการเตรียมข้อมูลดังต่อไปนี้ใช้กับข้อมูลสำหรับการเรียนรู้เท่านั้น

4.1 การเปลี่ยนรูปแบบข้อมูลแบบกลุ่มและตัวเลข

สำหรับการเปลี่ยนข้อมูลแบบกลุ่มและตัวเลขเป็นการเพิ่มประสิทธิภาพในการเรียนรู้ของแบบจำลอง การเปลี่ยนข้อมูลแบบกลุ่ม Categorical Data ให้อยู่ในรูปแบบตัวเลขหรือ Numerical Data เป็นขั้นตอนที่สำคัญเพื่อเพิ่มประสิทธิภาพแบบจำลองให้ค่าความน่าจะเป็นง่ายขึ้น ดังนั้นจึงพิจารณาทำการเปลี่ยนข้อมูลฟีเจอร์แบบกลุ่ม ได้แก่ FLAG_OWN_CAR , FLAG_OWN_REALTY , FLAG_MOBIL , FLAG_WORK_PHONE , FLAG_PHONE , FLAG_EMAIL , NAME_INCOME_TYPE , NAME_EDUCATION_TYPE , NAME_FAMILY_STATUS , NAME_HOUSING_TYPE และ OCCUPATION_TYPE และ ให้เป็นตัวเลข และสำหรับฟีเจอร์ที่เก็บข้อมูลเป็นตัวเลขคือฟีเจอร์ CNT_CHILDREN , AMT_INCOME_TOTAL , DAYS_BIRTH พบว่าข้อมูลโดยส่วนมากมีการกระจุกตัวของชุดข้อมูล ดังนั้นก่อนนำข้อมูลตัวเลขไปใช้ทดสอบแบบจำลองต้องมีการปรับเปลี่ยนช่วงของข้อมูลเพื่อเพิ่มประสิทธิภาพในการทำงานของแบบจำลอง

กระบวนการที่เลือกใช้ชื่อว่า Column Transformer เป็นกระบวนการที่สามารถทำการเปลี่ยนแปลงข้อมูลแบบกลุ่มเป็นข้อมูลตัวเลขหรือ One Hot Encoding และสามารถปรับเปลี่ยนช่วงของข้อมูลตัวเลขได้พร้อมกัน สาเหตุที่ต้องทำพร้อมกันเนื่องจากไม่สามารถปรับช่วงข้อมูลตัวเลขชุดข้อมูลได้ และอาจเกิดปัญหาข้อมูลในชุดทดสอบรั่วไหลหรือว่า Data Leakage ดังนั้นจึงมีการใช้ Column Transformer ร่วมกับ Pipeline เพื่อให้สามารถระบุฟีเจอร์ที่ต้องการทำกระบวนการเปลี่ยนแปลงข้อมูลแบบกลุ่มเป็นข้อมูลตัวเลขและปรับช่วงข้อมูลได้อย่างปลอดภัย

สำหรับการเปลี่ยนแปลงข้อมูลกลุ่มเป็นข้อมูลตัวเลขหรือ One Hot Encoding คือการเพิ่มฟีเจอร์ขึ้นมาจากค่าข้อมูล เช่น ฟีเจอร์ CODE_GENDER มีทั้งหมด 2 ค่าคือ M และ F เมื่อผ่านการเปลี่ยนแปลงข้อมูลจะได้ฟีเจอร์ใหม่ขึ้นมา 2 ฟีเจอร์คือ CODE_GENDER==M และ CODE_GENDER==F หากลูกหนี้คนใดเป็นเพศชายจะเก็บค่า 1 ที่ฟีเจอร์ CODE_GENDER==M และเก็บค่า 0 ที่ CODE_GENDER==F โดยทำกระบวนการนี้ในทุกข้อมูลประเภทฟีเจอร์แบบกลุ่ม

การพิจารณาฟีเจอร์ที่เก็บข้อมูลเป็นตัวเลข พบว่าข้อมูลบางมีความสอดคล้องไปในทิศทางเดียวกัน (Correlation Data) ดังนั้นจึงทำการตัดฟีเจอร์

กระบวนการที่เลือกใช้คือ StandardScaler เป็นการปรับค่าข้อมูลตัวเลขโดยคำนวณจากค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐาน ดังสมการ (8)

$$x\_scaled_i = \frac{x_i - \mu}{\sigma} \quad (8)$$

โดยที่

x_scaled_i คือข้อมูล x ในชุดข้อมูลสำหรับเรียนรู้ตัวที่ i ผ่านการปรับค่าของข้อมูล

x_i คือข้อมูล x ชุดข้อมูลสำหรับเรียนรู้ตัวที่ i ที่ยังไม่ผ่านการปรับข้อมูล

μ คือค่าเฉลี่ยชุดข้อมูลสำหรับการเรียนรู้

σ คือส่วนเบี่ยงเบนมาตราฐานของชุด ข้อมูลสำหรับการเรียนรู้

4.2 การแก้ปัญหาข้อมูลที่ไม่สมดุล

จากการสำรวจข้อมูลข้างต้นพบว่าข้อมูลลาเบลในชุดข้อมูลนี้ เกิดปัญหาข้อมูลที่ไม่สมดุลหรือ Imbalanced Data หมายความว่าจำนวนสถานะลูกหนี้มีจำนวนไม่เท่ากัน ซึ่งปัญหานี้อาจส่งผลกระทบต่อประสิทธิภาพในการเรียนรู้ของแบบจำลองได้ กล่าวคือแบบจำลองอาจทำนายโดยให้ผลลัพธ์การทำนายลูกหนี้ใหม่โดยอ้างอิงจากชุดข้อมูลกลุ่มมาก ซึ่งวิธีการสังเคราะห์ข้อมูลจะเพิ่มจำนวนข้อมูลของกลุ่มน้อยให้เพิ่มขึ้น

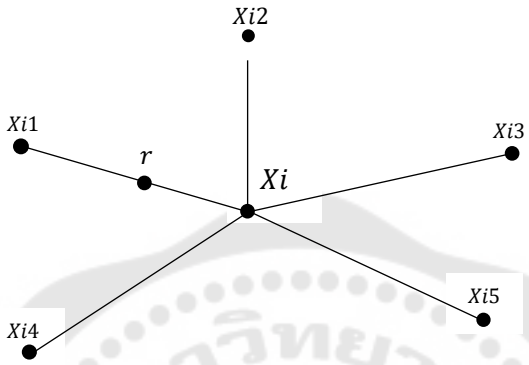
วิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling : SMOTE) จากภาพแสดงที่ 29 แทนที่จะสุ่มเพิ่มข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมที่มีอยู่โดยใช้หลักการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) โดยใช้ค่าตั้งต้นของฟังก์ชั่น SMOTE ที่กำหนดค่า K ให้เท่ากับ 5 โดยขั้นตอนในการสังเคราะห์ข้อมูลใหม่มีขั้นตอนดังนี้คือระบุเพื่อนบ้านที่ใกล้เคียงที่สุดจากค่า K จึงได้ข้อมูลที่ใกล้เคียงของ Xi จะได้ $Xi1$, $Xi2$, $Xi3$, $Xi4$, $Xi5$ เมื่อทำการระบุข้อมูลใกล้เคียงแล้วจะได้ข้อมูลใหม่คือ r

Xi คือ ข้อมูลที่ถูกเลือกแบบสุ่มจากกลุ่มข้อมูลที่น้อย

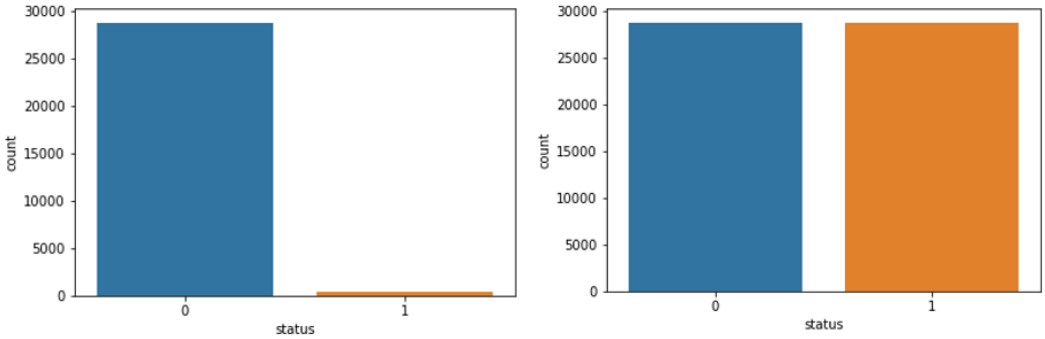
$Xi1$ ถึง $Xi5$ คือ ข้อมูลที่มีใกล้เคียงจากข้อมูลที่ถูกเลือก

มีทั้งหมด 5 จุดจากการกำหนดค่า K

r คือ ข้อมูลที่ถูกสังเคราะห์



ในชุดข้อมูลสำหรับทดสอบแบบจำลองพบว่าข้อมูลเลเบลมีจำนวนไม่เท่ากันแสดงดังภาพประกอบที่ 29 โดยมีข้อมูลลูกหนี้สถานะปกติอยู่ที่ 28,774 แถว และข้อมูลลูกหนี้สถานะผิดนัดชำระ 391 แถว หลังจากทำการจัดการกับข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE จะได้ชุดข้อมูลทั้งสองเท่ากันคือ 28,774 แถว ดังภาพประกอบที่ 29 ดังนั้นจำนวนข้อมูลทั้งหมดที่แบบจำลองใช้สำหรับการเรียนรู้คือ 57,548 แถว เมื่อข้อมูลมีความสมดุลกันแล้ว จะช่วยลดปัญหาที่แบบจำลองทำนายกลุ่มลูกหนี้ที่มีสถานะมากกว่า



ภาพประกอบ 29 แสดงจำนวนข้อมูลของ STATUS ก่อนและหลังโดยผ่านกระบวนการแก้ปัญหาข้อมูลที่ไม่สมดุลด้วยวิธีการสังเคราะห์ข้อมูล (SMOTE)

ที่มา (Tierney, 2020)

4.3 การสร้างแบบจำลองเพื่อจัดประเภทลูกหนี้

เมื่อผ่านกระบวนการเตรียมข้อมูลลูกหนี้แล้ว ขั้นตอนต่อไปคือนำข้อมูลเข้าสู่แบบจำลอง โดยใช้ข้อมูลในการเรียนรู้ทั้งหมด 29,165 แถวและข้อมูลสำหรับการทดสอบทั้งหมด 7,292 แถว และได้ใช้ Cross Validation ที่ 10 fold เพื่อเลือกชุดข้อมูลที่ดีที่สุดในการเรียนรู้ โดยใช้ชุดข้อมูลที่ผ่านกระบวนการแก้ปัญหาข้อมูลที่ไม่สมดุลด้วย SMOTE ซึ่งได้ข้อมูลสำหรับเรียนรู้ทั้งหมด 57,548 แถว โดยทำการทดลองตามแบบจำลองที่เลือกใช้คือ Logistic Regression, Extreme Gradient Boosting (XGBoost), CatBoost ตามลำดับร่วมกับการปรับจูนพารามิเตอร์

สำหรับแบบจำลอง Logistic Regression ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 2 ค่าดังนี้คือ

1. C คือส่วนกลับของค่าคงที่ที่กำหนดขนาดของพจน์ Penalty เช่น จากพจน์ของ Penalty ดังสมการที่ 9

$$\alpha||w||^2 \tag{9}$$

โดยที่

α คือค่าคงที่ที่กำหนดขนาดของ $||w||^2$

w คือความชันของแบบจำลอง

$||w||^2$ คือพจน์ L2 Regularization

หากค่า α มากหมายความว่ามีการให้ความสำคัญที่พจน์ Penalty มาก ทำให้แบบจำลองลดความยึดติดกับชุดข้อมูลในการเรียนรู้ ในทางตรงกันข้ามหากค่า α น้อยหมายความว่าให้ความสำคัญกับพจน์ Penalty น้อยทำให้แบบจำลองยึดติดกับชุดข้อมูลในการเรียนรู้ ในขณะที่ค่า C แสดงดังสมการที่ 10

$$C = \frac{1}{\alpha} \tag{10}$$

คือส่วนกลับของค่าที่ α ส่งผลให้ค่า C ที่ได้ค่าน้อยลง ทำให้ความสำคัญกับพจน์ Penalty ลดลงไปด้วย

2. Penalty คือพจน์ในการช่วยทำให้แบบจำลองลดการเกิดเหตุการณ์ Overfitting กับชุดข้อมูลสำหรับการเรียนรู้ โดยทำการปรับจูนระหว่าง L1 Regularization ดังสมการ 10หรือ L2 Regularization ดังสมการ 11

$$\alpha||w|| \quad (11)$$

โดยที่

α	คือค่าคงที่ที่กำหนดขนาดของ $||w||$
w	คือความชันของแบบจำลอง
$||w||$	คือพจน์ L1 Regularization

3. Solver คือชื่อเรียกอัลกอริทึมที่ต้องการเรียกใช้ในการทำงานของแบบจำลอง โดยในการวิจัยนี้เลือกปรับจูนระหว่าง Liblinear และ Lbfgs ซึ่งในการใช้อัลกอริทึมแต่ละอันมีข้อจำกัดเกี่ยวกับการเลือกใช้พจน์ Penalty คือสามารถเลือกใช้ได้บางตัวเท่านั้นดังตาราง 19 อ้างอิง (Zeyang Dou, 2019)

ตาราง 19 แสดงการเลือกใช้อัลกอริทึมและพจน์ Penalty ที่สามารถใช้ร่วมกันได้

พารามิเตอร์	**พจน์** Penalty
Liblinear	L1 หรือ L2 Regularization
Lbfgs	L2 Regularization หรือไม่ใช้พจน์ Penalty

เมื่อทำการปรับจูนพารามิเตอร์ครบทั้ง 3 ตัวแล้ว ได้ผลการปรับจูนออกมาดังตาราง 20

ตาราง 20 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง Logistic Regression

พารามิเตอร์	Parameter Tuning	Best Parameter
Penalty	L1 , L2	L1
C	100, 10, 1.0, 0.1, 0.01	1.0
Solver	Liblinear, Lbfgs	Liblinear

สำหรับแบบจำลอง Extreme Gradient Boosting ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 3 ค่าดังนี้

1. max_depth คือการกำหนดจำนวนชั้นของต้นไม้ในการตัดสินใจ
2. Criterion คือ สูตรที่ใช้ในการวัดคุณภาพของการแบ่งพาร์ทิชัน (Partition) ของต้นไม้ที่ใช้ในการตัดสินใจ โดยมีทั้งหมด 2 ค่าคือ

Gini คือค่าที่ใช้ในการวัดความสะอาดของพาร์ทิชั่นที่ถูกแบ่งโดยฟีเจอร์หนึ่งโดยฟีเจอร์ที่ให้ค่า Gini ต่ำหมายความว่ามีความสะอาดมาก

Entropy คือค่าวัดความไม่แน่นอนของข้อมูล ซึ่งความไม่แน่นอนหมายถึงจำนวนข้อมูลที่ทำนายผิด ดังนั้นเราต้องการฟีเจอร์ที่ให้ค่า Entropy ต่ำ

3. Leaning_rate คือค่าในการกำหนดน้ำหนักของการเปลี่ยนแปลงแบบจำลองใน 1 รอบ

เมื่อทำการปรับจูนพารามิเตอร์ทั้งหมดสำหรับแบบจำลอง XGBoost ทำให้ได้ผลออกมาดังตาราง 21

ตาราง 21 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง XGBoost

พารามิเตอร์	Parameter Tuning	Best Parameter
max_depth	[5, 10, 15]	15
leaning_rate	[0.001, 0.01, 0.1]	0.1
Criterion	[Entropy, Gini]	Gini

สำหรับแบบจำลอง CatBoost ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 2 ค่าดังนี้

1. max_depth คือการกำหนดจำนวนชั้นของต้นไม้ในการตัดสินใจ
2. Leaning_rate คือค่าในการกำหนดน้ำหนักของการเปลี่ยนแปลงแบบจำลองใน 1 รอบ

เมื่อทำการปรับจูนพารามิเตอร์ทั้งหมดสำหรับแบบจำลอง CatBoost ทำให้ได้ผลออกมาดังตาราง 22

ตาราง 22 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง CatBoost

พารามิเตอร์	Parameter Tuning	Best Parameter
Depth	[4, 5, 6, 7]	7
leaning_rate	[0.001, 0.01, 0.1]	0.1

เมื่อผ่านกระบวนการปรับจูนพารามิเตอร์ครบทุกแบบจำลองแล้ว ให้แบบจำลองได้เรียนรู้กับข้อมูลชุดข้อมูลเรียนรู้ต่อไป หลังจากทำการเรียนรู้สำเร็จจึงทำการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision Macro Avg, Recall Macro Avg, F1-Score Macro Avg และ ROC AUC แสดงผลการทำนายผิดถูกด้วยตาราง Confusion Matrix ทั้งนี้ผลการทำนายการจำแนกประเภทลูกหนี้ได้แก่ 1. ผลการทำนายลูกหนี้ปกติ 2. ผลการทำนายลูกหนี้ผิดนัดชำระ มีความสำคัญเท่ากันจึงได้ใช้ค่า Macro Avg ในการอภิปราย โดยสรุปความสำคัญ 2 ประเภท ได้ทั้งหมด 4 ข้อดังนี้

1. แบบจำลองทำนายเป็นลูกหนี้ปกติ และข้อมูลจริงเป็นลูกหนี้ปกติส่งผลให้ธนาคารเรียกเก็บเงินได้ครบกำหนดส่งผลให้ธนาคารได้กำไรจากดอกเบี้ย
2. แบบจำลองทำนายเป็นลูกหนี้ผิดนัดชำระ แต่ข้อมูลจริงลูกหนี้ปกติส่งผลให้ธนาคารขาดทุนเนื่องจากปฏิเสธลูกหนี้ที่มีความสามารถในการชำระหนี้
3. แบบจำลองทำนายเป็นลูกหนี้ผิดนัดชำระ และข้อมูลจริงลูกหนี้ผิดนัดชำระส่งผลให้ลดความเสี่ยงจากการปฏิเสธลูกหนี้ที่มีโอกาสผิดนัดชำระ
4. แบบจำลองทำนายเป็นลูกหนี้ปกติ แต่ข้อมูลจริงเป็นลูกหนี้ผิดนัดชำระ ส่งผลให้ธนาคารมีความเสี่ยงเนื่องจากอนุมัติสินเชื่อแก่ลูกหนี้ผิดนัดชำระ

ดังนั้นจึงสรุปได้ว่าเมื่อความสำคัญต่อการจำแนกประเภทมีความสำคัญเท่ากัน การวัดประสิทธิภาพด้วยค่า Macro Avg จะสามารถสะท้อนความเป็นจริงในการดำเนินธุรกิจธนาคาร หากแบบจำลองสามารถจำแนกประเภทลูกหนี้ได้อย่างถูกต้องส่งผลให้ธนาคารสามารถเพิ่มกำไรจากการอนุมัติสินเชื่อแก่ลูกหนี้ปกติ และลดความเสี่ยงจากการขาดทุนต่อการอนุมัติสินเชื่อแก่ลูกหนี้ผิดนัดชำระ

บทที่ 4
ผลการดำเนินการวิจัย

การวัดประสิทธิภาพแบบจำลองนั้น ผู้วิจัยใช้ชุดข้อมูลทดสอบทั้งหมด 7,292 แถว ประกอบด้วยข้อมูลลูกหนี้สถานะปกติ 7,194 แถว และสถานะผิดนัดชำระ 98 แถว จากนั้นวัดประสิทธิภาพแบบจำลองด้วยค่า Accuracy, Precision Macro Avg, Recall Macro Avg, F1-Score Macro Avg และ ROC ได้ตามตาราง 23

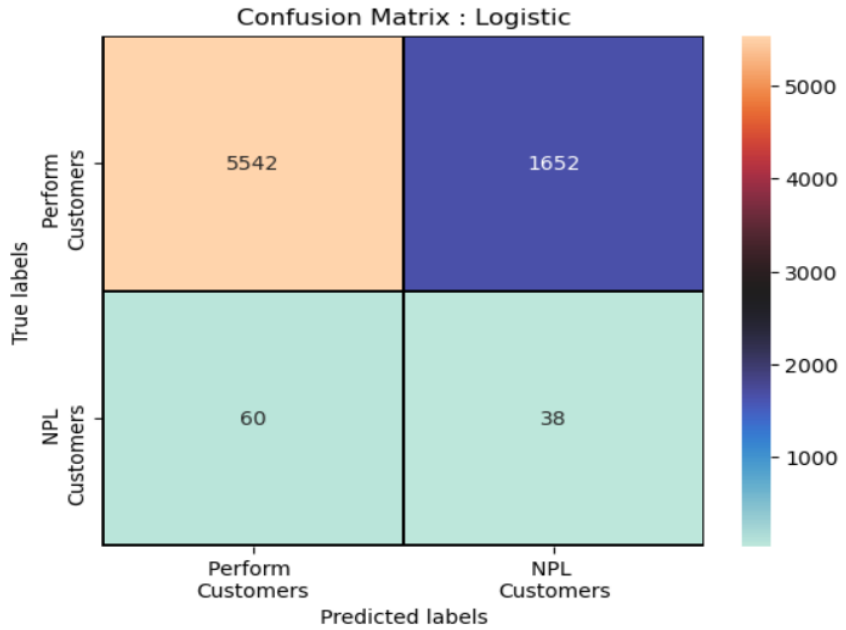
ตาราง 23 แสดงผลการวัดประสิทธิภาพแบบจำลองจากชุดข้อมูลทดสอบ

Model Name	Accuracy (%)	Precision Macro Avg (%)	Recall Macro Avg (%)	F1-Score Macro Avg (%)	ROC AUC (%)
Logistic Regression	77	51	58	46	60
XGBoost	98	97	92	95	93
CatBoost	97	63	83	71	92

แสดงผลการเปรียบเทียบ ระหว่างค่า Accuracy Train (Cross-Validation) เทียบกับ Accuracy Test พบว่าแบบจำลอง Logistic Regression เมื่อนำมาทดสอบให้ความถูกต้องสูงกว่าช่วงเวลาที่แบบจำลองทำการเรียนรู้ที่ 77% นอกจากนี้ทั้ง XGBoost และ CatBoost ให้ค่าความถูกต้องได้ใกล้เคียงกันทั้งช่วงเวลาที่แบบจำลองเรียนรู้ และเมื่อนำมาทดสอบ ได้ตามตาราง 24

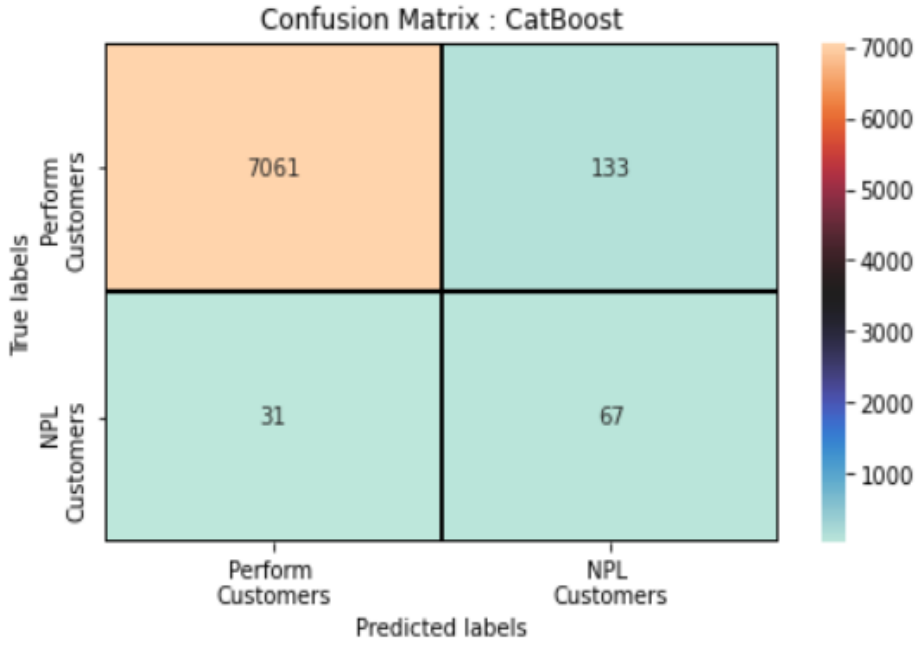
ตาราง 24 แสดงการเปรียบเทียบผลการทำ Cross-Validation ของทุกแบบจำลอง

Model Name	Accuracy Train Crossvalidation (%)	Accuracy Test (%)
Logistic Regression	63	77
XGBoost	99	98
CatBoost	98	97



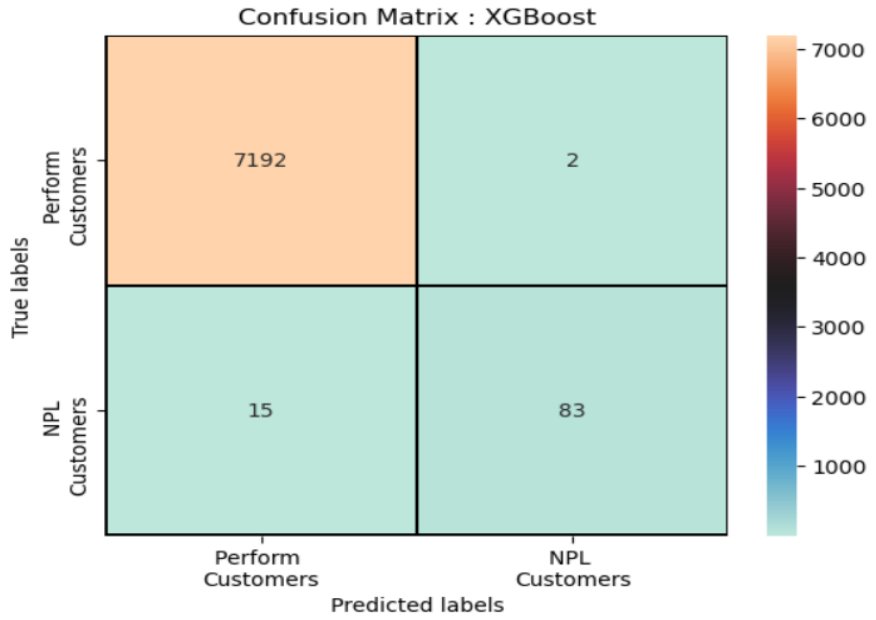
ภาพประกอบ 30 Confusion Matrix : Logistic Regression

จากรูปภาพที่ 30 แสดงการการจำแนกประเภทลูกหนี้เพื่ออนุมัติสินเชื่อด้วยอัลกอริทึม Logistic Regression พบว่าแบบจำลองอนุมัติสินเชื่อให้ลูกหนี้ที่สถานะปกติ 5,542 ราย และปฏิเสธลูกหนี้สถานะปกติ 1,652 ราย กล่าวคือธนาคารจะขาดรายได้จากการปฏิเสธลูกหนี้ปกติในการปล่อยกู้ถึง 1,652 รายการ ในส่วนของการปฏิเสธลูกหนี้ที่มีโอกาสผิดนัดชำระแบบจำลองปฏิเสธลูกหนี้ที่มีโอกกาสผิดนัดชำระ 38 รายและอนุมัติสินเชื่อแก่ลูกหนี้ที่มีโอกาสผิดนัดชำระ 60 รายกล่าวธนาคารมีโอกาสสูญเสียรายจากการอนุมัติสินเชื่อแก่ลูกหนี้ไม่ดีถึง 60 รายการ



ภาพประกอบ 31 แสดงตาราง Confusion Matric แบบจำลอง CatBoost

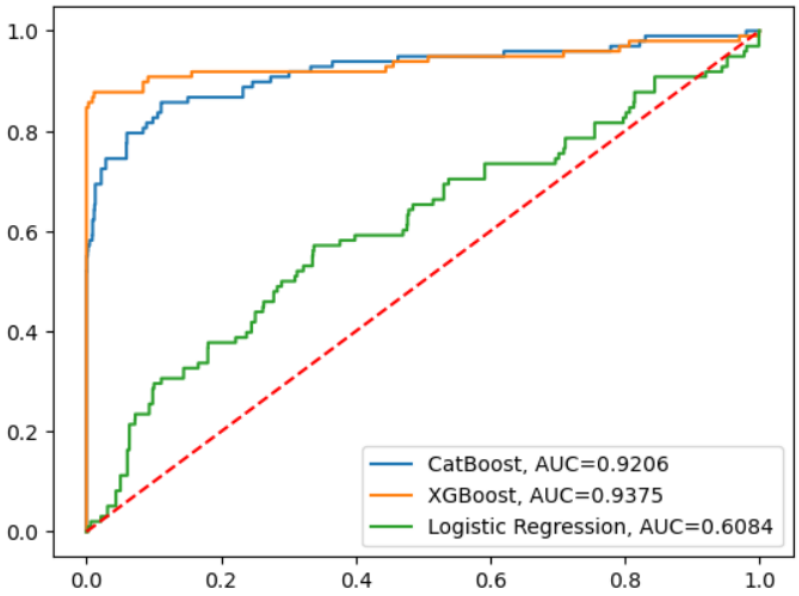
จากรูปภาพที่ 31 แสดงการการจำแนกประเภทลูกหนี้เพื่ออนุมัติสินเชื่อด้วยอัลกอริทึม CatBoost พบว่าแบบจำลองอนุมัติสินเชื่อให้ลูกหนี้ที่สถานะปกติ 7,061 ราย และปฏิเสธลูกหนี้สถานะปกติ 133 ราย กล่าวคือธนาคารจะขาดรายได้จากการปฏิเสธลูกหนี้ปกติในการปล่อยกู้ 133 รายการ ในส่วนของการปฏิเสธลูกหนี้ที่มีโอกาสผิดนัดชำระแบบจำลองปฏิเสธลูกหนี้ที่มีโอกกาสผิดนัดชำระ 67 รายและอนุมัติสินเชื่อแก่ลูกหนี้ที่มีโอกาสผิดนัดชำระ 31 รายกล่าวธนาคารมีโอกาสสูญเสียรายจากการอนุมัติสินเชื่อแก่ลูกหนี้ไม่ดีถึง 31 รายการ



ภาพประกอบ 32 Confusion Matrix : Extreme Gradient Boosting

จากรูปภาพที่ 32 แสดงการการจำแนกประเภทลูกหนี้เพื่ออนุมัติสินเชื่อด้วยอัลกอริทึม XGBoost พบว่าแบบจำลองอนุมัติสินเชื่อให้ลูกหนี้ที่สถานะปกติ 7,192 ราย และปฏิเสธลูกหนี้สถานะปกติ 2 ราย กล่าวคือธนาคารจะขาดรายได้จากการปฏิเสธลูกหนี้ปกติในการปล่อยกู้เพียง 2 รายการ ในส่วนของการปฏิเสธลูกหนี้ที่มีโอกาสผิดนัดชำระแบบจำลองปฏิเสธลูกหนี้ที่มีโอกกาสผิดนัดชำระ 83 รายและอนุมัติสินเชื่อแก่ลูกหนี้ที่มีโอกาสผิดนัดชำระ 15 รายกล่าวธนาคารมีโอกาสสูญเสียรายจากการอนุมัติสินเชื่อแก่ลูกหนี้ไม่ดีถึง 15 รายการ

จากการสำรวจผล Confusion Matrix แสดงให้เห็นถึงความถูกต้องในการอนุมัติสินเชื่อของแบบจำลอง เพื่อให้ธนาคารสามารถลดการขาดทุนค่าเสียโอกาสจากการปฏิเสธอนุมัติสินเชื่อแก่ลูกหนี้ปกติ และลดความเสี่ยงจากการอนุมัติสินเชื่อแก่ลูกหนี้ที่มีโอกาสผิดนัดชำระ โดยผู้วิจัยจะวิเคราะห์ผลของการทำนายทั้ง 2 กลุ่มเพื่อให้ธนาคารมีคุณภาพพอร์ตสินเชื่อบัตรเครดิตที่ดี เพราะมีลูกหนี้ที่มีความสามารถชำระหนี้ และลดความเสี่ยงจากหนี้เสียจากลูกหนี้ที่มีโอกาสผิดนัดชำระ

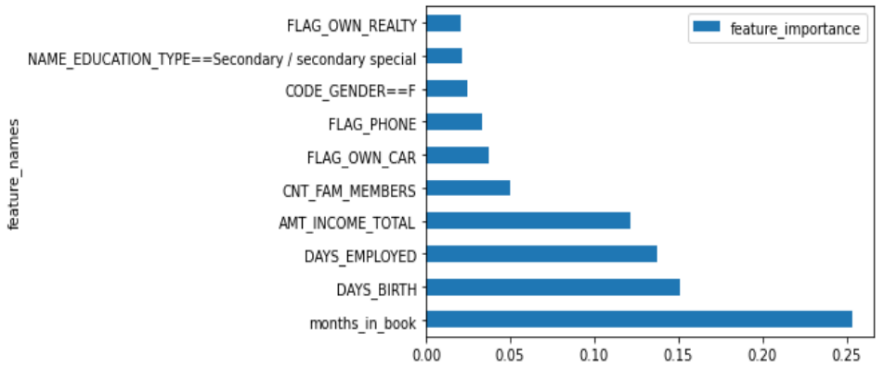


ภาพประกอบ 33 แสดง ROC CURVES ผลการวัดประสิทธิภาพแบบจำลองการจำแนกความถูกต้องทั้ง 3 แบบจำลอง

เมื่อทำการวัดประสิทธิภาพทั้ง 3 แบบจำลองเรียบร้อยและเพื่อให้สามารถเข้าใจถึงความสำคัญของฟีเจอร์ที่มีผลต่อการเรียนรู้ของแบบจำลอง ผู้วิจัยจึงได้จัดอันดับฟีเจอร์ที่มีความสำคัญต่อการเรียนรู้ของแบบจำลอง 10 อันดับแรก โดยพบว่าแบบจำลอง Logistic Regression ใช้ฟีเจอร์ months_in_book มากที่สุดและ DAY_EMPLOYED, AMT_INCOME_TOTAL, DAY_BIRTH ทั้ง 3 ฟีเจอร์มีความสำคัญใกล้เคียงกันดังภาพประกอบที่ 34

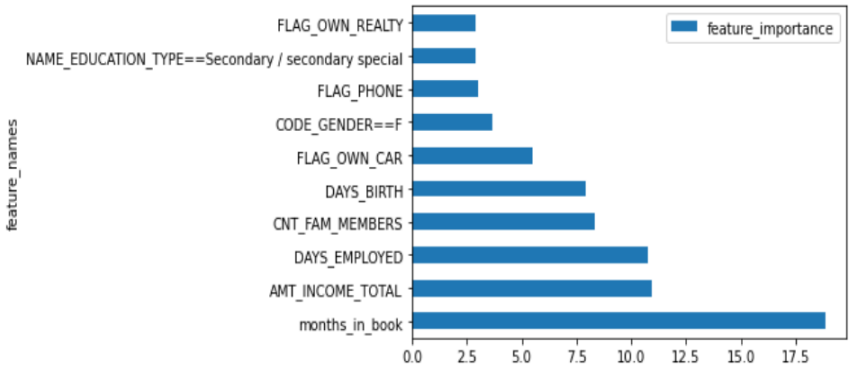
ภาพประกอบ 34 แสดง 10 อันดับ Feature Importance แบบจำลอง Logistic Regression

ตัวแปรที่มีผลต่อประสิทธิภาพการเรียนรู้แบบจำลอง CatBoost แสดงให้เห็นถึงความสำคัญตัวแปรที่มีผลต่อการเรียนรู้ของแบบจำลองพบว่าตัวแปร months_in_book มีความสำคัญมากที่สุด DAY_BIRTH และ DAY_EMPLOYED มีความสำคัญตามลำดับ ดังภาพประกอบที่ 35



ภาพประกอบ 35 แสดง 10 อันดับ Feature Importance แบบจำลอง CatBoost

ตัวแปรที่มีผลต่อประสิทธิภาพการเรียนรู้แบบจำลอง Extreme Gradient Boosting แสดงให้เห็นถึงความสำคัญตัวแปรที่มีผลต่อการเรียนรู้ของแบบจำลองพบว่าตัวแปร months_in_book มีความสำคัญมากที่สุด AMT_INCOME_TOTAL และ DAY_EMPLOYED มีความสำคัญใกล้เคียงกัน ดังภาพประกอบที่ 36

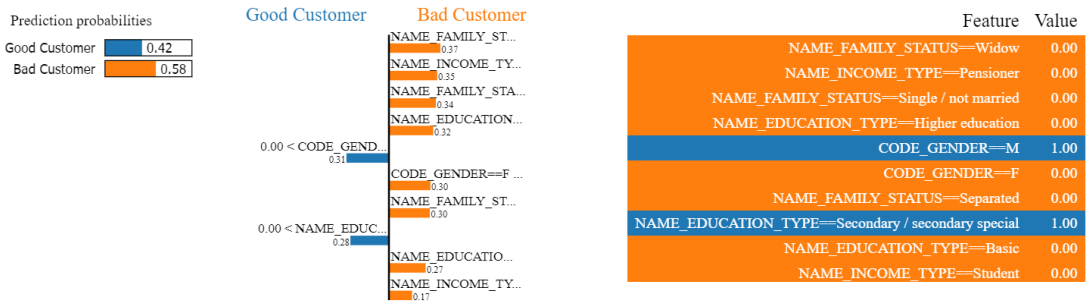


ภาพประกอบ 36 แสดง 10 อันดับ Feature Importance แบบจำลอง Extreme Gradient Boosting

สรุปความสำคัญของฟีเจอร์ที่มีผลต่อการเรียนรู้ของแบบจำลองทั้ง 3 พบว่าความสำคัญของฟีเจอร์ที่มีผลต่อการการเรียนรู้ของแบบจำลองมีความสอดคล้องกัน โดยอันดับ 1 คือ months_in_book และฟีเจอร์ทั้ง 4 ได้แก่ AMT_INCOME_TOTAL, DAYS_EMPLOYED, CNT_FAM_MEMBERS, DAY_BIRTH และฟีเจอร์อื่นๆก็มีผลต่อการเรียนรู้ต่อแบบจำลองที่คล้ายกัน โดยภาพรวมการการจัดอันดับความสำคัญต่อการเรียนรู้ของแบบจำลองทั้ง 10 อันดับนั้นมีความสอดคล้องกัน

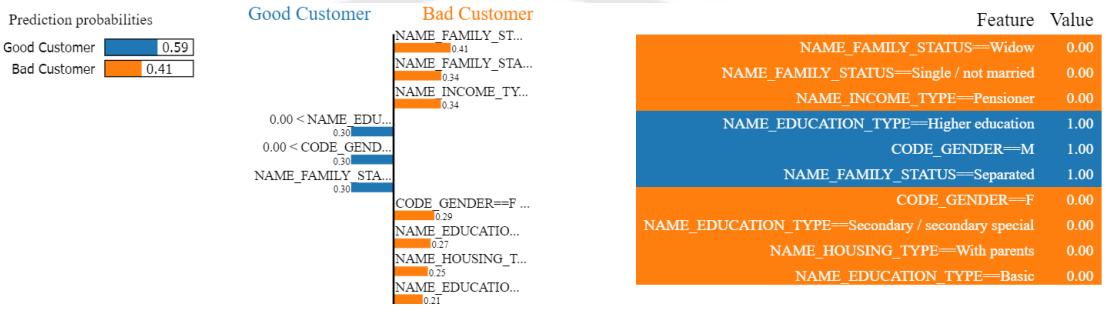
เมื่อทำการวิเคราะห์ภาพรวมความสำคัญของฟีเจอร์ทุกแบบจำลองแล้วนั้นพบว่าฟีเจอร์ที่สำคัญต่อแบบจำลอง Logistic Regression, XGBoost และ CatBoost คือ months_in_book

จากนั้นเพื่อให้เข้าใจถึงปัจจัยที่ส่งผลให้แบบจำลองทำนายผิดพลาด ทางผู้วิจัยจึงเลือกใช้ไลบารี่ Local Interpretable Model-agnostix Explanation (LIME) โดยการหาฟีเจอร์บนข้อมูล 1 ตัว และทำการเปรียบเทียบแต่ละแบบจำลองเพื่อหาว่าในการทำนายนั้นมีฟีเจอร์ใดบ้างที่ถูกนำมาพิจารณาโดยผู้วิจัยได้ทำการคัดเลือกข้อมูลจากชุดข้อมูลทดสอบ ได้ข้อมูลที่ 2 ซึ่งพบว่าการทำนายผิดเกิดขึ้นจากเลเบลของข้อมูลนี้คือลูกหนี้ปกติ แต่เมื่อตรวจสอบการทำงานของแบบจำลองที่ Logistic Regression ทำนายเป็นลูกหนี้ผิดนัดชำระ ด้วยความน่าจะเป็น 58% ในขณะที่ความน่าจะเป็นที่ทำนายได้ลูกหนี้ปกติ 42% เมื่อพิจารณาการฟีเจอร์กลุ่มลูกปกติพบว่าไม่เพียง 2 ฟีเจอร์ได้แก่ CODE_GENDER==F : 0 และ NAME_EDUCATION_TYPE== Secondary / secondary special : 1 หรือกล่าวคือลูกหนี้เพศชายที่มีระดับการศึกษาระดับชั้นมัธยมส่งผลทำให้แบบจำลองพิจารณาว่าเป็นลูกหนี้ผิดนัดชำระ โดยการใช้ฟีเจอร์และความน่าจะเป็นของแบบจำลอง Logistic Regression แสดงทั้งหมดดังภาพประที่ 37 ดังนี้



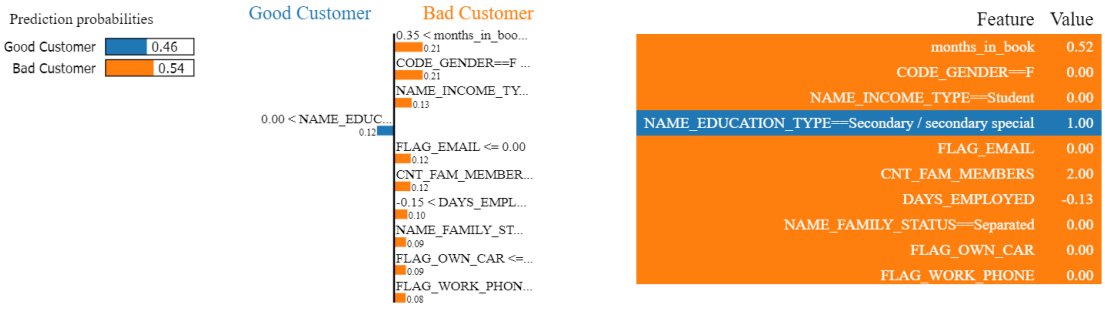
ภาพประกอบ 37 แสดงผลการใช้ LIME สำหรับแบบจำลอง Logistic Regression ข้อมูลที่ 2

นอกจากนี้ผู้วิจัยได้ทำการคัดเลือกข้อมูลชุดทดสอบที่ 252 ซึ่งเป็นข้อมูลที่แบบจำลองทำนายผิดพลาดจากเลเบลของข้อมูลนี้คือลูกหนี้ผิดนัดชำระ แต่เมื่อตรวจสอบการทำงานของแบบจำลองที่ Logistic Regression ทำนายเป็นลูกหนี้ปกติ ด้วยความน่าจะเป็น 59% ในขณะที่ความน่าจะเป็นที่ทำนายได้ลูกหนี้ผิดนัดชำระ 41% เมื่อพิจารณาการฟีเจอร์กลุ่มลูกหนี้ผิดนัดชำระพบว่าไม่เพียง 3 ฟีเจอร์ได้แก่ NAME_EDUCATION_TYPE==Higher education : 1, CODE_GENDER==M : 1 และ NAME_FAMILY_STATUS==Separate : 1 หรือกล่าวคือลูกหนี้เพศชาย ผ่านการหย่าร้าง มีระดับการศึกษาที่สูงทำให้แบบจำลองพิจารณาว่าเป็นลูกหนี้ปกติ โดยการใช้ฟีเจอร์และความน่าจะเป็นของแบบจำลอง Logistic Regression แสดงทั้งหมดดังภาพประที่ 38 ดังนี้



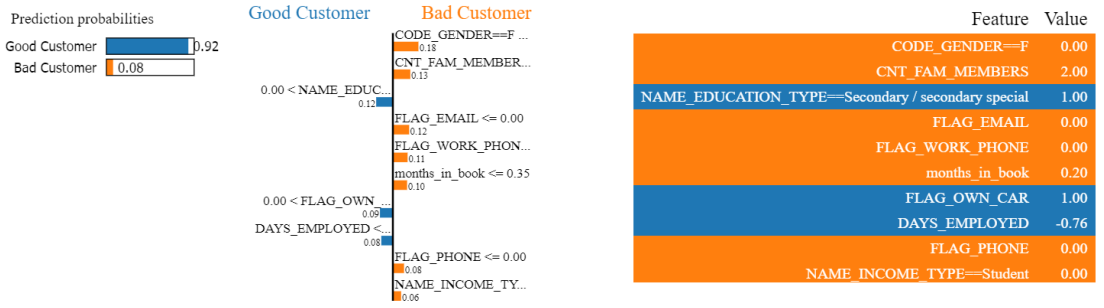
ภาพประกอบ 38 แสดงผลการใช้ LIME สำหรับแบบจำลอง Logistic Regression ข้อมูลที่ 252

สำหรับแบบจำลอง CatBoost พบว่าการทำนายผิดเกิดขึ้นจากเลเบลของข้อมูลนี้คือลูกหนี้ปกติ แต่เมื่อตรวจสอบการทำงานของแบบจำลองที่ CatBoost ทำนายเป็นลูกหนี้ผิดนัดชำระ ด้วยความน่าจะเป็น 54% ในขณะที่ความน่าจะเป็นที่ทำนายได้ลูกหนี้ปกติ 46% เมื่อพิจารณาการฟีเจอร์กลุ่มลูกหนี้ปกติพบว่ามีหลายฟีเจอร์ที่ส่งผลทำให้แบบจำลองทำนายลูกหนี้อยู่ในกลุ่มผิดนัดชำระได้แก่ months_in_book : 0.52, CODE_GENDER==F : 0, CNT_FAM_MEMBERS : 2 หมายความว่าลูกหนี้เป็นเพศชาย มีจำนวนสมาชิกครอบครัว 2 คน และผู้กู้เพิ่งกู้เนื่องจากมีเวลากู้น้อยทำให้แบบจำลองพิจารณาว่าเป็นลูกหนี้ผิดนัดชำระ โดยการใช้ฟีเจอร์และความน่าจะเป็นของแบบจำลอง CatBoost แสดงทั้งหมดดังภาพประที่ 39 ดังนี้



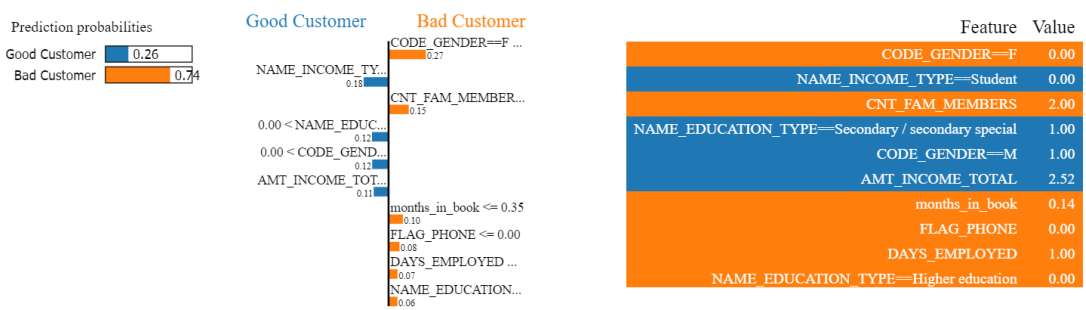
ภาพประกอบ 39 แสดงผลการใช้ LIME สำหรับแบบจำลอง CatBoost ข้อมูลที่ 88

นอกจากนี้ผู้วิจัยได้ทำการคัดเลือกข้อมูลชุดทดสอบที่ 597 ซึ่งเป็นข้อมูลที่แบบจำลองทำนายผิดพลาดจากเลเบลของข้อมูลนี้คือลูกหนี้ผิดนัดชำระ แต่เมื่อตรวจสอบการทำงานของแบบจำลองที่ CatBoost ทำนายเป็นลูกหนี้ปกติ ด้วยความน่าจะเป็นถึง 92% ในขณะที่ความน่าจะเป็นที่ทำนายได้ลูกหนี้ผิดนัดชำระเพียง 8% เมื่อพิจารณาการฟีเจอร์กลุ่มลูกหนี้ผิดนัดชำระพบว่าไม่เพียง 3 ฟีเจอร์ได้แก่ NAME_EDUCATION_TYPE==Secondary / secondary special : 1, FLAG_OWN_CAR : 1 และ DAYS_EMPLOYED : -0.76 หมายความว่าลูกหนี้มีระดับการศึกษามัธยม ไม่เคยทำงาน แต่เนื่องจากมีทรัพย์สินเป็นรถยนต์ แบบจำลองจึงพิจารณาว่าเป็นลูกหนี้ปกติ โดยการใช้ฟีเจอร์และความน่าจะเป็นของแบบจำลอง CatBoost แสดงทั้งหมดดังภาพประที่ 40 ดังนี้



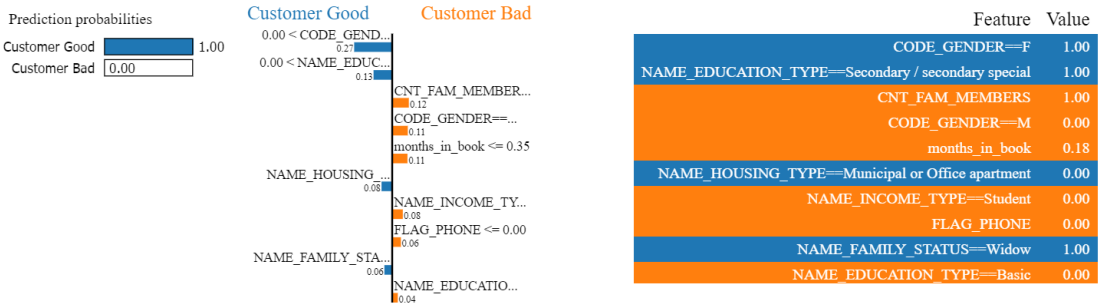
ภาพประกอบ 40 แสดงผลการใช้ LIME สำหรับแบบจำลอง CatBoost ข้อมูลที่ 597

สำหรับแบบจำลอง XGBoost พบว่าการทำนายผิดเกิดขึ้นจากเลเบลของข้อมูลนี้คือลูกหนี้ปกติ แต่เมื่อตรวจสอบการทำงานของแบบจำลองที่ XGBoost ทำนายเป็นลูกหนี้ผิดนัดชำระด้วยความน่าจะเป็นถึง 74% ในขณะที่ความน่าจะเป็นที่ทำนายได้ลูกหนี้ปกติเพียง 26% เมื่อพิจารณาการฟีเจอร์กลุ่มลูกหนี้ผิดนัดชำระพบว่ามีฟีเจอร์ได้แก่ CODE_GENDER==F : 0, CNT_FAM_MEMBERS : 2, months_in_book : 0.14 เป็นต้น ซึ่งสามารถอธิบายลักษณะของลูกหนี้ดังกล่าวคือเป็นลูกหนี้เพศชาย มีสมาชิกครอบครัวจำนวน 2 คนและเพิ่งมีการกู้สินเชื่อตามรายละเอียดฟีเจอร์จึงเป็นสาเหตุทำให้แบบจำลองพิจารณาว่าเป็นลูกหนี้ผิดนัดชำระ โดยการใช้ฟีเจอร์และความน่าจะเป็นของแบบจำลอง CatBoost แสดงทั้งหมดดังภาพประที่ 41 ดังนี้



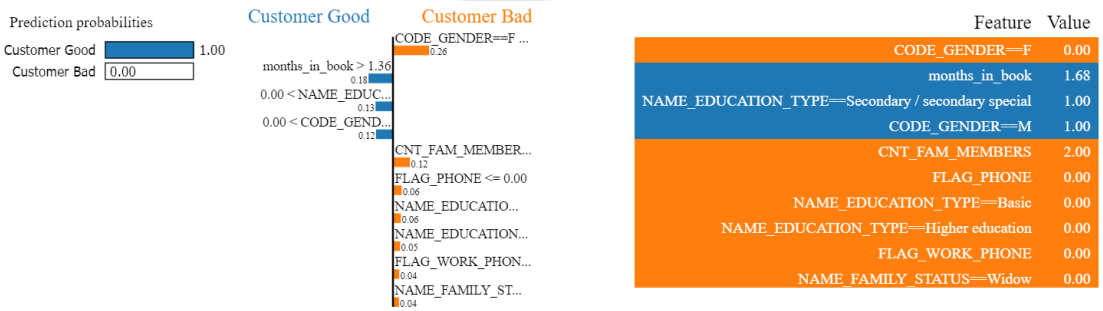
ภาพประกอบ 41 แสดงผลการใช้ LIME สำหรับแบบจำลอง XGBoost ข้อมูลที่ 10

นอกจากนี้ผู้วิจัยได้ทำการคัดเลือกข้อมูลชุดทดสอบที่ 288 ซึ่งเป็นข้อมูลที่แบบจำลองทำนายผิดพลาดจากเลเบลของข้อมูลนี้คือลูกหนี้ผิดนัดชำระ แต่เมื่อตรวจสอบการทำงานของแบบจำลองที่ XGBoost ทำนายเป็นลูกหนี้ปกติ ด้วยความน่าจะเป็นถึง 100% ในขณะที่ความน่าจะเป็นที่ทำนายได้ลูกหนี้ผิดนัดชำระ 0% เมื่อพิจารณาการฟีเจอร์กลุ่มลูกปกติพบว่ามี ฟีเจอร์ได้แก่ CODE_GENDER==F : 1, NAME_EDUCATION_TYPE==Secondary / secondary special : 1, NAME_FAMILY_STATUS==Widow : 1 กล่าวคือ ลูกหนี้เป็นเพศหญิง ระดับการศึกษามัธยม สถานะสมรสเป็นหม้าย ทำให้แบบจำลองทำนายลูกหนี้ดังกล่าวเป็นลูกหนี้ปกติ โดยการใช้ฟีเจอร์และความน่าจะเป็นของแบบจำลอง XGBoost แสดงทั้งหมดดังภาพประที่ 42 ดังนี้



ภาพประกอบ 42 แสดงผลการใช้ LIME สำหรับแบบจำลอง XGBoost ข้อมูลที่ 288

เมื่อเข้าในถึงผลการทำนายที่ผิดพลาดของแบบจำลองด้วยการใช้เครื่องมือ LIME แล้วนั้น ทางผู้วิจัยได้ทำการพิจารณาผลจากค่าความสำคัญของฟีเจอร์ (Feature Importance) ว่ามีความสอดคล้องกับ LIME ด้วยหรือไม่ โดยเลือกแบบจำลองที่มีการวัดประสิทธิภาพที่ดีที่สุดคือ XGBoost ดังนั้นผู้วิจัยได้ทำการเลือกแบบจำลองทำนายถูกจากข้อมูลชุดทดสอบมาแสดงเพิ่มเติม ซึ่งจากผลค่าความสำคัญของฟีเจอร์สำหรับแบบจำลอง XGBoost นั้นพบว่าฟีเจอร์ที่ให้ค่าสูงสุดคือ months_in_book หรือระยะเวลาการกู้ โดยข้อมูลที่เลือกมาคือข้อมูลที่ 2 ดังภาพประกอบที่ 43 พบว่ามีความสอดคล้องกับค่าความสำคัญของฟีเจอร์ตามการคาดการณ์



ภาพประกอบ 43 แสดงผลการใช้ LIME สำหรับแบบจำลอง XGBoost ข้อมูลที่ 2

บทที่ 5
สรุปผลการวิจัย อภิปราย และข้อเสนอแนะ

ปัจจุบันการแข่งขันระหว่างสถาบันการเงินมีความรุนแรงมากขึ้น เนื่องจากลูกหนี้มีความต้องการกู้สินเชื่อโดยมีความคาดหวังบริการที่รวดเร็ว ซึ่งกระบวนการในการอนุมัติสินเชื่อนั้นมีความซับซ้อนในการวิเคราะห์ข้อมูลลูกหนี้ หากธนาคารไม่มีความแม่นยำอาจเกิดความเสี่ยงที่ธนาคารจะขาดทุน อันเกิดจากธนาคารอนุมัติสินเชื่อต่อลูกหนี้ที่มีโอกาสเป็นหนี้สูญ

ผู้วิจัยมองเห็นถึงความสำคัญของการจำแนกประเภทลูกหนี้เพื่อให้ธุรกิจมีความเข้าใจและสามารถเข้าถึงลูกหนี้ได้อีกยังสามารถลดความเสี่ยงที่อาจส่งผลต่อการขาดทุนต่อธุรกิจ จึงได้จัดทำวิจัยนี้ขึ้นมาเพื่อศึกษาและเปรียบเทียบการทำงานของแบบจำลองเพื่อทำนายกลุ่มลูกหนี้สินเชื่อบัตรเครดิตจากสถาบันการเงินแห่งหนึ่งโดยใช้ข้อมูลประเภทประชากร ซึ่งคาดหวังว่าแบบจำลองจะสามารถนำไปใช้ในการดำเนินธุรกิจในการอนุมัติสินเชื่อ โดยทำให้ความสูญเสียจากลูกหนี้ไม่สามารถชำระหนี้น้อยที่สุด ผู้วิจัยได้ทำการสอนแบบจำลองและวัดประสิทธิภาพการทำงานของแบบจำลองและทำการสรุปผลแบ่งตามหัวข้อดังนี้

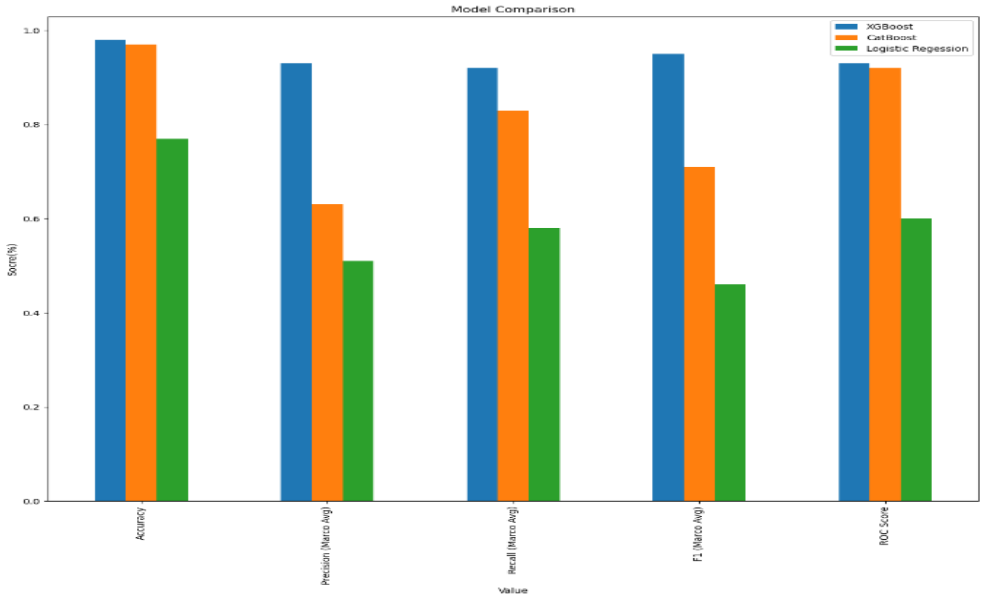
5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการวิจัย

5.3 ข้อเสนอแนะ

1. สรุปผลการวิจัย

งานวิจัยนี้ทำการศึกษาการจำแนกประเภทลูกหนี้สินเชื่อบัตรเครดิตจากสถาบันการเงินแห่งหนึ่งซึ่งข้อมูลที่ใช้คือ ข้อมูลประเภทลูกหนี้และข้อมูลรายการธุรกรรม โดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) และทำการเปรียบเทียบประสิทธิภาพการทำงานของแบบจำลองทั้งหมด 3 แบบจำลองได้แก่ Logistic Regression, Extreme Gradient Boosting, CatBoost ซึ่งเป็นแบบจำลองสำหรับการแยกประเภทข้อมูลร่วมกับการปรับจูนพารามิเตอร์ด้วย GridSearchCV และแก้ปัญหาข้อมูลที่ไม่สมดุล (Imbalanced Data) ด้วย SMOTE เมื่อวัดประสิทธิภาพแบบจำลองด้วยค่า Accuracy, Precision Macro Avg, Recall Macro Avg, F1-Score Macro Avg และ ROC AUC พบว่าแบบจำลอง Extreme Gradient Boosting (XGBoost) ให้ผลลัพธ์ดีที่สุดที่ Accuracy 98%, Precision Macro Avg 97%, Recall Macro Avg 92%, F1-Score Macro Avg 95% และ ROC AUC 93% เมื่อทำการเปรียบเทียบค่าวัดประสิทธิภาพทั้งหมดกับทุกแบบจำลองดังภาพประกอบที่ 44 และจากการพิจารณา Confusion Matrix พบว่าผลการจำแนกประเภทลูกหนี้ปกติมากถึง 7,192 ราย จากลูกหนี้ปกติทั้งหมด 7,194 และลูกหนี้ผิดนัดชำระ 83 รายจากลูกหนี้ผิดนัดชำระทั้งหมด 98 ราย แสดงให้เห็นว่าแบบจำลองสามารถจำแนกประเภทลูกหนี้ได้อย่างถูกต้องส่งผลให้ธนาคารสามารถอนุมัติสินเชื่อเพื่อลดความเสี่ยงที่อาจเกิดขึ้นจากลูกหนี้ผิดนัดชำระ และได้ประโยชน์จากการรับชำระหนี้จากการอนุมัติสินให้ลูกหนี้ปกติ



ภาพประกอบ 44 แสดงการเปรียบเทียบค่า Accuracy, Precision Macro Avg, Recall Macro Avg, F1-Score Macro Avg และ ROC ของทุกแบบจำลอง

2. อภิปรายผลการวิจัย

งานวิจัยนี้ทำการศึกษาการจำแนกประเภทลูกหนี้สินเชื่อบัตรเครดิตจากสถาบันการเงินแห่งหนึ่งโดยข้อมูลที่ใช้คือข้อมูลลูกหนี้และข้อมูลรายการธุรกรรม ซึ่งกำหนดคุณลักษณะของลูกหนี้ด้วยปัจจัยเชิงปริมาณจากแหล่งอ้างอิงธนาคารแห่งประเทศไทย กล่าวคือจะใช้ข้อมูลสถานะลูกหนี้เพื่อกำหนดประเภทลูกหนี้ 2 ประเภท คือ 1. ลูกหนี้ปกติมีจำนวนวันค้างน้อยกว่า 30 วัน 2. ลูกหนี้ผิดนัดชำระมีจำนวนวันค้างตั้งแต่ 30 วัน หรือ ตัดหนี้สูญ (Write-Offs) จากนั้นทำการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด 3 แบบจำลองคือ Logistic Regression, Extreme Gradient Boosting (XGBoost) และ CatBoost ซึ่งเป็นแบบจำลองเพื่อใช้ในการจำแนกประเภทของข้อมูล (Classification) ผู้วิจัยได้เลือกทั้งแบบเชิงเส้นและไม่เชิงเส้นเพื่อนำมาวิเคราะห์เปรียบเทียบประสิทธิภาพ จากผลการทดสอบพบว่าแบบจำลอง Extreme Gradient Boosting ให้ประสิทธิภาพดีที่สุด จากนั้นทำการพิจารณา Confusion Matrix เพื่อสังเกตผลการทำนายถูกและผิดของแบบจำลอง ซึ่งพบว่าลูกหนี้ปกติ เกิดการทำนายถูกมากที่สุด ซึ่งจากการตรวจสอบข้อมูลพบว่าจำนวนข้อมูลลูกหนี้ปกติมีจำนวนข้อมูลแต่ละกลุ่มใดข้อมูลทดสอบพบว่ามีข้อมูลลูกหนี้ปกติมากที่สุด อาจเป็นปัจจัยส่งผลให้แบบจำลองทำนายลูกหนี้ปกติถูกสูงที่สุด ในขณะที่กลุ่มที่เกิดการทำนายคือกลุ่มลูกหนี้ผิดนัดชำระ ดังนั้นผู้วิจัยจึงนำเทคนิค SMOTE เพื่อจัดการปัญหาข้อมูลที่ไม่สมดุลและพบว่าแบบจำลองมีประสิทธิภาพในการจำแนกกลุ่มได้สมดุลทั้งกลุ่มลูกหนี้ปกติและลูกหนี้ผิดนัดชำระ

ผู้วิจัยทำการแสดงค่าความสำคัญของแต่ละฟีเจอร์ที่ส่งผลในการเรียนรู้ของแบบจำลอง (Feature Importance) เพื่อทำความเข้าใจการทำงานแบบจำลองมากขึ้น โดยแบบจำลอง Logistic Regression สามารถดูความสำคัญของฟีเจอร์ได้จากค่าสัมประสิทธิ์ของฟีเจอร์ โดยฟีเจอร์ที่มีอิทธิพลต่อการทำนายแบบจำลองที่สำคัญคือ months_in_book และสำหรับแบบจำลอง XGBoost และ CatBoost นั้นฟีเจอร์ months_in_book มีอิทธิพลต่อการจำแนกกลุ่มมากที่สุดเช่นกัน ซึ่งสรุปได้ว่าฟีเจอร์ months_in_book มีนัยสำคัญส่งผลต่อการทำนายแบบจำลองทั้ง 3 เป็นอย่างมาก

นอกจากนี้ผู้วิจัยได้นำเครื่องมือ LIME เป็นเครื่องมือที่ช่วยวิเคราะห์การเลือกใช้ฟีเจอร์ในการทำนายของแบบจำลอง โดยการเลือกข้อมูลจากชุดทดสอบขึ้นมา 1 ข้อมูลที่แบบจำลองทำนายผิดพลาด เป้าหมายเพื่อศึกษาฟีเจอร์ใดเป็นสาเหตุส่งผลให้แบบจำลองทำนายผิดพลาด ซึ่งได้ทำการศึกษาข้อมูลการทำนายผิดพลาดของทั้ง 2 กลุ่มคือ

1. ข้อมูลที่เลเบลเป็นลูกหนี้ปกติแต่แบบจำลองทำนายเป็นลูกหนี้ผิดนัดชำระ
2. ข้อมูลที่เลเบลเป็นลูกหนี้ผิดนัดชำระแต่แบบจำลองทำนายเป็นลูกหนี้ปกติ

ผู้วิจัยได้ทำการศึกษารูปแบบข้อมูลกับทั้ง 3 แบบจำลองและพบว่าฟีเจอร์ NAME_EDUCATION_TYPE เป็นฟีเจอร์ที่ส่งผลให้แบบจำลองทำนายผิดพลาดเนื่องจากพบในทุกข้อมูลที่แบบจำลองทำนายผิดพลาด

เพื่อให้มั่นใจว่าค่าความสำคัญของฟีเจอร์ (Feature Importance) และ LIME มีความสอดคล้องกัน ผู้วิจัยจึงได้เลือกข้อมูลที่แบบจำลองทำนายถูก จากแบบจำลอง XGBoost เพิ่มเติม เนื่องจากเป็นแบบจำลองที่ให้ประสิทธิภาพดีที่สุดเพื่อสำรวจฟีเจอร์ พบว่าข้อมูลที่ถูกเลือกมีความสอดคล้องกันคือจากการใช้ LIME พบฟีเจอร์ month_in_book เป็นฟีเจอร์ที่ให้ค่ามากที่สุดเมื่อตรวจสอบค่าความสำคัญของฟีเจอร์และข้อมูลส่วนใหญ่หากแบบจำลองให้น้ำหนักต่อฟีเจอร์ดังกล่าวจะมีโอกาสที่แบบจำลองจะทำนายถูกอีกด้วย

สำหรับการนำเครื่องมือ LIME มาใช้นั้นนอกจากช่วยให้ผู้อ่านผลสามารถเข้าใจการทำงานของแบบลองได้มากขึ้นแล้วนั้น ยังช่วยสร้างความน่าเชื่อถือของแบบจำลอง อีกทั้งหากนำไปใช้ต่อในด้านธุรกิจ LIME ยังสามารถทำให้ธนาคารเข้าใจถึงรายละเอียดข้อมูลการอนุมัติและปฏิเสธการกู้สินเชื่อ ส่งผลให้ธนาคารสามารถจัดรูปแบบกลยุทธ์ทางการตลาด จัดการด้านความเสี่ยงที่อาจจะเกิดขึ้น เพื่อประโยชน์สูงสุดต่อธนาคาร

3. ข้อเสนอแนะ

1. จากผลการวิจัยพบว่าค่าที่ได้จากการวัดประสิทธิภาพแบบจำลองนั้นมีค่าที่ค่อนข้างสูง แต่อย่างไรก็ตามเนื่องจากข้อมูลในแต่ละช่วงเวลาอาจส่งผลให้แบบจำลองเกิดการทำนายผิดพลาดได้ เนื่องจากข้อมูลมีความผันผวนจากสถานะการทางเศรษฐกิจ เช่น สถานะการณ์โควิดหรือการช่วยเหลือจากมาตรการรัฐบาลซึ่งอาจส่งผลให้พฤติกรรมการชำระเงินลูกหนี้เปลี่ยนไป ดังนั้นผู้พัฒนาแบบจำลองควรมีการติดตามแบบจำลองอย่างสม่ำเสมอเพื่อให้แบบจำลองสะท้อนถึงลักษณะของลูกหนี้ได้อย่างเหมาะสม

2. เนื่องจากปัจจุบันในกระบวนการอนุมัติสินเชื่อจำเป็นต้องเดินทางไปยังสาขาเพื่อขอสินเชื่อและยังมีระยะเวลาในการอนุมัติสินเชื่อที่ยาวนาน ซึ่งหากนำแบบจำลองมาใช้ร่วมกับแอพพลิเคชั่นในการวิเคราะห์และอนุมัติสินเชื่อที่เหมาะสมแก่ลูกหนี้นั้น จะทำให้ลูกหนี้ได้รับความสะดวกที่ไม่ต้องเดินทางมาใช้บริการด้วยตัวเอง อีกทั้งยังลดระยะเวลาในการอนุมัติสินเชื่ออีกด้วย

บรรณานุกรม

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal, 12*(2), 1545-1556.

Kandel, I., & Castelli, M. (2020). Transfer Learning with Convolutional Neural Networks for Diabetic Retinopathy Image Classification. A Review. *Applied Sciences, 10*(6).

Le, T., Vo, M. T., Vo, B., Lee, M. Y., & Baik, S. W. (2019). A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction. *Complexity, 2019*, 8460934.

Melo, F. (2013). Area under the ROC Curve W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota *Encyclopedia of Systems Biology* (pp. 38-39). New York, NY: Springer New York.

SATPATHY, S. (2020). Overcoming Class Imbalance using SMOTE Techniques. www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques

Solanki, V. (2020). Explainable AI : The Next Level. https://medium.com/analytics-vidhya/explainable-ai-the-next-level-c6b4dadc240

Tierney, B. (2020). Managing imbalanced Data Sets with SMOTE in Python. https://oralytics.com/2019/07/01/managing-imbalanced-data-sets-with-smote-in-python/

Zeyang Dou, M. S., Kun Gao, Zeqiang Jiang. (2019). *Image Smoothing via Truncated Total Variation*. Paper presented at the Recent Advantages of Computer Vision based on Chinese Conference on Computer Vision (CCCV) 2017.

ธนาคารแห่งประเทศไทย. (2016). ประกาศ ธปท. เรื่อง หลักเกณฑ์การจัดชั้นและการกันเงินสำรองของสถาบันการเงินเฉพาะกิจ.

ทองคำ, ส. (2565). MACHINE LEARNING MODELS FOR CREDIT CARD DEFAULT PREDICTION. Paper presented at the Proceeding of the Data Science Conference (DSCon), มหาวิทยาลัยศรีนครินทรวิโรฒ. https://msds.science.swu.ac.th/wp-

content/uploads/2022/04/4_63199130122_Sakulkran-Thongkham_37_48.pdf

Ke, L., Li, C., Zhong, T., Cai, Z., Wen, J., Wang, R., . . . Tang, H. (2021). *Loan Repayment Behavior Prediction of Provident Fund Users Using a Stacking-Based Model*. Paper presented at the 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA).

Lai, L. (2020). *Loan Default Prediction with Machine Learning Techniques*. Paper presented at the 2020 International Conference on Computer Communication and Network Security (CCNS).

Meer, K. (2021). *Machine learning models for mortgage default prediction in Pakistan*. Paper presented at the 2021 International Conference on Artificial Intelligence (ICAI).

Shaheen, S. K., & ElFakharany, E. (2018). *Predictive analytics for loan default in banking sector using machine learning techniques*. Paper presented at the 2018 28th International Conference on Computer Theory and Applications (ICCTA).

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020, 2-4 July 2020). *An Approach for Prediction of Loan Approval using Machine Learning Algorithm.* Paper presented at the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC).

Sun, X. (2020, 25-27 Sept. 2020). *Prediction of the Borrowers' Payback to the Loan with Lending Club Data.* Paper presented at the 2020 International Conference on Modern Education and Information Management (ICMEIM).

Yu, Y. (2020). *The Application of Machine Learning Algorithms in Credit Card Default Prediction*. Paper presented at the 2020 International Conference on Computing and Data Science (CDS).

สำนักงานสภาพัฒนาเศรษฐกิจและสังคมแห่งชาติ. (2021). แผนภาพอัตราการว่างงาน. https://www.nesdc.go.th/main.php?filename=index

ประวัติผู้เขียน