การใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้ เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคาร

USING INTERPRETABLE MACHINE LEARNING METHODS FOR ANALYZING BANK TELEMARKETING DATA

ชิน เลิศวิภาดา

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2565

การใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้ เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคาร

ชิน เลิศวิภาดา

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2565

USING INTERPRETABLE MACHINE LEARNING METHODS FOR ANALYZING BANK TELEMARKETING DATA

CHIN LERTVIPADA

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

2022

สารนิพนธ์

เรื่อง

การใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้ เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคาร

ของ

ชิน เลิศวิภาดา

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)
คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก
(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... ประธาน
(อาจารย์ ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ อาจารย์นุวีย์ วิวัฒนวัฒนา)

ชื่อเรื่อง	การใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้ เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคาร
ผู้วิจัย	ชิน เลิศวิภาดา
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2565
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

สถาบันการเงินมีบทบาทต่อการขับเคลื่อนเศรษฐกิจโดยมีผลิตภัณฑ์การออมเงินเป็นแหล่งเงินทุนหลัก และแม้ว่าจะมีระบบการเงินแบบดิจิทัลแล้ว แต่การติดต่อสื่อสารผ่านทางโทรศัพท์เพื่อนำเสนอผลิตภัณฑ์แก่ลูกค้ายังคงได้รับความนิยม ซึ่งหากไม่มีการวิเคราะห์ข้อมูลก่อนทำการติดต่ออาจส่งผลให้สิ้นเปลืองค่าใช้จ่ายเวลา และก่อให้เกิดประสบการณ์ที่ไม่ดีแก่ลูกค้า จุดประสงค์ของงานวิจัยนี้มุ่งเน้นไปยังการใช้การเรียนรู้ด้วยเครื่องแบบอธิบายได้เพื่อช่วยคัดเลือกคุณลักษณะในการพัฒนาแบบจำลองสำหรับจำแนกกลุ่มลูกค้าที่มีแนวโน้มในการสมัครผลิตภัณฑ์เปรียบเทียบกับวิธีการแบบดั้งเดิม โดยใช้งานชุดข้อมูล ‘Bank Marketing Data Set’ ซึ่งเป็นชุดข้อมูลสาธารณะจาก University of California, Irvine ที่เก็บรวบรวมเกี่ยวกับการนำเสนอผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์ของธนาคารแห่งหนึ่งในประเทศโปรตุเกส ซึ่งชุดข้อมูลมีความไม่สมดุลกันสูงจึงมีการจัดการด้วยเทคนิค Class Weight, Random Under Sampling และ SMOTE ร่วมกับการสร้างแบบจำลอง Logistic Regression, Random Forest, LightGBM และ XGBoost สำหรับการทำนาย รวมไปถึงการคัดเลือกคุณลักษณะด้วยวิธี F-Value, Recursive Feature Elimination และเทคนิคการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธีการแบบ SHAP โดยการประเมินประสิทธิภาพจะเน้นไปยังการตรวจจับ (Recall) กลุ่มลูกค้าที่สมัครผลิตภัณฑ์ โดยค่าประสิทธิภาพอื่นๆ เช่น ความแม่นยำ (Accuracy) ยังอยู่ในเกณฑ์ที่เหมาะสม และการนำ SHAP มาใช้งานสามารถช่วยอธิบายการทำงานของแบบจำลองรวมถึงการทำนายในระดับรายบุคคลได้อย่างชัดเจน โดยจากวิธีการคัดเลือกเพื่อค้นหาคุณลักษณะที่สำคัญสูงสุดสองอันดับ 6 วิธีการพบว่าคุณลักษณะที่ปรากฏบ่อยครั้งที่สุดสองอันดับแรก ได้แก่ ‘อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน’ และ ‘จำนวนพนักงานรายไตรมาส’ ซึ่งแบบจำลองที่มีการใช้งานเพียงสองคุณลักษณะนี้สามารถให้ค่าประสิทธิภาพ Recall ของกลุ่มลูกค้าที่สมัครผลิตภัณฑ์ที่ 71% และ Accuracy ที่ 72% เทียบเท่าแบบจำลองที่ใช้งานทุกคุณลักษณะ นอกจากนี้การวิเคราะห์ความผิดพลาดของแบบจำลองแสดงให้เห็นว่าการทำนายที่ผิดพลาดเนื่องมาจากลักษณะของข้อมูลมีความใกล้เคียงกับข้อมูลอีกกลุ่มอย่างมากจนไม่สามารถจำแนกกลุ่มได้อย่างชัดเจน

คำสำคัญ : การเรียนรู้ด้วยเครื่อง, การเรียนรู้ด้วยเครื่องแบบอธิบายได้, การทำนายการสมัครผลิตภัณฑ์ของธนาคาร, การคัดเลือกคุณลักษณะ, SHAP

Title	USING INTERPRETABLE MACHINE LEARNING METHODS FOR ANALYZING BANK TELEMARKETING DATA
Author	CHIN LERTVIPADA
Degree	MASTER OF SCIENCE
Academic Year	2022
Thesis Advisor	Assistant Professor Dr. Sirisup Laohakiat

Financial institutions play a vital role in driving the economy, with savings products as the primary source of funding. Despite the digital financial systems and products offering products via phone to customers remains popular. Without analysis of prior customer contacting, there is a risk of time-consuming, wasteful expenses, and dissatisfied customers. This research focuses on using interpretable machine learning to discover important features in model development for classifying customers who are likely to apply for a product, compared to traditional methods. The public dataset known as the Bank Marketing Data Set, collected information about offering deposit products by a phone call from a bank in Portugal, was used. The dataset is extremely imbalanced, so Class Weight, Random Undersampling, and SMOTE techniques were implemented along with creating models such as Logistic Regression, Random Forest, LightGBM, and XGBoost for prediction, as well as F-Value, Recursive Feature Elimination and SHAP (SHapley Additive exPlanations) for feature selection. Performance evaluation focuses on the detection of customers who applied for the product (recall), with other metrics such as accuracy and remaining adequate. Using SHAP is able to explain the operations of the model and to clarify individual-level predictions. Regarding the six feature selection techniques, the two most important features frequently appeared were found to be 'euribor3m' and 'nr. employed'. The experiment revealed that classification models with only these two features were able to reach the same capability level, with a recall of positive class at 71% and accuracy at 72%, of models with completed features. Furthermore, error analysis proves that the similarity of instance characteristics is able to mislead the classification models and resulting in inaccurate behavior.

Keyword : Machine Learning, Interpretable Machine Learning, Bank Telemarketing, Feature Selection, SHAP

กิตติกรรมประกาศ

การจัดทำวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากการสนับสนุน ความรู้ ความช่วยเหลือ คำแนะนำ ตลอดจนแนวทางในการทำวิจัยและจัดทำสารนิพนธ์ของ ผศ.ดร.ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษา และคณาจารย์ทุกท่านในภาควิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ การสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

ชิน เลิศวิภาดา

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของการวิจัย

สถาบันการเงินเป็นองค์กรหลักซึ่งมีบทบาทและความสำคัญต่อการพัฒนาและขับเคลื่อนเศรษฐกิจตั้งแต่ระดับรายบุคคลไปจนถึงในระดับประเทศ เนื่องจากเงินทุนเป็นปัจจัยสำคัญทางเศรษฐกิจซึ่งมีอยู่จำกัด สถาบันการเงินจึงต้องจัดสรรเงินทุนเหล่านั้นให้เพียงพอกับความต้องการของหน่วยธุรกิจและประชาชน โดยสถาบันการเงินมีหน้าที่หลักๆ ในการระดมทุนและจัดสรรเงินทุนให้แก่ภาคเศรษฐกิจ รับดำเนินการชำระสินค้าและค่าบริการ การให้บริการการออมเงินและสินเชื่อ การบริหารความเสี่ยง รวมไปถึงการให้ข้อมูลทางการเงินเพื่อประกอบการตัดสินใจต่างๆ

สถาบันการเงินเป็นตัวกลางระหว่างผู้ฝากเงินกับผู้กู้เงิน ซึ่งจะเป็นผู้รับภาระความเสี่ยงแทนเนื่องจากสถาบันการเงินจะนำเงินของผู้ฝากไปปล่อยสินเชื่อแก่ผู้กู้เพื่อสร้างผลกำไร ดังนั้นสถาบันการเงินจึงต้องรับภาระความเสี่ยงแทนผู้ฝากเงิน นอกจากนี้สถาบันการเงินยังเป็นผู้สร้างสภาพคล่องทางการเงิน โดยเป็นผู้ค้ำประกันการออกเช็ค ตั๋วสัญญาใช้เงิน บัตรเครดิต ฯลฯ ให้กับบริษัท ห้างร้าน และประชาชน เพื่อให้ตราสารดังกล่าวได้รับการยอมรับและมีความน่าเชื่อถือ ทำให้การซื้อขายสินค้าต่างๆ ระหว่างผู้ซื้อกับผู้ขายเป็นไปอย่างสะดวกรวดเร็ว

การออมเงินหรือการฝากเงิน ถือเป็นส่วนสำคัญส่วนหนึ่งที่ทำให้ชีวิตสามารถบรรลุเป้าหมายในสิ่งที่มุ่งหวังได้ เช่น การออมเงินเพื่อซื้อที่อยู่อาศัย การออมเงินสำหรับการเดินทางท่องเที่ยว เป็นต้น นอกจากนี้การออมเงินไว้เพื่อใช้ยามฉุกเฉินสามารถลดปัญหาการขาดสภาพคล่องทางการเงินในสถานการณ์ต่างๆ ได้ เช่น เงินสำรองเมื่อเกิดการเจ็บป่วย เงินฉุกเฉินสำหรับซ่อมแซมยานพาหนะ เป็นต้น ดังที่กล่าวมาข้างต้นจะเห็นได้ว่าการออมเงินนั้นถือได้ว่ามีความสำคัญกับชีวิตของทุกคน ซึ่งสามารถทำให้ชีวิตบรรลุเป้าหมายหรือสามารถลดการสร้างภาระหนี้สินได้

ในแง่ของสถาบันการเงิน การออมเงินหรือการฝากเงินถือได้ว่ามีความสำคัญอย่างสูง เนื่องจากเงินฝากถือเป็นแหล่งเงินทุนหลักของสถาบันการเงินสำหรับนำไปใช้ในการดำเนินธุรกิจ ไม่ว่าจะเป็นการนำไปลงทุนประเภทต่างๆ ทั้งในประเทศและต่างประเทศเพื่อให้เกิดผลตอบแทน การปล่อยกู้ยืมสินเชื่อต่อทั้งภาครัฐ ภาคเอกชน และระดับรายย่อย โดยสถาบันการเงินจะได้ผลตอบแทนเป็นดอกเบี้ยจากการกู้ยืม ดังนั้นสถาบันการเงินต่างๆ จึงต้องมีการแข่งขันกันในด้านของผลิตภัณฑ์เงินฝาก ผลตอบแทน การโฆษณา และการตลาด เพื่อโน้มน้าวและดึงดูดให้ลูกค้าเกิดความสนใจในการลงทุนเงินฝากกับสถาบันการเงินนั้นๆ

ผลิตภัณฑ์เงินฝากของสถาบันการเงินมีอยู่เป็นจำนวนมาก ซึ่งอาจจะมีความแตกต่างกันไปในแต่ละสถาบัน ผลิตภัณฑ์เงินฝากที่เป็นที่นิยมส่วนใหญ่ ได้แก่

1. บัญชีเงินฝากกระแสรายวัน (บัญชีเดินสะพัด) เป็นบัญชีเงินฝากที่เน้นความสะดวกสบายสำหรับการดำเนินธุรกิจ เน้นกลุ่มลูกค้าในระดับกลางหรือระดับสูงที่มีเครดิตดี เนื่องจากบัญชีเงินฝากกระแสรายวันสามารถเบิกถอนเงินเกินบัญชีได้ ซึ่งสถาบันการเงินจะมีการคิดอัตราดอกเบี้ยรายวันซึ่งเบิกเกินบัญชี โดยการใช้จ่ายเงินในบัญชีกระแสรายวันจะอยู่ในรูปแบบของการเช็นเช็ค นอกจากนี้บัญชีเงินฝากกระแสรายวันจะไม่มีการจ่ายดอกเบี้ยให้แก่ผู้ฝาก เนื่องจากลูกค้าได้รับความสะดวกสบายในการใช้จ่ายเป็นการตอบแทน

2. บัญชีเงินฝากออมทรัพย์ เป็นบัญชีเงินฝากที่สถาบันการเงินให้บริการโดยเน้นกลุ่มลูกค้ารายย่อย ซึ่งจะมีความสะดวกสบายในการฝากเงินและถอนเงิน โดยไม่มีการกำหนดจำนวนเงินขั้นต่ำในการฝากและถอน บัญชีเงินฝากออมทรัพย์มีการจ่ายดอกเบี้ยเงินฝากให้แก่ผู้ฝากเงินโดยส่วนใหญ่จะมีการจ่ายดอกเบี้ยปีละ 2 ครั้งซึ่งมีการคำนวณเป็นรายวัน แต่อัตราดอกเบี้ยจะค่อนข้างต่ำเนื่องจากไม่ได้มีการกำหนดเวลาในการฝากและถอนเงิน ธนาคารจึงไม่สามารถนำเงินไปลงทุนต่อได้อย่างเต็มประสิทธิภาพ นอกจากนี้ลูกค้าส่วนใหญ่มักจะไม่นำฝากเงินจำนวนมากในบัญชีออมทรัพย์ ส่งผลให้มีจำนวนเงินที่ไม่มากพอในการนำไปลงทุนอย่างมีประสิทธิภาพ

3. บัญชีเงินฝากประจำ เป็นบัญชีเงินฝากที่มีการกำหนดระยะเวลาและจำนวนเงินขั้นต่ำในการฝาก โดยสถาบันการเงินจะจ่ายดอกเบี้ยในอัตราที่สูงกว่าบัญชีออมทรัพย์เนื่องจากการฝากเงินในบัญชีเงินฝากประจำจะมีระยะเวลาในการฝากที่ยาวนานกว่าบัญชีเงินฝากแบบออมทรัพย์ ซึ่งทำให้สถาบันการเงินสามารถนำเงินไปลงทุนต่อยอดได้อย่างเต็มประสิทธิภาพ เช่น การนำเงินไปปล่อยสินเชื่อเงินกู้ การนำเงินไปลงทุนในสินทรัพย์ประเภทต่างๆ เป็นต้น ดังนั้นบัญชีเงินฝากประจำจึงเหมาะสำหรับลูกค้าที่มีเงินสดซึ่งไม่มีความจำเป็นเร่งด่วนที่จะต้องใช้งานและต้องการผลตอบแทนเงินฝากซึ่งมีมูลค่าสูง นอกจากนี้การลงทุนประเภทเงินฝากถือเป็นการลงทุนที่มีความเสี่ยงน้อยหรือไม่มีความเสี่ยง จึงเหมาะสมกับทุกกลุ่มลูกค้าโดยเฉพาะอย่างยิ่งกลุ่มลูกค้าที่สามารถรับความเสี่ยงได้น้อย

จากข้อมูลข้างต้นจะสามารถเห็นได้ว่าผลิตภัณฑ์เงินฝากประจำมีความสำคัญเป็นอย่างมากต่อสถาบันการเงินสำหรับการนำไปใช้ประโยชน์ในการดำเนินธุรกิจเพื่อสร้างผลกำไร

ในสังคมยุคปัจจุบันได้มีการนำสื่อดิจิทัลจำนวนมากเข้ามาเป็นส่วนหนึ่งของการดำรงชีวิตของผู้คน มีกระบวนการในการนำเทคโนโลยีสมัยใหม่มาสร้าง ปรับปรุง หรือเปลี่ยนแปลงแทนที่การดำเนินธุรกิจแบบเก่าในอดีต เพื่อให้ธุรกิจหรือองค์กรสามารถรองรับกับยุคดิจิทัลซึ่งมีการ

เปลี่ยนแปลงอย่างรวดเร็วอยู่ตลอดเวลา หรือที่เรียกว่า Digital Transformation ซึ่งส่งผลดีหลายประการต่อธุรกิจหรือองค์กร ไม่ว่าจะเป็นความสามารถในการเก็บรวบรวมข้อมูลเชิงลึกได้มีประสิทธิภาพมากขึ้น ซึ่งส่งผลดีในด้านของความสามารถในการทำความเข้าใจลูกค้า ทำให้สามารถวางแผนกลยุทธ์ทางธุรกิจเพื่อเพิ่มโอกาสในการประสบความสำเร็จ หรือในด้านของความสามารถในการเพิ่มประสิทธิภาพความคล่องตัวต่อธุรกิจหรือองค์กร เช่น พนักงานขององค์กรสามารถทำงานจากสถานที่ใดก็ได้โดยใช้พื้นที่จัดเก็บข้อมูลแบบบนกลุ่มเมฆ (Cloud Storage) เพื่อจัดเก็บเอกสารแบบดิจิทัล (Soft File) แทนการใช้งานเอกสารแบบกระดาษ ทำให้ผู้มีส่วนเกี่ยวข้องสามารถเข้าถึงเอกสารได้อย่างรวดเร็ว นอกจากนี้ยังมีข้อดีอีกหลายประการที่ยังไม่ถูกกล่าวถึง

ในสังคมยุคดิจิทัลสถาบันการเงินต่างๆ ก็ต้องมีการปรับตัวเพื่อให้มีความสามารถในการแข่งขันและดำเนินธุรกิจในยุคปัจจุบัน โดยการเปลี่ยนแปลงหลักๆ ทางดิจิทัลของสถาบันการเงินประกอบไปด้วยการเปลี่ยนแปลงในระดับภายในองค์กรซึ่งจะเกี่ยวข้องกับการดำเนินธุรกิจและพัฒนาผลิตภัณฑ์ของสถาบันการเงินนั้นๆ เช่น การใช้งานเอกสารดิจิทัลแทนที่การใช้เอกสารแบบกระดาษ การใช้งานโปรแกรมสื่อสารออนไลน์เพื่อดำเนินการประชุม การพัฒนาโปรแกรมโดยใช้งานเทคโนโลยีสมัยใหม่ที่มีประสิทธิภาพสูงเป็นต้น และการเปลี่ยนแปลงต่อประสบการณ์การใช้งานของลูกค้า เช่น การติดต่อสื่อสารกับลูกค้าด้วยจดหมายอิเล็กทรอนิกส์แทนจดหมายแบบกระดาษ การใช้งานการเรียนรู้ด้วยเครื่องเพื่อศึกษาทำความเข้าใจลูกค้าแต่ละรายเพื่อจัดกลุ่มประเภทของลูกค้าเป็นต้น

ตัวอย่างของการวางโครงสร้างพื้นฐานสู่ระบบการเงินดิจิทัล เช่น แผนยุทธศาสตร์ธนาคารแห่งประเทศไทย พ.ศ. 2563-2565 (ธนาคารแห่งประเทศไทย, 2563) โดยมีการยกประเด็น ‘ระบบการเงินเข้าสู่โลกการเงินดิจิทัลอย่างรวดเร็ว’ มาเป็นส่วนหนึ่งในการทำงานของธนาคารแห่งประเทศไทย ซึ่งให้ความสำคัญอย่างมากกับการวางโครงสร้างพื้นฐานสู่บริการทางการเงินดิจิทัล ที่จะช่วยยกระดับคุณภาพชีวิตของประชาชนและศักยภาพของภาคธุรกิจ โดยเฉพาะอย่างยิ่งวิสาหกิจขนาดกลางและขนาดย่อม (SMEs) ตัวอย่างโครงสร้างพื้นฐานที่สำคัญในระบบการเงินดิจิทัลของไทย เช่น

- ระบบชำระเงินอิเล็กทรอนิกส์ (E-Payment) ที่คนไทยคุ้นเคยอย่างพร้อมเพย์ (PromptPay)

- ระบบการพิสูจน์และยืนยันตัวตนทางดิจิทัลข้ามหน่วยงานผ่านแพลตฟอร์ม National Digital ID (NDID) ซึ่งเริ่มดำเนินการโดยภาคธนาคารสำหรับการเปิดบัญชีเงินฝาก

ออนไลน์ ส่งผลให้ลูกค้าไม่จำเป็นต้องเข้าไปดำเนินการที่สาขาของธนาคารหากเคยมีการพิสูจน์และยืนยันตัวตนไว้กับธนาคารหนึ่งแล้ว ซึ่งระบบนี้จะมีการขยายการใช้งานไปยังภาคส่วนอื่นๆ ทั้งบริการทางการเงินของภาครัฐและภาคเอกชน เช่น การยื่นภาษีเงินได้บุคคลธรรมดา การเปิดบัญชีซื้อขายหลักทรัพย์ และการซื้อกรมธรรม์ประกันภัย

ซึ่งระบบเหล่านี้จะเป็นรากฐานสำคัญที่จะช่วยให้ระบบการเงินไทยเข้าสู่ยุคดิจิทัลอย่างแท้จริง

อย่างไรก็ตามแม้ว่าสถาบันการเงินต่างๆ จะพยายามปรับตัวเพื่อเข้าสู่ระบบการเงินแบบดิจิทัลแล้ว แต่เราจะยังเห็นได้ว่าการติดต่อสื่อสารกับลูกค้าส่วนใหญ่ของทางธนาคารยังคงไม่ได้มีการเปลี่ยนแปลงไปจากเดิมมากนัก โดยเฉพาะอย่างยิ่งการติดต่อสื่อสารกับลูกค้าผ่านทางโทรศัพท์ ไม่ว่าจะเป็นการแจ้งข้อมูลข่าวสารทั่วไป การขอยืนยันข้อมูล การแจ้งเตือนต่างๆ การนำเสนอผลิตภัณฑ์ของสถาบันการเงิน เป็นต้น ซึ่งการติดต่อสื่อสารผ่านทางโทรศัพท์ประกอบด้วยข้อดีหลายประการ เช่น

- การได้รับการตอบสนองในทันที เพราะเป็นการติดต่อสื่อสารกับปลายทางโดยตรงและสามารถยืนยันได้ว่าปลายทางได้รับสารข้อมูล

- ความรวดเร็วและความสะดวกสบายในการติดต่อสื่อสาร โดยสามารถติดต่อสื่อสารได้ทันที ไม่มีช่วงระยะเวลาในการรอระหว่างการติดต่อสื่อสาร เช่น จดหมาย

- ความมีประสิทธิภาพในการสื่อสาร การพูดคุยสามารถลดความสับสนหรือซับซ้อน ผู้รับสารสามารถซักถามเพื่อความละเอียดชัดเจนได้

ด้วยเหตุผลต่างๆ ข้างต้น ส่งผลให้สถาบันการเงินทั้งหลายยังคงนิยมใช้งานการติดต่อสื่อสารกับลูกค้าผ่านทางโทรศัพท์ โดยเฉพาะอย่างยิ่งการติดต่อสื่อสารกับลูกค้าเพื่อนำเสนอผลิตภัณฑ์ต่างๆ ของสถาบันการเงินเพื่อโน้มน้าวจูงใจให้ลูกค้าทำการสมัครใช้งาน ซึ่งผลิตภัณฑ์หลักที่มีคุณค่าแก่สถาบันการเงินสูงคือผลิตภัณฑ์บัญชีเงินฝากประจำ ดังนั้นสถาบันการเงินส่วนใหญ่จึงมุ่งความสนใจไปยังการนำเสนอและโน้มน้าวลูกค้าในการสมัครผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์

แม้ว่าการติดต่อสื่อสารเพื่อนำเสนอผลิตภัณฑ์เงินฝากประจำแก่ลูกค้าผ่านทางโทรศัพท์จะมีความจำเป็นและมีข้อดีหลากหลายประการ แต่หากไม่มีการวิเคราะห์หรือพิจารณาความเป็นไปได้ในการสมัครใช้งานผลิตภัณฑ์ของลูกค้าอย่างเหมาะสม ก็อาจจะส่งผลเสียมากกว่าผลดี ตัวอย่างเช่น ค่าใช้จ่ายในการติดต่อสื่อสารกับลูกค้าผ่านทางโทรศัพท์ ซึ่งหากไม่มีการวิเคราะห์ความเป็นไปได้ในการสมัครผลิตภัณฑ์ของลูกค้าก็จะส่งผลให้เกิดค่าใช้จ่ายสิ้นเปลืองจำนวนมาก

ในการติดต่อกับกลุ่มลูกค้าซึ่งไม่มีความสนใจในการสมัครผลิตภัณฑ์เงินฝากประจำ นอกจากนี้ยังเป็นการสิ้นเปลืองในด้านของเวลาที่สถาบันการเงินต้องใช้ติดต่อกับลูกค้าที่ไม่ได้มีความสนใจในผลิตภัณฑ์ ผลเสียที่สำคัญอีกหนึ่งอย่างคือในด้านประสบการณ์การใช้งานของลูกค้า ซึ่งการติดต่อสื่อการในเรื่องที่ไม่ได้อยู่ในความสนใจของลูกค้าหรือการติดต่อสื่อสารที่มีความถี่ที่มากเกินไปก็อาจส่งผลให้ลูกค้าเกิดความรู้สึกถูกรบกวนหรืออาจก่อสร้างความรำคาญแก่ลูกค้าได้ ดังนั้นการวิเคราะห์เพื่อหาแนวโน้มของลูกค้าที่มีโอกาสในการสมัครผลิตภัณฑ์เงินฝากประจำจึงเป็นสิ่งสำคัญที่สถาบันการเงินส่วนใหญ่ให้ความสำคัญ

ในการวิจัยนี้ได้มุ่งเน้นไปยังการใช้งานการเรียนรู้ด้วยเครื่อง (Machine Learning : ML) เพื่อสร้างแบบจำลองสำหรับทำนายแนวโน้มของลูกค้าในการสมัครผลิตภัณฑ์เงินฝากประจำด้วยการนำเสนอผ่านทางโทรศัพท์ และการใช้งานเทคนิคการเรียนรู้ด้วยเครื่องแบบอธิบายได้ (Interpretable Machine Learning : Interpretable ML) สำหรับการอธิบายแบบจำลองทั้งในระดับชุดข้อมูลและระดับรายบุคคล เพื่อให้สามารถเข้าใจได้ว่าคุณลักษณะใดที่ส่งผลต่อโอกาสในการสมัครหรือปฏิเสธผลิตภัณฑ์ของลูกค้า นอกจากนี้ยังมุ่งเน้นในส่วนของการใช้ประโยชน์จากการเรียนรู้ด้วยเครื่องแบบอธิบายได้เพื่อนำมาใช้งานในกระบวนการคัดเลือกคุณลักษณะ (Feature Selection) สำหรับการใช้สร้างแบบจำลอง รวมถึงการใช้วิธีการคัดเลือกคุณลักษณะแบบอื่นๆ เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพในการทำนายดีที่สุด

1.2 วัตถุประสงค์ของการวิจัย

ในการทดลองวิจัยครั้งนี้ได้ทำการตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อศึกษาปัจจัยในการสมัครผลิตภัณฑ์เงินฝากประจำที่ทางสถาบันการเงินมีการนำเสนอผ่านทางโทรศัพท์โดยประยุกต์ใช้เทคนิคต่างๆ เพื่อสร้างแบบจำลองในการทำนายว่าลูกค้าจะทำการสมัครผลิตภัณฑ์หรือไม่

2. เพื่อศึกษาหลักการการทำงานของการเรียนรู้ด้วยเครื่องแบบอธิบายได้สำหรับการนำมาใช้ในการอธิบายแบบจำลองที่ใช้ในการทำนายเพื่อให้สามารถเข้าใจถึงคุณลักษณะที่ส่งผลต่อการทำนายการสมัครผลิตภัณฑ์ทั้งในระดับชุดข้อมูลและระดับรายบุคคล

3. เพื่อศึกษาหลักการการทำงานของการเรียนรู้ด้วยเครื่องแบบอธิบายได้สำหรับการนำมาประยุกต์ใช้ในขั้นตอนการคัดเลือกคุณลักษณะที่ส่งผลต่อการทำนายสูงเพื่อใช้ในสร้างแบบจำลองสำหรับการทำนาย

4. เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในการคัดเลือกคุณลักษณะแบบต่างๆ รวมถึงการประยุกต์ใช้งานเทคนิคการเรียนรู้ด้วยเครื่องแบบอธิบายได้เพื่อการคัดเลือกคุณลักษณะ

5. เพื่อศึกษาหลักการการทำงานของเทคนิคในการแก้ไขปัญหาความไม่สมดุลกันของข้อมูล (Imbalance Data) ในการจำแนกประเภทข้อมูลแบบสองกลุ่มสำหรับใช้ในการสังเคราะห์ข้อมูลที่มีความคล้ายคลึงกับข้อมูลต้นฉบับเพื่อใช้ในการฝึกฝนในขั้นตอนการสร้างแบบจำลอง

1.3 ขอบเขตของการวิจัย

1.3.1 กลุ่มตัวอย่างประชากรที่ใช้ในการวิจัย

ในการดำเนินการวิจัยมีการใช้งานชุดข้อมูล ‘Bank Marketing Data Set’ (https://archive.ics.uci.edu/ml/datasets/bank+marketing) (Moro, Cortez, & Rita, 2014) ซึ่งเป็นชุดข้อมูลสาธารณะแบบเปิดจาก University of California, Irvine โดยเป็นชุดข้อมูลที่เก็บรวบรวมเกี่ยวกับการนำเสนอผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์ของธนาคารแห่งหนึ่งในประเทศโปรตุเกส เพื่อใช้ในการทำนายการตอบรับการสมัครหรือไม่สมัครผลิตภัณฑ์ของลูกค้าธนาคาร โดยมีการเก็บรวบรวมข้อมูลตั้งแต่พฤษภาคม 2551 จนถึงพฤศจิกายน 2553 ซึ่งประกอบด้วยข้อมูลทั้งหมด 41,188 ตัวอย่าง โดยข้อมูลภายในชุดข้อมูลไม่สามารถระบุกลับไปถึงตัวตนของลูกค้าแต่ละรายได้ ชุดข้อมูลประกอบด้วยกลุ่มข้อมูลเบื้องต้นของลูกค้า กลุ่มข้อมูลเกี่ยวกับการติดต่อสื่อสารกับลูกค้า กลุ่มข้อมูลทั่วไป กลุ่มข้อมูลทางเศรษฐศาสตร์โดยรวม และข้อมูลของการสมัครหรือไม่สมัครผลิตภัณฑ์เงินฝากประจำ

1.3.2 คุณลักษณะและตัวแปรของชุดข้อมูลในการวิจัย

คุณลักษณะของตัวอย่างหรือตัวแปรอิสระ ประกอบด้วยข้อมูลดังตารางที่ 1

ตาราง 1 ข้อมูลคุณลักษณะของชุดข้อมูลในการดำเนินงานวิจัย

ลำดับ	คุณลักษณะ	ประเภทข้อมูล	ค่าที่ปรากฏ
1	age (อายุ)	Numerical (เชิงตัวเลข)	
2	job (อาชีพ)	Categorical (หมวดหมู่)	‘admin.’ 'blue-collar' 'entrepreneur' 'housemaid' 'management'

ตาราง 1 (ต่อ)

ลำดับ	คุณลักษณะ	ประเภทข้อมูล	ค่าที่ปรากฏ
			'retired'
			'self-employed'
			'services'
			'student'
			'technician'
			'unemployed'
			'unknown'
3	marital (สถานภาพสมรส)	Categorical (หมวดหมู่)	'divorced'
			'married'
			'single'
			'unknown'
4	education (ระดับการศึกษา)	Categorical (หมวดหมู่)	'basic.4y'
			'basic.6y'
			'basic.9y'
			'high.school'
			'illiterate'
			'professional.course'
			'university.degree'
			'unknown'
5	default (ปรากฏการผิดนัดชำระหนี้)	Categorical (หมวดหมู่)	'no'
			'yes'
			'unknown'
6	housing (ปรากฏสินเชื่อเคหสถาน)	Categorical (หมวดหมู่)	'no'
			'yes'
			'unknown'
7	loan	Categorical	'no'

ตาราง 1 (ต่อ)

ลำดับ	คุณลักษณะ	ประเภทข้อมูล	ค่าที่ปรากฏ
	(ปรากฏสินเชื่อส่วนบุคคล)	(หมวดหมู่)	'yes' 'unknown'
8	contact (ประเภทของการติดต่อ)	Categorical (หมวดหมู่)	'cellular' 'telephone'
9	month (เดือนที่ติดต่อครั้งล่าสุด)	Categorical (หมวดหมู่)	'jan' 'feb' 'mar' ‘apr’ ’may’ ‘jun’ ‘jul’ ‘aug’ ‘sep’ ‘oct’ 'nov' 'dec'
10	day_of_week (วันในสัปดาห์ที่ติดต่อครั้งล่าสุด)	Categorical (หมวดหมู่)	'mon' 'tue' 'wed' 'thu' 'fri'
11	duration (ระยะเวลาเป็นวินาทีของการติดต่อครั้งสุดท้าย)	Numerical (เชิงตัวเลข)	ข้อสังเกต : ข้อมูลมาจากการติดต่อครั้งสุดท้ายเพื่อรับทราบผลการสมัครผลิตภัณฑ์ซึ่งมีอิทธิพลต่อตัวแปร

ตาราง 1 (ต่อ)

ลำดับ	คุณลักษณะ	ประเภทข้อมูล	ค่าที่ปรากฏ
			ผลลัพธ์ค่อนข้างสูง เช่น เมื่อ duration มีค่าเป็น 0 ตัวแปรผลลัพธ์จะมีค่าเป็น ‘no’ หรือ ‘ไม่สมัครผลิตภัณฑ์’ เสมอ ดังนั้นจึงควรถูกนำออกจากการสร้างแบบจำลอง
12	campaign (จำนวนครั้งการติดต่อเพื่อนำเสนอผลิตภัณฑ์นี้)	Numerical (เชิงตัวเลข)	
13	pdays (ระยะเวลาเป็นวันจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า)	Numerical (เชิงตัวเลข)	
14	previous (จำนวนครั้งการติดต่อก่อนการนำเสนอผลิตภัณฑ์นี้)	Numerical (เชิงตัวเลข)	
15	poutcome (ผลลัพธ์จากการนำเสนอผลิตภัณฑ์ก่อนหน้า)	Categorical (หมวดหมู่)	'failure', 'nonexistent', 'success'
16	emp.var.rate (อัตราการจ้างงานรายไตรมาส)	Numerical (เชิงตัวเลข)	
17	cons.price.idx (ดัชนีราคาผู้บริโภครายเดือน)	Numerical (เชิงตัวเลข)	
18	cons.conf.idx (ดัชนีความเชื่อมั่นผู้บริโภครายเดือน)	Numerical (เชิงตัวเลข)	

ตาราง 1 (ต่อ)

ลำดับ	คุณลักษณะ	ประเภทข้อมูล	ค่าที่ปรากฏ
19	euribor3m (อัตราดอกเบี้ยกู้ยืมรายวันระหว่างธนาคารภายในยุโรป)	Numerical (เชิงตัวเลข)	
20	nr.employed (จำนวนพนักงานรายไตรมาส)	Numerical (เชิงตัวเลข)	

ผลลัพธ์ของตัวอย่างหรือตัวแปรทำนายหรือตัวแปรตาม ได้แก่ผลของการสมัครหรือไม่สมัครผลิตภัณฑ์เงินฝากประจำของธนาคาร ดังตารางที่ 2

ตาราง 2 ข้อมูลผลลัพธ์ของชุดข้อมูลในการดำเนินงานวิจัย

ลำดับ	คุณลักษณะ	ประเภทข้อมูล	ค่าที่ปรากฏ
1	Y (ผลการสมัครผลิตภัณฑ์เงินฝากประจำ)	Categorical (หมวดหมู่)	‘yes’ ‘no’

1.3.3 กรอบแนวคิดของการวิจัย

การวิจัยการใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคารในครั้งนี้ ประกอบด้วยขั้นตอนหลักๆ จำนวน 8 ขั้นตอน ดังนี้

1. กระบวนการรวบรวมและสำรวจข้อมูล (Exploratory Data Analysis : EDA) เป็นกระบวนการในการคัดเลือกชุดข้อมูลเพื่อนำมาใช้สำหรับดำเนินการวิจัย รวมถึงการนำเข้าข้อมูล การวิเคราะห์ข้อมูลทางด้านสถิติและความสัมพันธ์ของข้อมูลในด้านอื่นๆ

2. กระบวนการเตรียมข้อมูล (Feature Engineering and Data Pre-Processing) เป็นกระบวนการซึ่งประกอบด้วยการทำความสะอาดข้อมูล (Data Cleansing) การแปลงข้อมูลให้อยู่ในรูปแบบซึ่งสามารถนำไปใช้ในการสร้างแบบจำลองได้ (Data Transformation) การแบ่งข้อมูลสำหรับการเรียนรู้และการทดสอบ (Data Train-Test Split) โดยการแบ่งข้อมูลจะมีการใช้เทคนิคในการแบ่งข้อมูลเพื่อให้สามารถแบ่งข้อมูลได้อย่างมีประสิทธิภาพ

3. กระบวนการคัดเลือกคุณลักษณะ (Feature Selection) เป็นกระบวนการในการคัดเลือกคุณลักษณะที่ส่งผลหรือมีความสัมพันธ์ต่อการทำนายผลลัพธ์ในการสมัครผลิตภัณฑ์ เพื่อลดการรบกวนของตัวแปรที่ไม่เกี่ยวข้องหรือมีความผันผวนสูง (Noise) และเป็นการเพิ่มประสิทธิภาพของการคำนวณในขั้นตอนการสร้างแบบจำลอง โดยสามารถช่วยลดความซับซ้อน (Complexity) และเวลาที่ใช้ในการสร้างแบบจำลองได้ ซึ่งในการวิจัยครั้งนี้มีการทดลองใช้งานเทคนิค F-Score หรือ F-Value และ Recursive Feature Elimination (RFE)

4. กระบวนการจัดการความไม่สมดุลของข้อมูล (Imbalance Data Handling) เป็นกระบวนการในการจัดการข้อมูลซึ่งมีความไม่เท่ากันของจำนวนข้อมูลในแต่ละกลุ่ม เนื่องจากชุดข้อมูลมีความไม่สมดุลกันของข้อมูลที่ค่อนข้างสูง โดยมีจำนวนข้อมูลของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์น้อยกว่ากลุ่มลูกค้าที่ไม่ได้สมัครผลิตภัณฑ์อยู่เป็นจำนวนมาก หากไม่มีการดำเนินการแก้ไขความไม่สมดุลกันของข้อมูลจะทำให้เกิดความโน้มเอียง (Bias) ในขั้นตอนการสร้างแบบจำลอง ดังนั้นจึงต้องมีการจัดการความไม่สมดุลของข้อมูลเพื่อป้องกันการมีอิทธิพลของกลุ่มข้อมูลส่วนใหญ่ที่จะส่งผลให้แบบจำลองเกิดความโน้มเอียงและความผิดพลาดในการทำนายได้ โดยจะมีโอกาสทำนายลูกค้าที่ทำการสมัครผลิตภัณฑ์เป็นไม่ทำการสมัครผลิตภัณฑ์สูง ในการวิจัยครั้งนี้มีการใช้งานเทคนิค Class Weight, Random Undersampling และ SMOTE (Synthetic Minority Oversampling Technique) ในการจัดการความไม่สมดุลกันของข้อมูล

5. กระบวนการสร้างแบบจำลองและประเมินประสิทธิภาพ (Model Creation and Evaluation) ในการวิจัยครั้งนี้มีการเลือกใช้แบบจำลองต่างๆ ซึ่งมีความน่าสนใจ มีประสิทธิภาพสูงและเป็นที่ยอมรับในยุคปัจจุบันจำนวน 4 แบบจำลอง ประกอบด้วย Logistic Regression (LR) สำหรับการใช้งานเป็นฐาน (Baseline) ของการประเมินประสิทธิภาพ, Random Forest (RF), Light Gradient Boosting Machine (LightGBM) และ eXtreme Gradient Boosting (XGBoost) ซึ่งแบบจำลองแต่ละแบบจะมีหลักของการทำงานที่แตกต่างกันไป

5.1 ขั้นตอนการสร้างแบบจำลองจะมีการใช้งานเทคนิค ‘GridSearchCV’ ซึ่งเป็นเทคนิคเพื่อช่วยค้นหาตัวแปร (Model Tuning) ที่ดีที่สุดของแต่ละแบบจำลอง โดยข้อมูลสำหรับการเรียนรู้จะถูกแบ่งย่อยเป็นข้อมูลสำหรับการฝึกสอน (Training Set) และข้อมูลสำหรับการตรวจสอบ (Validation Set) ซึ่งจะมีการใช้งานเทคนิคการทำงานแบบสายท่อ (Pipeline) เพื่อจัดลำดับขั้นตอนการทำงานให้เป็นระบบและป้องกันการเกิดการรั่วไหลของข้อมูล (Data Leakage) โดยขั้นตอนที่ถูกบรรจุอยู่ในการทำงานแบบสายท่อประกอบด้วย การเตรียมข้อมูล การ

คัดเลือกคุณลักษณะ การจัดการความไม่สมดุลของข้อมูล และการสร้างแบบจำลองด้วยตัวแปรต่างๆ

5.2 ขั้นตอนการประเมินผลประสิทธิภาพของแบบจำลองมีการใช้งานค่าความแม่นยำโดยรวมของแบบจำลอง (Accuracy), ค่าความครบถ้วนสมบูรณ์ในการตรวจจับข้อมูลที่สนใจ (Recall), ค่าความถูกต้องในการตรวจจับข้อมูลที่สนใจ (Precision) และค่าเฉลี่ยระหว่าง Recall กับ Precision (F1-Score) สำหรับกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ ซึ่งเป็นการประเมินประสิทธิภาพที่มุ่งเน้นไปยังการตรวจจับกลุ่มลูกค้าที่เราสนใจ

6.กระบวนการอธิบายแบบจำลองด้วยการเรียนรู้ของเครื่องแบบอธิบายได้ด้วยเทคนิค Shapley Additive Explanations (SHAP) ซึ่งเป็นกระบวนการในการอธิบายแบบจำลองที่สร้างขึ้นมาเพื่อให้ทราบถึงคุณลักษณะที่สำคัญซึ่งส่งผลต่อการทำนายในแต่ละแบบจำลอง และยังสามารถบอกระดับของอิทธิพลในการส่งผลต่อการทำนายได้ โดยมีการใช้งานเทคนิค SHAP เพื่ออธิบายแบบจำลองทั้งในระดับชุดข้อมูลและในระดับรายบุคคล

7. กระบวนการเปรียบเทียบประสิทธิภาพของกระบวนการคัดเลือกคุณลักษณะแบบต่างๆ รวมทั้งการประยุกต์ใช้การเรียนรู้ของเครื่องแบบอธิบายได้ในการคัดเลือกคุณลักษณะเป็นขั้นตอนในการประยุกต์ใช้เทคนิคการอธิบายแบบจำลองด้วย SHAP มาเพื่อใช้สำหรับกระบวนการคัดเลือกคุณลักษณะสำหรับทำการสร้างแบบจำลอง โดยเปรียบเทียบผลลัพธ์ของการคัดเลือกคุณลักษณะแบบดั้งเดิมซึ่งประกอบด้วยเทคนิค F-Value และ Recursive Feature Elimination (RFE) กับผลลัพธ์ของการคัดเลือกคุณลักษณะจากการประยุกต์ใช้งานการเรียนรู้ของเครื่องแบบอธิบายได้ด้วยเทคนิค SHAP ว่ามีลักษณะที่คล้ายคลึงหรือแตกต่างกัน จากนั้นเปรียบเทียบประสิทธิภาพของแบบจำลองที่สร้างจากการคัดเลือกคุณลักษณะที่แตกต่างกัน

8. กระบวนการวิเคราะห์ความผิดพลาดของแบบจำลอง เป็นขั้นตอนในการวิเคราะห์ถึงสาเหตุที่ส่งผลให้แบบจำลองทำงานผิดพลาด โดยมีการใช้งานเทคนิคต่างๆ เพื่อทำการวิเคราะห์และทำความเข้าใจเพื่อให้สามารถแยกแยะความผิดพลาดในการทำนายได้ เช่น เป็นความผิดพลาดเนื่องจากเป็นข้อมูลที่ไม่เคยมีมาก่อนในกระบวนการเรียนรู้ของแบบจำลอง หรือเป็นความผิดพลาดเนื่องจากแบบจำลองไม่มีประสิทธิภาพเพียงพอ เป็นต้น เพื่อให้สามารถนำไปใช้ปรับปรุงและพัฒนาแบบจำลองหรือกระบวนการวิจัยให้มีประสิทธิภาพสูงขึ้น

1.4 สมมุติฐานในการวิจัย

1. แบบจำลองทั้ง 4 แบบซึ่งประกอบด้วย Logistic Regression, Random Forest, LightBGM และ XGBoost จะมีประสิทธิภาพในการทำนายที่ใกล้เคียงกัน โดยแบบจำลอง XGBoost จะมีประสิทธิภาพดีที่สุด

2. การใช้งานเทคนิคการจัดการความไม่สมดุลของข้อมูลแบบ Class Weight, Random Undersampling และ SMOTE จะช่วยให้แบบจำลองมีประสิทธิภาพในการทำนายที่ดีขึ้น โดยสามารถลดความโน้มเอียงในการทำนายของแบบจำลองได้ โดยเทคนิค SMOTE จะมีประสิทธิภาพที่ดีกว่า Random Undersampling

3. การใช้งานกระบวนการการคัดเลือกคุณลักษณะ จะช่วยเพิ่มประสิทธิภาพในขั้นตอนการสร้างแบบจำลองในแง่ของทรัพยากรและเวลาในการเรียนรู้ โดยประสิทธิภาพในการทำนายของแบบจำลองจะลดลงเพียงเล็กน้อย

4. คุณลักษณะที่มีอิทธิพลต่อการทำนายซึ่งได้มาจากการใช้งานเทคนิคการเรียนรู้ด้วยเครื่องแบบอธิบายได้ จะมีผลลัพธ์ใกล้เคียงกับคุณลักษณะที่ได้มาจากการใช้งานกระบวนการการคัดเลือกคุณลักษณะแบบดั้งเดิม

5. การใช้งานเทคนิคการเรียนรู้ด้วยเครื่องแบบอธิบายได้สามารถนำมาใช้งานการคัดเลือกคุณลักษณะได้ ซึ่งประสิทธิภาพของแบบจำลองจะเทียบเท่าหรือสูงกว่าการใช้งานกระบวนการการคัดเลือกคุณลักษณะแบบดั้งเดิม

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สามารถนำแบบจำลองไปใช้ประโยชน์ในการทำนายการนำเสนอผลิตภัณฑ์เงินฝากประจำของธนาคารผ่านทางโทรศัพท์ได้ว่าลูกค้ามีแนวโน้มจะทำการสมัครผลิตภัณฑ์หรือไม่

2. ทำให้ทราบถึงคุณลักษณะหรือปัจจัยที่มีอิทธิพลหรือส่งผลต่อการสมัครหรือไม่สมัครผลิตภัณฑ์ เพื่อให้สามารถระบุกลุ่มของลูกค้าที่มีแนวโน้มในการสมัครผลิตภัณฑ์ได้

3. ข้อมูลผลลัพธ์จากการทำนายสามารถนำไปปรับปรุงประสิทธิภาพในการนำเสนอผลิตภัณฑ์แก่ลูกค้าที่มีแนวโน้มสนใจในผลิตภัณฑ์ เป็นการลดค่าใช้จ่ายและลดการรบกวนในการติดต่อกับลูกค้ากลุ่มที่มีแนวโน้มไม่สนใจในผลิตภัณฑ์

4. เนื่องจากมีการใช้งานเทคนิคการเรียนรู้ด้วยเครื่องแบบอธิบายได้ ทำให้สามารถวิเคราะห์แนวโน้มในการสมัครผลิตภัณฑ์ของลูกค้าธนาคารได้แบบรายบุคคล โดยสามารถบ่งบอกถึงอิทธิพลของแต่ละคุณลักษณะได้

5. สามารถนำไปพัฒนาเพื่อต่อยอดในการสร้างแบบจำลองสำหรับการนำเสนอผลิตภัณฑ์อื่นๆ ของธนาคารได้

โดยโครงสร้างของงานวิจัยเล่มนี้จะประกอบไปด้วยเนื้อหาต่างๆ ดังนี้ บทที่ 2 ทบทวนวรรณกรรม บทที่ 3 วิธีการดำเนินการวิจัย บทที่ 4 ผลการดำเนินการวิจัย และบทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

บทที่ 2
ทบทวนวรรณกรรม

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่มีความเกี่ยวข้องกับการดำเนินการวิจัยและได้มีการนำเสนอตามหัวข้อต่อไปนี้

1. ทฤษฎีเกี่ยวกับการเรียนรู้ด้วยเครื่อง (Machine Learning)
2. ทฤษฎีเกี่ยวกับแบบจำลองสำหรับงานในการจำแนกประเภท (Classification Model)
3. ทฤษฎีเกี่ยวกับการจัดการข้อมูลที่ไม่สมดุล (Imbalance Data Handling)
4. ทฤษฎีเกี่ยวกับวิศวกรรมคุณลักษณะข้อมูล (Feature Engineering)
5. ทฤษฎีเกี่ยวกับการคัดเลือกคุณลักษณะ (Feature Selection)
6. ทฤษฎีเกี่ยวกับการประเมินผลประสิทธิภาพของแบบจำลอง (Model Evaluation)
7. ทฤษฎีเกี่ยวกับการเรียนรู้ด้วยเครื่องแบบอธิบายได้ (Interpretable Machine Learning)
8. งานวิจัยที่เกี่ยวข้อง (Literature Review)

2.1 ทฤษฎีเกี่ยวกับการเรียนรู้ด้วยเครื่อง (Machine Learning)

การเรียนรู้ด้วยเครื่องเป็นศาสตร์แขนงหนึ่งซึ่งได้รับอิทธิพลมาจากการศึกษาวิจัยในหลากหลายสาขาวิชารวบรวมมาไว้ด้วยกัน เช่น สาขาวิชาวิทยาการคอมพิวเตอร์ (Computer Science), สาขาวิชาทางสถิติ (Statistics), สาขาวิชาชีววิทยา (Biology) และสาขาวิชาจิตวิทยา (Psychology) โดยการเรียนรู้ของเครื่องคือการศึกษาทางวิทยาศาสตร์เกี่ยวกับอัลกอริทึม (Algorithm) และแบบจำลองทางสถิติ (Statistic Model) ซึ่งมุ่งเน้นไปที่การพัฒนาโปรแกรมคอมพิวเตอร์เพื่อให้สามารถเข้าถึงข้อมูลและนำข้อมูลไปใช้ในการเรียนรู้ด้วยตนเองสำหรับการทำงานเฉพาะอย่างให้เกิดประสิทธิภาพโดยไม่ต้องมีการออกคำสั่งที่ชัดเจนจากผู้ใช้งาน โดยจะอาศัยการคำนวณในรูปแบบของการอนุมานและการแทนที่

การเรียนรู้ด้วยเครื่องสามารถถือได้ว่าเป็นส่วนหนึ่งของเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence : AI) โดยวัตถุประสงค์หลักของการเรียนรู้ด้วยเครื่องคือการพยายามสอนให้คอมพิวเตอร์สามารถที่จะเรียนรู้ในการนำข้อมูลเก่าที่เคยเกิดขึ้นมาแล้วหรือประสบการณ์ต่างๆ ที่เคยพบมาก่อน เพื่อนำมาใช้ในการทำนาย จำแนกประเภท หรือแก้ไขปัญหาบางสิ่งบางอย่างได้อย่างมีประสิทธิภาพ (Manghani, 2017) เช่น การจำแนกประเภทของจดหมายอิเล็กทรอนิกส์ (Email) ว่าเป็นจดหมายขยะ (Junk Mail) หรือจดหมายรบกวน (Spam) หรือไม่, การวิเคราะห์

ข้อมูลการขายสินค้าย้อนหลังสำหรับงานในการทำนายพฤติกรรมการซื้อของผู้บริโภค และการตรวจจับการฉ้อโกง (Fraud Detection) เป็นต้น

นอกจากนี้การเรียนรู้ด้วยเครื่องยังถูกจำกัดคำนิยามและความหมายจากสถาบันการศึกษา หรือบริษัทชั้นนำในกลุ่มธุรกิจอุสาหกรรมทางด้านเทคโนโลยีสารสนเทศไว้อย่างน่าสนใจ ซึ่งส่วนใหญ่จะมีคำจำกัดความที่ใกล้เคียงกัน ดังตัวอย่างต่อไปนี้

การเรียนรู้ด้วยเครื่องเป็นแขนงย่อยหนึ่งของนวัตกรรมปัญญาประดิษฐ์และวิทยาการคอมพิวเตอร์ที่มุ่งเน้นไปยังการใช้ประโยชน์จากข้อมูลและชุดคำสั่งขั้นตอนวิธีหรืออัลกอริทึม เพื่อลอกเลียนแบบกระบวนการเรียนรู้ของมนุษย์เพื่อเพิ่มความแม่นยำของเครื่อง

การเรียนรู้ด้วยเครื่องถือเป็นองค์ประกอบที่สำคัญในการเจริญเติบโตของสาขาวิทยาการคอมพิวเตอร์ อัลกอริทึมของการเรียนรู้ด้วยเครื่องได้รับการฝึกฝนผ่านการใช้งานขั้นตอนกระบวนการทางสถิติสำหรับนำไปใช้ในการจัดกลุ่มของข้อมูล ทำนายหรือคาดการณ์ผลลัพธ์บางอย่าง รวมไปถึงการค้นหาข้อมูลเชิงลึกในงานเกี่ยวกับการทำเหมืองของข้อมูล ซึ่งประโยชน์จากสิ่งที่ได้เหล่านี้จะเป็นตัวช่วยตอบคำถาม ค้นหารูปแบบกลยุทธ์ ขับเคลื่อนหรือสนับสนุนกระบวนการตัดสินใจ ของภาคธุรกิจหรือองค์กรเพื่อให้เกิดความเจริญเติบโตและเกิดประโยชน์ที่สูงที่สุดแก่องค์กร (IBM, 2020a)

การเรียนรู้ด้วยเครื่องเป็นกระบวนการในการใช้งานแบบจำลองทางสถิติซึ่งเกี่ยวข้องกับข้อมูลเพื่อนำมาช่วยให้คอมพิวเตอร์สามารถเรียนรู้ได้ด้วยตนเองโดยไม่จำเป็นต้องป้อนชุดคำสั่งที่ชัดเจนลงไป ซึ่งถือเป็นส่วนย่อยหนึ่งของปัญญาประดิษฐ์ โดยการเรียนรู้ด้วยเครื่องจะใช้งานอัลกอริทึมเพื่อค้นหารูปแบบบางอย่างจากชุดข้อมูลที่ได้รับ จากนั้นจะสร้างแบบจำลองของข้อมูลจากรูปแบบที่ค้นพบเพื่อใช้ทำนายหรือคาดการณ์บางสิ่งบางอย่าง โดยเมื่อมีข้อมูลที่มากขึ้นและการเรียนรู้ที่ยาวนานขึ้นจะทำให้ประสิทธิภาพของการเรียนรู้ด้วยเครื่องมีความแม่นยำมากขึ้น ซึ่งมีความคล้ายคลึงกับการฝึกฝนเรียนรู้ของมนุษย์

ความสามารถในการปรับปรุงประสิทธิภาพของตัวเองในการเรียนรู้ด้วยเครื่องเป็นสิ่งที่มีประโยชน์ต่อสถาณการณ์ต่างๆ เช่น เมื่อข้อมูลมีจำนวนมากและเปลี่ยนแปลงอย่างรวดเร็วอยู่เสมอๆ เมื่อลักษณะของงานหรือความต้องการมีการเปลี่ยนแปลงตลอดเวลา หรือเมื่อการค้นหาคำตอบด้วยวิธีแบบดั้งเดิมเป็นไปได้ยาก (Microsoft)

ปัญญาประดิษฐ์เป็นคำจำกัดความอย่างกว้างสำหรับอ้างอิงถึงระบบที่ลอกเลียนความสามารถในการคิดและเรียนรู้ของมนุษย์ โดยปัญญาประดิษฐ์และการเรียนรู้ของเครื่องมักจะถูกอภิปรายร่วมกันอยู่เสมอๆ และในบางครั้งบางคราวมักถูกนำมาใช้เรียกแทนกันแม้โดยพื้นฐาน

แล้วทั้งสองกระบวนการจะมีความแตกต่างกันอยู่ โดยความแตกต่างที่สำคัญคือการเรียนรู้ด้วยเครื่องทุกประเภทถือเป็นแขนงย่อยของปัญญาประดิษฐ์ แต่มีเพียงปัญญาประดิษฐ์เพียงบางชนิดที่ถือเป็นการเรียนรู้ด้วยเครื่อง

ในปัจจุบันการเรียนรู้ด้วยเครื่องถูกใช้งานกันอย่างแพร่หลายในธุรกิจต่าง ๆ ไม่ว่าจะเป็นการเงินการธนาคาร การซื้อขายสินค้าทางออนไลน์ หรือการใช้งานสื่อเครือข่ายทางสังคมต่าง ๆ (Social Networks) โดยการเรียนรู้ด้วยเครื่องเข้ามามีบทบาทสำคัญในการช่วยพัฒนาเพื่อยกระดับความมีประสิทธิภาพ ความปลอดภัย และความสะดวกสบายแก่ผู้ใช้งาน (Oracle)

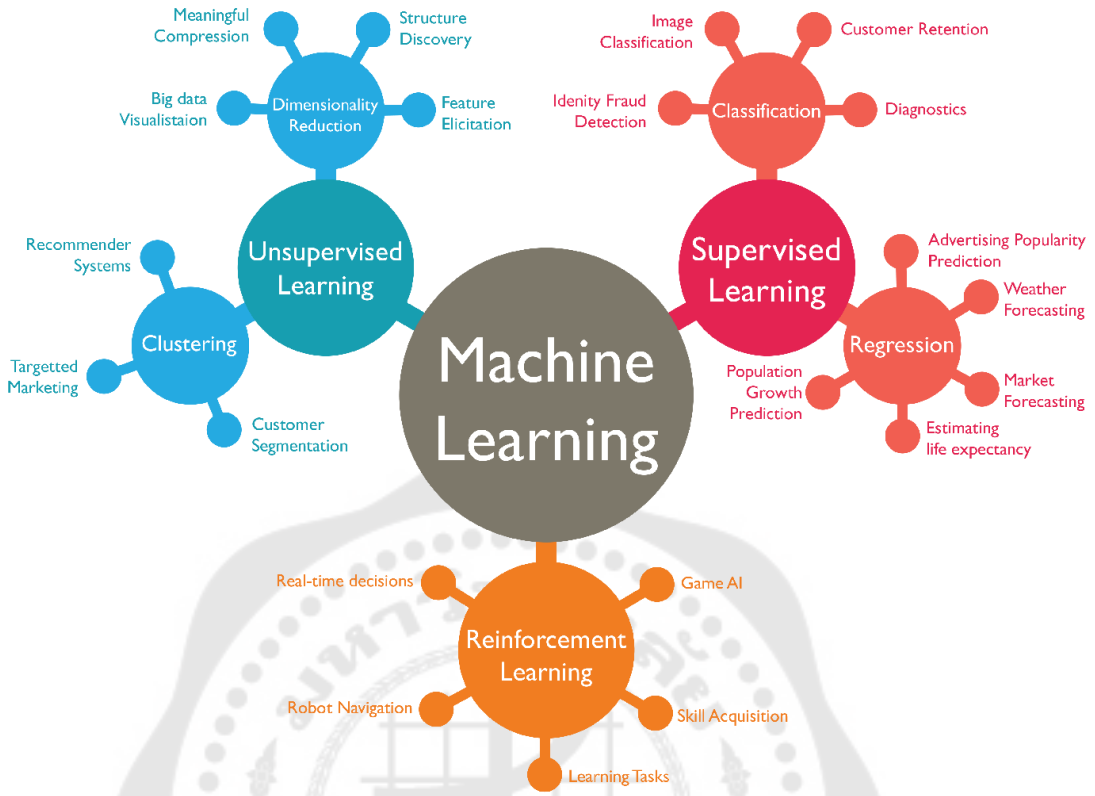
การเรียนรู้ด้วยเครื่องเป็นส่วนย่อยหนึ่งของเทคโนโลยีปัญญาประดิษฐ์ซึ่งมุ่งเน้นการสอนให้คอมพิวเตอร์เรียนรู้จากข้อมูลและปรับปรุงความสามารถจากประสบการณ์ แทนการตั้งโปรแกรมหรือชุดคำสั่งที่ชัดเจน โดยในการเรียนรู้ด้วยเครื่องอัลกอริทึมจะมีการฝึกฝนเพื่อค้นหารูปแบบและความสัมพันธ์ภายในชุดข้อมูลขนาดใหญ่ สำหรับทำงานที่เกี่ยวข้องกับการตัดสินใจและการคาดการณ์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด การเรียนรู้ของเครื่องจะยิ่งแม่นยำมากขึ้นเมื่อสามารถเข้าถึงข้อมูลที่มีจำนวนมากขึ้น ซึ่งเราสามารถพบการใช้งานการเรียนรู้ด้วยเครื่องรอบๆ ตัวเราอยู่ตลอดเวลา เช่น ภายในที่พักอาศัย การซื้อขายสินค้า สื่อบันเทิง และการดูแลเกี่ยวกับสุขภาพ (SAP)

การเรียนรู้ด้วยเครื่องเป็นรูปแบบหนึ่งของการวิเคราะห์ข้อมูลด้วยการใช้อัลกอริทึมที่เรียนรู้จากชุดข้อมูลอย่างต่อเนื่อง การเรียนรู้ด้วยเครื่องสามารถทำให้คอมพิวเตอร์เรียนรู้ ค้นหาและจดจำรูปแบบบางอย่างที่ซ่อนอยู่ในชุดข้อมูลโดยไม่จำเป็นต้องมีการระบุชุดคำสั่งที่แน่นอน ประโยชน์สำคัญของการเรียนรู้ด้วยเครื่องคือแบบจำลองที่มีการเรียนรู้และถูกสร้างขึ้นมาสามารถนำไปปรับใช้เพื่อทำนายหรือคาดการณ์ผลลัพธ์กับข้อมูลใหม่ๆ ที่ยังไม่เคยพบเจอมาก่อนได้อย่างมีประสิทธิภาพและเป็นที่น่าพอใจ (Manghani, 2017)

ประเภทของการเรียนรู้ด้วยเครื่องสามารถแบ่งออกเป็นประเภทหลักๆ ได้ดังนี้

1. การเรียนรู้ด้วยเครื่องแบบมีผู้สอน (Supervised Learning)
2. การเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน (Unsupervised Learning)
3. การเรียนรู้ด้วยเครื่องแบบเสริมกำลัง (Reinforcement Learning)

โดยในการวิจัยครั้งนี้เราจะมุ่งเน้นไปยังทฤษฎีของการเรียนรู้ด้วยเครื่องแบบมีผู้สอนเป็นหลัก



ภาพประกอบ 1 แสดงประเภทของการเรียนรู้ด้วยเครื่อง

ที่มา: (Khadka, 2017)

2.1.1 **การเรียนรู้ด้วยเครื่องแบบมีผู้สอน** (Supervised Learning)

การเรียนรู้ด้วยเครื่องแบบมีผู้สอนจะเป็นการค้นหารูปแบบและสร้างแบบจำลองโดยใช้งานข้อมูลที่มีป้ายกำกับประเภท (Label) ซึ่งเป็นผลลัพธ์ที่แท้จริงของตัวอย่างข้อมูลนั้นๆ ยกตัวอย่างเช่น อีเมล หรือ จดหมายอิเล็กทรอนิกส์สามารถถูกจัดหมวดหมู่ให้อยู่ในประเภทจดหมายทั่วไปหรือจดหมายขยะได้ โดยในขั้นตอนการเรียนรู้ของแบบจำลองจะมีการใช้งานชุดข้อมูลซึ่งประกอบด้วยคุณลักษณะของแต่ละข้อมูลรวมไปถึงป้ายกำกับของแต่ละตัวอย่างข้อมูล ซึ่งการเรียนรู้ของเครื่องจะดำเนินการแบบวนซ้ำไปเรื่อยๆ เพื่อลดความผิดพลาดและเพิ่มประสิทธิภาพของแบบจำลอง โดยทำการวนซ้ำตามจำนวนรอบที่กำหนดหรือวนซ้ำจนได้ประสิทธิภาพของแบบจำลองที่เป็นที่น่าพอใจ เมื่อเสร็จสิ้นกระบวนการเรียนรู้จะเป็นขั้นตอนในการประเมินประสิทธิภาพของแบบจำลองโดยใช้งานชุดข้อมูลสำหรับการประเมินผลซึ่งมีป้ายกำกับที่ถูกต้องของข้อมูลอยู่เช่นกัน แต่จะไม่ถูกนำไปใช้ในขั้นตอนการทำนายของแบบจำลอง จากนั้นจะนำผลการทำนายของแบบจำลองมาเปรียบเทียบกับป้ายกำกับที่แท้จริงของข้อมูลนั้นๆ เพื่อ

คำนวณประสิทธิภาพของแบบจำลอง โดยสามารถนำความผิดพลาดกลับไปปรับปรุงแบบจำลองเพื่อให้มีประสิทธิภาพที่สูงขึ้นจนถึงเกณฑ์ที่ยอมรับได้

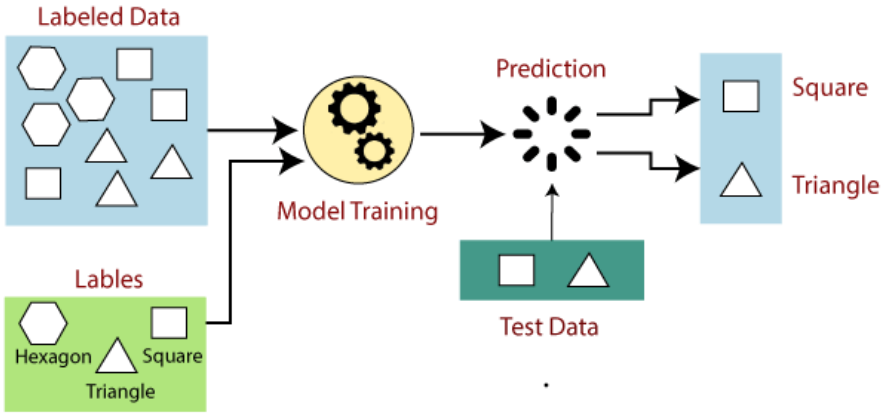
เป้าหมายสำคัญของการเรียนรู้ด้วยเครื่องแบบมีผู้สอนคือการสร้างแบบจำลองสำหรับการใช้งานในการทำนายหรือจำแนกประเภทของข้อมูล ซึ่งแต่ละกลุ่มของผลลัพธ์จากการจำแนกประเภทจะถูกกำหนดไว้โดยชื่อประเภทของข้อมูลนั้นๆ เพื่อใช้เป็นป้ายกำกับสำหรับข้อมูลแต่ละตัวที่ทำการทดสอบ โดยในขั้นตอนการจำแนกประเภทตัวแบบจำลองจะทราบถึงคุณลักษณะของข้อมูลนั้นๆ แต่จะไม่ทราบถึงป้ายกำกับของแต่ละข้อมูล

โดยส่วนมากการเรียนรู้ด้วยเครื่องแบบมีผู้สอนมักจะถูกนำมาใช้กับงานที่เรียนรู้จากข้อมูลในอดีตเพื่อทำการทำนายผลลัพธ์ในอนาคต โดยงานในการวิเคราะห์การถดถอย (Regression Analysis) และงานในการวิเคราะห์การจัดกลุ่ม (Classification Analysis) ถือเป็นงานที่พบได้ทั่วไปสำหรับการเรียนรู้ด้วยเครื่องแบบมีผู้สอน (Manghani, 2017)

การเรียนรู้ด้วยเครื่องแบบมีผู้สอนถูกจำกัดความเนื่องมาจากการใช้งานชุดข้อมูลซึ่งมีป้ายกำกับสำหรับการฝึกฝนอัลกอริทึมหรือแบบจำลองเพื่อใช้ในการทำนายผลลัพธ์ที่แม่นยำ โดยในขั้นตอนการเรียนรู้จะมีการปรับปรุงค่าน้ำหนัก (Weight) ต่างๆ ภายในตัวแบบจำลองจนกระทั่งได้ประสิทธิภาพที่เหมาะสม ซึ่งการเรียนรู้เหล่านี้เป็นส่วนหนึ่งของกระบวนการที่เรียกว่า ‘Cross Validation’ เพื่อให้มั่นใจได้ว่าแบบจำลองจะไม่เกิดการ Overfitting (การที่แบบจำลองเรียนรู้กับชุดข้อมูลฝึกจนแม่นยำมากเกินไป ทำให้ไม่สามารถทำนายข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างแม่นยำ) หรือ Underfitting (การที่แบบจำลองไม่มีประสิทธิภาพเนื่องจากมีการเรียนรู้ที่น้อยจนเกินไป) (IBM, 2020a)

ในอัลกอริทึมของการเรียนรู้ด้วยเครื่องแบบมีผู้สอน คอมพิวเตอร์จะฝึกฝนเรียนรู้จากข้อมูลตัวอย่าง แบบจำลองของการเรียนรู้ด้วยเครื่องแบบมีผู้สอนประกอบไปด้วยคู่ของข้อมูลขาเข้าและข้อมูลขาออก ซึ่งเป็นผลลัพธ์หรือป้ายกำกับของข้อมูลขาเข้า ตัวอย่างเช่น แบบจำลองสำหรับการแยกดอกเดซี (Daisies) และดอกแพนซี (Pansies) ซึ่งข้อมูลขาเข้าจะเป็นรูปของดอกไม้ ส่วนผลลัพธ์ที่ต้องการคือการทำนายว่ารูปนั้นคือดอกไม้ชนิดใด โดยแบบจำลองจะถูกฝึกฝนแบบวนซ้ำกับชุดข้อมูลเพื่อค้นหาสิ่งที่สามารถระบุความใกล้เคียงและความแตกต่างกันของดอกไม้ทั้งสองประเภทจนกระทั่งสามารถแยกประเภทได้อย่างถูกต้องมีประสิทธิภาพ (SAP) ซึ่งหลักการนี้จะคล้ายคลึงกับการที่เด็กเรียนรู้ในการระบุชนิดของผลไม้จากการจำมาจากหนังสือภาพ (Oracle)

ขั้นตอนการทำงานของการเรียนรู้ด้วยเครื่องแบบมีผู้สอนสามารถแสดงเป็นภาพอย่างง่ายได้ดังภาพประกอบที่ 2



ภาพประกอบ 2 แสดงขั้นตอนการทำงานของการเรียนรู้ด้วยเครื่องแบบมีผู้สอน

ที่มา: (Javatpoint)

จากภาพประกอบที่ 2 ชุดข้อมูลมีวัตถุที่แตกต่างกันอยู่ 3 ชนิด ได้แก่ วัตถุหกเหลี่ยม (Hexagon) วัตถุสามเหลี่ยม (Triangle) และวัตถุสี่เหลี่ยม (Square) ในขั้นตอนแรกต้องมีการเรียนรู้ของแบบจำลองจากชุดข้อมูลและป้ายกำกับ โดยวัตถุที่มีหกด้านจะถูกติดป้ายกำกับว่าเป็นวัตถุหกเหลี่ยม วัตถุที่มีสามด้านจะถูกติดป้ายกำกับว่าเป็นวัตถุสามเหลี่ยม และวัตถุที่มีสี่ด้านจะถูกติดป้ายกำกับว่าเป็นวัตถุสี่เหลี่ยม

หลังจากเสร็จสิ้นขั้นตอนการเรียนรู้เราจะทำการทดสอบประสิทธิภาพของแบบจำลองในการระบุชนิดของวัตถุที่ไม่เคยเห็นมาก่อนในขั้นตอนการเรียนรู้ซึ่งประกอบด้วยวัตถุสี่เหลี่ยมและวัตถุสามเหลี่ยม โดยแบบจำลองจะค้นรูปแบบเพื่อเปรียบเทียบกับข้อมูลที่ได้เรียนรู้มาเพื่อค้นหาว่าข้อมูลใหม่ที่ได้รับมาเป็นวัตถุประเภทใด จากนั้นจะทำการทำนายผลลัพธ์ชนิดของวัตถุที่แบบจำลองคาดการณ์ว่ามีความใกล้เคียงมากที่สุดออกมา

การเรียนรู้ด้วยเครื่องแบบมีผู้สอนสามารถแบ่งประเภทย่อยลงไปได้อีกสองประเภทหลักๆ ตามปัญหาของงาน ได้แก่งานเกี่ยวกับการวิเคราะห์การถดถอย (Regression) และงานเกี่ยวกับการจำแนกประเภท (Classification)

งานแบบ Regression จะถูกใช้สำหรับหาความสัมพันธ์ระหว่างข้อมูลขาเข้าและข้อมูลขาออก มักใช้ในการคาดการณ์หรือทำนายข้อมูลที่เป็นเลขจำนวนจริงซึ่งมีความต่อเนื่อง เช่น การคาดการณ์อุณหภูมิล่วงหน้า การทำนายแนวโน้มราคาสินค้าในตลาด

งานแบบ Classification จะถูกใช้สำหรับงานที่ผลลัพธ์ของแบบจำลองเป็นแบบชนิดหรือประเภท เช่น ใช่หรือไม่ใช่ เพศชายหรือเพศหญิง ถูกหรือผิด พันธุ์ของสุนัข (Javatpoint)

2.1.2 การเรียนรู้ด้วยเครื่องแบบไม่มีมีผู้สอน (Unsupervised Learning)

การเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะเป็นการใช้งานอัลกอริทึมสำหรับการวิเคราะห์แยกแยะ และจัดกลุ่ม (Clustering) ของข้อมูลซึ่งไม่มีป้ายกำกับ โดยการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะค้นหารูปแบบที่ซ่อนอยู่ในชุดข้อมูลเพื่อแบ่งข้อมูลออกเป็นกลุ่มๆ โดยไม่จำเป็นต้องพึ่งพาการแทรกแซงของมนุษย์ในการเรียนรู้ ความสามารถของการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนคือการค้นหาหรือตรวจจับความเหมือนและความแตกต่างของแต่ละข้อมูล ซึ่งมีประโยชน์อย่างสูงต่องานการสำรวจและวิเคราะห์ข้อมูล การนำเสนอผลิตภัณฑ์เสริมที่เกี่ยวข้องกับผลิตภัณฑ์หลัก (Cross-Selling Strategies) เช่น การนำเสนอผลิตภัณฑ์เกี่ยวกับการประกันภัยเนื่องจากอุบัติเหตุเมื่อลูกค้ามีการดำเนินการกู้ยืมเงินสำหรับการซื้อรถยนต์เป็นต้น งานในการจัดกลุ่มของลูกค้า (Customer Segmentation) เช่น กลุ่มลูกค้าสัญจร กลุ่มลูกค้าที่มียอดใช้จ่ายสูงเป็นต้น และงานทางด้านรูปภาพและการจำแนกรูปแบบ (Image and Pattern Recognition)

นอกจากนี้การเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนยังสามารถใช้ประโยชน์ในการลดจำนวนคุณลักษณะหรือตัวแปรต้นของข้อมูลในงานเกี่ยวกับการลดมิติของข้อมูล (Dimensionality Reduction) ซึ่งวิธีการที่เป็นที่นิยมได้แก่ Principal Component Analysis (PCA) และ Singular Value Decomposition (SVD) นอกจากนี้ยังมีอัลกอริทึมอื่นๆ ของการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนอีกมากมาย เช่น Neural Networks, K-Mean Clustering เป็นต้น (IBM, 2020a)

ในงานการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะไม่มีคำตอบหรือผลลัพธ์ที่ถูกต้องและชัดเจนแบบงานการเรียนรู้ด้วยเครื่องแบบมีผู้สอน คอมพิวเตอร์หรือเครื่องจะเรียนรู้จากข้อมูลขาเข้าซึ่งไม่มีป้ายกำกับและโครงสร้างที่ชัดเจนเพื่อจะค้นหาและระบุรูปแบบและความสอดคล้องของข้อมูล โดยในหลายๆ ครั้งการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะมีขั้นตอนในการสร้างแบบจำลองในลักษณะที่คล้ายคลึงกับการสังเกตมนุษย์ ซึ่งมนุษย์จะจะมีการใช้งานประสบการณ์ความรู้และสามัญสำนึกในการจัดกลุ่มของสิ่งต่างๆ เมื่อยิ่งมีประสบการณ์ที่สูงมากขึ้นในสาขาหรือหัวข้อนั้นๆ ก็จะยิ่งเพิ่มความสามารถในการแยกประเภทและจัดกลุ่มของสิ่งต่างๆ ได้มีประสิทธิภาพสูงขึ้น โดยในด้านของคอมพิวเตอร์ ประสบการณ์จะหมายถึงจำนวนของข้อมูลขาเข้าที่ถูกนำมาใช้ในการเรียนรู้ ตัวอย่างของงานแบบการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนที่เป็นที่นิยม เช่น งานเกี่ยวกับการจดจำใบหน้า งานเกี่ยวกับการวิเคราะห์ลำดับพันธุกรรม งานการวิเคราะห์ตลาดซื้อขาย และงานทางด้านความปลอดภัยไซเบอร์ เป็นต้น (SAP)

การเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนเป็นการเรียนรู้ในรูปแบบที่คอมพิวเตอร์สามารถเรียนรู้ด้วยตนเองเพื่อที่จะค้นหารูปแบบผ่านขั้นตอนและกระบวนการที่ซับซ้อนโดยอาศัยมนุษย์เข้า

มาเกี่ยวข้องเพียงเล็กน้อย โดยการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะเป็นการเรียนรู้จากชุดข้อมูลซึ่งไม่มีป้ายกำกับหรือผลลัพธ์ที่เฉพาะเจาะจงที่ถูกระบุไว้ล่วงหน้า

การเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะมีความคล้ายคลึงกับการเรียนรู้ของเด็กเล็กในการแยกกลุ่มประเภทของผลไม้จากกระบวนการสังเกตเฉดสีและรูปร่างแทนการจดจำชื่อชนิดของผลไม้ โดยเด็กจะสังเกตความคล้ายคลึงกันจากรูปภาพและทำการจัดกลุ่มผลไม้ตามความคล้ายคลึง จากนั้นจึงทำการระบุป้ายกำกับให้กับแต่ละกลุ่ม (Oracle)

ในการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนจะไม่มีการระบุผลลัพธ์ที่ถูกต้องหรือชัดเจน แต่จะทำงานโดยอัลกอริทึมซึ่งต้องทำการเรียนรู้ วิเคราะห์ข้อมูล และสังเกตรูปแบบหรือโครงสร้างในข้อมูลนั้นๆ เอง การเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอนมักจะได้ประสิทธิภาพที่ดีกับข้อมูลประเภทรายธุรกรรม (Transactional Data) ซึ่งอาจใช้สำหรับการจัดกลุ่มของพฤติกรรมของผู้ซื้อที่คล้ายคลึงกันสำหรับใช้กำหนดเพื่อเป็นกลุ่มเป้าหมายสำหรับสิทธิพิเศษทางการตลาด ตัวอย่างอัลกอริทึมของการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน เช่น การจำแนกกลุ่มแบบ K-Means (K-Means Clustering) , Principal and Independent Component Analysis (PCA) และ Association Rules เป็นต้น (Manghani, 2017)

2.1.3 **การเรียนรู้ด้วยเครื่องแบบเสริมกำลัง** (Reinforcement Learning)

การเรียนรู้ด้วยเครื่องแบบเสริมกำลังคือหนึ่งในรูปแบบของการเรียนรู้ด้วยเครื่อง เมื่อเปรียบเทียบกับการเรียนรู้ด้วยเครื่องแบบมีผู้สอนซึ่งคอมพิวเตอร์จะเรียนรู้การหาความสัมพันธ์จากคุณลักษณะของข้อมูลกับผลลัพธ์ที่ถูกต้องจากชุดข้อมูลที่มีป้ายกำกับที่ชัดเจน แต่การเรียนรู้ด้วยเครื่องแบบเสริมกำลังจะมีความแตกต่างกันในส่วนของการเรียนรู้ซึ่งข้อมูลจะไม่มีผลลัพธ์หรือป้ายกำกับที่ถูกต้องแน่นอน แต่จะเรียนรู้จากข้อมูลขาเข้าซึ่งเป็นกลุ่มของข้อมูลเกี่ยวกับทางเลือกหรือการกระทำที่เป็นไปได้ กฎเกณฑ์ต่างๆ และเป้าหมายหรือจุดสิ้นสุดที่เป็นไปได้ โดยเมื่อเป้าหมายหรือจุดสิ้นสุดที่อัลกอริทึมต้องการมีค่าเป็นแบบคงที่หรือแบบไบนารี (Binary) จะทำให้คอมพิวเตอร์หรือเครื่องสามารถเรียนรู้จากข้อมูลตัวอย่างได้ แต่ในกรณีที่เป้าหมายหรือจุดสิ้นสุดที่ต้องการมีความไม่แน่นอน สามารถเปลี่ยนแปลงได้ ระบบจะต้องมีการเรียนรู้จากประสบการณ์และการได้รับรางวัลตอบแทน ในการเรียนรู้ด้วยเครื่องแบบเสริมกำลัง ‘รางวัล’ จะหมายถึงคะแนนที่คอมพิวเตอร์จะต้องพยายามเรียนรู้เพื่อค้นหาและเก็บรวบรวมไว้ให้ได้มากที่สุด

หากยกตัวอย่างเพื่อนำมาเปรียบเทียบให้เข้าใจง่ายขึ้น การเรียนรู้ด้วยเครื่องแบบเสริมกำลังจะมีความคล้ายคลึงกับการสอนมนุษย์ให้เล่นหมากรุก ซึ่งแน่นอนว่าเป็นไปไม่ได้ที่จะสามารถสอนให้คอมพิวเตอร์เรียนรู้ทุกกลยุทธ์การเดินหมากที่มีประสิทธิภาพได้ ดังนั้นจะมีการ

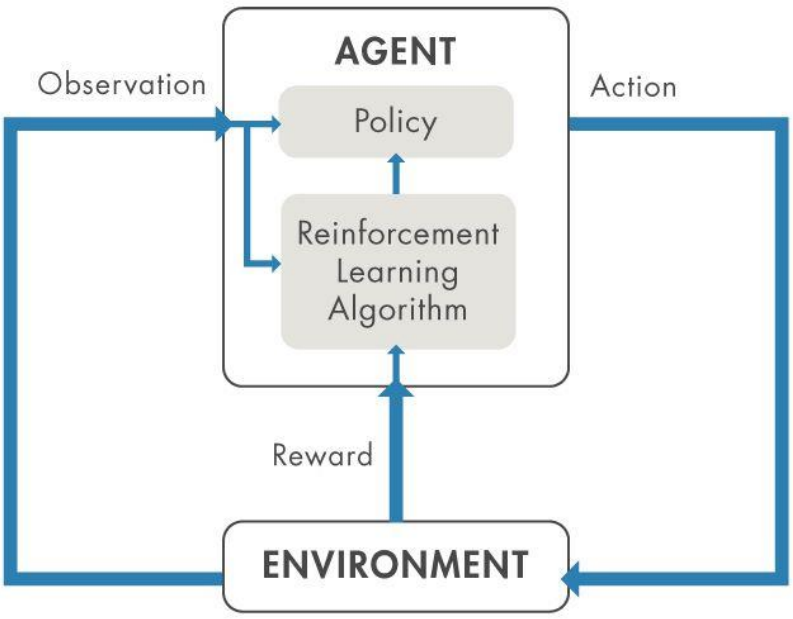
สอนคอมพิวเตอร์เฉพาะในส่วนของกฎการเล่นหมากรุกเพียงเท่านั้นและให้คอมพิวเตอร์เรียนรู้ด้วยตัวเองผ่านการฝึกฝนเพื่อพัฒนาทักษะ ในส่วนของรางวัลนั้นนอกจากการชนะการแข่งขันแล้ว ยังสามารถได้รับคะแนนจากการกำจัดหมากของคู่แข่งได้อีกด้วย ตัวอย่างของการประยุกต์ใช้การเรียนรู้ด้วยเครื่องแบบเสริมกำลังนั้น เช่น การประมูลราคาแบบอัตโนมัติสำหรับผู้ซื้อโฆษณาออนไลน์ การพัฒนาเกมส์คอมพิวเตอร์ และการเทรดหุ้นที่มีความเสี่ยงสูง (SAP)

การเรียนรู้ด้วยเครื่องแบบเสริมกำลังคือการเรียนรู้ของเครื่องซึ่งมีวิธีการที่คล้ายคลึงกับการเรียนรู้ด้วยเครื่องแบบไม่มีผู้สอน แต่มีความแตกต่างกันตรงที่อัลกอริทึมของการเรียนรู้ด้วยเครื่องแบบเสริมกำลังไม่ได้เรียนรู้จากชุดข้อมูลตัวอย่าง แต่จะมีการเรียนรู้จากการลองผิดลองถูกแทน ซึ่งผลลัพธ์ในแต่ละรอบที่ประสบความสำเร็จจะถูกรวบรวมเพื่อนำมาปรับปรุงแบบจำลองเพื่อให้เกิดคำแนะนำหรือหลักการที่ดีที่สุดสำหรับการแก้ไขปัญหานั้น (IBM, 2020a)

การเรียนรู้ด้วยเครื่องแบบเสริมกำลังจะใช้วิธีการเรียนรู้จากการลองผิดลองถูกโดยอัลกอริทึมจะค้นหากลยุทธ์หรือวิธีการที่ทำให้ได้มาได้มาซึ่งรางวัลที่สูงที่สุด โดยการเรียนรู้ด้วยเครื่องแบบเสริมกำลังจะประกอบด้วยองค์ประกอบพื้นฐานหลักๆ อยู่ 3 องค์ประกอบ ได้แก่

1. ผู้กระทำ (Agent) ซึ่งเป็นผู้ตัดสินใจในการเลือกกลยุทธ์
2. สิ่งแวดล้อมหรือทรัพยากรโดยรอบซึ่งผู้กระทำสามารถมีปฏิสัมพันธ์ได้
3. การกระทำที่ผู้กระทำตัดสินใจเลือก

วัตถุประสงค์ของการเรียนรู้ด้วยเครื่องแบบเสริมกำลังคือการที่ตัวแทนตัดสินใจเลือกกระทำการบางสิ่งบางอย่างที่ส่งผลให้เกิดรางวัลที่สูงที่สุด หรือการตัดสินใจกระทำบางสิ่งบางอย่างที่ทำให้เกิดผลลัพธ์ที่เหมาะสมมากที่สุด การเรียนรู้ด้วยเครื่องแบบเสริมกำลังมักถูกนิยมนำมาปรับใช้ในงานเกี่ยวกับการพัฒนาเกมส์คอมพิวเตอร์หรืองานทางด้านวิศวกรรมหุ่นยนต์ นอกจากนี้ยังสามารถพบเห็นการใช้งานการเรียนรู้ด้วยเครื่องแบบเสริมกำลังได้ในงานที่เกี่ยวข้องกับการค้นหาค่าที่เหมาะสมที่สุด (Optimization Techniques)



ภาพประกอบ 3 แสดงวิธีการทำงานโดยคร่าวของการเรียนรู้ด้วยเครื่องแบบเสริมกำลัง

ที่มา: (Tzorakoleftherakis, 2019)

2.2 ทฤษฎีเกี่ยวกับแบบจำลองสำหรับงานในการจำแนกประเภท (Classification Model)

การจำแนกประเภทเป็นหนึ่งในงานที่สำคัญและแพร่หลายอย่างมากของการเรียนรู้ด้วยเครื่องแบบมีผู้สอน ซึ่งเป็นเทคนิคที่จะใช้ในการทำนายผลลัพธ์หรือประเภทของตัวอย่างข้อมูลใหม่ๆ ด้วยการใช้งานแบบจำลองที่ผ่านการเรียนรู้จากชุดข้อมูลสำหรับฝึกสอน การจำแนกประเภทของข้อมูลประกอบด้วยขั้นตอนจำนวน 2 ขั้นตอน โดยขั้นตอนแรกคือการเรียนรู้เพื่อสร้างแบบจำลองจากชุดข้อมูลสำหรับการเรียนรู้ และขั้นตอนที่สองคือการทดสอบใช้งานการจำแนกประเภทเพื่อประเมินประสิทธิภาพ โดยจะมีการใช้งานชุดข้อมูลสำหรับการทดสอบซึ่งแบบจำลองไม่เคยเห็นข้อมูลเหล่านี้มาก่อน โดยแบบจำลองที่เกี่ยวกับการจำแนกประเภทประกอบด้วยอัลกอริทึมจำนวนมากซึ่งจะมีความเหมาะสมและให้ประสิทธิภาพที่แตกต่างกันตามการใช้งานกับชุดข้อมูลหรือวัตถุประสงค์ที่แตกต่างกันไป (Sarker, Kayes, & Watters, 2019)

ผลลัพธ์ของงานในการจำแนกประเภทสามารถแบ่งย่อยได้เป็นสองประเภทหลักๆ ตามจุดประสงค์ของการใช้งาน การจำแนกประเภทแบบแรกคือแบบผลลัพธ์ไบนารี ซึ่งผลลัพธ์ของข้อมูลตัวอย่างจะประกอบด้วยป้ายกำกับสองประเภทหลัก เช่น 0 หรือ 1, เพศชายหรือเพศหญิง, ซื้อหรือไม่ซื้อ เป็นต้น การจำแนกประเภทแบบที่สองคือแบบผลลัพธ์ความน่าจะเป็น ซึ่งป้ายกำกับ

ของข้อมูลตัวอย่างจะมีมากกว่าสองประเภทขึ้นไป โดยผลลัพธ์จากแบบจำลองจะได้ออกมาเป็นค่าความน่าจะเป็นของแต่ละป้ายกำกับ ซึ่งข้อมูลตัวอย่างจะถูกจำแนกประเภทไปยังป้ายกำกับที่มีค่าความน่าจะเป็นสูงมากที่สุด (Dreiseitl & Ohno-Machado, 2002)

ดังนั้นในการเลือกใช้งานแบบจำลองและอัลกอริทึมจึงต้องมีการเลือกใช้ให้เหมาะสมกับแต่ละบริบท โดยในงานวิจัยนี้มีการใช้งานอัลกอริทึมในการจำแนกประเภทข้อมูล ดังนี้

2.2.1 **ทฤษฎีอัลกอริทึม** Logistic Regression

การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์ข้อมูลที่มีหลายคุณลักษณะซึ่งมีวัตถุประสงค์เพื่อประมาณค่าหรือทำนายผลลัพธ์บางสิ่งบางอย่างภายใต้ปัจจัยของคุณลักษณะตัวอย่างข้อมูล แบบจำลองโลจิสติกจะประกอบด้วยตัวแปรตามหรือผลลัพธ์ และตัวแปรต้นหรือคุณลักษณะของตัวอย่างข้อมูล

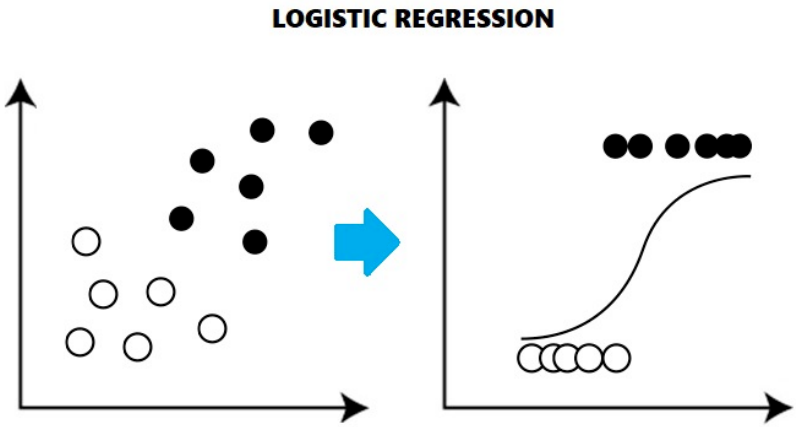
Logistic Regression สามารถแบ่งออกเป็น 3 ประเภทย่อยได้ตามลักษณะของผลลัพธ์ดังนี้ (Wikipedia, 2013)

1. Binary Logistic Regression คือการประมวลผลกับข้อมูลซึ่งมีความน่าจะเป็นของผลลัพธ์เพียงสองกลุ่ม เช่น 0 หรือ 1 เพศชายหรือเพศหญิง ซื้อสินค้าหรือไม่ซื้อเป็นต้น

2. Multinomial Logistic Regression คือการประมวลผลกับข้อมูลซึ่งมีความน่าจะเป็นของข้อมูลมากกว่าสองกลุ่มขึ้นไปซึ่งไม่มีความสัมพันธ์กันทางลำดับข้อมูล เช่น พันธุ์ของสุนัข รุ่นของโทรศัพท์มือถือ เป็นต้น

3. Ordinal Logistic Regression คือการประมวลผลกับข้อมูลซึ่งมีความน่าจะเป็นของข้อมูลมากกว่าสองกลุ่มขึ้นไปซึ่งมีความสัมพันธ์กันทางลำดับข้อมูล เช่น ระดับการศึกษา (ประถม, มัธยม, ปริญญา) หรือ ความพึงพอใจ (แย่, ปานกลาง, ดี, ดีมาก) เป็นต้น

Logistic Regression เป็นแบบจำลองที่ใช้สำหรับงานในการทำนายซึ่งผลลัพธ์ของการทำนายเป็นค่าแบบไม่ต่อเนื่องกัน เช่น การทำนายว่าป่วยหรือไม่ป่วย



ภาพประกอบ 4 แสดงการเปรียบเทียบก่อนและหลังการทำงานของ Logistic Regression

ที่มา: (Equiskill, 2018)

Logistic Regression เป็นแบบจำลองที่ถูกใช้งานอย่างแพร่หลายสำหรับงานจำแนกประเภทแบบไบนารี แบบจำลองจะทำนายความน่าจะเป็นของแต่ละป้ายกำกับจากการใช้งาน Logit Function จากนั้นจะมีการใช้งาน Sigmoid Activation Function เพื่อแปลงจากค่าความน่าจะเป็นให้กลายเป็นผลลัพธ์แบบป้ายกำกับ

Sigmoid Function (สมการที่ 1) เป็นฟังก์ชันทางคณิตศาสตร์ที่มีพฤติกรรมซึ่งสามารถแปลงเลขจำนวนเต็มใดๆ ให้กลายเป็นค่าระหว่างช่วง 0 ถึง 1 ภายใต้กราฟที่มีลักษณะโค้งเป็นรูปตัว ‘S’ ซึ่ง Sigmoid สามารถเรียกได้อีกชื่อหนึ่งว่า Logistic Function

$$Y = \frac{1}{1+ e^{-z}} \qquad (1)$$

การทำงานของ Sigmoid Function คือเมื่อค่าที่นำมาเข้าฟังก์ชันมีค่าเป็นบวกแบบไร้ขอบเขตจะได้ผลลัพธ์ของฟังก์ชันออกมาเป็น 1 ในทางตรงกันข้ามหากค่าที่นำมาคำนวณเป็นค่าติดลบแบบไร้ขอบเขตจะได้ผลลัพธ์ของฟังก์ชันออกมาเป็น 0 และค่าอื่นๆ ที่อยู่ระหว่างนี้จะได้ผลลัพธ์จากฟังก์ชันออกมาระหว่าง 0 ถึง 1

หากผลลัพธ์ที่ได้จากฟังก์ชันมีค่ามากกว่า 0.5 จะทำให้แบบจำลองทำนายผลลัพธ์ของป้ายกำกับออกมาเป็นกลุ่ม 1 หรือกลุ่มบวก (Positive) และหากผลลัพธ์ที่ได้จากฟังก์ชันมีค่า

น้อยกว่า 0.5 จะทำให้แบบจำลองทำนายผลลัพธ์ของป้ายกำกับออกมาเป็นกลุ่ม 0 หรือกลุ่มลบ (Negative) (Kumawat, 2019)

Logistic Regression เป็นแบบจำลองที่มีความซับซ้อนน้อย โดยเฉพาะอย่างยิ่งเมื่อคุณลักษณะของข้อมูลตัวอย่างมีความเป็นอิสระต่อกันสูง และมีการปรับปรุงข้อมูลให้อยู่ในรูปแบบมาตรฐานที่พร้อมใช้งาน (Variable Transformation) ซึ่งจะทำให้ลดโอกาสของการเกิดปัญหา Overfitting (ปัญหาซึ่งแบบจำลองสามารถทำงานได้ดีกับชุดข้อมูลสำหรับการเรียนรู้ แต่มีประสิทธิภาพที่ลดลงอย่างมากเมื่อนำมาใช้งานกับชุดข้อมูลสำหรับการทดสอบหรือเมื่อนำมาใช้งานจริง)

นอกจากนี้กระบวนการคัดเลือกคุณลักษณะยังสามารถช่วยลดความซับซ้อนของแบบจำลองและลดโอกาสของการเกิด Overfitting ลงไปได้อีกด้วย (Dreiseitl & Ohno-Machado, 2002)

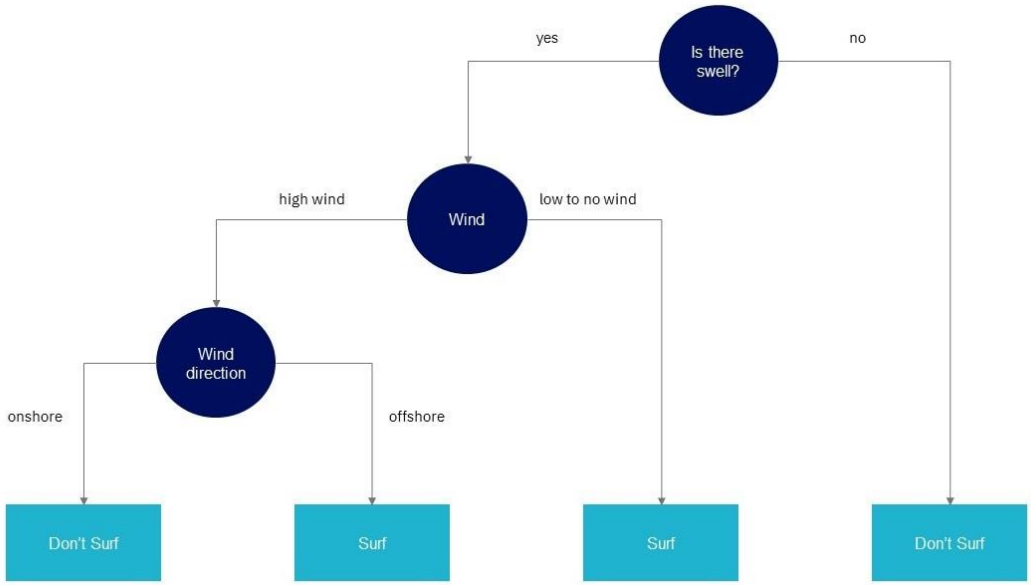
2.2.2 ทฤษฎีอัลกอริทึม Random Forest

Random Forest หรือป่าต้นไม้ตัดสินใจ เป็นอัลกอริทึมที่ถูกใช้งานอย่างแพร่หลายสำหรับการเรียนรู้ด้วยเครื่องซึ่งถูกคิดค้นโดย Leo Breiman และ Adele Cutler ซึ่งการทำงานของ Random Forest จะเป็นการรวบรวมผลลัพธ์จาก Decision Tree หรือต้นไม้ตัดสินใจหลายๆ ต้น เพื่อนำมาคำนวณสำหรับการได้มาซึ่งผลลัพธ์สุดท้าย โดย Random Forest ค่อนข้างมีความยืดหยุ่นและสะดวกสบายในการเรียกใช้งาน เนื่องจากสามารถใช้ได้กับงานทั้งแบบการจำแนกประเภท (Classification) และแบบการวิเคราะห์เชิงถดถอย (Regression)

Decision Tree ถือเป็นพื้นฐานเริ่มต้นสำหรับ Random Forest โดย Decision Tree จะมีแนวคิดคือการสร้างชุดลำดับของคำถามขึ้นมาเพื่อใช้แบ่งข้อมูลออกเป็นสองส่วนไปเรื่อยๆ ซึ่งคำถามต่างๆ ในชุดลำดับจะส่งผลให้เกิดจุดตัดสินใจ (Decision Node) ขึ้นมาบนต้นไม้ โดยตัวอย่างข้อมูลที่เป็นไปตามเงื่อนไขจะเคลื่อนต่อไปยังกิ่งของต้นไม้แบบ ‘ใช่’ (Yes) และข้อมูลที่ไม่เป็นไปตามเงื่อนไขก็จะเคลื่อนไปยังกิ่งของต้นไม้ฝั่ง ‘ไม่ใช่’ (No) โดยการผ่านจุดตัดสินใจไปเรื่อยๆ จะสามารถนำพาตัวอย่างข้อมูลไปสู่คำตอบของคำถามบางอย่างได้

ตัวอย่างของคำถามง่ายๆ สำหรับต้นไม้ตัดสินใจ เช่น ‘ควรออกไปเล่นกระดานโต้คลื่นหรือไม่’ ซึ่งคำตอบสามารถเป็นไปได้ 2 ป้ายกำกับ คือ ‘เล่น’ และ ‘ไม่เล่น’ จากคำถามนี้จะสามารถสร้างชุดลำดับของคำถามขึ้นมาเพื่อใช้หาคำตอบได้ เช่น ‘ในทะเลมีคลื่นน้ำหรือไม่’ ,’มีลมพัดแรงหรือไม่’ ,’ทิศของลมพัดเข้าสู่หรือออกจากฝั่ง’ ซึ่งคำถามต่างๆ ในชุดลำดับจะเป็นตัวนำพาตัวอย่าง

ของข้อมูลเพื่อไปยังผลลัพธ์สุดท้ายของการตัดสินใจ หรือที่เรียกว่า Leaf Node ได้ ดังแสดงในภาพประกอบที่ 9 (IBM, 2020b)



ภาพประกอบ 5 แสดงต้นไม้ตัดสินใจของการทำนายการออกไปเล่นกระดานโต้คลื่น

ที่มา: (IBM, 2020b)

ปัญหาหลักของ Decision Tree คือสามารถเกิดการ Overfitting กับข้อมูลได้ง่าย ซึ่งมีวิธีแก้ไขด้วยการใช้งาน Decision Tree หลายๆ ต้นเพื่อสร้างเป็น Random Forest สำหรับการช่วยกันตัดสินใจ โดยต้นไม้แต่ละต้นใน Random Forest จะทำงานอย่างเป็นอิสระจากกัน ซึ่งจะช่วยเพิ่มประสิทธิภาพความแม่นยำของการทำนายผลลัพธ์ได้

Random Forest เป็นอัลกอริทึมที่มีการใช้งานของวิธีการ Bagging (หรือ Bootstrap Aggregation) ร่วมกับวิธีการ Feature Randomness สำหรับการสร้างต้นไม้ตัดสินใจแต่ละต้นซึ่งเป็นอิสระจากกันภายใต้ป่าต้นไม้ตัดสินใจ โดย Bagging จะเป็นกระบวนการในการสุ่มเลือกข้อมูลจากชุดข้อมูลสำหรับการเรียนรู้เพื่อนำมาใช้สร้างต้นไม้ตัดสินใจแต่ละต้นด้วยวิธีการสุ่มแบบใส่คืน ซึ่งทำให้แต่ละตัวอย่างของข้อมูลสามารถถูกสุ่มเลือกขึ้นมาซ้ำได้ ส่วน Feature Randomness จะเป็นกระบวนการในการสุ่มเลือกเซตของคุณลักษณะเพื่อนำมาใช้สร้างต้นไม้ตัดสินใจแต่ละต้น

การนำวิธีการ Bagging มาใช้ร่วมกับวิธีการ Feature Randomness จะช่วยให้ต้นไม้ตัดสินใจแต่ละต้นเป็นอิสระต่อกันมากขึ้น มีความเกี่ยวข้องกันน้อยลง ซึ่งถือเป็นข้อแตกต่างสำคัญ

ระหว่าง Random Forest และ Decision Tree โดย Decision Tree จะมีการใช้งานคุณลักษณะทั้งหมดที่เป็นไปได้และชุดข้อมูลสำหรับการเรียนรู้ทั้งหมดในการสร้างต้นไม้

หลักการทำงานของ Random Forest คือต้นไม้ตัดสินใจแต่ละต้นจะทำงานอย่างเป็นอิสระจากกัน โดยแต่ละต้นไม้ตัดสินใจจะถูกสร้างขึ้นมาจากชุดข้อมูลย่อยและเซตย่อยของคุณลักษณะที่แตกต่างกัน ซึ่งแม้จะเป็นตัวอย่างข้อมูลเดียวกันแต่ต้นไม้ตัดสินใจแต่ละต้นก็อาจจะให้ผลลัพธ์ออกมาที่แตกต่างกันได้ ซึ่งผลลัพธ์สุดท้ายจากการจำแนกประเภทของ Random Forest จะเกิดจากเสียงส่วนใหญ่ของต้นไม้ในป่าต้นไม้ตัดสินใจ หรือที่เรียกว่า Majority Vote

Random Forest ประกอบด้วยข้อดีหลายประการ ซึ่งข้อดีหลักๆ ที่มีความน่าสนใจได้แก่ (IBM, 2020b)

- ลดโอกาสในการเกิด Overfitting : โดยเมื่อต้นไม้ตัดสินใจมีจำนวนมากถึงจุดหนึ่ง การจำแนกประเภทจะไม่เกิดการ Overfit เนื่องจากค่าเฉลี่ยของต้นไม้ตัดสินใจซึ่งอิสระจากกันจะลดโอกาสของการผันแปร (Variance) และความผิดพลาดโดยรวมได้

- ความยืดหยุ่น : เนื่องจาก Random Forest สามารถใช้ได้กับงานทั้งแบบ Classification และ Regression อย่างมีประสิทธิภาพ และเนื่องจาก Feature Randomness เป็นการสุ่มใช้งานเซตย่อยของคุณลักษณะ จึงมีประสิทธิภาพในการทำงานกับข้อมูลที่ค่าขาดหายไป (Missing Value)

- ประเมินความสำคัญของคุณลักษณะได้ง่าย : Random Forest สามารถประเมินความสำคัญหรืออิทธิพลของแต่ละคุณลักษณะที่ส่งผลต่อแบบจำลองได้ง่ายโดยการใช้งานวิธีการ เช่น Gini Importance หรือ Mean Decrease in Impurity (MDI) ซึ่งเป็นการประเมินจากการลดลงของความแม่นยำของแบบจำลองเมื่อมีการนำแต่คุณลักษณะออก

2.2.3 ทฤษฎีอัลกอริทึม XGBoost (eXtreme Gredient Boosting)

XGBoost หรือ Extreme Gradient Boosting เป็นชุดอัลกอริทึมแบบ Optimized Distributed Gradient Boosting ที่ออกแบบมาให้มีประสิทธิภาพ ความยืดหยุ่น และสะดวกสบายในการใช้งาน ซึ่งมีการประยุกต์มาจากการเรียนรู้ด้วยเครื่องภายใต้แนวคิดแบบ Gradient Boosting โดย XGBoost จะมีการใช้งาน Tree Boosting แบบขนาน (Parallel Tree Boosting) หรือที่รู้จักในชื่อ Gradient-Boosted Decision Trees (GBDT) ซึ่งสามารถใช้ในงานเกี่ยวกับการแก้ปัญหาทางวิทยาศาสตร์ข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพ (XGBoost, 2021)

XGBoost เป็น Open-Source Software (ซอฟต์แวร์ที่เปิดเผยหลักการหรือแหล่งที่มาของเทคโนโลยีของซอฟต์แวร์นั้นให้บุคคลภายนอกสามารถใช้งานได้ภายใต้เงื่อนไขบาง

ประการ) ซึ่งใช้งานอัลกอริทึมการเรียนรู้ด้วยเครื่องแบบ Optimized Distributed Gradient Boosting ภายใต้กรอบแนวคิดของ Gradient Boosting โดย XGBoost ถือเป็นอัลกอริทึมของการเรียนรู้ด้วยเครื่องชั้นนำสำหรับงานการจำแนกประเภท (Classification) งานการวิเคราะห์การถดถอย (Regression) และงานแบบจัดลำดับ (Ranking)

อัลกอริทึมแบบ XGBoost มีแนวคิดพื้นฐานหลักๆ มาจากการเรียนรู้ด้วยเครื่องแบบมีผู้สอน (Supervised Machine Learning) ต้นไม้ตัดสินใจ (Decision Tree) การเรียนรู้แบบเป็นกลุ่ม (Ensemble Learning) และ Gradient Boosting

การเรียนรู้ด้วยเครื่องแบบมีผู้สอนเป็นการใช้งานอัลกอริทึมเพื่อเรียนรู้ในการค้นหารูปแบบที่ซ่อนอยู่ในคุณลักษณะของข้อมูลและสร้างแบบจำลองขึ้นมาสำหรับใช้ในการจำแนกประเภทของข้อมูล

ต้นไม้ตัดสินใจเป็นแบบจำลองในการจำแนกประเภทของข้อมูลโดยการใช้ Rule-Based ในการคัดแยกข้อมูล

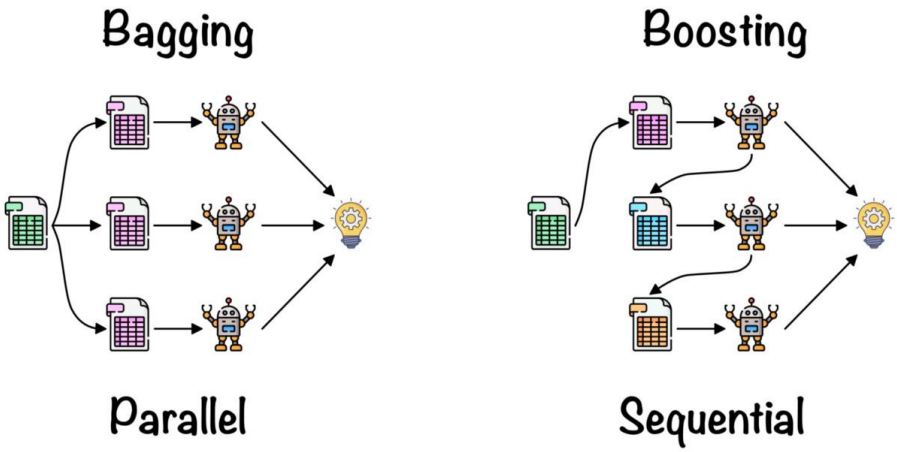
Gradient Boosting Decision Tree (GBDT) เป็นการเรียนรู้แบบเป็นกลุ่มของต้นไม้ตัดสินใจ (Decision Tree Ensemble Learning Algorithm) ซึ่งมีความคล้ายคลึงกับป่าต้นไม้ตัดสินใจ หรือ Random Forest โดย Ensemble Learning Algorithm จะมีการรวบรวมหลากหลายอัลกอริทึมของการเรียนรู้ด้วยเครื่องมาใช้งานร่วมกันเพื่อให้เกิดแบบจำลองที่มีประสิทธิภาพสูงที่สุด

ป่าต้นไม้ตัดสินใจและ GBDT จะมีความคล้ายคลึงกันในการสร้างแบบจำลองจากต้นไม้ตัดสินใจหลายๆ ต้น แต่ทั้งสองอัลกอริทึมมีความแตกต่างกันในขั้นตอนของกระบวนการสร้างต้นไม้ตัดสินใจและการรวมต้นไม้ตัดสินใจเข้าด้วยกัน

ป่าต้นไม้ตัดสินใจจะใช้งานวิธีการ Bagging เพื่อสร้างต้นไม้ตัดสินใจแต่ละต้นแบบขนานไปพร้อมๆ กัน และผลการจำแนกประเภทหรือการทำนายจะมาจากการหาค่าเฉลี่ยของการทำนายของทุกต้นไม้ตัดสินใจ

Gradient Boosting มีแนวคิดมาจากคำว่า ‘Boosting’ (การส่งเสริมหรือการเพิ่มขึ้น) หรือการปรับปรุงแบบจำลองโดยค่อยๆ พัฒนาแบบจำลองที่มีประสิทธิภาพต่ำหลายๆ แบบจำลองตามลำดับเพื่อสร้างแบบจำลองที่มีประสิทธิภาพในท้ายที่สุด โดย Gradient Boosting จะนำผลลัพธ์จากการทำนายของแบบจำลองไปใช้งานเป็นข้อมูลขาเข้าของแบบจำลองถัดไปโดยมีจุดประสงค์เพื่อพยายามลดค่าของความผิดพลาด

GBDT เป็นการเรียนรู้ของต้นไม้ตัดสินใจที่มีขนาดเล็กจำนวนมากแบบวนซ้ำหลายรอบ โดยในแต่ละรอบของการเรียนรู้จะใช้ค่าความผิดพลาดจากแบบจำลองก่อนหน้ามาใช้งานในการเรียนรู้ของแบบจำลองถัดไปตามลำดับ ผลลัพธ์สุดท้ายจากการทำนายจะถูกคิดค่าน้ำหนักจากผลลัพธ์ของทุกๆ ต้นไม้ตัดสินใจย่อยในแบบจำลอง โดยป่าต้นไม้ตัดสินใจที่ใช้งานวิธีการแบบ Bagging จะเป็นการลด Variance และ Overfitting ส่วน GBDT ที่ใช้งานวิธีการแบบ Boosting จะเป็นการลด Bias และ Underfitting



ภาพประกอบ 6 แสดงการเปรียบเทียบการทำงานระหว่างวิธีการแบบ Bagging และ Boosting

ที่มา: (López, 2021)

XGBoost เป็นการนำ Gradient Boosting มาปรับปรุงให้มีประสิทธิภาพและความแม่นยำที่สูงขึ้นซึ่งเป็นการพัฒนาขีดจำกัดในการประมวลผลสำหรับอัลกอริทึมแบบ Boosted Tree โดยมีจุดประสงค์เพื่อกระตุ้นประสิทธิภาพของการเรียนรู้ด้วยเครื่องและความเร็วในการประมวลผล ซึ่งการสร้างต้นไม้ตัดสินใจของ XGBoost จะเป็นการสร้างแบบขนานไปพร้อมๆ กันแทนที่การสร้างแบบเรียงตามลำดับของ GBDT

XGBoost ประกอบด้วยข้อดีหลายประการดังนี้

1. มีการใช้งานอย่างกว้างขวางรวมไปถึงการใช้งานในการแก้ปัญหาการวิเคราะห์การถดถอย การจำแนกประเภท การจัดลำดับ

2. การใช้งานชุดโปรแกรม (Library) มีความสะดวกสบายและยืดหยุ่น โดยสามารถทำงานได้บนหลายระบบปฏิบัติการ (Operating System : OS) ได้แก่ OS X, Windows และ Linux

3. รองรับการใช้งานร่วมกับการประมวลผลแบบกลุ่มเมฆ (Cloud Ecosystem) ไม่ว่าจะเป็น AWS, Azure, Yarn clusters และ Ecosystem อื่นๆ (NVIDIA, 2022)

2.2.4 อัลกอริทึม LightGBM (Light Gradient Boosting Machine)

LightGBM มีโครงสร้างแบบ Gradient Boosting ซึ่งใช้งานอัลกอริทึมการเรียนรู้แบบต้นไม้ โดยมีข้อดีและประสิทธิภาพ ดังนี้ (Microsoft, 2021b)

1. การเรียนรู้ของแบบจำลองเป็นไปอย่างรวดเร็วและมีประสิทธิภาพสูง

2. มีการใช้งานหน่วยความจำของคอมพิวเตอร์ในจำนวนน้อย

3. มีความแม่นยำสูง

4. รองรับการเรียนรู้แบบขนาน (Parallel) กระจาย (Distributed) และรวมไปถึงการเรียนรู้โดยใช้งานหน่วยประมวลผลของแผงวงจรสำหรับการแสดงผล (Graphic Processing Unit : GPU)

5. สามารถจัดการกับชุดข้อมูลที่มีขนาดใหญ่ได้

โดยทั่วไปแบบจำลอง Boosting ส่วนใหญ่มักมีการใช้งานอัลกอริทึมแบบ Pre-sort-Based เช่น อัลกอริทึมพื้นฐานของ XGBoost สำหรับการเรียนรู้ของต้นไม้ตัดสินใจ ซึ่งค่อนข้างยากในการปรับแต่งเพื่อให้ได้ประสิทธิภาพที่สูงที่สุด ส่วนในแบบจำลอง LightGBM มีการใช้งานอัลกอริทึมแบบ Histogram-Based ซึ่งมีคุณสมบัติต่อไปนี้

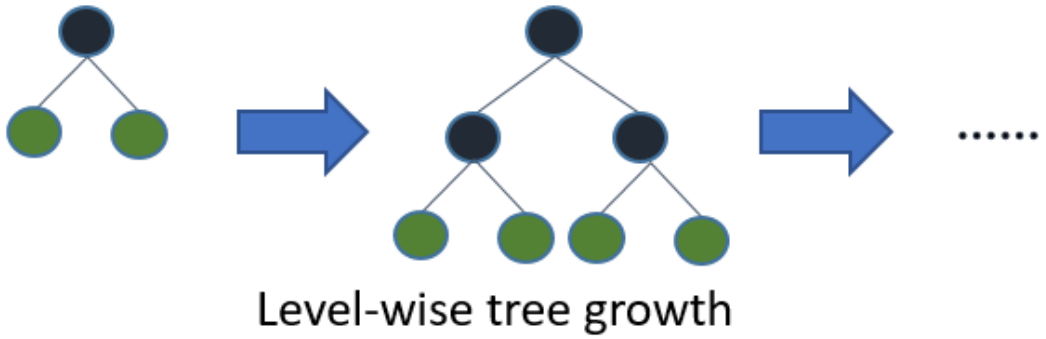
1. ลดค่าใช้จ่ายและทรัพยากรในการคำนวณ โดยอัลกอริทึมแบบ Histogram-Based มี Time Complexity ที่น้อยกว่าออักอริทึมแบบ Pre-Sort-Based

2. การใช้งาน Histogram Subtraction เพื่อเพิ่มความเร็วในการคำนวณ โดยทำการสร้าง Histrogram สำหรับเพียง 1 Leaf Node และทำการเรียกใช้งาน Histrogram ข้างเคียงด้วยวิธี Histogram Subtraction

3.กาลดการใช้งานหน่วยความจำ โดยการแทนค่าแบบต่อเนื่อง (Continuous Value) ด้วยช่วงข้อมูลแบบไม่ต่อเนื่อง (Discrete Bin) และไม่ต้องมีการจัดเก็บค่าข้อมูลคุณลักษณะสำหรับ Pre-Sorting

4. การลดค่าใช้จ่ายในการติดต่อสื่อสารสำหรับการเรียนรู้แบบกระจาย

อัลกอริทึมแบบต้นไม้ตัดสินใจส่วนใหญ่จะมีการขยายขนาดของต้นไม้แบบ Level (Depth)-Wise ซึ่งมีลักษณะตามภาพประกอบที่ 7

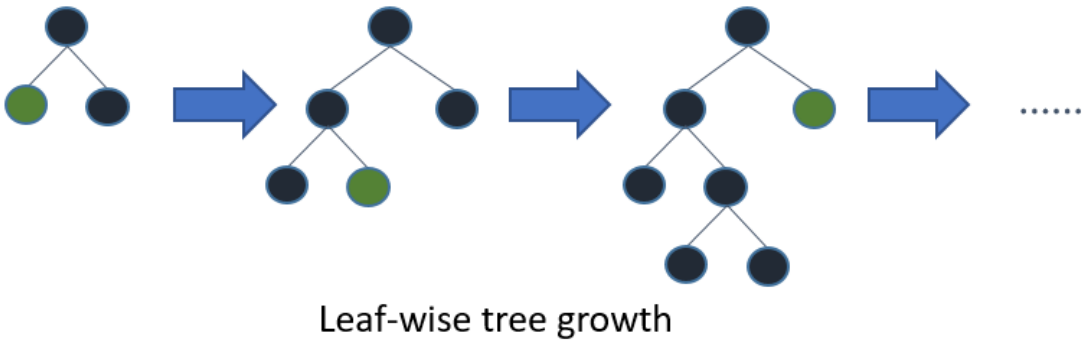


ภาพประกอบ 7 แสดงวิธีการขยายขนาดของต้นไม้ด้วยวิธีแบบ Level-Wise

ที่มา: (Microsoft, 2021a)

ส่วน LightGBM มีกระบวนการในการขยายขนาดของต้นไม้ตัดสินใจแบบ Leaf-Wise (Best-First) โดยจะมีการเลือก Leaf Node ที่จะขยายขนาดต่อไปด้วยวิธี Max Delta Loss ซึ่งอัลกอริทึมแบบ Leaf-Wise มีแนวโน้มที่จะได้รับความผิดพลาด (Loss) ที่น้อยกว่าอัลกอริทึมแบบ Level-Wise

กรณีที่ชุดข้อมูลมีขนาดเล็กหรือมีจำนวนของข้อมูลที่ไม่เพียงพอ อาจส่งผลให้อัลกอริทึมแบบ Leaf-Wise สามารถเกิดปัญหา Overfitting ได้ จึงเป็นสาเหตุที่ LightGBM มีการนำพารามิเตอร์หรือตัวแปร ‘max_depth’ มาใช้งานเพื่อกำหนดขอบเขตความลึกของชั้นต้นไม้



ภาพประกอบ 8 แสดงวิธีการขยายขนาดของต้นไม้ด้วยวิธีแบบ Leaf-Wise

ที่มา: (Microsoft, 2021a)

ความแตกต่างระหว่าง LightGBM และ XGBoost คือ LightGBM มีการประมวลผลที่รวดเร็วกว่า XGBoost ค่อนข้างมาก โดยยังคงความมีประสิทธิภาพที่ใกล้เคียงกับ XGBoost ไว้ได้ นอกจากนี้ยังมีความแตกต่างหลักอื่นๆ คือ วิธีการขยายขนาดของต้นไม้ (Leaf Growth), การจัดการกับข้อมูลคุณลักษณะชนิดประเภท (Categorical Feature Handling), การจัดการข้อมูลที่ขาดหายไป (Missing Value Handling) และวิธีการเกี่ยวกับความสำคัญของคุณลักษณะของข้อมูล (Feature Importance Method)

LightGBM มีการประมวลผลที่รวดเร็วโดยที่ยังสามารถรักษาประสิทธิภาพความแม่นยำไว้ได้เนื่องจากการขยายขนาดของต้นไม้ (Leaf Growth) มีการใช้งานวิธีการ 2 วิธีการ ได้แก่ Gradient-Based One-Side Sampling (GOSS) และ Exclusive Feature Bunding (EFB)

feature1	feature2	feature_bundle
0	2	6
0	1	5
0	2	6
1	0	1
2	0	2
3	0	3
4	0	4

ภาพประกอบ 9 แสดงวิธีการแบบ EFB ในการจัดชุด Feature1 และ Feature2 เข้าไว้ด้วยกันที่ ‘feature_bundle’ เพื่อลดจำนวนของคุณลักษณะ

ที่มา: (Sharma, 2018)

ทั้ง LightGBM และ XGBoost ยอมรับเฉพาะข้อมูลคุณลักษณะที่เป็นตัวเลขเท่านั้น จึงต้องมีการจัดการคุณลักษณะที่เป็นข้อมูลชนิดประเภท หรือ Nominal ให้กลายเป็นข้อมูลชนิดตัวเลข กรณีตัวอย่าง เช่น การทำนายป้ายกำกับที่มีประเภท ‘ร้อน’, ‘หนาว’, ‘ไม่ทราบ’ หรือ 2, 1, 0 ซึ่งเป็นป้ายกำกับแบบชนิดประเภทแบบไม่เรียงลำดับ โดยอัลกอริทึมต้องสร้างเงื่อนไขและรูปแบบที่มีความเท่าเทียมกันโดยหลีกเลี่ยงการเปรียบเทียบลำดับของตัวเลข

โดยปกติมาตรฐานของ XGBoost จะมีการดำเนินการแบบเรียงลำดับกับข้อมูลชนิดตัวเลขซึ่งไม่ใช่การทำงานที่เหมาะสมจึงต้องมีการจัดการข้อมูลด้วยวิธีการแบบ One-Hot Encoding เพื่อให้สามารถทำงานได้อย่างถูกต้องเหมาะสม แต่หากชุดข้อมูลมีขนาดที่ใหญ่การจัดการข้อมูลด้วยวิธีการ One-Hot Encoding จะมีการทำงานที่ช้าลง ในทางกลับกัน LightGBM

จะสามารถตรวจสอบชนิดของคุณลักษณะในชุดข้อมูลก่อนได้เพื่อดำเนินการจัดการกับคุณลักษณะชนิดประเภทอย่างเท่าเทียมโดยไม่ต้องจัดการด้วยวิธี One-Hot Encoding

ในส่วนของการจัดการข้อมูลที่ขาดหายไป ทั้ง LightGBM และ XGBoost จะมีการดำเนินการเติมข้อมูลที่ขาดหายที่ทำให้เกิดความผิดพลาด (Loss) ที่น้อยที่สุดในแต่ละการขยายขนาดของต้นไม้

การดำเนินการเกี่ยวกับความสำคัญของคุณลักษณะของ LightGBM ประกอบด้วย 2 วิธีการ ซึ่งแตกต่างจาก XGBoost ที่มีด้วยกัน 3 วิธีการ ได้แก่ (Saha, 2022)

1. Gain (LightGBM และ XGBoost) แต่ละคุณลักษณะภายในชุดข้อมูลจะมีความสำคัญ (Importance หรือ Weightage) ที่ช่วยในการสร้างแบบจำลองที่แม่นยำ มีประสิทธิภาพ ซึ่ง Gain จะอ้างอิงถึงอิทธิพลของคุณลักษณะที่มีความสำคัญในแต่ละต้นไม้ตัดสินใจที่แตกต่างกัน ซึ่งเมื่อนำมาประมวลผลร่วมกันจะทำให้แบบจำลองมีประสิทธิภาพที่ดีขึ้น

2. Split (LightGBM) หรือ Frequency/Weight (XGBoost) เป็นการคำนวณจำนวนครั้งในการที่คุณลักษณะของข้อมูลถูกนำไปใช้งานสำหรับเป็นเงื่อนไขในการแบ่งข้อมูลของต้นไม้ตัดสินใจ โดยคุณลักษณะชนิดประเภทซึ่งมีประเภทของข้อมูลจำนวนมากอาจส่งผลให้เกิดความโน้มเอียง (Bias) ได้

3. Coverage (XGBoost) เป็นการคำนวณจำนวนครั้งในการปรากฏของค่าในแต่ละคุณลักษณะ

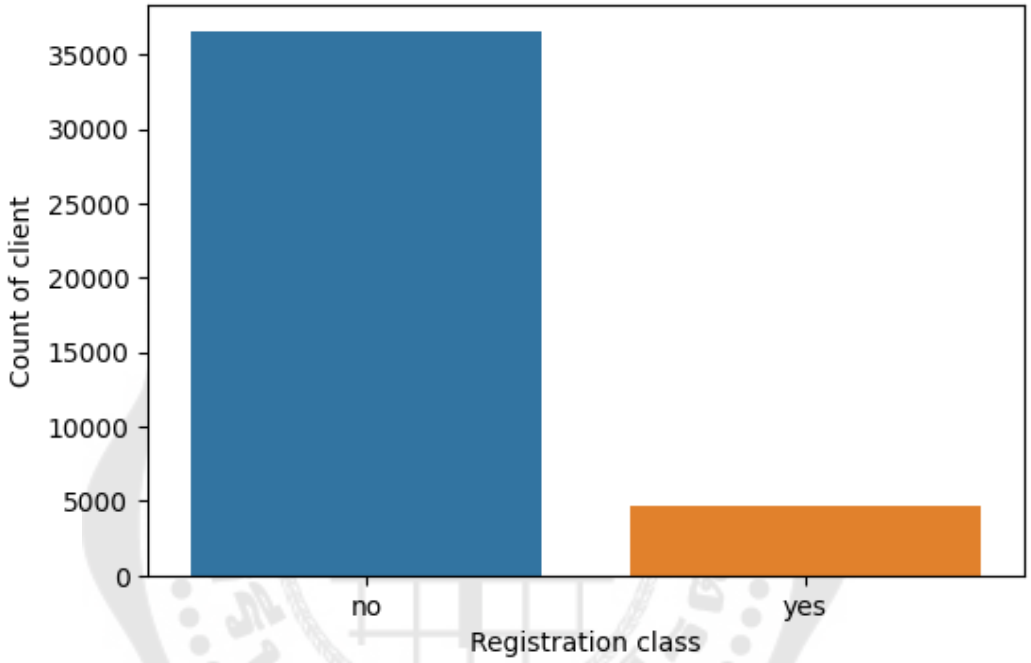
LightGBM เป็น Open-Source แบบอัลกอริทึม Gradient Boost Decision Tree (GBDT) ที่ถูกคิดค้นและพัฒนาขึ้นมาโดยบริษัท Microsoft ซึ่งมีการใช้อัลกอริทึม Histogram-Based เพื่อเพิ่มความเร็วของการเรียนรู้ของเครื่อง ลดปริมาณการใช้หน่วยความจำ และสามารถดำเนินการเรียนรู้แบบขนานที่เรียกว่า Parallel Voting DT ได้

LightGBM มีความแตกต่างจากแบบจำลอง GDBT อื่นๆ ในส่วนของการคำนวณการเพิ่มขึ้นของค่าความแปรปรวน (Variance) โดยการใช้งานวิธีการแบบ Leaf-Wise ในการขยายกิ่งของต้นไม้ตัดสินใจ โดยจะเลือกจากกิ่งที่มีการเพิ่มขึ้นของค่าความแปรปรวนที่สูงที่สุดมาใช้ในการขยาย (Machado, Karray, & Sousa, 2019)

2.3 ทฤษฎีเกี่ยวกับการจัดการข้อมูลที่ไม่สมดุล (Imbalance Data Handling)

ความไม่สมดุลกันของข้อมูลถือเป็นปัญหาที่สามารถพบได้อย่างกว้างขวางในชุดข้อมูลสำหรับการเรียนรู้ด้วยเครื่อง ซึ่งความรุนแรงของความไม่สมดุลกันของข้อมูลสามารถคิดได้จากสัดส่วนของจำนวนตัวอย่างข้อมูลในแต่ละกลุ่ม โดยในงานการจำแนกประเภทแบบไบนารี

(Binary) หรือสองกลุ่มข้อมูลสามารถคำนวณความรุนแรงของความไม่สมดุลได้จากจำนวนตัวอย่างข้อมูลในกลุ่มส่วนน้อยเทียบกับจำนวนตัวอย่างข้อมูลในกลุ่มส่วนใหญ่ ซึ่งแบบจำลองในการจำแนกประเภทของข้อมูลส่วนใหญ่มักจะประสบปัญหาความโน้มเอียง (Bias) ในการทำนายเนื่องจากขาดแคลนข้อมูลที่สามารถนำเสนอรูปแบบของกลุ่มข้อมูลส่วนน้อยได้ (Saripuddin, Suliman, Sameon, & Jorgensen, 2021)



ภาพประกอบ 10 แสดงความไม่สมดุลกันของข้อมูลซึ่งมีสัดส่วนของจำนวนกลุ่มข้อมูลประเภท no อยู่จำนวนมาก

ในงานซึ่งเกี่ยวข้องกับการจำแนกประเภทมักจะมีปัญหาในเรื่องความไม่สมดุลกันของข้อมูลในชุดข้อมูล โดยมักจะมีป้ายกำกับหรือกลุ่มประเภทข้อมูลหนึ่งที่มีจำนวนของข้อมูลมากกว่าข้อมูลป้ายกำกับประเภทอื่นๆ ในงานการจำแนกประเภทแบบสองกลุ่ม (Binary Classification) กลุ่มของข้อมูลที่เราให้ความสนใจ (Positive) มักจะเป็นกลุ่มข้อมูลซึ่งมีจำนวนน้อย

ความไม่สมดุลกันของข้อมูลมักจะส่งผลให้การเรียนรู้ด้วยเครื่องมีประสิทธิภาพที่ไม่ค่อยดีนักเมื่อทำงานกับตัวอย่างข้อมูลที่เป็นข้อมูลส่วนน้อย โดยมักจะมีความผิดพลาดในการจำแนกข้อมูลในกลุ่มส่วนน้อย ดังนั้นในงานเกี่ยวกับการเรียนรู้ด้วยเครื่องจึงต้องมีการจัดการกับความไม่สมดุลกันของข้อมูลเพื่อเพิ่มประสิทธิภาพในการจำแนกประเภทเมื่อข้อมูลกลุ่มหนึ่งมีจำนวนน้อยมากๆ

วิธีแก้ปัญหาที่พื้นฐานที่สุดคือการเพิ่มตัวอย่างของข้อมูลไม่ว่าจะด้วยวิธีการสุ่มหรือแบบมีหลักการ เพื่อให้ข้อมูลเกิดความสมดุลและปรับปรุงการกระจายตัวของข้อมูล โดยแม้ว่าจะมีวิธีการในการใช้เพิ่มตัวอย่างของข้อมูลที่หลากหลายวิธีแต่ก็ไม่สามารถระบุได้อย่างแน่ชัดว่าวิธีการใดที่ส่งผลดีที่สุดในการใช้งาน

มีงานวิจัยจำนวนมากที่มุ่งเน้นในด้านเกี่ยวกับการจัดการกับความไม่สมดุลกันของข้อมูล เพื่อค้นหาว่าวิธีการเพิ่มจำนวนของตัวอย่างข้อมูลวิธีใดที่ให้ประสิทธิภาพที่ดีที่สุด เช่น ในงานวิจัย (Hulse, Khoshgoftaar, & Napolitano, 2007) มีการใช้งานชุดข้อมูลในการประเมินประสิทธิภาพมากถึง 35 ชุดข้อมูล มีการใช้งานวิธีการในการเพิ่มตัวอย่างของข้อมูลมากถึง 7 วิธี และมีการใช้งานร่วมกับการเรียนรู้ด้วยเครื่องมากถึง 7 แบบจำลอง โดยงานวิจัยนี้มุ่งเน้นไปยังการจัดการข้อมูลที่มีความไม่สมดุลซึ่งสามารถจำแนกได้สองประเภท

ชุดข้อมูลในงานวิจัยที่มีความไม่สมดุลมากที่สุดมีจำนวนข้อมูลของกลุ่มที่น้อยสุดอยู่ที่ 1.33% ซึ่งมีความไม่สมดุลที่สูงมาก และชุดข้อมูลที่ความไม่สมดุลน้อยที่สุดมีจำนวนข้อมูลของกลุ่มที่น้อยที่สุดอยู่ที่ 35% โดยชุดข้อมูลที่มีขนาดเล็กที่สุดในการทดลองประกอบด้วยข้อมูลจำนวนเพียง 214 ตัวอย่าง และชุดข้อมูลที่มีขนาดใหญ่ที่สุดในการทดลองมีสองชุดซึ่งแต่ละชุดประกอบด้วยข้อมูลจำนวน 20,000 ตัวอย่าง

ในงานวิจัยมีการใช้งานวิธีการจัดการข้อมูลที่ไม่สมดุลหลากทั้งหมด 7 วิธีการ ประกอบด้วย Random Undersampling (RUS), Random Oversampling (ROS), One-side Selection (OSS), Cluster-based Oversampling (CBOS), Wilson's Editing (WE), SMOTE (SM) และ Boderline-SMOTE (BSM)

วิธีการที่ค่อนข้างเป็นที่นิยมในการใช้จัดการกับชุดข้อมูลที่ไม่สมดุล ได้แก่ Random Oversampling และ Random Undersampling โดย Random Oversampling จะมีการสุ่มทำซ้ำตัวอย่างข้อมูลเพื่อเพิ่มจำนวนของข้อมูลในกลุ่มที่มีจำนวนน้อยกว่า ในทางกลับกัน Random Undersampling จะเป็นการสุ่มตัดทิ้งตัวอย่างข้อมูลเพื่อลดจำนวนของข้อมูลในกลุ่มที่มีจำนวนมากกว่า

โดยผลการวิจัยสามารถแสดงให้เห็นถึงข้อสรุปโดยคร่าว เช่น วิธีการเพิ่มตัวอย่างข้อมูลแต่ละวิธีจะมีประสิทธิภาพที่ดีที่สุดแตกต่างกันไปเมื่อทำงานร่วมกับแบบจำลองที่แตกต่างกัน เช่น วิธีการ Random Undersampling จะให้ประสิทธิภาพสูงที่สุดเมื่อทำงานร่วมกับแบบจำลองป่าต้นไม้ตัดสินใจ (Random Forest) หรือ วิธีการ Random Oversampling เมื่อทำงานร่วมกับแบบจำลอง Logistic Regression จะให้ประสิทธิภาพที่ดีที่สุด

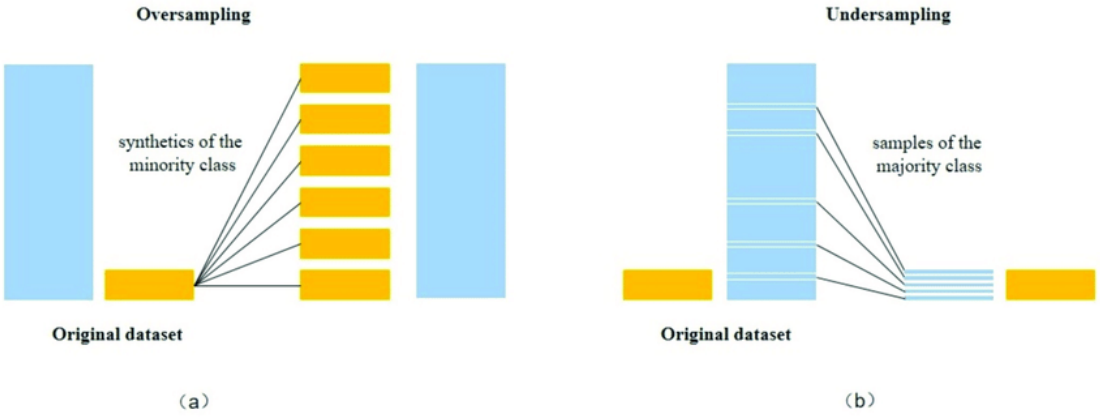
ซึ่งโดยภาพรวม Random Undersampling ค่อนข้างให้ประสิทธิภาพโดยรวมที่ดีกับชุดข้อมูลและแบบจำลองส่วนใหญ่ในการทดลอง ซึ่งวิธีการ Random Undersampling สรุปได้ว่าเป็นวิธีการจัดการกับความไม่สมดุลของข้อมูลที่ดีที่สุดการวิจัย โดยวิธีการ Random Oversampling เป็นวิธีการที่มีประสิทธิภาพโดยรวมสูงสุดเป็นอันดับรองลงมา จากนั้นจึงตามมาด้วยวิธีการ SMOTE และ Boderline-SMOTE ตามลำดับ (Hulse et al., 2007)

2.3.1 เทคนิค Random Undersampling

Random Undersampling มักถูกใช้กับงานในการจำแนกประเภทแบบไบนารี (Binary Classification) ซึ่งมีกลุ่มของการจำแนกอยู่สองกลุ่ม เพื่อลดจำนวนตัวอย่างข้อมูลของกลุ่มที่มีข้อมูลส่วนใหญ่ให้ลดลงมาเท่ากับกลุ่มข้อมูลที่มีจำนวนน้อย โดยสามารถปรับลดจนกระทั่งทั้งสองกลุ่มมีสัดส่วนของข้อมูลที่เท่ากันได้ ในการลดจำนวนของตัวอย่างข้อมูลจะเป็นการดำเนินการคัดออกแบบสุ่ม (Hasanin & Khoshgoftaar, 2018)

การใช้งานวิธีการ Random Undersampling มีข้อดีกว่าการใช้งานวิธีแบบ Oversampling ในแง่ของการไม่เกิดปัญหา Overfitting เนื่องจาก Random Undersampling จะไม่มีการวนซ้ำเพื่อเรียนรู้ในการหารูปแบบของข้อมูลกับตัวอย่างข้อมูลเดิมๆ ที่ถูกสุ่มทำซ้ำขึ้นมา โดยวิธีการนี้เป็นวิธีการที่มีความรวดเร็ว มีประสิทธิภาพและสามารถเชื่อถือได้ในการนำมาใช้จัดการกับชุดข้อมูลที่มีความไม่สมดุล (Saripuddin et al., 2021)

แต่ในขณะเดียวกันแม้ว่าวิธีการ Random Undersampling จะเป็นวิธีการที่ง่ายและมีความสะดวกรวดเร็วในการจัดการความไม่สมดุลของชุดข้อมูล แต่ก็อาจจะส่งผลต่อความแปรปรวน (Variance) ของแบบจำลองได้เนื่องจากอาจจะมีการสูญเสียข้อมูลที่มีความสำคัญไป (Wikipedia, 2017b)



ภาพประกอบ 11 แสดงการเปรียบเทียบหลักการทำงานของวิธีการแบบ Random Oversampling และ Random Undersampling

ที่มา: (Xia et al., 2019)

2.3.2 เทคนิค SMOTE (Synthetic Minority Oversampling Technique)

เทคนิค SMOTE หรือ Synthetic Minority Oversampling เกิดขึ้นเนื่องจากมีผลของการวิจัยที่ระบุว่าการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling แบบใส่คืน (Oversampling with Replacement) ไม่ได้ส่งผลในการช่วยให้แบบจำลองมีประสิทธิภาพที่ดีมากนักในการเรียนรู้กับข้อมูลกลุ่มที่มีประชากรน้อย ซึ่งจากการทดลองกับต้นไม้ตัดสินใจแสดงให้เห็นว่าการเพิ่มข้อมูลแบบ Oversampling จะไปเพิ่มขอบเขตของการตัดสินใจในฝั่งของกลุ่มข้อมูลที่มีประชากรน้อย โดยไม่ได้ส่งผลต่อขอบเขตของการตัดสินใจในฝั่งของกลุ่มข้อมูลที่มีประชากรมากกว่า จึงส่งผลให้มีความเฉพาะเจาะจงเพิ่มมากขึ้นและสามารถนำมาซึ่งการแตกกิ่งเพื่อขยายขนาดของต้นไม้ในฝั่งของกลุ่มข้อมูลที่มีประชากรน้อยกว่า ซึ่งสามารถก่อให้เกิดปัญหา Overfitting ได้

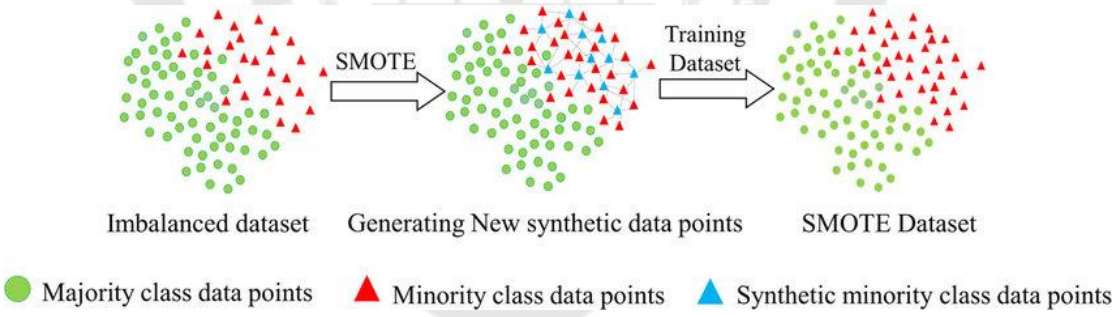
SMOTE เป็นการ Oversampling ด้วยการสังเคราะห์ข้อมูลขึ้นมาในมิติคุณลักษณะของข้อมูล (Feature Space) โดยมีการคำนวณจากข้อมูลจริงที่ใกล้ที่สุดจำนวน k ตัวที่ปรากฏในชุดข้อมูล โดยข้อมูลที่ถูกสร้างขึ้นมาใหม่จะไม่ทับซ้อนกับจุดข้อมูลเดิมที่มีอยู่ในชุดข้อมูล ซึ่งจุดข้อมูลใหม่ที่ถูกสังเคราะห์ขึ้นจะส่งผลในการเพิ่มหรือขยายขอบเขตของการตัดสินใจ (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)

ในการจัดการกับชุดข้อมูลที่ไม่สมดุลในงานการจำแนกประเภทด้วยวิธีการเพิ่มตัวอย่างของข้อมูล หรือ Oversampling ประกอบด้วยวิธีการที่หลากหลายโดย SMOTE เป็นหนึ่ง

ในวิธีการที่ได้รับความนิยม โดยวิธีการคือจะมีการสร้างมิติของคุณลักษณะของข้อมูล (Feature Space) ตัวอย่างเช่น งานในการจำแนกประเภทของนก จะมีการสร้างมิติของคุณลักษณะของข้อมูลกลุ่มที่มีจำนวนประชากรน้อยซึ่งประกอบด้วย ความยาวของจะงอยปาก ความยาวของปีก และน้ำหนักของนก เป็นต้น หลังจากขั้นตอนในการสร้างมิติของคุณลักษณะจะเป็นขั้นตอนในการสร้างข้อมูลสังเคราะห์โดยอ้างอิงจากข้อมูลดั้งเดิมที่ใกล้ที่สุดจำนวน k ตัว ที่ปรากฏในมิติของคุณลักษณะ ซึ่ง k จะเป็นค่าตัวแปรที่สามารถกำหนดได้ จากนั้นทำการเพิ่มจุดข้อมูลสังเคราะห์ที่ได้ลงไปในมิติของคุณลักษณะข้อมูล และทำซ้ำกระบวนการจนได้จำนวนของข้อมูลในกลุ่มที่มีประชากรน้อยในสัดส่วนตามที่กำหนดหรือต้องการ (Wikipedia, 2017b)

SMOTE เป็นวิธีการจัดการความไม่สมดุลของข้อมูลที่มีพื้นฐานหรือหลักการมาจากการ Oversampling เพื่อเพิ่มข้อมูลในกลุ่มที่มีจำนวนน้อยโดยใช้กระบวนการสังเคราะห์ข้อมูลตัวอย่างแบบมีหลักการแทนที่การใช้เพิ่มตัวอย่างข้อมูลแบบสุ่ม (Hulse et al., 2007)

SMOTE เป็นการสร้างจุดข้อมูลใหม่โดยอาศัยการอ้างอิงจากข้อมูลดั้งเดิมที่ปรากฏในชุดข้อมูลซึ่งมีความท้าทายในแง่ของการตรวจสอบตัวอย่างข้อมูลที่ถูกสร้างขึ้นมาใหม่มีความถูกต้องหรือไม่ โดยการจัดการความไม่สมดุลของข้อมูลด้วยวิธีการแบบ SMOTE อาจก่อให้เกิดปัญหาในด้านของ Overfitting (Saripuddin et al., 2021)



ภาพประกอบ 12 แสดงการทำงานของวิธีการ SMOTE ในการจัดการความไม่สมดุลกันของข้อมูล

ที่มา: (Aldraimli et al., 2020)

2.4 ทฤษฎีเกี่ยวกับวิศวกรรมคุณลักษณะข้อมูล (Feature Engineering)

การทำวิศวกรรมคุณลักษณะข้อมูลมีชื่อเรียกที่หลากหลาย เช่น Feature Engineering หรือ Feature Extraction หรือ Feature Discovery ซึ่งเป็นกระบวนการซึ่งมีการใช้งานองค์ความรู้ทางสาขาวิชานั้นๆ ในการนำมาใช้สกัดคุณลักษณะ ปรับปรุงแก้ไข ทำความสะอาด และเปลี่ยนแปลงข้อมูลให้อยู่ในรูปที่พร้อมนำไปใช้งาน โดยมีเป้าหมายเพื่อเพิ่มประสิทธิภาพของการ

เรียนรู้ด้วยเครื่อง ซึ่งจะมีประสิทธิภาพมากกว่าการเรียนรู้จากชุดข้อมูลดิบที่ไม่ได้มีการทำวิศวกรรม (Wikipedia, 2017a)

การทำวิศวกรรมคุณลักษณะข้อมูลเป็นขั้นตอนที่มีความสำคัญในการเรียนรู้ด้วยเครื่อง เนื่องจากว่าคุณภาพของข้อมูลจะส่งผลโดยตรงต่อประสิทธิภาพในการเรียนรู้ของแบบจำลอง โดยกระบวนการหลักๆ ในการทำวิศวกรรมข้อมูลมีดังต่อไปนี้

2.4.1 การจัดการกับข้อมูลค่าว่าง (Handling Null Value)

ในการเรียนรู้ด้วยเครื่องไม่ว่าจะเป็นงานแบบการจำแนกประเภท หรืองานแบบการวิเคราะห์การถดถอย หรืองานอื่นๆ ในชุดข้อมูลที่จะนำมาใช้งานมักจะปรากฏข้อมูลที่มีค่าว่างอยู่เสมอ ซึ่งแบบจำลองส่วนมากไม่สามารถจัดการกับข้อมูลค่าว่างเหล่านี้ได้อย่างมีประสิทธิภาพ ดังนั้นจึงต้องมีการจัดการกับข้อมูลเหล่านี้อย่างเหมาะสม การจัดการกับข้อมูลค่าว่างที่เป็นที่นิยม เช่น การคัดออกตัวอย่างข้อมูลหรือคุณลักษณะที่มีข้อมูลค่าว่างปรากฏ ซึ่งเป็นวิธีการที่เหมาะกับชุดข้อมูลที่มีข้อมูลค่าว่างเป็นสัดส่วนในจำนวนที่น้อย ทำให้การคัดออกไม่ส่งผลกระทบต่อจำนวนของข้อมูลมากนัก หรือการเติมข้อมูล (Imputation) ให้กับข้อมูลที่มีค่าว่าง ซึ่งเป็นการเติมข้อมูลลงไปให้กับข้อมูลที่มีค่าว่างโดยสามารถใช้งานฟังก์ชันที่มีการกำหนดขึ้นมาเอง หรือใช้งานฟังก์ชันสำเร็จรูปซึ่งมีเทคนิคที่สามารถแบ่งย่อยได้จำนวนมาก เช่น การเติมค่าว่างด้วยค่าเฉลี่ย หรือการเติมค่าว่างด้วยค่า 0

2.4.2 การจัดมาตรฐานของข้อมูล (Standardization)

เป็นขั้นตอนที่มีความสำคัญสำหรับการเรียนรู้ด้วยเครื่อง เป็นการเปลี่ยนแปลงข้อมูลชนิดตัวเลขหรือ Numerical Feature ซึ่งจะทำให้แต่ละคุณลักษณะมีค่าเฉลี่ยของข้อมูล (Mean) เป็น 0 และมีส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) เป็น 1

ตัวอย่างเช่น ชุดข้อมูลประกอบด้วยข้อมูลของอายุและเงินเดือน ซึ่งทั้งสองคุณลักษณะจะมีหน่วยวัด (Metric) และขนาด (Scale) ที่แตกต่างกันมาก ส่งผลให้เมื่อแบบจำลองทำการเรียนรู้จะให้ความสำคัญกับเงินเดือนซึ่งมีขนาดของข้อมูลที่ใหญ่กว่า ซึ่งไม่ใช่การทำงานที่เหมาะสมของแบบจำลอง ดังนั้นจึงต้องมีการจัดการมาตรฐานของข้อมูลให้อยู่ในขนาดเดียวกันเพื่อป้องกันปัญหาที่กล่าวไป

โดยหลักของการดำเนินการอย่างง่ายคือคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของทั้งชุดข้อมูล จากนั้นคำนวณโดยนำแต่ละจุดข้อมูลมาทำการลบด้วยค่าเฉลี่ยแล้วทำการหารด้วยส่วนเบี่ยงเบนมาตรฐาน ดังแสดงในสมการที่ 2 ซึ่งจะมีฟังก์ชันสำเร็จรูปให้สามารถเรียกใช้งานได้ โดยไม่ต้องทำการคำนวณเอง เช่น ฟังก์ชัน ‘StandardScaler’ ของไลบรารี Scikit-Learn

$$Z = \frac{x_i - \mu}{\sigma} \qquad (2)$$

2.4.3 การจัดการข้อมูลชนิดประเภท (Handling Categorical Feature)

ข้อมูลชนิดประเภทเป็นข้อมูลแบบไม่ต่อเนื่องกัน โดยข้อมูลชนิดประเภทสามารถแบ่งออกได้เป็น 3 รูปแบบ คือ

1. ข้อมูลชนิดประเภทแบบมีลำดับ (Ordinal Categorical Feature) ซึ่งข้อมูลสามารถเรียงลำดับได้ เช่น ระดับการศึกษา

2. ข้อมูลชนิดประเภทแบบไม่มีลำดับ (Nominal Categorical Feature) ซึ่งข้อมูลไม่สามารถเรียงลำดับได้ เช่น สีของวัตถุ

3. ข้อมูลชนิดประเภทที่เป็นผลลัพธ์ของการจำแนกประเภทหรือป้ายกำกับ (Label Categorical Feature)

การจัดการข้อมูลชนิดประเภทแบบมีลำดับและข้อมูลชนิดประเภทแบบไม่มีลำดับ จะใช้วิธีการที่แตกต่างกัน ดังนี้

การจัดการข้อมูลชนิดประเภทแบบมีลำดับจะมีการแปลงข้อมูลออกมาเป็นตัวเลข โดยจะยังคงลำดับของข้อมูลตามความมากน้อย เช่น ระดับความพึงพอใจซึ่งสามารถเรียงจากน้อยไปมากดังนี้ แย่, ดี, ดีมาก และยอดเยี่ยม ตามลำดับ โดยเมื่อผ่านการจัดการแล้วข้อมูลจะออกมาเป็นลำดับ 1 (แย่), 2 (ดี), 3 (ดีมาก) และ 4 (ยอดเยี่ยม) ตามลำดับ ดังแสดงในภาพประกอบที่ 13

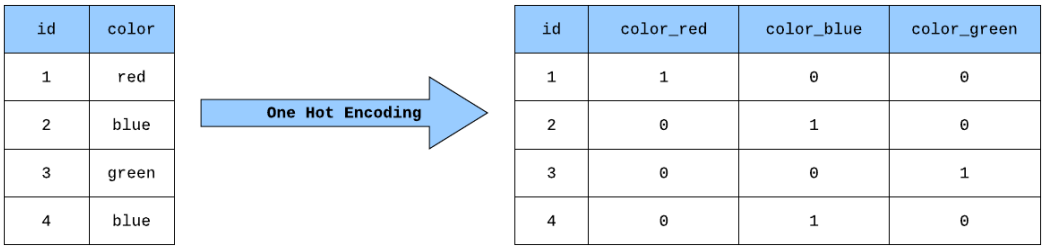
Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

ภาพประกอบ 13 แสดงการจัดการข้อมูลด้วยวิธีการแบบ Ordinal Encoding กับข้อมูลความพึงพอใจ

ที่มา: (SagarDhandare, 2021)

การจัดการข้อมูลชนิดประเภทแบบไม่มีลำดับต้องใช้วิธีการที่ไม่ทำให้ข้อมูลหลังการเปลี่ยนแปลงมีการเรียงลำดับของข้อมูล ซึ่งมักจะเป็นข้อผิดพลาดในการดำเนินการที่พบเจอได้บ่อย โดยการจัดการที่ถูกต้องควรใช้วิธีการที่เรียกว่า One-Hot Encoding (McGinnis, 2022) ซึ่ง

จะมีการสร้างคุณลักษณะเพิ่มเติมขึ้นมา (Dummy Feature) เท่ากับจำนวนของค่าที่เป็นไปได้ทั้งหมดของคุณลักษณะนั้น ตัวอย่างเช่น คุณลักษณะสีของวัตถุ ที่ประกอบด้วยค่าที่เป็นไปได้คือ แดง ฟ้า และเขียว เมื่อผ่านกระบวนการจัดการแล้วจะได้ข้อมูลออกมาเป็น 3 คุณลักษณะคือ คุณลักษณะสีแดง คุณลักษณะสีฟ้า และคุณลักษณะสีเขียว โดยตัวอย่างข้อมูลที่มีค่าเป็นสีใดจะส่งผลให้คุณลักษณะของสีใหม่ที่เกิดขึ้นมีค่าเป็น 1 และคุณลักษณะที่เกิดใหม่อื่นๆ จะมีค่าเป็น 0 เช่น ข้อมูลมีค่าเป็นสีฟ้า เมื่อผ่านการจัดการข้อมูลจะได้คุณลักษณะสีฟ้าที่มีค่าเป็น 1 ส่วนคุณลักษณะสีแดงและคุณลักษณะสีเขียวจะมีค่าเป็น 0 ดังแสดงในภาพประกอบที่ 14 (Kumar, 2018)



ภาพประกอบ 14 แสดงการจัดการข้อมูลด้วยวิธีการแบบ One-Hot Encoding กับข้อมูลสี

ที่มา: (Novack, 2020)

การจัดการข้อมูลชนิดประเภทที่เป็นผลลัพธ์ของการจำแนกประเภทหรือป้ายกำกับมักนิยมใช้งานวิธีการแบบ Label Encoding โดยหลักการทำงานจะคล้ายคลึงกับการทำงานของ Ordinal Encoding ซึ่งจะไม่มีการขยายจำนวณของคุณลักษณะใหม่ออกมา แต่จะเป็นการเปลี่ยนแปลงค่าที่ปรากฏทั้งหมดให้กลายเป็นตัวเลขซึ่งจะมีจำนวนตัวเลขเท่ากับค่าที่แตกต่างกันที่ปรากฏในป้ายกำกับ โดยตัวเลขหลังการจัดการจะมีลำดับต่อกันเช่นเดียวกับการทำ Ordinal Encoding ดังนั้นจึงนิยมนำมาใช้งานเฉพาะกับคุณลักษณะที่เป็นผลลัพธ์ของการจำแนกประเภทเท่านั้น ตัวอย่างเช่น แบบจำลองในการทำนายทวีป ซึ่งจะมีผลลัพธ์ของการทำนายที่เป็นไปได้ทั้งหมด 6 ทวีป เมื่อทำการจัดการข้อมูลด้วยวิธี Label Encoding แล้ว จะได้ข้อมูลหลังการจัดการออกมาเป็นค่า 1 ถึง 6 โดยไม่ได้มีการเพิ่มขึ้นใหม่ของคุณลักษณะ

2.5 ทฤษฎีเกี่ยวกับการคัดเลือกคุณลักษณะ (Feature Selection)

การทำงานกับข้อมูลซึ่งมีมิติของคุณลักษณะของข้อมูลแบบหลายมิติ (High-Dimensional) ถือเป็นความท้าทายของนักวิจัยในสาขาวิชาการเรียนรู้ด้วยเครื่องและการทำเหมืองข้อมูล (Data Mining) โดยกระบวนการคัดเลือกคุณลักษณะจะเข้ามามีบทบาทสำคัญในการช่วยแก้ไขปัญหาต่างๆ เช่น การตัดคุณลักษณะที่เป็นข้อมูลรบกวนหรือไม่มีความเกี่ยวข้องออก หรือการตัดคุณลักษณะที่มีความซ้ำซ้อนต่อกัน ซึ่งจะช่วยให้การทำงานของการเรียนรู้ด้วยเครื่องมีประสิทธิภาพมากขึ้น ใช้เวลาในการดำเนินการน้อยลง ช่วยเพิ่มความแม่นยำ และช่วยให้การทำความเข้าใจกับแบบจำลองสามารถทำได้ง่ายและสะดวกมากขึ้น

ในเทคโนโลยียุคปัจจุบันซึ่งมีการใช้งานทั้งคอมพิวเตอร์ อินเตอร์เน็ต และอุปกรณ์ต่างๆ ทำให้จำนวนของข้อมูลมีเพิ่มสูงขึ้นอย่างรวดเร็ว ซึ่งข้อมูลส่วนมากในยุคปัจจุบันมันมีมิติของข้อมูลที่หลากหลาย ซึ่งส่งผลโดยตรงต่องานที่เกี่ยวกับการวิเคราะห์ข้อมูลหรือกระบวนการตัดสินใจ (Decision-Making Process) โดยกระบวนการคัดเลือกคุณลักษณะได้ถูกพิสูจน์แล้วว่าสามารถช่วยเพิ่มประสิทธิภาพแก่การเรียนรู้ด้วยเครื่อง

กระบวนการคัดเลือกคุณลักษณะเป็นกระบวนการซึ่งคัดเลือกชุดย่อยคุณลักษณะของข้อมูลที่มีความสำคัญออกมาจากคุณลักษณะทั้งหมดของชุดข้อมูล ซึ่งจะมีลักษณะเหมือนกับการบีบให้การดำเนินการกับข้อมูลมีขั้นตอนที่เล็กลงโดยการนำคุณลักษณะที่มีความซ้ำซ้อนสูง หรือคุณลักษณะที่ไม่มีความเกี่ยวข้องกับชุดข้อมูลออก โดยการใช้งานวิธีการคัดเลือกคุณลักษณะที่เหมาะสมกับชุดข้อมูลสามารถส่งผลให้แบบจำลองมีประสิทธิภาพที่สูงขึ้น สามารถประหยัดเวลาสำหรับใช้ในการเรียนรู้ด้วยเครื่อง และทำให้แปรผลแบบจำลองได้ง่ายมากขึ้น

กระบวนการคัดเลือกคุณลักษณะแบบมีผู้สอน (Supervised Feature Selection) มักถูกนำเสนอและนำมาใช้งานกับปัญหาประเภทการจำแนกข้อมูลโดยใช้การค้นหาความเกี่ยวข้องหรือความสัมพันธ์ระหว่างคุณลักษณะกับป้ายกำกับของข้อมูล โดยแบบจำลองของการเรียนรู้ด้วยเครื่องแบบมีผู้สอนจะทำการค้นหาชุดย่อยของคุณลักษณะ ที่ทำให้การจำแนกข้อมูลมีประสิทธิภาพความแม่นยำที่สูงที่สุด (Cai, Luo, Wang, & Yang, 2018)

กระบวนการคัดเลือกคุณลักษณะของข้อมูลสามารถแบ่งออกเป็นสองแบบย่อย ได้แก่ Wrapper และ Filter โดยวิธีการแบบ Wrapper จะต้องมีการระบุตัวจำแนกประเภท (Classifier) ไว้ล่วงหน้าสำหรับนำมาเพื่อใช้ในการประเมินประสิทธิภาพของแต่ละคุณลักษณะ ส่วนวิธีการแบบ Filter จะไม่มีนำตัวจำแนกประเภทมาเกี่ยวข้องในการประเมินประสิทธิภาพของแต่ละคุณลักษณะ โดย F-Score เป็นเทคนิคหนึ่งในการคัดเลือกคุณลักษณะซึ่งเป็นวิธีการแบบ Filter

วิธีการแบบ Wrapper และ Filter มีทั้งข้อดีและข้อเสียที่แตกต่างกันไป โดยวิธีการแบบ Filter จะมีกระบวนการในการคำนวณที่ไม่ซับซ้อนและประหยัดแต่ก็จะมีประสิทธิภาพความแม่นยำที่ด้อยกว่า ส่วนวิธีการแบบ Wrapper สามารถให้ประสิทธิภาพที่ดีกว่าแต่ก็จะมีการใช้พลังงานในการคำนวณที่สูงขึ้นและความยืดหยุ่นในการนำไปปรับใช้ค่อนข้างน้อย

แนวทางของกระบวนการคัดเลือกคุณลักษณะ อัลกอริทึมจะมีวิธีการดำเนินการแบ่งได้สองแนวทาง วิธีการแรกเป็นการหาความสำคัญของคุณลักษณะรายตัว วิธีการที่สองเป็นการหาความสำคัญของคุณลักษณะแบบชุดย่อย

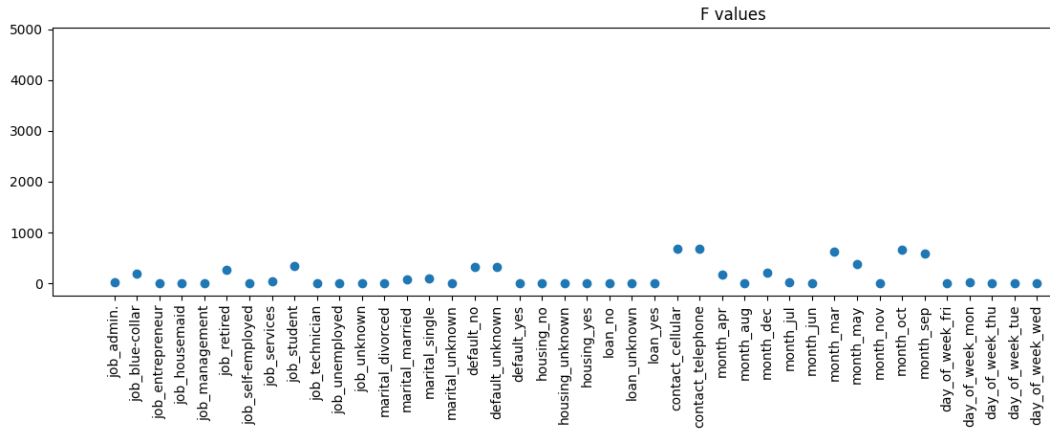
วิธีการหาความสำคัญของคุณลักษณะรายตัวจะมีการจัดลำดับความสำคัญโดยใช้วิธีการคำนวณทางสถิติและจะมีการกำจัดคุณลักษณะที่มีความสำคัญน้อยออกไป วิธีการนี้เป็นที่นิยมในการใช้งานเนื่องจากมีความซับซ้อนน้อย มีความยืดหยุ่นสูง แต่ต้องมีการกำหนดเส้นแบ่ง (Threshold) สำหรับระบุจำนวนของคุณลักษณะที่ต้องการนำมาใช้งาน และยังมีข้อจำกัดในการทำงานกับการจำแนกประเภทข้อมูลที่มีลักษณะหลายมิติ

วิธีการที่สองเป็นการหาความสำคัญของคุณลักษณะแบบชุดย่อย จะมีการค้นหาชุดย่อยของคุณลักษณะที่ทำให้มีประสิทธิภาพดี โดยคุณลักษณะจะค่อยๆ ถูกจับกลุ่มเพิ่มขึ้นเรื่อยๆ เพื่อค้นหาชุดย่อยของคุณลักษณะที่ให้ประสิทธิภาพสูงที่สุดออกมา แต่จะส่งผลทำให้การประมวลผลต้องใช้พลังงานและเวลาที่สูงขึ้น (Song, Jiang, & Liu, 2017)

2.5.1 F-Score (F-Value)

การคัดเลือกคุณลักษณะด้วยวิธี F-Score หรือ F-Value เป็นวิธีที่มีความซับซ้อนน้อยแต่มีประสิทธิภาพ ซึ่งในปัจจุบันมีงานวิจัยมากมายที่ประสบความสำเร็จในการคัดเลือกคุณลักษณะด้วยวิธีนี้ ค่าของ F-Score ที่สูงบ่งบอกถึงความมีอิทธิพลของคุณลักษณะนั้นๆ ต่อความสามารถในการใช้ระบุถึงป้ายกำกับหรือประเภทของตัวอย่างข้อมูล ซึ่งกลุ่มของคุณลักษณะที่มีค่าของ F-Score สูงมักจะถูกนำไปใช้ในการทำงานสร้างแบบจำลองต่อไป

F-Score มีประโยชน์ในการช่วยลดจำนวนของคุณลักษณะเพื่อใช้ในการสร้างแบบจำลอง ส่งผลการเรียนรู้ของเครื่องทำงานได้รวดเร็วและมีประสิทธิภาพมากขึ้น นอกจากนี้ยังมีความสำคัญในการช่วยเพิ่มประสิทธิภาพความแม่นยำของการจำแนกประเภทของข้อมูล (Gao et al., 2014)



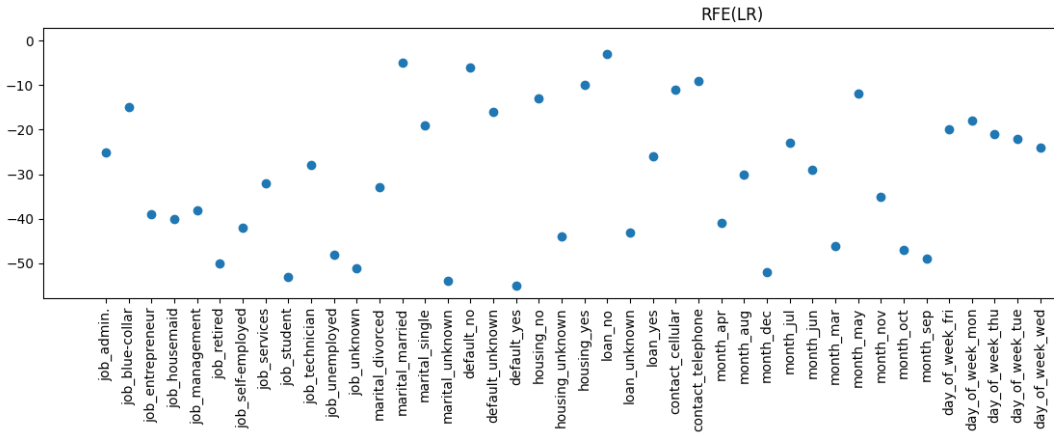
ภาพประกอบ 15 แสดง F-Score ของคุณลักษณะผ่านการเรียกใช้งานไลบรารี Scikit-Learn

มีงานวิจัยจำนวนมากที่มีการใช้ F-Score ในกระบวนการคัดเลือกคุณลักษณะ เช่น การใช้สำหรับการคัดเลือกคุณลักษณะที่มีค่า F-Score ในระดับสูงเพื่อนำมาใช้งาน หรือการใช้ F-Score สำหรับคัดเลือกคุณลักษณะในงานตรวจจับการโกหกและเพิ่มประสิทธิภาพความแม่นยำของการจำแนกประเภท (Song et al., 2017)

2.5.2 Recursive Feature Elimination (RFE)

Recursive Feature Elimination เป็นหนึ่งในวิธีการคัดเลือกคุณลักษณะที่มักถูกนำมาใช้งานกับชุดข้อมูลที่มีขนาดไม่ใหญ่มากนัก โดย RFE มีการประยุกต์ใช้งานประโยชน์จากลักษณะโดยทั่วไป (Generalization Capability) ของ Support Vector Machine (SVM) ซึ่งจะพยายามลบคุณลักษณะที่มีความซ้ำซ้อนต่อกัน (Redundant) และเป็นคุณลักษณะที่อ่อนแอ (Weak Feature) และพยายามเก็บคุณลักษณะที่มีความเป็นอิสระไว้

RFE มีหลักการทำงานโดยจะพยายามคัดออกคุณลักษณะที่อ่อนแอ ซึ่งมีอิทธิพลต่ำในการจำแนกประเภทของข้อมูลหรือระบุป้ายกำกับ ซึ่งจะทำให้เหลือแต่ชุดของคุณลักษณะที่มีประสิทธิภาพในการช่วยเพิ่มความแม่นยำในการสร้างแบบจำลอง โดยในการคัดคุณลักษณะออกจะคำนวณจากอิทธิพลของคุณลักษณะที่ส่งผลต่อค่าความผิดพลาดระหว่างการเรียนรู้ (Training Errors) โดยจะนำคุณลักษณะที่ส่งผลต่อค่าความผิดพลาดระหว่างการเรียนรู้น้อยที่สุดออก (Chen & Jeong, 2007)



ภาพประกอบ 16 แสดงความสำคัญของคุณลักษณะด้วยวิธีการ RFE กับแบบจำลอง Logistic Regression ผ่านการเรียกใช้งานไลบรารี Scikit-Learn

RFE เป็นอัลกอริทึมแบบ Greedy ซึ่งมีการใช้งานเทคนิคการจัดลำดับของคุณลักษณะ โดยจะเริ่มจากการใช้งานชุดของคุณลักษณะทั้งหมด แล้วจึงค่อยๆ คัดออกคุณลักษณะที่มีประสิทธิภาพต่ำสุดทีละคุณลักษณะเพื่อให้เหลือเพียงชุดของคุณลักษณะที่มีประสิทธิภาพสูง (Zhou, Zhou, Zhou, Yang, & Luo, 2014)

ข้อดีอื่นๆ ของ RFE นอกจากการลดจำนวนของคุณลักษณะซึ่งส่งผลต่อประสิทธิภาพของแบบจำลองและการลดเวลาในการเรียนรู้ด้วยเครื่อง ยังมีประโยชน์ในการใช้งานกับชุดข้อมูลที่มีขนาดเล็กหรือมีตัวอย่างของข้อมูลจำนวนไม่มากโดยยังสามารถให้ประสิทธิภาพที่ดี (Johannes et al., 2010)

2.6 ทฤษฎีเกี่ยวกับการประเมินผลประสิทธิภาพของแบบจำลอง (Model Evaluation)

ในงานเกี่ยวกับการจำแนกประเภทนั้นการประเมินผลประสิทธิภาพของแบบจำลองถือเป็นส่วนสำคัญในกระบวนการเรียนรู้ด้วยเครื่องเพื่อให้แบบจำลองสามารถทำงานอยู่ในระดับที่กำหนดได้ ดังนั้นการคัดเลือกวิธีการในการนำมาใช้เพื่อประมวลผลจึงมีความสำคัญอย่างยิ่งเพื่อให้ได้มาซึ่งแบบจำลองที่มีประสิทธิภาพดี

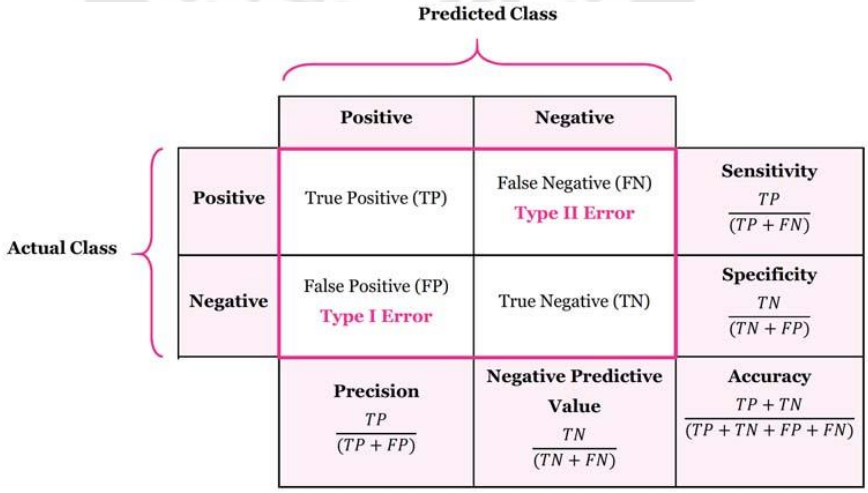
การประเมินประสิทธิภาพของแบบจำลองสามารถแบ่งออกได้เป็น 2 ระยะ ระยะแรกคือในขั้นตอนของการเรียนรู้เพื่อสร้างแบบจำลอง โดยการประเมินผลในระยะนี้จะถูกใช้เพื่อนำไปปรับปรุงแบบจำลองเพื่อพัฒนาให้มีประสิทธิภาพที่ดียิ่งขึ้น และระยะที่สองคือในขั้นตอนการประเมินผลกับข้อมูลสำหรับการทดสอบที่ไม่เคยเห็นมาก่อนในระหว่างการเรียนรู้ด้วยเครื่อง

โดยส่วนใหญ่แบบจำลองเกี่ยวกับการจำแนกประเภทมักถูกประเมินผลด้วยค่าความแม่นยำ (Accuracy) ในระหว่างการเรียนรู้เพื่อทำการสร้างแบบจำลอง ซึ่งในบางครั้งค่าความแม่นยำไม่สามารถแสดงถึงข้อมูลที่เพียงพอในการบ่งบอกประสิทธิภาพของแบบจำลอง โดยเฉพาะอย่างยิ่งค่าความแม่นยำมักจะมีความโน้มเอียงไปยังกลุ่มของข้อมูลที่เป็นประชากรส่วนใหญ่ในชุดข้อมูล ดังนั้นจึงต้องมีการใช้งานค่าประสิทธิภาพอื่นๆ ร่วมด้วยในการประเมินประสิทธิภาพของแบบจำลอง

2.6.1 Confusion Matrix

สำหรับการจำแนกประเภทแบบสองกลุ่มหรือไบนารี การประเมินผลประสิทธิภาพของแบบจำลองสามารถใช้งาน Confusion Matrix สำหรับการประเมินผลได้ โดยข้อมูลในแกนแนวตั้ง (Column) จะแสดงถึงข้อมูลจากการทำนายของแบบจำลอง ส่วนข้อมูลในแกนแนวนอน (Row) จะแสดงถึงข้อมูลจากกลุ่มของข้อมูลจริงหรือป้ายกำกับ

จาก Confusion Matrix ค่าของ True Positive (TP) และ True Negative (TN) จะบ่งบอกถึงจำนวนของข้อมูลในกลุ่มบวก (Positive) หรือกลุ่มข้อมูลที่สนใจ และจำนวนของข้อมูลในกลุ่มลบ (Negative) หรือกลุ่มข้อมูลที่ไม่สนใจซึ่งแบบจำลองสามารถจำแนกประเภทได้ถูกต้องตามลำดับ ในขณะที่ False Positive (FP) และ False Negative (FN) จะบ่งบอกถึงจำนวนของข้อมูลในกลุ่มบวก (Positive) หรือกลุ่มข้อมูลที่สนใจ และจำนวนของข้อมูลในกลุ่มลบ (Negative) หรือกลุ่มข้อมูลที่ไม่สนใจซึ่งแบบจำลองจำแนกประเภทผิดตามลำดับ ดังแสดงในภาพประกอบที่ 17



ภาพประกอบ 17 แสดง Confusion Matrix และมาตรวัดอื่นๆ ที่นิยมใช้งาน

ที่มา: (RapidMiner)

ซึ่งสามารถนำไปทำการคำนวณค่าอื่นๆ ที่น่าสนใจและเป็นที่นิยมสำหรับการบ่งชี้ประสิทธิภาพของแบบจำลองได้ ดังแสดงในตารางที่ 3 โดยงานในการจำแนกประเภทแบบสองกลุ่มหรือไบนารีค่าของ F1-Score สามารถใช้ประเมินผลประสิทธิภาพของแบบจำลองได้ดีกว่าค่าความแม่นยำ

ตาราง 3 มาตรวัดที่นิยมสำหรับการวัดประสิทธิภาพของแบบจำลองในการจำแนกประเภทจากการใช้งาน Confusion Matrix

มาตรวัด	การคำนวณ	การใช้งาน
ความแม่นยำ (Accuracy : Acc)	$\frac{TP + TN}{TP + FP + TN + FN}$	บ่งบอกสัดส่วนจำนวนของการจำแนกประเภทได้อย่างถูกต้องต่อข้อมูลทั้งหมดที่นำมาจำแนกประเภท
ความผิดพลาด (Error Rate : Err)	$\frac{FP + FN}{TP + FP + TN + FN}$	บ่งบอกสัดส่วนจำนวนของการจำแนกประเภทที่ผิดต่อข้อมูลทั้งหมดที่นำมาจำแนกประเภท
ความถูกต้องของกลุ่มข้อมูลบวก (Precision : P)	$\frac{TP}{TP + FP}$	บ่งบอกสัดส่วนของการจำแนกข้อมูลกลุ่มบวกได้อย่างถูกต้องต่อการจำแนกข้อมูลกลุ่มบวกทั้งหมด
ความครบถ้วนของกลุ่มข้อมูลบวก (Recall : R หรือ Sensitivity)	$\frac{TP}{TP + FN}$	บ่งบอกสัดส่วนของความครบถ้วนในการจำแนกข้อมูลกลุ่มบวกต่อข้อมูลกลุ่มบวกทั้งหมด
ค่าเฉลี่ยของ Precision และ Recall (F-Measure : FM หรือ F1-Score)	$\frac{2 * P * R}{P + R}$	บ่งบอกถึงค่าเฉลี่ยแบบ harmonic ระหว่างค่าของ Precision และ Recall

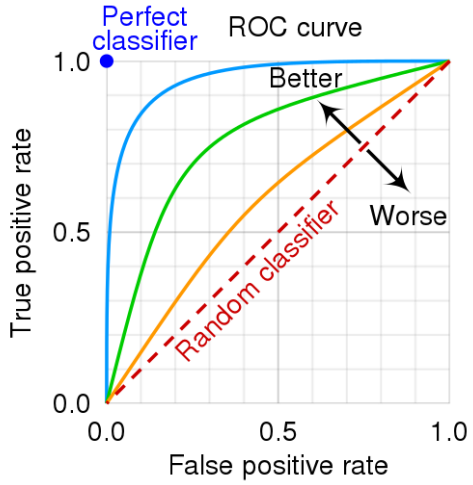
2.6.2 **พื้นที่ใต้กราฟ** (Area Under the Curve) Receiver Operating Characteristic (AUC-ROC)

พื้นที่ใต้กราฟ Receiver Operating Characteristic เป็นหนึ่งในการประเมินประสิทธิภาพของแบบจำลองที่เป็นที่นิยม และมักถูกนำมาใช้ในการเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองต่างๆ ที่สร้างขึ้น โดยค่าทางสถิติที่นำมาใช้คำนวณและวาดกราฟประกอบด้วย 2 ค่า คือค่า Sensitivity (ค่าความอ่อนไหว) หรือ True Positive Rate (TPR) หรือ Recall และค่า Specificity (ค่าความจำเพาะ) หรือ True Negative Rate (TNR)

โดย Sensitivity จะแสดงอัตราส่วนของจำนวนการจำแนกประเภทของกลุ่มบวกได้ถูกต้องต่อจำนวนของข้อมูลกลุ่มบวกทั้งหมด ส่วน Specificity จะแสดงอัตราส่วนของจำนวนการจำแนกประเภทของกลุ่มลบได้ถูกต้องต่อจำนวนของข้อมูลกลุ่มลบทั้งหมด

กราฟจะถูกวาดและแสดงโดยแกนในแนวตั้งหรือแกน Y จะแสดงถึงค่าของ Sensitivity ส่วนแกนในแนวนอนหรือแกน X จะแสดงถึงค่าของ 1-Specificity หรือ False Positive Rate (FPR) โดยแบบจำลองที่มีพื้นที่ใต้กราฟสูงสามารถบ่งบอกถึงความถูกต้องและความน่าเชื่อถือของแบบจำลองนั้นได้ (เบญจพร เอี่ยมประโคน, 2560)

โดยกราฟ Receiver Operating Characteristic จะมีจุดอุดมคติอยู่ที่มุมซ้ายบนของกราฟ ซึ่งหมายถึงมีค่าของ True Positive Rate เป็น 1 และมีค่าของ False Positive Rate เป็น 0 หรือสามารถจำแนกข้อมูลกลุ่มบวกได้อย่างครบถ้วนและถูกต้องทั้งหมด ซึ่งส่งผลให้แบบจำลองที่มีพื้นที่ใต้กราฟที่สูงสามารถบ่งบอกถึงความมีประสิทธิภาพของแบบจำลองนั้นๆ ได้ (scikit-learn)



ภาพประกอบ 18 แสดงการตีความหมายของพื้นที่ใต้กราฟแบบ ROC

ที่มา: (Commons, 2021)

2.7 ทฤษฎีเกี่ยวกับการเรียนรู้ด้วยเครื่องแบบอธิบายได้ (Interpretable Machine Learning)

ความสามารถในการอธิบาย (Interpretability) คือระดับของความเข้าใจของมนุษย์ในการรับรู้ถึงสาเหตุหรือเหตุผลของการตัดสินใจของแบบจำลอง หรือระดับความสอดคล้องในการที่มนุษย์สามารถคาดการณ์ถึงผลลัพธ์ของแบบจำลองได้ สาเหตุที่แบบจำลองควรมีความสามารถในการอธิบายได้สูงเนื่องจากผลลัพธ์ที่ถูกต้องสามารถแก้ไขปัญหาได้เพียงบางส่วน ซึ่งหลังจากนั้นในการแก้ไขปัญหาเราต้องการการเข้าใจถึงสาเหตุของปัญหาและกระบวนการในการตัดสินใจ

ตัวอย่างความสำคัญของความสามารถในการอธิบายได้ เช่น เราสามารถทำนายได้ว่าลูกค้าคนใดจะบอกเลิกการใช้บริการ (Churn Prediction) หากไร้ซึ่งความสามารถในการอธิบายได้เราจะรับรู้ได้เพียงแค่ลูกค้าคนใดมีแนวโน้มที่จะบอกเลิกการใช้บริการ แต่จะไม่สามารถรู้ถึงสาเหตุหรือเหตุผลที่แบบจำลองใช้ในการทำนายหรือคาดการณ์ ส่งผลให้ไม่สามารถกลับไปแก้ไขยังต้นตอที่แท้จริงของปัญหา

ในงานบางประเภทอาจไม่ต้องการความสามารถในการอธิบายที่สูงมากนักเนื่องจากเป็นงานที่มีความเสี่ยงที่ค่อนข้างต่ำ ไม่มีผลกระทบมากนักเมื่อเกิดความผิดพลาดขึ้น แต่ในทางกลับกันหากเป็นงานที่มีความเสี่ยงสูงซึ่งอาจเกี่ยวกับชีวิตหรือทรัพย์สิน การที่แบบจำลองสามารถอธิบายได้จะช่วยให้สามารถรับรู้ถึงปัญหาหรือสาเหตุที่แท้จริงพร้อมทั้งสามารถเพิ่มความเชื่อมั่นในการใช้งานแบบจำลองเพื่อดำเนินการแก้ไขปัญหา (Molnar, 2022)

แบบจำลองที่ยิ่งมีประสิทธิภาพความแม่นยำสูงก็จะยิ่งมีความซับซ้อนที่เพิ่มสูงขึ้นตามไปด้วยซึ่งส่งผลให้แบบจำลองถูกอธิบายหรือแปลผลออกมาได้ยากแม้แต่โดยผู้เชี่ยวชาญ ก่อให้เกิดความตึงเครียดระหว่างประสิทธิภาพความแม่นยำและความสามารถในการอธิบายได้สำหรับการเรียนรู้ด้วยเครื่อง ดังนั้นในปัจจุบันจึงมีวิธีการจำนวนมากซึ่งสามารถนำมาใช้ในการช่วยอธิบายแบบจำลองที่มีความซับซ้อนสูงซึ่งจะทำให้สามารถใช้งานแบบจำลองเหล่านั้นได้อย่างมั่นใจและเต็มประสิทธิภาพ (S. M. Lundberg & Lee, 2017)

2.7.1 Local Interpretable Model-Agnostic Explanations (LIME)

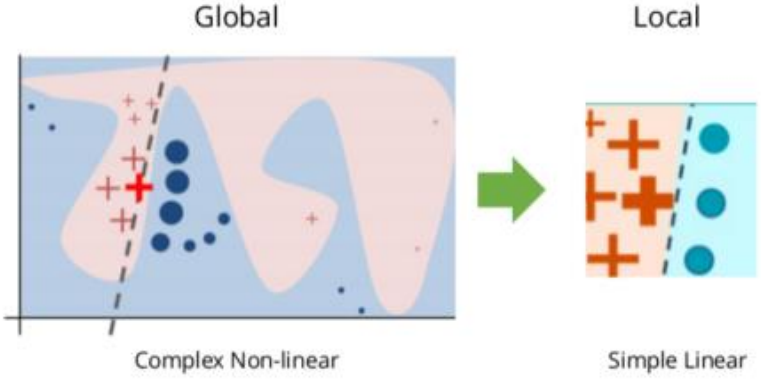
ในยุคปัจจุบันมีการนำการเรียนรู้ด้วยเครื่องไปใช้ในการทำงานหรือแก้ไขปัญหาอย่างกว้างขวางในหลากหลายธุรกิจและอุตสาหกรรมซึ่งสามารถทำงานได้อย่างมีประสิทธิภาพ แต่อย่างไรก็ตามการเรียนรู้ด้วยเครื่องส่วนใหญ่มักเป็นสิ่งที่ทำความเข้าใจได้ยากหรืออาจไม่สามารถทำความเข้าใจเกี่ยวกับเบื้องหลังการทำงานได้เลยโดยมนุษย์ จึงมักถูกเปรียบเสมือนกับกล่องดำ (Black Box) ซึ่งการนำการเรียนรู้ด้วยเครื่องไปใช้งานในการแก้ไขปัญหาจะมีข้อที่ต้องคำนึงถึงคือ ‘ผู้ใช้งานมักจะไม่ใช้แบบจำลองหากไม่สามารถเชื่อมั่นในการทำงานหรือการทำนายได้’ ดังนั้นการ

ทำความเข้าใจการทำงานของแบบจำลองจึงมีความสำคัญในการสร้างความมั่นใจและเชื่อใจแก่ผู้ใช้งานในการใช้ประโยชน์จากแบบจำลอง

สิ่งสำคัญในการอธิบายแบบจำลองคือต้องสามารถอธิบายออกมาได้อย่างมีคุณภาพ สามารถทำให้เข้าใจความสัมพันธ์ระหว่างตัวแปรต้นและผลลัพธ์ โดยต้องคำนึงถึงความสามารถที่จะเข้าใจได้ของมนุษย์ด้วย เช่น แบบจำลองที่มีคุณลักษณะของข้อมูลมากกว่า 100 ตัวที่ส่งผลต่อการทำนาย ซึ่งเป็นไปได้ยากที่จะมีผู้ใช้งานซึ่งสามารถเข้าใจแบบจำลองได้ ดังนั้นการอธิบายแบบจำลองจึงต้องทำให้อยู่ในรูปแบบที่ยืดหยุ่นและง่ายในการทำความเข้าใจ

Local Interpretable Model-Agnostic Explanations หรือ LIME (Ribeiro, Singh, & Guestrin, 2016) ถือเป็นวิธีในการอธิบายแบบจำลองที่มีความยืดหยุ่น โดยสามารถใช้งานร่วมกับแบบจำลองที่หลากหลายไม่ว่าจะเป็นการเรียนรู้ด้วยเครื่องแบบมีผู้สอน เช่น ป่าต้นไม้ตัดสินใจ หรือโครงข่ายประสาทเทียม (Neuron Networks) ในงานที่เกี่ยวกับรูปภาพ

LIME เป็นเทคนิคหรือวิธีการหนึ่งที่มีความน่าเชื่อถือในการใช้สำหรับอธิบายแบบจำลองในการจำแนกประเภท โดยใช้การเรียนรู้จากข้อมูลบริเวณโดยรอบของจุดข้อมูลที่กำลังทำการทำนายหรือจำแนกประเภท ซึ่งจะเป็นการอธิบายแบบจำลองในระดับรายตัวอย่างข้อมูลกับตัวอย่างข้อมูลอื่นๆ ในบริเวณใกล้เคียง (Local Interpretation) การอธิบายการทำงานในระดับแบบจำลองค่อนข้างทำได้ยากและอาจไม่มีประสิทธิภาพหรือแม่นยำไม่เพียงพอ ดังนั้นการทำงานของ LIME จึงเน้นไปที่ความมีประสิทธิภาพหรือแม่นยำในการอธิบายในระดับเฉพาะตำแหน่ง (local fidelity) ซึ่งจะค่อนข้างแม่นยำและเชื่อถือได้ในระดับจุดข้อมูลที่ทำการอธิบาย



ภาพประกอบ 19 แสดงการทำงานของ LIME ในการสร้าง Surrogate Model แบบเฉพาะพื้นที่สำหรับการอธิบายจุดข้อมูลแบบ Local

ที่มา: (AI)

การทำงานของ LIME โดยหลักการทำงานจะเริ่มจากการสร้างชุดข้อมูลใหม่ภายในบริเวณโดยรอบจุดข้อมูลที่ต้องการอธิบายและคำนวณความใกล้เคียงของจุดข้อมูลโดยรอบเทียบกับจุดข้อมูลที่ต้องการอธิบาย จากนั้นจะมีการสร้างแบบจำลองที่มีความซับซ้อนน้อย ยืดหยุ่นสูง และสามารถเข้าใจได้ง่าย เช่น แบบจำลองชนิดเส้นตรง (Linear Model) หรือแบบจำลองต้นไม้ตัดสินใจ (Decision Tree Model) มาทำการเรียนรู้ชุดข้อมูลย่อยที่สร้างขึ้นมาโดยพยายามให้การจำแนกประเภทของแบบจำลองใหม่มีผลลัพธ์เหมือนกับแบบจำลองหลักที่ใช้งาน จากนั้นจึงนำความสำคัญของแต่ละคุณลักษณะจากแบบจำลองที่มีความซับซ้อนน้อยมาแสดง ซึ่งสามารถเขียนได้ตามสมการที่ 3 (Molnar, 2022)

$$explanation(x) = argmin_{e \in G}\ [L(f, g, \pi_{x(z)}) + \ \Omega(\mathrm{g})] \qquad (3)$$

โดยที่

g : แบบจำลองสำหรับใช้ในการอธิบายแบบจำลองหลักในการจำแนกประเภท เช่น แบบจำลองแบบเส้นตรง (linear model)

G : เซตของแบบจำลองสำหรับใช้ในการอธิบายแบบจำลองหลักในการจำแนกประเภท

$\Omega(\mathrm{g})$: ความซับซ้อนของแบบจำลองสำหรับใช้อธิบายแบบจำลองหลักในการจำแนกประเภท (ตรงข้ามกับความสามารถในการอธิบายได้) เช่น ความลึก (Depth) ของแบบจำลองต้นไม้ตัดสินใจ

x : จุดข้อมูลที่ต้องการทำการอธิบาย

f(x) : แบบจำลองหลักในการจำแนกประเภทซึ่งมีความซับซ้อน

z : จุดข้อมูลบริเวณโดยรอบของ x

$\pi_{x(z)}$: ระยะทางหรือความใกล้เคียงระหว่าง x และ z โดยรอบ

$L(f, g, \pi_x)$: ปริมาณความไม่สอดคล้องกันระหว่างแบบจำลองสำหรับใช้ในการอธิบายและแบบจำลองหลักในการจำแนกประเภท

เป้าหมายของ LIME คือการเน้นความสามารถในการอธิบายได้และเข้าใจง่ายต่อผู้ใช้ โดยการอธิบายต้องมีความแม่นยำและสอดคล้องกับการทำนายของแบบจำลองหลักที่ใช้งาน ซึ่ง

ส่งผลให้เราต้องการลดความไม่สอดคล้อง หรือ $L(f, g, \pi_x)$ และความซับซ้อนของแบบจำลองในการอธิบาย หรือ $\Omega(g)$ ให้มีค่าน้อยที่สุด (Argmin)

2.7.2 SHapley Additive exPlanations (SHAP)

ในการอธิบายแบบจำลองที่มีความซับซ้อนสูงซึ่งเป็นที่นิยมในปัจจุบัน โดยส่วนมากมักจะใช้งานวิธีการในการสร้างแบบจำลองที่มีความซับซ้อนน้อย สามารถทำความเข้าใจได้ง่าย เพื่อนำมาใช้สำหรับการอธิบายแบบจำลองหลัก

Shapley Additive Explanations หรือ SHAP ใช้หลักการของทฤษฎีเกมส์ (Game Theory) มาใช้ในการอธิบายผลลัพธ์ของแบบจำลองจากการเรียนรู้ด้วยเครื่อง โดยมีการใช้งานค่าของ Shapley Value (SHAP Value) ซึ่งถูกใช้อย่างแพร่หลายในทฤษฎีเกมส์แบบร่วมมือกัน (Cooperative game theory) (S. M. Lundberg & Lee, 2017)

ภาพประกอบ 20 การทำงานของ SHAP ในการอธิบายความสำคัญของแต่ละคุณลักษณะทั้งในเชิงบวกและเชิงลบ

ที่มา: (S. Lundberg, 2018)

ทฤษฎีเกมส์เป็นทฤษฎีที่เกี่ยวข้องกับกลยุทธ์ในสถานการณ์ที่ต้องการเอาชนะผู้เล่นคนอื่นๆ โดยมีหลักการเกี่ยวกับการหาจุดที่ได้เปรียบสูงที่สุดสำหรับทำการตัดสินใจบางอย่าง ซึ่งทฤษฎีเกมส์มีจุดกำเนิดจากคณิตศาสตร์ชื่อ John von Neumann และ John Nash และนักเศรษฐศาสตร์ชื่อ Oskar Morgenstern

ทฤษฎีเกมส์จะพยายามอธิบายสถานการณ์ต่างๆ ออกมาในรูปแบบทางพจน์ของคณิตศาสตร์และทำการค้นหารูปแบบต่างๆ ที่เป็นไปได้ เมื่อผู้เล่นแต่ละคนต่างคิดและตัดสินใจอย่างมีเหตุผลเพื่อให้ตนเองได้รับรางวัลสูงที่สุด

ทฤษฎีเกมส์แบบร่วมมือกันจะถูกสมมติว่าผู้เล่นหลายๆ คนซึ่งอยู่ในกลุ่มหรือทีมเดียวกันต้องทำการคิดและตัดสินใจร่วมกัน และในบางครั้งอาจมีการแข่งกันกันภายในกลุ่มหรือการบังคับพฤติกรรมบางอย่างของผู้เล่นคนอื่นๆ เพื่อให้เกิดผลประโยชน์ที่สูงที่สุดแก่ทีม

ตัวอย่างของทฤษฎีเกมส์แบบร่วมมือกันและหลักการของ SHAP Value เช่น การค้นหาค่าใช้จ่ายในการรับประทานอาหารร่วมกันของกลุ่มเพื่อนที่มีจำนวน 3 คน (Choudhary, 2019) ได้แก่ นายหนึ่ง นายสอง และนายสาม ซึ่งแต่ละคนมีปริมาณในการทานอาหารที่แตกต่างกันไป จึงทำให้คำนวณออกมาได้ยาก โดยในการคำนวณจะมีข้อมูลเบื้องต้นมาให้ดังนี้

1. นายหนึ่งทานอาหารคนเดียว มีค่าใช้จ่าย 800 บาท
2. นายสองทานอาหารคนเดียว มีค่าใช้จ่าย 560 บาท
3. นายสามทานอาหารคนเดียว มีค่าใช้จ่าย 700 บาท
4. นายหนึ่งและนายสองทานอาหารร่วมกัน มีค่าใช้จ่าย 800 บาท
5. นายหนึ่งและนายสามทานอาหารร่วมกัน มีค่าใช้จ่าย 850 บาท
6. นายสองและนายสามทานอาหารร่วมกัน มีค่าใช้จ่าย 720 บาท
7. ทุกคนทานอาหารร่วมกัน มีค่าใช้จ่าย 900 บาท

โดยเมื่อทราบข้อมูลที่จำเป็นเบื้องต้นครบทุกสถานการณ์หรือความเป็นไปได้แล้วซึ่งรวมถึงเหตุการณ์ที่ทุกคนทานอาหารร่วมกันและเกิดค่าใช้จ่าย 900 บาท จะทำให้สามารถคำนวณเพื่อหาค่าใช้จ่ายตามสัดส่วนของแต่ละบุคคลต่อไปได้โดยใช้หลักการคำนวณของการเรียงสับเปลี่ยนลำดับ (Permutation)

ตัวอย่างของการคำนวณ เช่น เมื่อนายหนึ่งทานอาหารคนเดียว มีค่าใช้จ่าย 800 บาท เมื่อนายหนึ่งกับนายสองทานอาหารร่วมกัน มีค่าใช้จ่าย 800 บาทเช่นกัน ดังนั้นเมื่อนายหนึ่งและนายสองทานอาหารร่วมกันจะทำให้ค่าใช้จ่ายของนายสองเป็น 0 บาท และเมื่อทุกคนทานอาหารร่วมกัน มีค่าใช้จ่าย 900 บาท ดังนั้นค่าใช้จ่ายของนายสามจะเป็น 100 บาท

โดยเมื่อคำนวณครบทุกการเรียงสับเปลี่ยนลำดับแล้ว จะได้ผลของการคำนวณในแต่ละครั้งออกมาดังตารางที่ 4

ตาราง 4 แสดงการคำนวณของทฤษฎีเกมส์แบบร่วมมือกันในการคำนวณค่าอาหาร

ลำดับการเรียงสับเปลี่ยน	บุคคลหลักในการคำนวณและค่าใช้จ่าย	บุคคลรองในการคำนวณและค่าใช้จ่าย	บุคคลเติมเต็มในการคำนวณและค่าใช้จ่าย	รวม
รอบที่ 1	นายหนึ่ง 800 บาท	นายสอง 0 บาท	นายสาม 100 บาท	900 บาท

ตาราง 4 (ต่อ)

ลำดับการเรียงสับเปลี่ยน	บุคคลหลักในการคำนวณและค่าใช้จ่าย	บุคคลรองในการคำนวณและค่าใช้จ่าย	บุคคลเติมเต็มในการคำนวณและค่าใช้จ่าย	รวม
รอบที่ 2	นายหนึ่ง 800 บาท	นายสาม 50 บาท	นายสอง 50 บาท	900 บาท
รอบที่ 3	นายสอง 560 บาท	นายหนึ่ง 240 บาท	นายสาม 100 บาท	900 บาท
รอบที่ 4	นายสอง 560 บาท	นายสาม 160 บาท	นายหนึ่ง 180 บาท	900 บาท
รอบที่ 5	นายสาม 700 บาท	นายหนึ่ง 150 บาท	นายสอง 50 บาท	900 บาท
รอบที่ 6	นายสาม 700 บาท	นายสอง 20 บาท	นายหนึ่ง 180 บาท	900 บาท

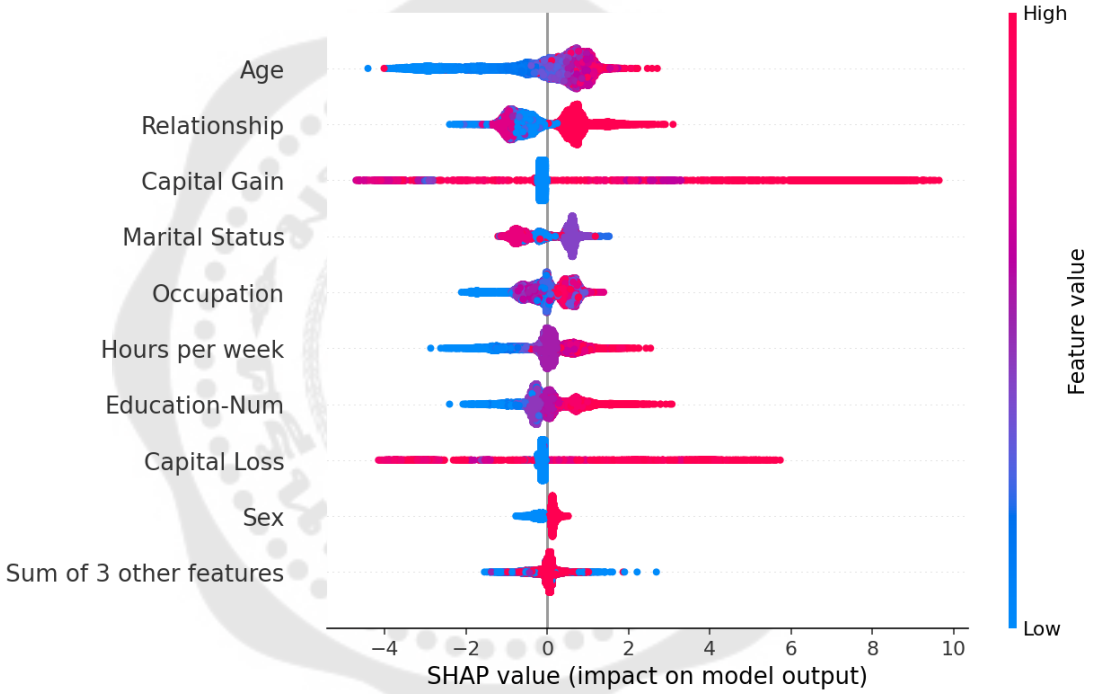
เมื่อได้ข้อมูลครบถ้วนตามตารางที่ 4 จะสามารถนำมาคำนวณ SHAP Value หรือค่าใช้จ่ายเฉลี่ยตามสัดส่วนของแต่ละคนได้ เช่น ค่าใช้จ่ายเฉลี่ยตามสัดส่วนของนายหนึ่งคำนวณได้จาก

$$\text{Shapley Value} = \frac{(800 + 800 + 240 + 180 + 150 + 180)}{6} = 391.67$$

ในทางเดียวกันจะสามารถคำนวณค่าใช้จ่ายเฉลี่ยตามสัดส่วนของนายสองและนายสามได้เป็น 206.67 และ 301.67 บาท ตามลำดับ ซึ่งเมื่อนำค่าใช้จ่ายเฉลี่ยตามสัดส่วนมารวมกันจะได้เป็น 900 บาท

จากตัวอย่างข้างต้นจะเห็นได้ว่าทฤษฎีเกมส์แบบร่วมมือกันและการคำนวณ SHAP Value สามารถนำมาประยุกต์ใช้ในการอธิบายแบบจำลองจากการเรียนรู้ด้วยเครื่องได้ โดย SHAP

Value จะสามารถเปรียบเทียบได้กับความสำคัญหรือความมีอิทธิพลของคุณลักษณะต่อการจำแนกประเภทหรือการทำนายของแบบจำลอง ซึ่งเพื่อนแต่ละคนในกลุ่มสามารถเปรียบได้กับคุณลักษณะแต่ละตัว และการทานอาหารคนเดียวหรือการจับคู่ทานอาหารไปจนถึงการทานอาหารร่วมกันทุกคนสามารถเปรียบได้กับในการเรียนรู้ของแบบจำลองแต่ละรอบที่สามารถปรับเปลี่ยนชุดของคุณลักษณะที่ใช้ในการสร้างแบบจำลองได้ และค่าใช้จ่ายที่เกิดขึ้นในแต่ละสถานการณ์สามารถเปรียบได้กับประสิทธิภาพของแบบจำลองในแต่ละการใช้ชุดคุณลักษณะย่อย ซึ่งเมื่อทำการคำนวณเป็นค่าเฉลี่ยออกมาแล้วจะทำให้สามารถทราบถึงความสำคัญหรือความมีอิทธิพลของแต่ละคุณลักษณะในแบบจำลองนั้นๆ ได้



ภาพประกอบ 21 แสดงค่าของความสัมพันธ์ระหว่างค่าของคุณลักษณะ และ SHAP Value ทั้งในทางเชิงบวกและเชิงลบ

ที่มา: (S. Lundberg, 2018)

2.8 **งานวิจัยที่เกี่ยวข้อง** (Literature Review)

ในการวิจัยครั้งนี้ได้มีการศึกษาค้นคว้าและทบทวนวรรณกรรมงานวิจัยอื่นๆ ที่มีความเกี่ยวข้องและเป็นประโยชน์กับงานวิจัย ดังต่อไปนี้

2.8.1 **บทความวิจัยเรื่อง** Application of interpretable machine learning for early prediction of prognosis in acute kidney injury

โดย Chang Hua, Qing Tan, Qinran Zhang, Yiming Li, Fengyun Wang, Xiufen Zou และ Zhiyong Peng (Hu et al., 2022)

ในงานวิจัยนี้ผู้วิจัยต้องการสร้างแบบจำลองในการทำนายการเสียชีวิตของผู้ป่วยด้วยโรคไตวายเฉียบพลัน (Acute Kidney Injury : AKI) ที่มีการเข้ารักษาตัวในห้องผู้ป่วยวิกฤต (Intensive Care Unit : ICU) และมุ่งเน้นในการพัฒนาการอธิบายแบบจำลองมีความน่าเชื่อถือและสามารถเข้าใจได้ เพื่อให้บุคลากรทางการแพทย์สามารถนำผลลัพธ์ไปใช้ประกอบการตัดสินใจในการดำเนินการรักษากับผู้ป่วยได้มีประสิทธิภาพ

ปัญหาสำคัญที่นำมาสู่การวิจัยในครั้งนี้เนื่องจากว่า จากการเก็บข้อมูลของผู้ป่วยที่ต้องเข้ารับการรักษาในห้องผู้ป่วยวิกฤติพบว่ามีจำนวนผู้ป่วยจากโรคไตวายเฉียบพลันมากกว่า 50% และการรักษาเบื้องต้นมักไม่ได้ประสิทธิภาพเท่าที่ควร นอกจากนี้ยังประกอบกับการขาดแคลนบุคลากรในการให้การรักษาแก่ผู้ป่วย ทำให้การเรียนรู้ด้วยเครื่องเข้ามามีบทบาทในการช่วยคาดการณ์ความรุนแรงของโรคหรืออัตราการเสียชีวิตของผู้ป่วย และแม้ว่าแบบจำลองในงานวิจัยอื่นๆ จะมีประสิทธิภาพเป็นที่น่าพอใจแต่ยังไม่มีงานวิจัยใดที่มีการอธิบายการทำงานของแบบจำลอง ซึ่งก่อให้เกิดความไม่น่าเชื่อถือสำหรับนำไปใช้งานประกอบการรักษาผู้ป่วย

ชุดข้อมูลที่นำมาใช้ในการวิจัยมีการนำมาจาก Medical Information Mart for Intensive Care IV (MIMIC-IV) ซึ่งมีการเก็บข้อมูลของผู้ป่วยด้วยโรคไตวายเฉียบพลันของประเทศสหรัฐอเมริกาในระหว่างปี พ.ศ. 2551 ถึงปี พ.ศ. 2562 (ค.ศ. 2008 - 2019) ซึ่งได้รับการรับรองทางด้านจริยธรรมจาก Massachusetts Institute of Technology (Cambridge, MA) และ Beth Israel Deaconess Medical Center (Boston, MA) แล้ว โดยชุดข้อมูลประกอบด้วยตัวอย่างข้อมูลมากกว่า 33,000 ตัวอย่างข้อมูล และคุณลักษณะของข้อมูลอีกจำนวนมาก เช่น ข้อมูลของผู้ป่วย ข้อมูลโรคประจำตัว ข้อมูลจากห้องปฏิบัติการ เป็นต้น

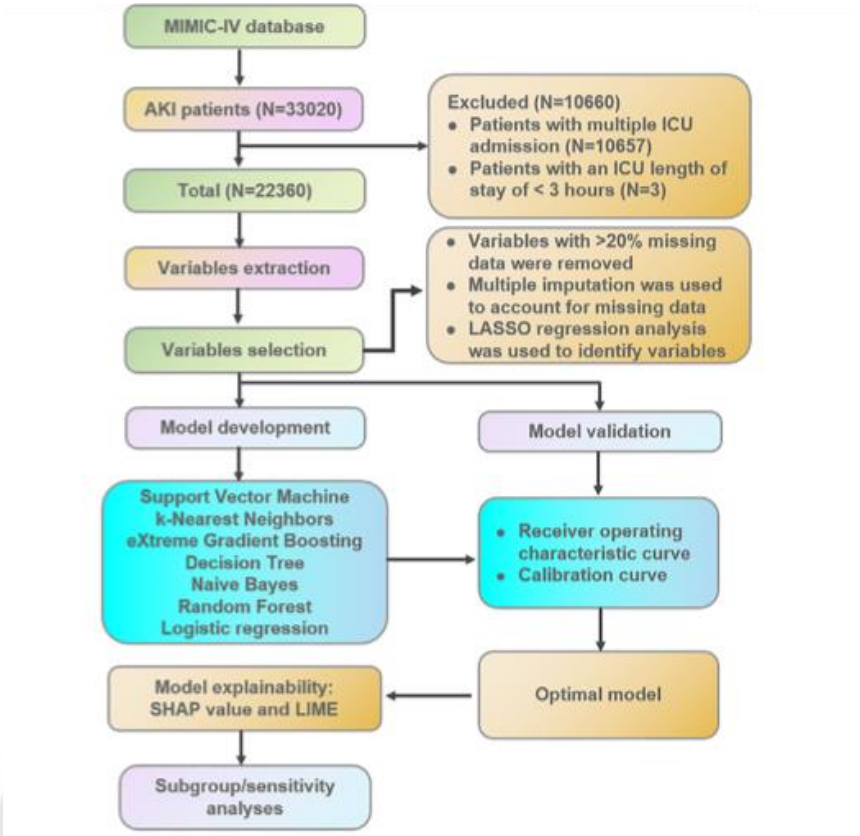
ประชากรที่ใช้ในการวิจัยประกอบด้วยผู้ป่วยที่มีอายุตั้งแต่ 18 ปีขึ้นไป และมีการเข้ารักษาตัวที่ห้องผู้ป่วยวิกฤติตั้งแต่ 3 ชั่วโมงขึ้นไป และเป็นการเข้ารักษาในห้องผู้ป่วยวิกฤติครั้งแรกในการนอนโรงพยาบาลครั้งนี้ของผู้ป่วยรายนั้นๆ ซึ่งมีจำนวนข้อมูลที่เข้าเงื่อนไขอยู่ที่ประมาณ 22,000 ตัวอย่าง โดยคุณลักษณะที่มีข้อมูลที่เป็นค่าว่างสูงกว่า 20% จะถูกคัดออก และมีการคัดเลือกคุณลักษณะเพื่อนำไปใช้สร้างแบบจำลองด้วยวิธีการแบบ ‘LASSO’ ซึ่งจะทำเหลือคุณลักษณะที่มีความสำคัญในการใช้สร้างแบบจำลองอยู่ที่ 29 คุณลักษณะ

จากการสำรวจข้อมูลเบื้องต้นพบว่าอายุเฉลี่ยของผู้ป่วยอยู่ที่ 69.5 ปี และเป็นสัดส่วนของผู้ป่วยหญิงอยู่ที่ 42.8% เพศชาย 57.2% โรคประจำตัวที่พบสูงสุดสามอันดับ ได้แก่ โรคความดันโลหิตสูง 41.3% โรคเบาหวาน 32.5% และโรคภาวะหัวใจล้มเหลว 30.5%

ในการวิจัยได้มีการใช้งานแบบจำลองของการเรียนรู้ด้วยเครื่องแบบมีผู้สอนเพื่อทำนายการเสียชีวิตของผู้ป่วยทั้งหมด 7 แบบจำลอง ประกอบด้วย Support Vector Machine (SVM), K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGBoost), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF) และ Logistic Regression (LR) โดยการวัดประสิทธิภาพของแบบจำลองมีการใช้งานความแม่นยำ (Accuracy), พื้นที่ใต้กราฟ (Area Under the Receiving Operating Characteristic Curve : AUC-ROC), ค่าความอ่อนไหว (Sensitivity) และค่าความจำเพาะ (Specificity)

การอธิบายแบบจำลองจะใช้สำหรับอธิบายแบบจำลองที่มีประสิทธิภาพสูงที่สุด โดยใช้วิธีการอธิบายแบบจำลองด้วย SHAP ในการค้นหาว่าคุณลักษณะใดในข้อมูลที่ส่งผลหรือมีอิทธิพลต่อความเสี่ยงในการเสียชีวิตของผู้ป่วย โดยจะเป็นการอธิบายในระดับแบบจำลอง จากนั้นมีการใช้งานวิธีการแบบ LIME ในการอธิบายผลของการทำนายการเสียชีวิตในผู้ป่วยระดับรายบุคคล

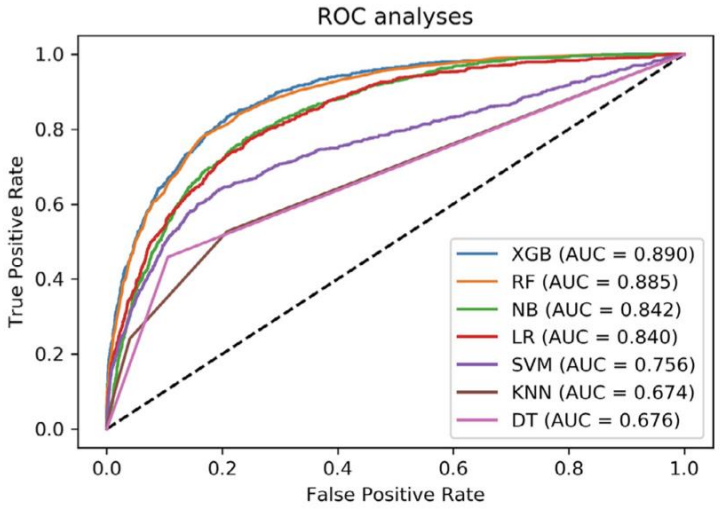
ขั้นตอนในการดำเนินการวิจัยเป็นไปดังแสดงในภาพประกอบที่ 22



ภาพประกอบ 22 ขั้นตอนในการดำเนินงานวิจัย Application of interpretable machine learning for early prediction of prognosis in acute kidney injury

ที่มา: (Hu et al., 2022)

ผลของการวิจัยพบว่าแบบจำลอง XGBoost ให้ประสิทธิภาพที่ดีที่สุด โดยมีค่าความแม่นยำอยู่ที่ 89% ซึ่งมีประสิทธิภาพสูงกว่างานวิจัยอื่นๆ ในอดีต ส่วนแบบจำลองอื่นๆ มีค่าความแม่นยำตามลำดับดังต่อไปนี้ RF 88.5%, NB 84.2%, LR 84%, SVM 75.6%, KNN 67.4%, DT67.6%

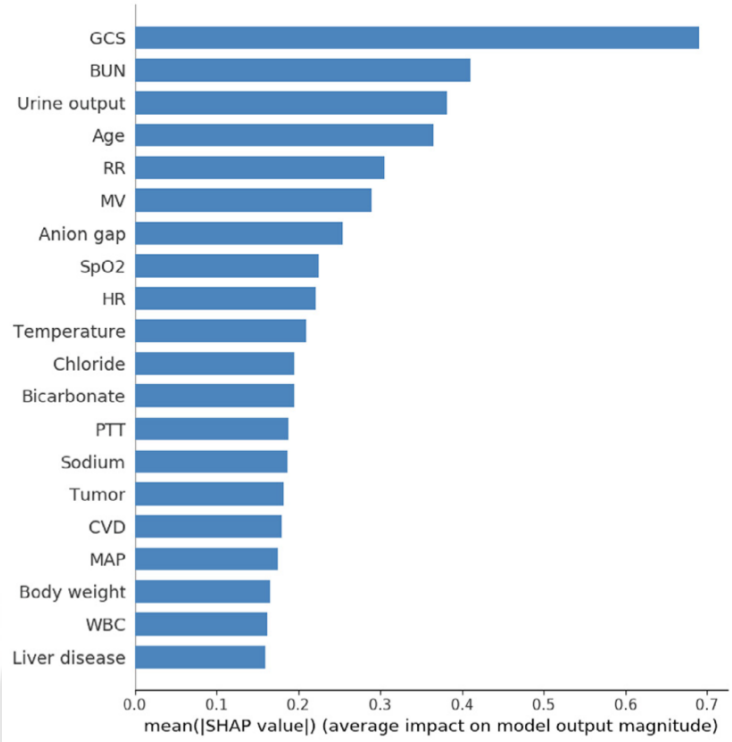


ภาพประกอบ 23 แสดงประสิทธิภาพของแต่ละแบบจำลองในงานวิจัยโดยใช้พื้นที่ใต้กราฟ ROC

ที่มา: (Hu et al., 2022)

ในการอธิบายแบบจำลองได้นำแบบจำลอง XGBoost ซึ่งมีประสิทธิภาพสูงที่สุดมาทำการอธิบาย โดยใช้วิธีการแบบ SHAP ในการค้นหาคุณลักษณะที่มีความสำคัญในการทำนายของแบบจำลอง ซึ่งคุณลักษณะที่มีความสำคัญสูงที่สุดสี่อันดับแรกประกอบด้วย

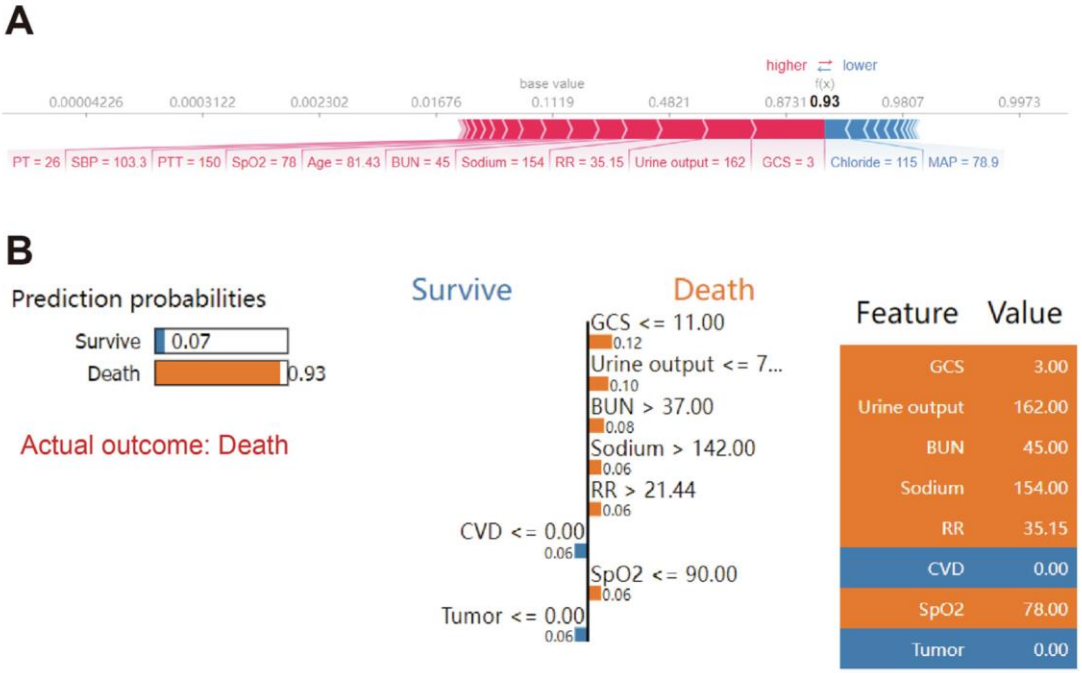
1. Glasgow Coma Scale (GCS): ระดับความรู้สึกตัว (3 วิกฤต - 15 เบา)
2. Blood Urea Nitrogen (BUN): ปริมาณไนโตรเจนจากยูเรียที่รั่วมาในกระแสเลือด
3. Cumulative Urine Output on Day 1: ปริมาณปัสสาวะวันแรกของการเข้า ICU
4. Age: อายุ



ภาพประกอบ 24 แสดงความสำคัญของคุณลักษณะที่ได้จากการอธิบายแบบจำลองด้วย SHAP

ที่มา: (Hu et al., 2022)

จากนั้นใช้วิธีการแบบ SHAP Force Analysis และ LIME เพื่อใช้ในการอธิบายผู้ป่วยรายบุคคล โดยจะพบว่าในการอธิบายแบบรายบุคคลระหว่างวิธี SHAP Force Analysis และ LIME จะให้ผลลัพธ์ที่ต่างกันเล็กน้อยในการระบุคุณลักษณะที่มีความสำคัญต่อการทำนาย แต่อย่างไรก็ตามผลลัพธ์จากการทำนายและคุณลักษณะที่มีความสำคัญโดยรวมจะเป็นไปในทิศทางที่สอดคล้องกัน



ภาพประกอบ 25 เปรียบเทียบการอธิบายแบบจำลองระหว่างวิธีการแบบ SHAP และ LIME ในระดับรายบุคคลสำหรับการทำนายโอกาสในการเสียชีวิตของผู้ป่วยโรคไตวายเฉียบพลัน

ที่มา: (Hu et al., 2022)

2.8.2 บทความวิจัยเรื่อง Diagnosis of Parkinson's disease based on SHAP value feature selection

โดย Yuchun Liu, Zhihui Liu, Xue Luo, Hongjingtian Zhao (Liu, Liu, Luo, & Zhao, 2022)

ในงานวิจัยนี้ผู้วิจัยต้องการสร้างแบบจำลองที่ใช้ในการทำนายผู้ที่มีโอกาสป่วยเป็นโรคของระบบประสาทแบบมีอาการสั่นของอวัยวะ หรือพาร์กินสัน (Pakinson's Disease) โดยมีการใช้งานการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธี SHAP มาช่วยในการคัดเลือกคุณลักษณะสำหรับการสร้างแบบจำลอง

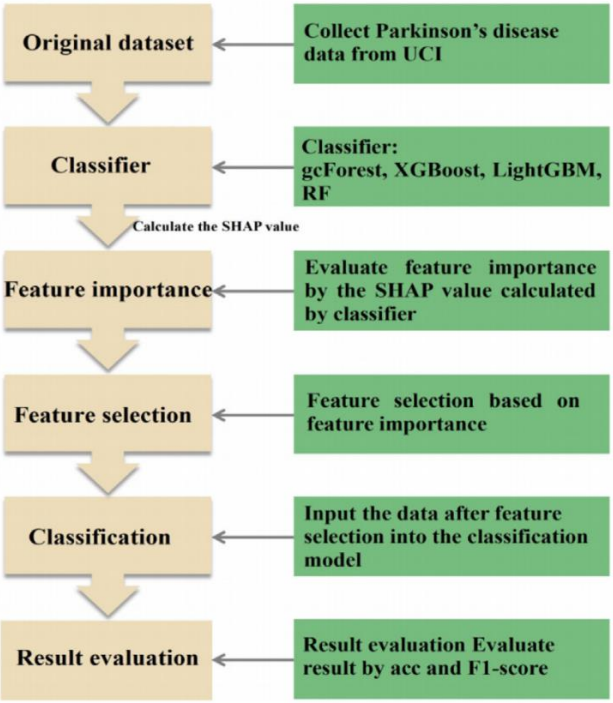
เนื่องจากโรคพาร์กินสันถือเป็นอาการป่วยที่ส่งผลกระทบต่อการใช้ชีวตประจำวันของผู้ป่วยซึ่งมีอยู่ประมาณ 7-10 ล้านคนทั่วโลก และมีแนวโน้มที่จะพบผู้ป่วยด้วยโรคนี้เพิ่มขึ้นในทุกๆ ปี โดยส่วนใหญ่มักจะเกิดกับผู้สูงอายุตั้งแต่ 60 ปีขึ้นไป ซึ่งการวินิจฉัยด้วยแพทย์แบบดั้งเดิมต้องเสียเวลาจำนวนมากในการวินิจฉัยเพื่อบ่งชี้การป่วยเป็นโรค ในปัจจุบันเริ่มมีงานวิจัยที่เกี่ยวข้องกับการบ่งชี้อาการป่วยด้วยโรคพาร์กินสันด้วยการเรียนรู้ด้วยเครื่อง แต่งานวิจัยส่วนใหญ่มักจะมุ่งเน้น

ไปยังการทำให้แบบจำลองมีประสิทธิภาพในการทำนายที่แม่นยำทำให้มีการใช้คุณลักษณะจำนวนมากซึ่งอาจจะเป็นข้อมูลรบกวนหรือข้อมูลที่ซ้ำซ้อนไปใช้ในการสร้างแบบจำลอง ดังนั้นผู้วิจัยจึงต้องการนำการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธีการแบบ SHAP มาใช้งานเพื่อช่วยลดมิติของข้อมูลเพื่อให้แบบจำลองมีขนาดที่เล็กลง ยืดหยุ่นมากขึ้น และสามารถเรียนรู้ได้รวดเร็วโดยยังมีประสิทธิภาพที่สูงหรือสูงขึ้น

ในงานวิจัยที่เกี่ยวกับชีวการแพทย์ (Biomedical) มีการนำการเรียนรู้ด้วยเครื่องแบบอธิบายได้มาใช้งานในการอธิบายแบบจำลองอยู่บ้าง แต่ในงานวิจัยซึ่งเกี่ยวข้องกับโรคพาร์กินสันยังไม่ปรากฏการนำเสนอการใช้งาน SHAP เพื่อนำมาใช้คัดเลือกคุณลักษณะสำหรับใช้ในการสร้างแบบจำลอง โดยในงานวิจัยนี้มีการใช้วิธีการแบบ SHAP ร่วมกับแบบจำลอง gcForest, XGBoost, LightGBM และ Random Forest ในการทำนายผู้ป่วยที่เป็นโรคพาร์กินสัน โดยข้อมูลที่นำมาใช้งานวิจัยมาจากฐานข้อมูลสาธารณะของ University of California, Irvine (UCI) ซึ่งเป็นข้อมูลที่เกี่ยวกับลักษณะเสียงของผู้ป่วยด้วยโรคพาร์กินสันเก็บรวมรวมโดย Department of Neurology at Istanbul University Medical School

ในการคัดเลือกคุณลักษณะของข้อมูลที่จะนำไปใช้งาน ผู้วิจัยมีการใช้เทคนิคที่หลากหลายประกอบด้วย F-Score, Anova-F, Mutual Information (MI) และ SHAP ส่วนการประเมินผลประสิทธิภาพมีการใช้งานค่าความแม่นยำ (Accuracy) และค่า F1-Score

โดยขั้นตอนในการนำ SHAP มาใช้ในการคัดเลือกคุณลักษณะจะสามารถแบ่งย่อยได้เป็นสามขั้นตอนคือ ขั้นตอนแรกเป็นการสร้างแบบจำลองโดยไม่มีการคัดเลือกคุณลักษณะ ขั้นตอนที่สองจะเป็นการนำ SHAP มาคำนวณค่าความสำคัญหรือความมีอิทธิพลของแต่ละคุณลักษณะ แล้วทำการเรียงลำดับออกมาตามความสำคัญ ในขั้นตอนสุดท้ายจะเป็นการสร้างแบบจำลองใหม่อีกครั้งโดยใช้งานคุณลักษณะที่มีค่าความสำคัญสูงที่ได้จากการคำนวณโดย SHAP

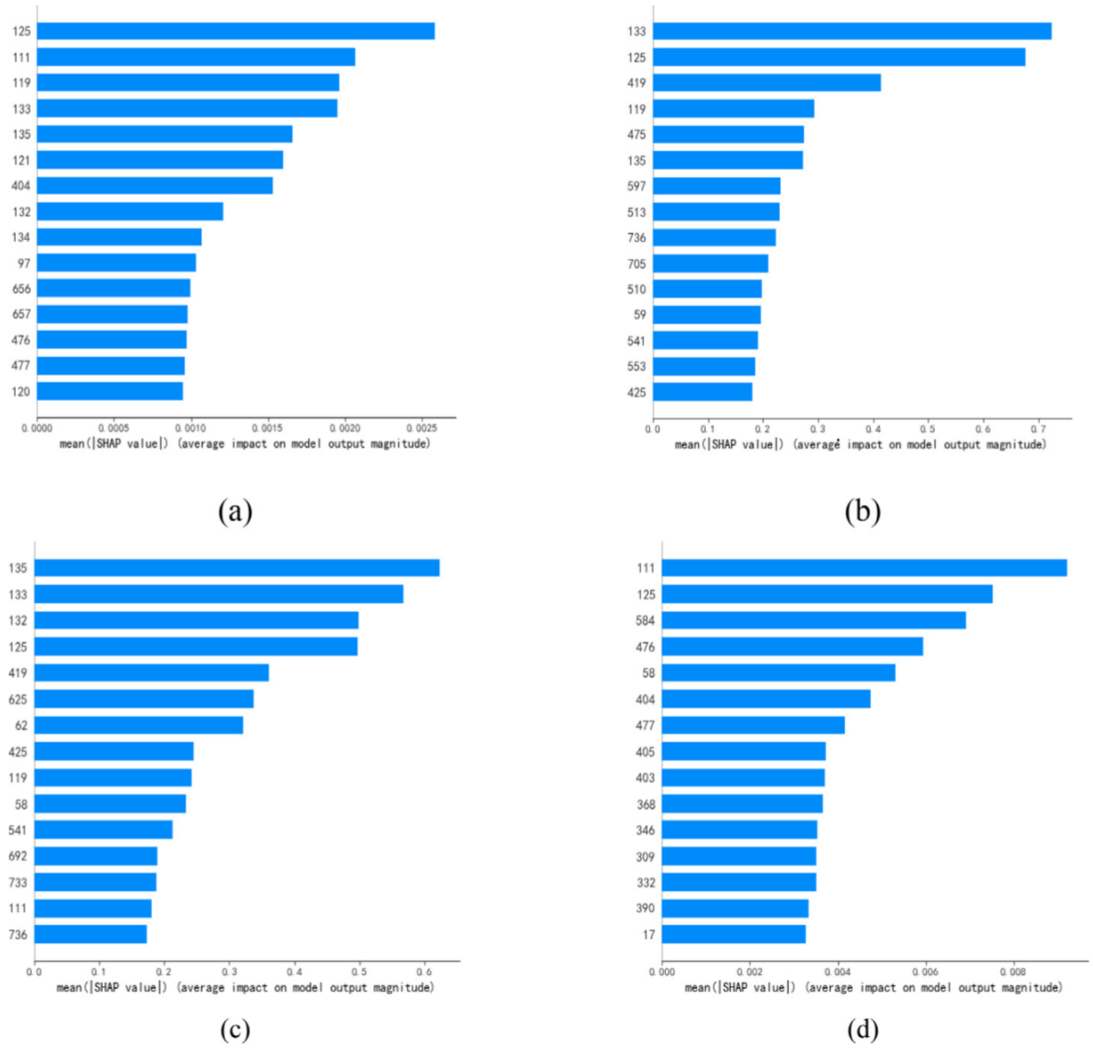


ภาพประกอบ 26 ขั้นตอนในการดำเนินงานวิจัย Diagnosis of Parkinson's disease based on SHAP value feature selection

ที่มา: (Liu et al., 2022)

จากการสำรวจข้อมูลเบื้องต้นชุดข้อมูลประกอบด้วยผู้ที่มารับการตรวจวินิจฉัยจำนวน 252 คน เป็นเพศชาย 130 คนและเพศหญิง 122 คน โดยประกอบด้วยผู้ป่วยเป็นโรคพาร์กินสันจำนวน 188 คน เป็นเพศชาย 107 คน และเพศหญิง 81 คน อายุเฉลี่ยของกลุ่มผู้ป่วยอยู่ที่ 65.1 ปี และกลุ่มผู้ที่ไม่ป่วยอยู่ที่ 61.1 ปี โดยชุดข้อมูลมีจำนวนคุณลักษณะที่จัดเก็บสูงถึง 753 คุณลักษณะ

ภาพประกอบที่ 27 แสดงถึงคุณลักษณะที่มีความสำคัญมากที่สุดจำนวน 15 อันดับแรก จากการใช้งาน SHAP ร่วมกับแบบจำลองทั้งสี่แบบ ได้แก่ gcForest (a), XGBoost (b), LightGBM (c) และ Random Forest (d) ซึ่งจะสังเกตเห็นว่าคุณลักษณะที่ 125 มีความสำคัญอยู่ในอันดับต้นๆ ของทั้งสี่แบบจำลอง ส่วนคุณลักษณะที่ 111, 119, 133 และ 135 ก็ปรากฏอยู่ในสามแบบจำลอง ซึ่งทำให้สามารถตั้งสมมติฐานได้ว่าคุณลักษณะเหล่านี้มีความสำคัญสูงต่อการทำนายผลลัพธ์

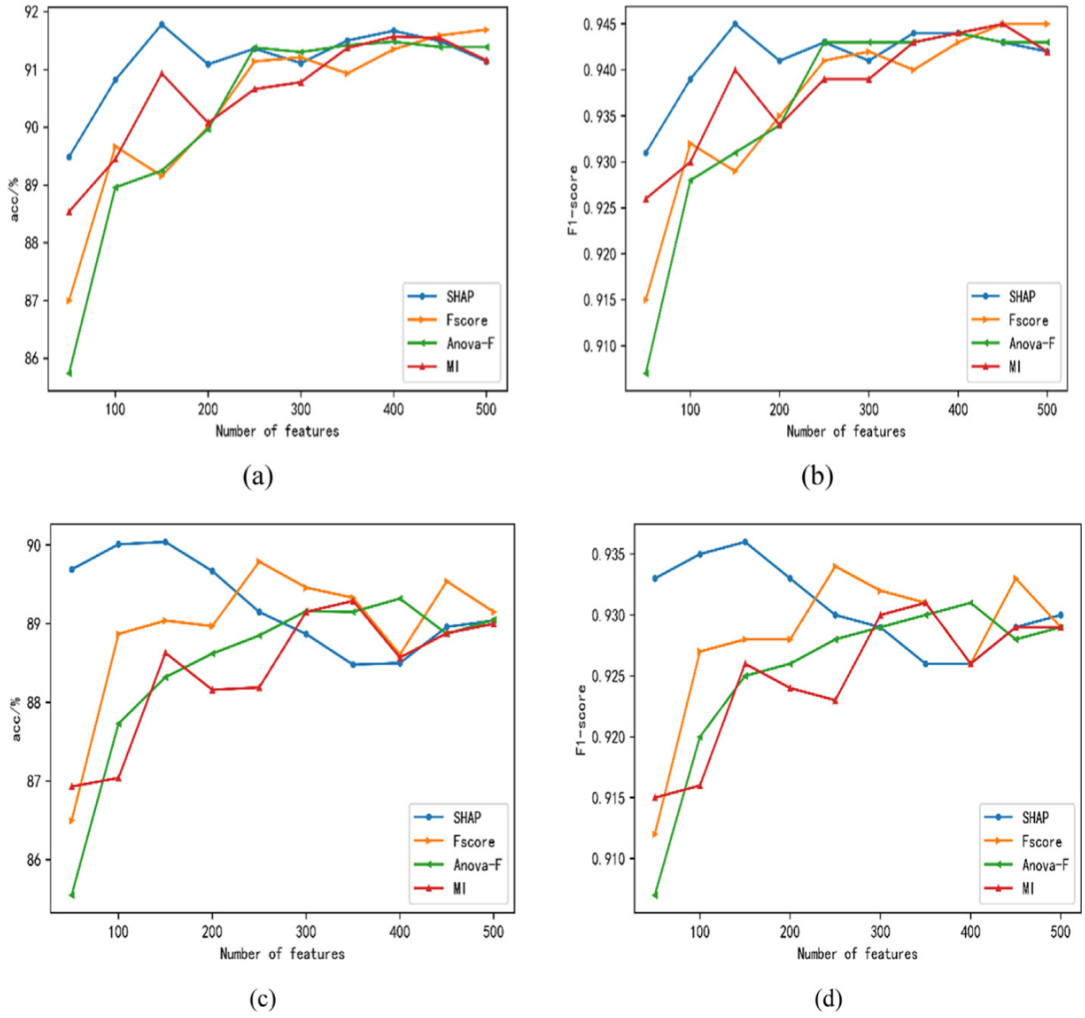


ภาพประกอบ 27 แสดงความสำคัญของคุณลักษณะที่ได้จากการอธิบายแบบจำลองต่างๆ ด้วย SHAP

ที่มา: (Liu et al., 2022)

เมื่อนำการใช้งานวิธีการคัดเลือกคุณลักษณะด้วยวิธีต่างๆ จะพบว่าวิธีการแบบ SHAP สามารถทำให้แบบจำลองมีประสิทธิภาพสูงที่สุดโดยใช้จำนวนของคุณลักษณะน้อยที่สุด เมื่อเทียบกับวิธีการแบบอื่นๆ จากภาพประกอบที่ 28 จะแสดงให้เห็นถึงค่าของความแม่นยำและ F1-Score ของการใช้งานวิธีการคัดเลือกคุณลักษณะแบบต่างๆ ร่วมกับแบบจำลอง gcForest (a และ b) ซึ่งจะเห็นได้ว่าการคัดเลือกคุณลักษณะแบบ SHAP ทำให้สามารถใช้งานคุณลักษณะเพียง 150 คุณลักษณะ ก็สามารถทำให้แบบจำลองมีประสิทธิภาพที่สูงที่สุดได้ ซึ่งจะส่งผลให้

แบบจำลองมีความซับซ้อนน้อยลงด้วย เช่นเดียวกับแบบจำลอง XGBoost (c และ d) ซึ่งมีผลลัพธ์ไปในทิศทางเดียวกัน



ภาพประกอบ 28 แสดงถึงการประเมินประสิทธิภาพความแม่นยำและ F1-Score ของแบบจำลองต่างๆ ในการใช้งานการคัดเลือกคุณลักษณะที่แตกต่างกัน

ที่มา: (Liu et al., 2022)

ในการประเมินประสิทธิภาพของแบบจำลองและเปรียบเทียบผลของการทดลอง ผู้วิจัยพบว่าแบบจำลองที่ให้ประสิทธิภาพของความแม่นยำและค่า F1-Score ที่สูงที่สุดคือแบบจำลอง gcForest เมื่อมีการใช้งานร่วมกับการคัดเลือกคุณลักษณะแบบ SHAP โดยเลือกใช้งานคุณลักษณะจำนวนทั้งสิ้น 150 คุณลักษณะ ซึ่งจะให้ค่าของความแม่นยำอยู่ที่ 91.78% และให้ค่าของ F1-Score อยู่ที่ 94.5% ซึ่งเมื่อเปรียบเทียบผลของการทดลองกับงานวิจัยอื่นๆ ก่อนหน้า

จะพบว่าเมื่อเปรียบเทียบกับงานวิจัยอื่นๆ ที่ไม่ได้มีการจัดการความไม่สมดุลกันของข้อมูล งานวิจัยนี้สามารถสร้างแบบจำลองได้มีประสิทธิภาพที่สูงกว่าแบบจำลองในงานวิจัยอื่นๆ ซึ่งมีความแม่นยำสูงที่สุดอยู่ที่ 91.6% ในขณะที่งานวิจัยที่มีการจัดการกับความไม่สมดุลกันของข้อมูลสามารถทำให้ประสิทธิภาพของแบบจำลองสูงขึ้นได้สูงสุดอยู่ที่ 94.89% และ F1-Score อยู่ที่ 94.9%

2.8.3 **บทความวิจัยเรื่อง** A data modeling approach for classification problems: application to bank telemarketing prediction

โดย Stéphane Cédric Koumetio, Walid Cherif และ Hassan Silkan (Tekouabou, Cherif, & Silkan, 2019)

ในงานวิจัยนี้ผู้วิจัยต้องการนำเสนอวิธีจัดการกับข้อมูลเบื้องต้นก่อนการนำข้อมูลไปใช้ในการสร้างแบบจำลองจากการเรียนรู้ด้วยเครื่องเพื่อใช้ในการทำนายการสมัครผลิตภัณฑ์เงินฝากประจำของลูกค้าธนาคารผ่านการนำเสนอทางโทรศัพท์ ซึ่งเป็นข้อมูลของธนาคารแห่งหนึ่งในประเทศโปรตุเกสซึ่งเก็บรวบรวมไว้ตั้งแต่ปี พ.ศ. 2551 ถึงปี พ.ศ. 2556 (ค.ศ. 2008 – 2013)

ในการโฆษณาเพื่อเผยแพร่ผลิตภัณฑ์ต่างๆ ของภาคธุรกิจหรืออุตสาหกรรมได้มีการพัฒนาต่อเนื่องมาอย่างยาวนานตั้งแต่ยุคสมัยที่เป็นการโฆษณาแบบทั่วไปที่ไม่ได้คำนึงกลุ่มของลูกค้าซึ่งมักจะไม่ได้ประสิทธิภาพเป็นที่น่าพอใจนัก จนมาถึงยุคสมัยที่เริ่มมีการโฆษณาโดยมุ่งเน้นกลุ่มลูกค้าที่มีความน่าจะเป็นที่จะสนใจในผลิตภัณฑ์ที่จะนำเสนอ ไม่ว่าจะเป็นการโฆษณาผ่านทางโทรศัพท์ การขายตรง หรือทางจดหมายอิเล็กทรอนิกส์ (Email) ซึ่งต้องมีการเก็บรวบรวมข้อมูลที่มากมายไม่ว่าจะเป็นข้อมูลส่วนตัวของลูกค้า ข้อมูลการติดต่อกับลูกค้า ข้อมูลของผลิตภัณฑ์ และข้อมูลอื่นๆ เพื่อนำมาใช้ในการวางแผนกลยุทธ์ในการนำเสนอผลิตภัณฑ์ ทำให้ได้ผลลัพธ์ที่ค่อนข้างมีประสิทธิภาพ

ด้วยความสำเร็จของการโฆษณาสินค้าโดยระบุกลุ่มเป้าหมาย จึงมีการนำการเรียนรู้ด้วยเครื่องเพื่อเข้ามาใช้งานในการช่วยเพิ่มประสิทธิภาพหรือแก้ไขปัญหาต่างๆ โดยสามารถนำมาใช้ในการทำนายล่วงหน้าได้ว่าลูกค้าคนใดจะทำการซื้อหรือสมัครผลิตภัณฑ์ที่มีการนำเสนอจากการใช้ข้อมูลที่เก็บรวบรวมไว้ โดยมีงานวิจัยอื่นๆ ก่อนหน้าที่มีการสร้างแบบจำลองสำหรับทำนายโอกาสที่ลูกค้าจะทำการตอบรับสมัครผลิตภัณฑ์เงินฝากประจำด้วยการนำเสนอผลิตภัณฑ์ผ่านทางโทรศัพท์ แต่เมื่อสำรวจงานวิจัยเหล่านั้นจะพบว่าแต่ละงานวิจัยจะมีการเลือกใช้คุณลักษณะที่ค่อนข้างแตกต่างกันไปในการสร้างแบบจำลอง จึงเป็นที่มาของงานวิจัยนี้ที่ต้องการนำเสนอวิธีการจัดการข้อมูลเบื้องต้นรูปแบบใหม่กับชุดข้อมูลนี้เพื่อคัดเลือกคุณลักษณะที่มีความสำคัญอย่างแท้จริงก่อนการนำไปใช้งาน

ข้อมูลที่นำมาใช้งานจะเป็นข้อมูลเกี่ยวกับการนำเสนอขายผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์ของธนาคารแห่งหนึ่งในประเทศโปรตุเกส โดยชุดข้อมูลต้นฉบับจะมีคุณลักษณะของข้อมูลสูงถึง 150 คุณลักษณะ แต่ในชุดข้อมูลที่มีการเปิดเผยเป็นสาธารณะจะมีคุณลักษณะในชุดข้อมูลอยู่ที่ 22 คุณลักษณะ ซึ่งชุดข้อมูลได้นำมาจากฐานข้อมูลสาธารณะของ University of California, Irvine (UCI) ในชุดข้อมูลประกอบด้วยตัวอย่างข้อมูลจำนวน 41,188 ตัวอย่าง โดยผลลัพธ์หรือป้ายกำกับแบ่งออกเป็นสองกลุ่ม ได้แก่ สมัครผลิตภัณฑ์ (yes) และไม่สมัครผลิตภัณฑ์ (no) และคุณลักษณะมีทั้งข้อมูลชนิดตัวเลขและข้อมูลชนิดประเภท

ในการดำเนินงานวิจัยผู้วิจัยได้แบ่งออกเป็นสามขั้นตอนประกอบด้วย ขั้นตอนการจัดการข้อมูลเบื้องต้น (Preprocessing) ขั้นตอนการจัดการข้อมูลให้เป็นมาตรฐาน (Normalization) และขั้นตอนในการสร้างแบบจำลองในการทำนาย (Classification) โดยในขั้นตอนการจัดการข้อมูลเบื้องต้นมีการนำเสนอเทคนิคที่ใช้จัดการกับข้อมูลชนิดประเภทแบบไม่เรียงลำดับ เช่น อาชีพหรือสถานภาพสมรส โดยจะอ้างอิงจากสัดส่วนของค่าในคุณลักษณะเปรียบเทียบกับจำนวนของกลุ่มข้อมูลเพื่อระบุค่า 0 หรือ 1 ตัวอย่างเช่น การจัดการคุณลักษณะสถานภาพสมรสที่มีค่าเป็น ‘married’ โดยมีชุดข้อมูลซึ่งประกอบด้วยกลุ่ม ‘no’ จำนวน 3 ตัวอย่าง ซึ่งมีค่า ‘married’ อยู่ 2 ตัวอย่าง และกลุ่ม ‘yes’ จำนวน 2 ตัวอย่าง ซึ่งมีข้อมูลที่มีค่า ‘married’ อยู่ 1 ตัวอย่าง ดังตารางที่ 5

ตาราง 5 ตัวอย่างข้อมูลในการจัดการสถานภาพสมรสซึ่งเป็นคุณลักษณะชนิดประเภทแบบไม่เรียงลำดับ

สถานภาพสมรส	**จำนวนข้อมูลกลุ่ม** ‘yes’	**จำนวนข้อมูลกลุ่ม** ‘no’	**รวม**
‘married’	1	2	3
อื่นๆ	1	1	2
รวม	2	3	5

ที่มา: (Tekouabou et al., 2019)

ซึ่งการคำนวณจะสัดส่วนออกมาตามสมการที่ 4 และ 5 ซึ่งจะเห็นได้ว่าสัดส่วนของสถานภาพสมรสที่มีค่าเป็น ‘married’ ที่เป็นกลุ่ม ‘no’ มีสัดส่วนที่มากกว่า ดังนั้นจะทำการระบุค่า 1 ให้กับคุณลักษณะนี้ให้กับทุกตัวอย่างข้อมูลที่เป็นกลุ่ม ‘no’ โดยไม่สนใจว่าจะมีค่าของสถานภาพ

สมรสเป็น ‘married’ หรือไม่ และเช่นเดียวกันในทุกตัวอย่างข้อมูลที่เป็นกลุ่ม ‘yes’ จะถูกระบุค่าเป็น 0 โดยไม่สนใจว่าจะมีค่าของสถานภาพสมรสเป็น ‘married’ หรือไม่

$$\frac{\text{no(married)}}{\text{no}} = \frac{2}{3} = 0.66 \qquad (4)$$

$$\frac{\text{yes(married)}}{\text{yes}} = \frac{1}{2} = 0.5 \qquad (5)$$

และในขั้นตอนการจัดการกับข้อมูลเบื้องต้นยังมีการนำเสนอเทคนิคในการจัดการกับข้อมูลค่าว่างที่เป็นประเภทตัวเลข (Numeric) หรือตัวเลขแบบลำดับขั้น (Ordinal) โดยใช้การคำนวณค่าเฉลี่ยจากค่าทั้งหมดที่อยู่ในกลุ่มเดียวกัน ตัวอย่างเช่น ข้อมูลการศึกษาที่มีการเปลี่ยนแปลงชนิดข้อมูลจากแบบประเภทให้อยู่ในรูปแบบของตัวเลขที่มีลำดับขั้นตามตารางที่ 6 ซึ่งจะเห็นได้ว่าตัวอย่างข้อมูลที่ 4 มีข้อมูลการศึกษาที่เป็นค่าว่างอยู่ ซึ่งสามารถคำนวณเพื่อเติมข้อมูลได้จากการนำข้อมูลของกลุ่ม ‘no’ ซึ่งเป็นข้อมูลในกลุ่มเดียวกันมาใช้ในการคำนวณ ซึ่งจะมีค่าเป็น 2.5 โดยจะสามารถคำนวณได้ตามสมการที่ 6

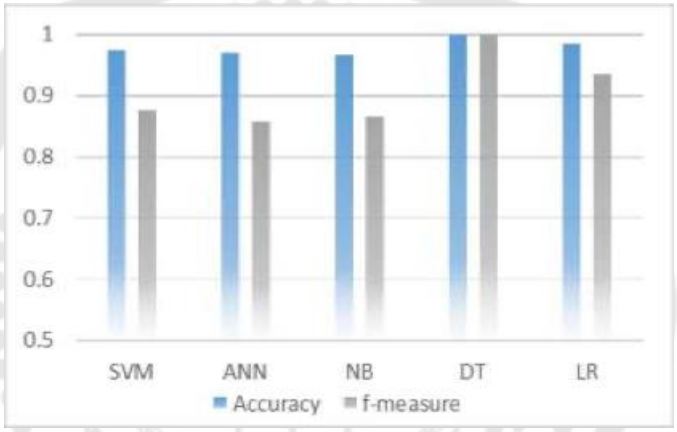
ตาราง 6 ตัวอย่างข้อมูลในการจัดการระดับการศึกษาซึ่งเป็นคุณลักษณะชนิดประเภทแบบเรียงลำดับ

หมายเลขตัวอย่างข้อมูล	คุณลักษณะระดับการศึกษา	ผลการสมัครผลิตภัณฑ์
1	3	‘no’
2	1	‘yes’
3	2	‘no’
4		‘no’
5	0	‘yes’

ที่มา: (Tekouabou et al., 2019)

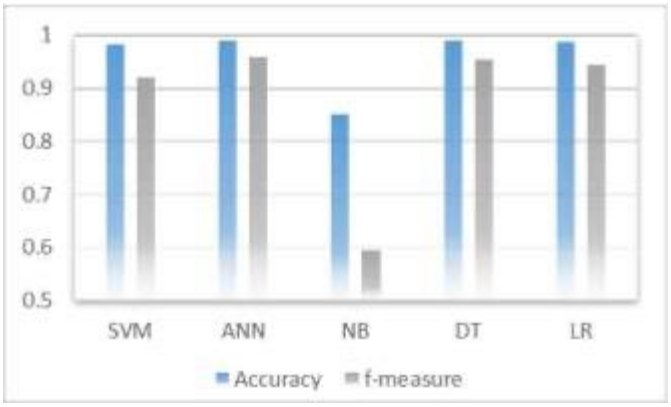
$$\frac{\sum \text{education(no)}}{\text{N(no)}} = \frac{3+2}{2} = 2.5 \qquad (6)$$

ในขั้นตอนของการสร้างแบบจำลองในการทำนายมีการใช้งานแบบจำลองทั้งหมดห้าประเภทประกอบด้วย Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Artificial Neural Network (ANN และ Support Vector Machine (SVM) โดยมีการใช้งานค่าของความแม่นยำ (Accuracy) และค่า F1-Score ในการประเมินผลประสิทธิภาพของแบบจำลอง โดยผลของการวิจัยพบว่าหากไม่มีการจัดการกับข้อมูลให้เป็นมาตรฐาน แบบจำลอง DT จะสามารถให้ค่าความแม่นยำได้สูงที่สุดที่ 100% และ LR รองลงมาที่ 98.61% แต่เมื่อมีการจัดการทำข้อมูลให้เป็นมาตรฐานจะทำให้แบบจำลอง ANN มีค่าความแม่นยำและค่า F1-Score สูงที่สุดที่ 99.07% และ 95.83% ตามลำดับ และแบบจำลอง DT จะมีค่าความแม่นยำและค่า F1-Score ตกลงมาอยู่ที่ 98.98% และ 95.45% ตามลำดับ



ภาพประกอบ 29 แสดงค่าความแม่นยำและ F1-Score ของแบบจำลองที่ไม่มีการทำข้อมูลให้เป็นมาตรฐาน

ที่มา: (Tekouabou et al., 2019)



ภาพประกอบ 30 แสดงค่าความแม่นยำและ F1-Score ของแบบจำลองที่มีการทำข้อมูลให้เป็นมาตรฐาน

ที่มา: (Tekouabou et al., 2019)

แบบจำลองที่มีประสิทธิภาพสูงที่สุดในงานวิจัยนี้มีประสิทธิภาพสูงกว่าแบบจำลองในงานวิจัยก่อนหน้าซึ่งแบบจำลองที่มีประสิทธิภาพสูงที่สุดมีค่าความแม่นยำอยู่ที่ 93.5% ซึ่งเป็นผลมาจากเทคนิคในการจัดการกับข้อมูลเบื้องต้นที่ผู้วิจัยได้ทำการนำเสนอ

บทที่ 3
วิธีการดำเนินงานวิจัย

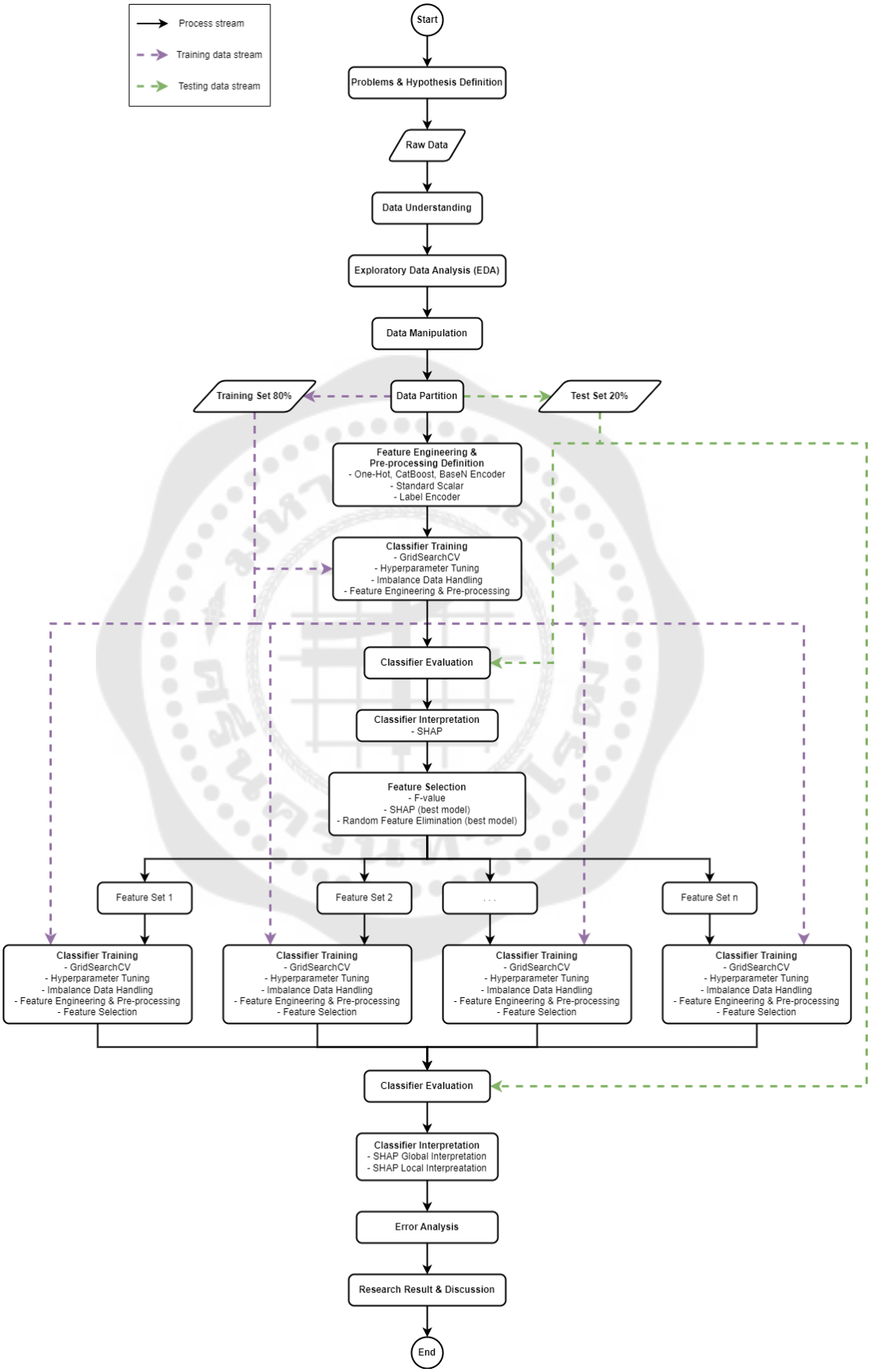
ในการวิจัยครั้งนี้ ผู้วิจัยได้มีการวางแผนขั้นตอนในการดำเนินงานวิจัยซึ่งมีรายละเอียดดังต่อไปนี้

1. การออกแบบขั้นตอนในการดำเนินงานวิจัย
2. การสำรวจและวิเคราะห์ข้อมูลเบื้องต้น
3. การจัดการกับข้อมูลเบื้องต้นและการแบ่งชุดข้อมูล
4. การทำวิศวกรรมคุณลักษณะและการจัดเตรียมข้อมูลเบื้องต้น
5. การสร้างแบบจำลองพร้อมจัดการความไม่สมดุลกันของข้อมูลและการประเมินผลประสิทธิภาพ
6. การอธิบายแบบจำลองโดยใช้การเรียนรู้ด้วยเครื่องแบบอธิบายได้สำหรับการคัดเลือกคุณลักษณะ
7. การคัดเลือกคุณลักษณะด้วยวิธีการต่างๆ รวมถึงการเรียนรู้ด้วยเครื่องแบบอธิบายได้
8. การสร้างแบบจำลองจากการคัดเลือกคุณลักษณะและการประเมินผลประสิทธิภาพ
9. การอธิบายแบบจำลองโดยใช้การเรียนรู้ด้วยเครื่องแบบอธิบายได้
10. การวิเคราะห์ความผิดพลาดของแบบจำลอง
11. การสรุปผลการวิจัยและการอภิปรายการวิจัย

โดยในการดำเนินการขั้นตอนต่างๆ จะมีการใช้งานไลบรารีของ Scikit-Learn เป็นหลัก โดย Scikit-Learn ถือเป็นไลบรารีที่นิยมใช้งานกันอย่างแพร่หลายในการดำเนินงานทางวิทยาศาสตร์ข้อมูลและการเรียนรู้ด้วยเครื่อง

3.1 การออกแบบขั้นตอนในการดำเนินงานวิจัย

การออกแบบขั้นตอนในการดำเนินงานวิจัยมีรายละเอียดดังภาพประกอบที่ 31


Process stream
Training data stream
Testing data stream
Start
Problems & Hypothesis Definition
Raw Data
Data Understanding
Exploratory Data Analysis (EDA)
Data Manipulation
Data Partition
Training Set 80%
Test Set 20%
Feature Engineering & Pre-processing Definition
- One-Hot, CatBoost, BaseN Encoder
- Standard Scalar
- Label Encoder
Classifier Training
- GridSearchCV
- Hyperparameter Tuning
- Imbalance Data Handling
- Feature Engineering & Pre-processing
Classifier Evaluation
Classifier Interpretation
- SHAP
Feature Selection
- F-value
- SHAP (best model)
- Random Feature Elimination (best model)
Feature Set 1
Feature Set 2
. . .
Feature Set n
Classifier Training
- GridSearchCV
- Hyperparameter Tuning
- Imbalance Data Handling
- Feature Engineering & Pre-processing
- Feature Selection
Classifier Evaluation
Classifier Interpretation
- SHAP Global Interpretation
- SHAP Local Interpreatation
Error Analysis
Research Result & Discussion
End

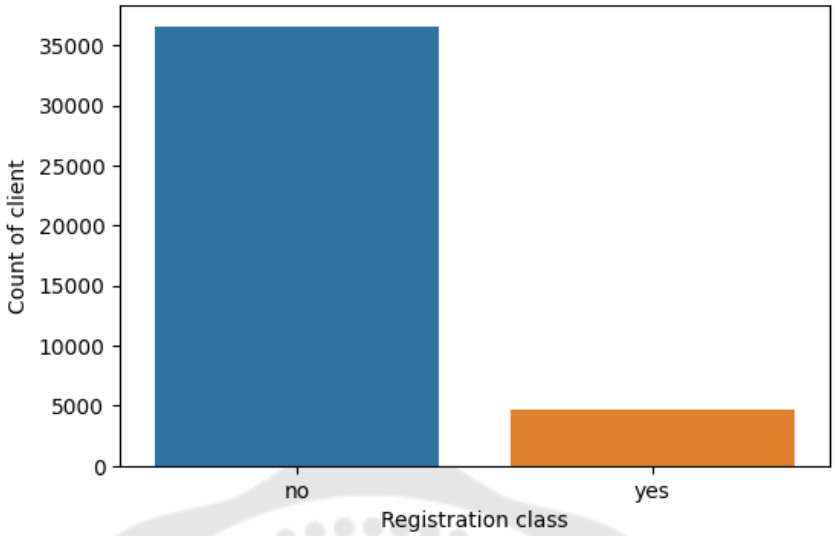
ภาพประกอบ 31 แสดงขั้นตอนในการดำเนินงานวิจัย

3.2 การสำรวจและวิเคราะห์ข้อมูลเบื้องต้น

ชุดข้อมูลที่นำมาใช้ในการวิจัยเป็นชุดข้อมูลสาธารณะจาก UCI Machine Learning Repository ซึ่งเกี่ยวกับการนำเสนอขายผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์ของธนาคารแห่งหนึ่งในประเทศโปรตุเกส ระหว่างเดือนพฤษภาคม พ.ศ. 2551 ถึงเดือนมิถุนายน พ.ศ. 2556 ซึ่งเก็บรวบรวมข้อมูลระหว่างการดำเนินงานวิจัย A Data-Driven Approach to Predict the Success of Bank Telemarketing (Moro et al., 2014)

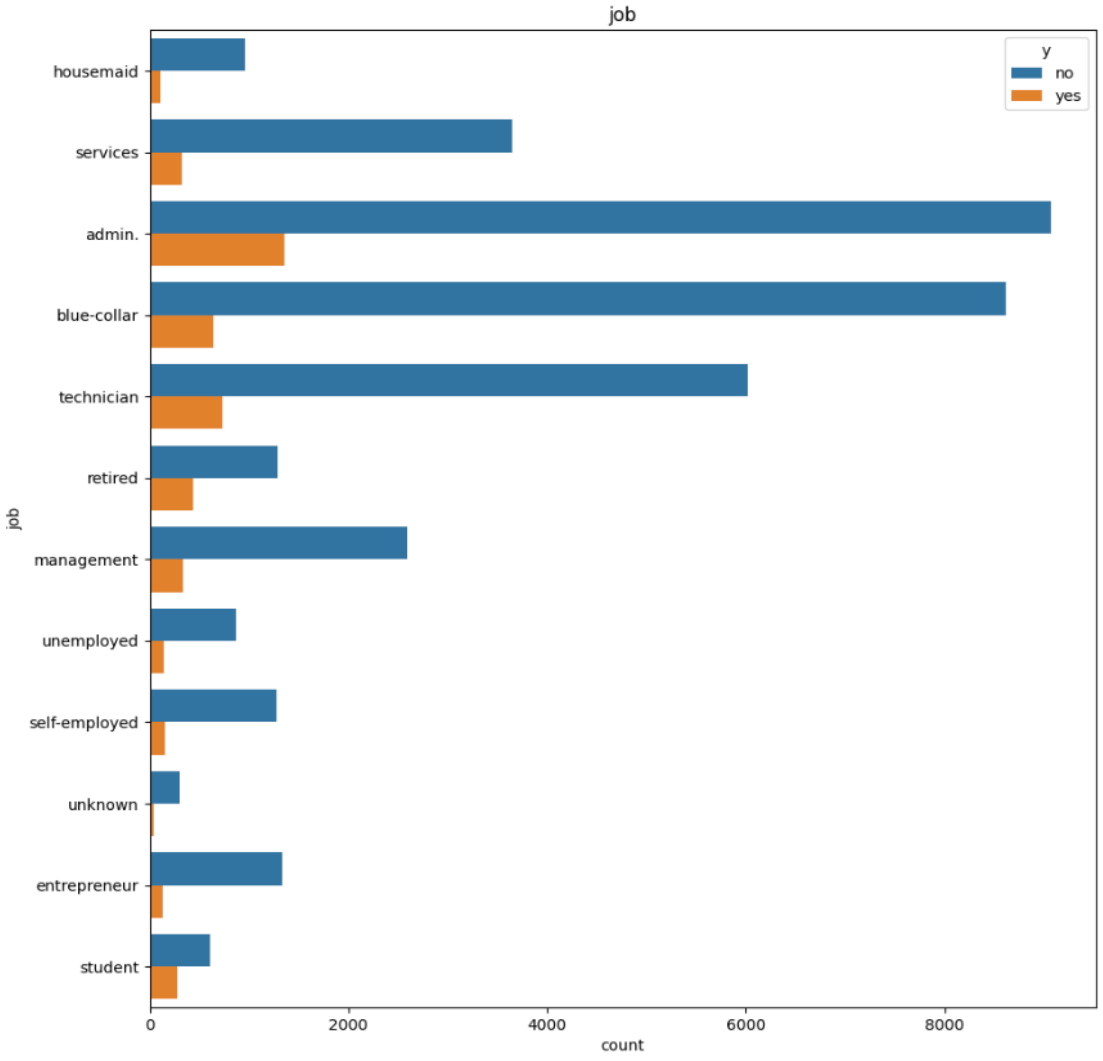
ชุดข้อมูลประกอบด้วยข้อมูลทั้งหมด 41,188 ตัวอย่างข้อมูล มีคุณลักษณะของข้อมูลทั้งหมด 21 คุณลักษณะ โดยสามารถแบ่งเป็นคุณลักษณะที่ใช้ในการจำแนกประเภทหรือตัวแปรต้นได้ 20 คุณลักษณะ ประกอบด้วย ‘อายุ’, ‘อาชีพ’, ‘สถานภาพการสมรส’, ‘ระดับการศึกษา’, ‘การปรากฏของการผิดนัดชำระหนี้’, ‘การปรากฏสินเชื่อเคหสถาน’, ‘การปรากฏสินเชื่อส่วนบุคคล’, ‘ประเภทของการติดต่อกับลูกค้า’, ‘เดือนที่ติดต่อลูกค้าครั้งล่าสุด’, ‘วันในสัปดาห์ที่ติดต่อลูกค้าครั้งล่าสุด’, ‘ระยะเวลาของการติดต่อลูกค้าครั้งสุดท้าย’, ’จำนวนครั้งการติดต่อเพื่อนำเสนอผลิตภัณฑ์นี้’, ‘จำนวนวันของระยะห่างจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า’, ’จำนวนครั้งการติดต่อก่อนนำเสนอผลิตภัณฑ์นี้’, ’ผลลัพธ์ของการนำเสนอผลิตภัณฑ์ก่อนหน้า’, ‘อัตราการจ้างงานรายไตรมาส’, ‘ดัชนีราคาผู้บริโภครายเดือน’, ‘ดัชนีความเชื่อมั่นผู้บริโภครายเดือน’, ’อัตราดอกเบี้ยกู้ยืมรายวันระหว่างธนาคารในยุโรป’ และ ‘จำนวนพนักงานรายไตรมาส’ ในส่วนของคุณลักษณะที่เป็นผลลัพธ์หรือตัวแปรตาม 1 คุณลักษณะคือผลการสมัครผลิตภัณฑ์ ซึ่งมีค่าเป็น ‘yes’ และ ‘no’

ชุดข้อมูลที่นำมาใช้งานมีความไม่สมดุลกันของข้อมูลที่ค่อนข้างสูง โดยประกอบด้วยข้อมูล Class ‘no’ สูงถึง 36,548 ตัวอย่างข้อมูล (~88.7%) ส่วนข้อมูล Class ‘yes’ มีจำนวนเพียง 4,640 ตัวอย่างข้อมูล (~11.3 %)



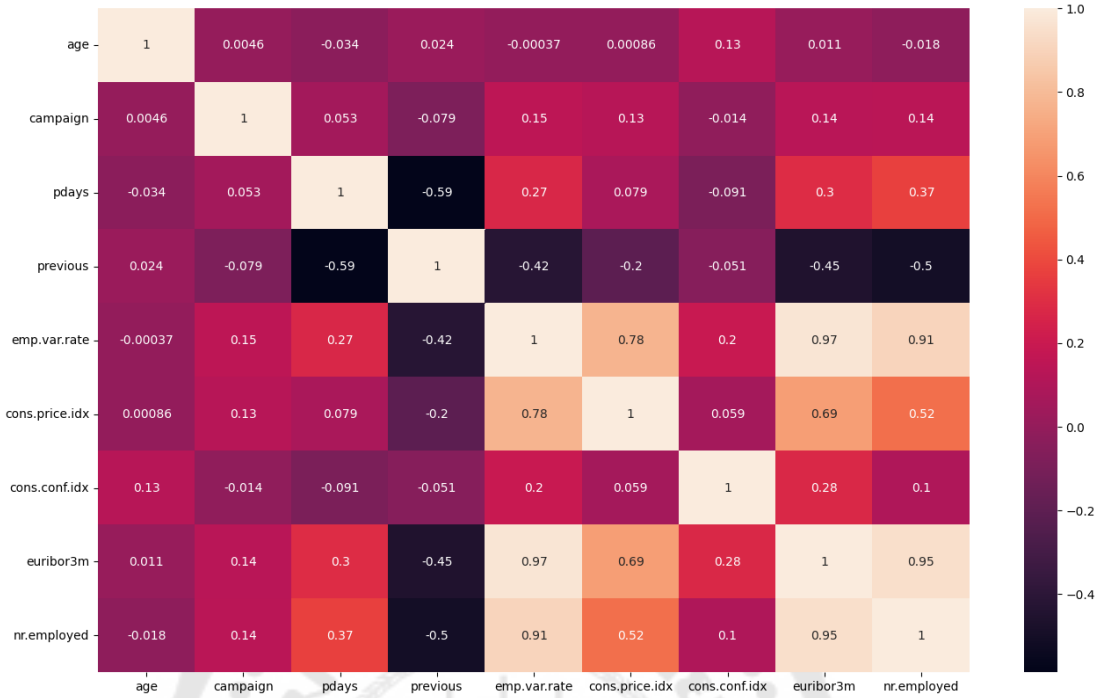
ภาพประกอบ 32 แสดงความไม่สมดุลกันของข้อมูลในชุดข้อมูลที่นำมาใช้ในการวิจัย

ในส่วนของการสำรวจข้อมูลค่าว่างปรากฏว่าไม่พบข้อมูลค่าว่างในชุดข้อมูล แต่เมื่อสำรวจค่าที่เป็นไปได้ของคุณลักษณะชนิดประเภทจะพบว่ามีค่าของข้อมูลที่เป็น ‘unknown’ ปะปนอยู่ภายในข้อมูลของบางคุณลักษณะ ซึ่งในการจัดการกับข้อมูลเหล่านี้เป็นเรื่องที่ต้องการความเข้าใจในกฎระเบียบรวมไปถึงบริบททางสังคมเฉพาะประเทศและช่วงเวลานั้นๆ ดังนั้นในการวิจัยครั้งนี้จึงจะดำเนินการกับค่า ‘unknown’ โดยถือเป็นค่าที่เป็นไปได้กลุ่มหนึ่งของข้อมูล



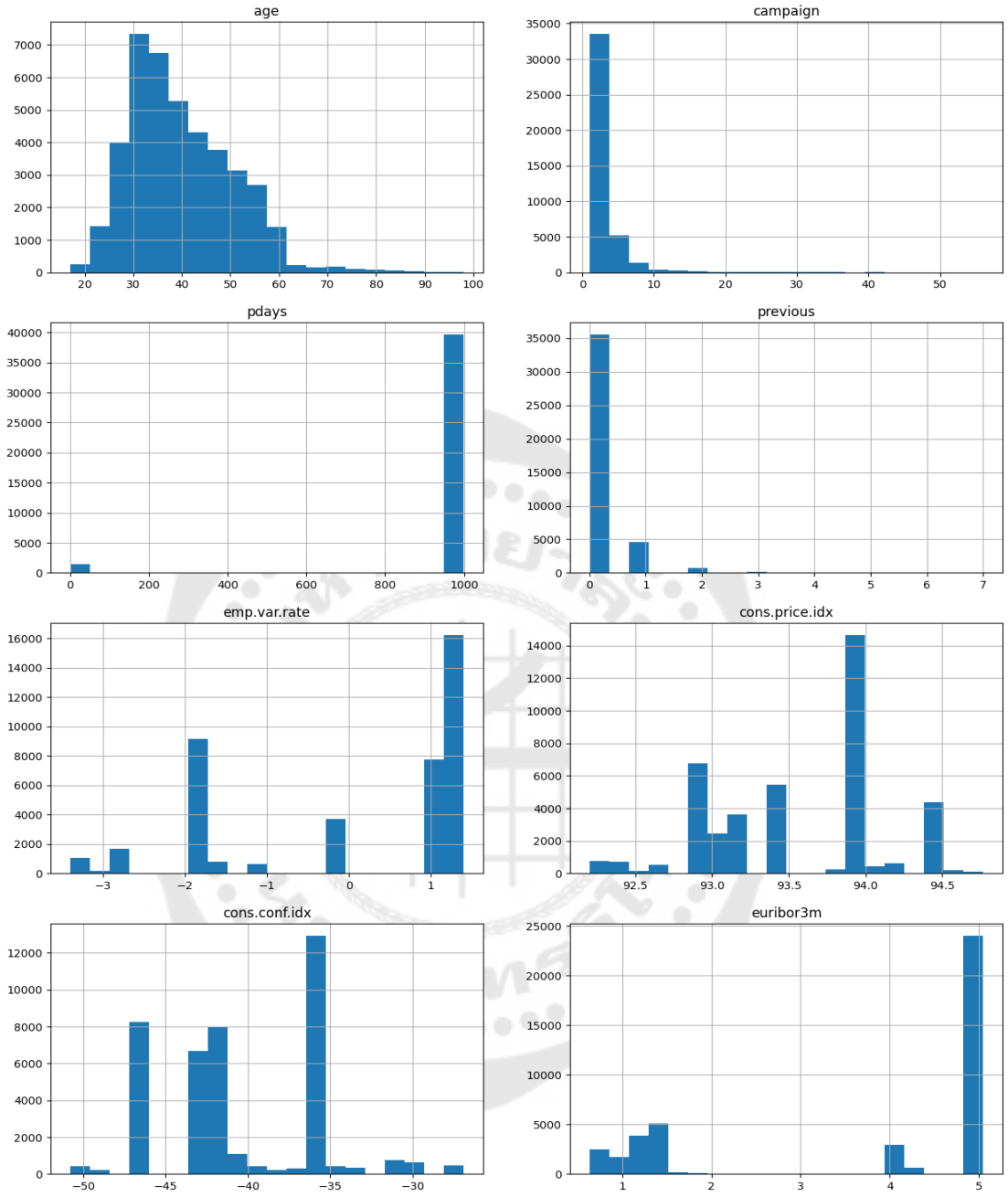
ภาพประกอบ 33 แสดงค่าที่เป็นไปได้ทั้งหมดที่ปรากฏในคุณลักษณะอาชีพโดยแบ่งตามผลลัพธ์ของการทำนาย

การสำรวจความมีปฏิสัมพันธ์ (Correlation) กันของคุณลักษณะที่เป็นแบบตัวเลข พบว่ามีบางคุณลักษณะที่มีปฏิสัมพันธ์ต่อกันสูงในทางบวก เช่น อัตราดอกเบี้ยกู้ยืมรายวันระหว่างธนาคารในยุโรป (euribor3m) และอัตราการจ้างงานรายไตรมาส (emp.var.rate) ซึ่งมีค่าปฏิสัมพันธ์กันสูงถึง 0.97 (97%) ซึ่งโดยปกติควรมีการกำจัดคุณลักษณะที่มีปฏิสัมพันธ์ต่อกันสูงออกเพื่อเพิ่มประสิทธิภาพในการเรียนรู้ด้วยเครื่องในแง่ของเวลาและทรัพยากรที่ใช้ในการดำเนินการ รวมไปถึงโอกาสในการเพิ่มประสิทธิภาพของแบบจำลองที่สูงขึ้นได้



ภาพประกอบ 34 แสดงค่าความมีปฏิสัมพันธ์กันของคุณลักษณะในชุดข้อมูล

นอกจากนี้เมื่อทำการสำรวจข้อมูลของคุณลักษณะชนิดตัวเลขจะพบว่ามีเพียง ‘อายุ’ ที่มีการกระจายตัวของข้อมูลแบบปกติ (Normal Distribution) ซึ่งจะต้องมีการดำเนินการจัดการกับคุณลักษณะชนิดตัวเลขทั้งหมดในภายหลังเพื่อให้มีการกระจายตัวแบบปกติและมีระดับของข้อมูล (Scale) อยู่ในระดับเดียวกัน

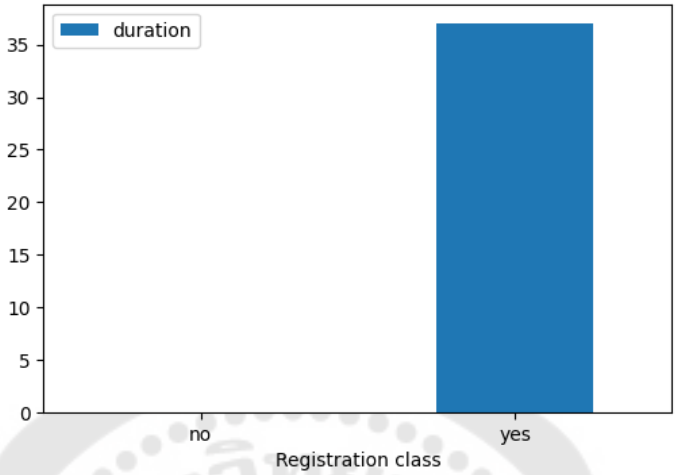


ภาพประกอบ 35 แสดงการกระจายตัวของข้อมูลในคุณลักษณะชนิดตัวเลข

3.3 การจัดการกับข้อมูลเบื้องต้นและการแบ่งชุดข้อมูล

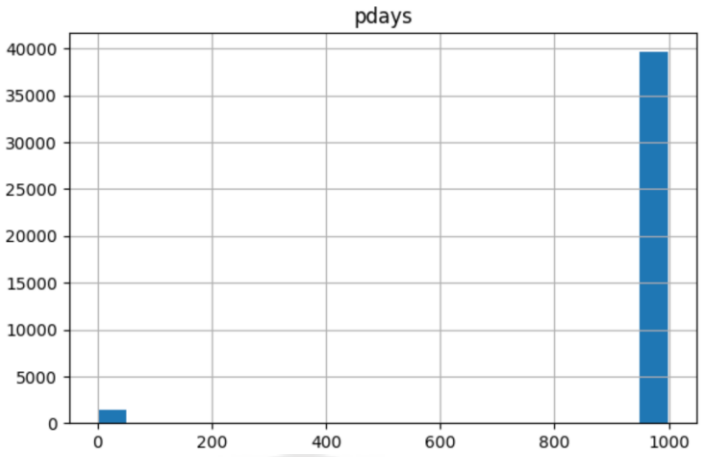
จากการสำรวจข้อมูลเบื้องต้นคุณลักษณะ ‘ระยะเวลาของการติดต่อลูกค้าครั้งสุดท้าย’ ควรถูกนำออกจากชุดข้อมูล เนื่องจากเป็นข้อมูลที่ได้มาจากการติดต่อลูกค้าครั้งสุดท้ายเพื่อทำการรับทราบผลลัพธ์ของการสมัครผลิตภัณฑ์ซึ่งอาจจะส่งผลกับการทำนายสูง เช่น จากการสำรวจ

ข้อมูลเพื่อหาค่าน้อยที่สุดที่ปรากฏของทั้งสอง Class พบว่าเมื่อ ‘ระยะเวลาของการติดต่อลูกค้าครั้งสุดท้าย’ มีค่าเป็น 0 จะสามารถระบุได้ทันทีว่าลูกค้าคนนั้นไม่ทำการสมัครผลิตภัณฑ์



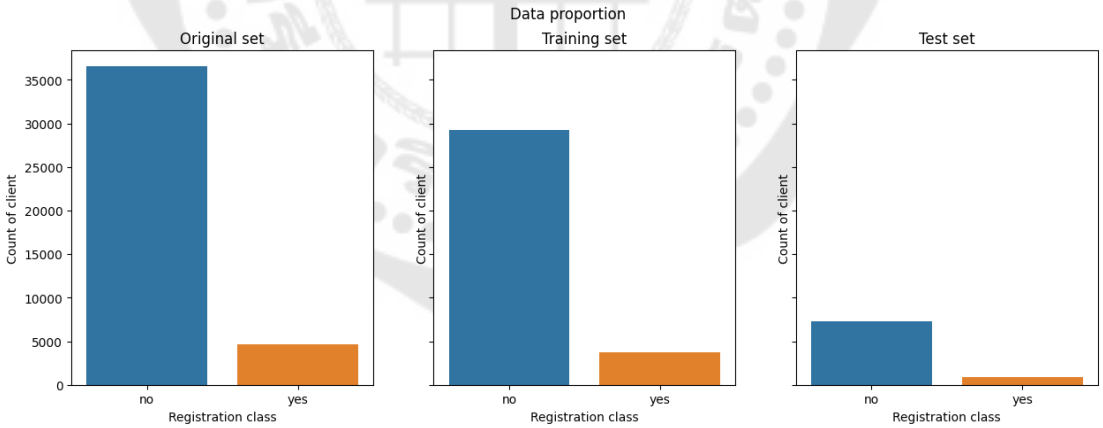
ภาพประกอบ 36 แสดงความสัมพันธ์ระหว่างคุณลักษณะการติดต่อลูกค้าครั้งสุดท้ายเพื่อรับทราบผลการสมัครและผลลัพธ์ของการทำนาย

ส่วนคุณลักษณะ ‘จำนวนวันของระยะห่างจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า’ (pdays) พบว่ามีข้อมูลที่มีค่าเป็น ‘999’ ซึ่งหมายถึงลูกค้าไม่เคยถูกติดต่อเพื่อนำเสนอผลิตภัณฑ์ครั้งก่อนหน้า จำนวนสูงถึง 39,673 ตัวอย่าง (~96.3%) และค่าสูงสุดที่ปรากฏซึ่งไม่ใช่ 999 คือ 27 วัน ดังนั้นจึงจะดำเนินการเปลี่ยนแปลงชนิดข้อมูลของคุณลักษณะนี้จากชนิดตัวเลขให้กลายเป็นชนิดประเภทสำหรับระบุการปรากฏของการติดต่อลูกค้าเพื่อนำเสนอผลิตภัณฑ์ครั้งก่อนหน้า โดยมีค่าที่เป็นไปได้คือ ‘yes’ สำหรับข้อมูลดั้งเดิมที่ไม่ใช่ ‘999’ และ ‘no’ สำหรับข้อมูลดั้งเดิมที่เป็น ‘999’



ภาพประกอบ 37 แสดงค่าที่เป็นไปได้ที่ปรากฏในคุณลักษณะจำนวนวันของระยะห่างจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า

จากนั้นทำการแบ่งชุดข้อมูลออกเป็นสองส่วน ได้แก่ ชุดข้อมูลสำหรับการเรียนรู้ซึ่งกำหนดสัดส่วนไว้ที่ 80% ของข้อมูลทั้งหมด และชุดข้อมูลสำหรับการทดสอบมีสัดส่วนของข้อมูลอยู่ที่ 20% โดยมีการใช้งานเทคนิค Stratify เพื่อให้การแบ่งข้อมูลมีสัดส่วนของกลุ่มผลลัพธ์ ‘yes’ และ ‘no’ เป็นไปในทางเดียวกับการแบ่งจำนวนของข้อมูล



ภาพประกอบ 38 แสดงการแบ่งข้อมูลสำหรับการวิจัย โดยสัดส่วนข้อมูลสำหรับการเรียนรู้ 80% และข้อมูลสำหรับการทดสอบ 20%

3.4 การทำวิศวกรรมคุณลักษณะและการจัดเตรียมข้อมูลเบื้องต้น

ชุดข้อมูลในการดำเนินงานวิจัยประกอบด้วยชนิดของข้อมูลหลักๆ อยู่ 3 ประเภท ได้แก่

1. ข้อมูลชนิดตัวเลข เช่น อายุ
2. ข้อมูลชนิดประเภทแบบไม่มีลำดับ เช่น สถานภาพการสมรส

3. ข้อมูลชนิดประเภทแบบมีลำดับ เช่น ระดับการศึกษา

4. ข้อมูลชนิดประเภทแบบไบนารี (ข้อมูลแบ่งได้ 2 กลุ่ม) เช่น ผลลัพธ์ของการนำเสนอผลิตภัณฑ์

ซึ่งข้อมูลแต่ละชนิดก็ต้องการการจัดการที่แตกต่างกันออกไป โดยการจัดการกับข้อมูลแต่ละประเภทจะใช้วิธีการต่างๆ ดังนี้

3.4.1 **ข้อมูลชนิดตัวเลข**

จะใช้วิธีการแบบ Standard Scaling ซึ่งเป็นการจัดการข้อมูลประเภทตัวเลขให้มีการกระจายตัวเป็นแบบปกติ (Normal Distribution) ซึ่งจะทำให้ข้อมูลมีคุณลักษณะที่มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 ซึ่งจะมีกราฟเป็นรูประฆังคว่ำ (Bell Curve) และมีการกำหนดให้ทำการแทนที่ข้อมูลที่เป็นค่าว่างด้วยค่าเฉลี่ยของคุณลักษณะ

3.4.2 **ข้อมูลชนิดประเภทแบบไม่มีลำดับ**

จะใช้วิธีการแบบ One-Hot Encoding ซึ่งจะมีการขยายจำนวนของคุณลักษณะหนึ่งๆ ออกไปเป็นจำนวนเทียบเท่ากับค่าทั้งหมดที่ปรากฏในข้อมูล และเมื่อค่าของข้อมูลตรงกับคุณลักษณะใหม่ใดๆ ที่ทำการขยายออกไปจะมีการแทนที่ค่าของข้อมูลด้วย 1 ส่วนคุณลักษณะอื่นๆ ที่ทำการขยายออกไปจะแทนที่ด้วย 0 ส่วนค่าของข้อมูลที่ไม่เคยพบเห็นมาก่อนจะให้ทำการละเลยค่าของข้อมูลนั้นไป และมีการกำหนดให้ทำการแทนที่ข้อมูลที่เป็นค่าว่างด้วยค่าของข้อมูลที่ปรากฏมากที่สุดในคุณลักษณะ

มีการใช้งานวิธีการแบบ CatBoost Encoding ซึ่งเป็นวิธีการในการจัดการคุณลักษณะชนิดประเภทซึ่งพัฒนาต่อยอดจากหลักการของ Target Encoding ซึ่งจะดำเนินการแทนที่คุณลักษณะชนิดประเภทด้วยค่าที่คำนวณจากความน่าจะเป็นของผลลัพธ์ที่สอดคล้องกับค่าของข้อมูลนั้นๆ ในแต่ละคุณลักษณะ ร่วมกับค่าความน่าจะเป็นของผลลัพธ์ในชุดข้อมูล ซึ่ง CatBoost Encoding มีการปรับปรุงในส่วนของการใช้งานเพียงแค่ตัวอย่างข้อมูลก่อนหน้าที่เคยเห็นมาแล้วมาใช้สำหรับการคำนวณแทนที่การใช้งานทั้งชุดข้อมูล เพื่อเป็นการลดการเกิดการรั่วไหลของข้อมูลของ Target Encoding

นอกจากนี้ยังมีการใช้งานวิธีการแบบ BaseN Encoding ซึ่งเป็นวิธีการในการจัดการคุณลักษณะชนิดประเภทซึ่งมีความคล้ายคลึงกับการใช้งานวิธีการแบบ Binary Encoding ซึ่งจะทำการแปลงค่าของข้อมูลที่แตกต่างกันในแต่ละคุณลักษณะให้เป็น Bit String (0 และ 1) โดยจะทำการเพิ่ม Dummy Feature เพื่อขยายความยาวของ Bit String เมื่อมีขนาดไม่เพียงพอสำหรับการรองรับค่าที่เป็นไปได้ทั้งหมดในคุณลักษณะนั้นๆ ส่วนวิธีการแบบ BaseN จะสามารถกำหนดจำนวนของค่าที่เพิ่มขึ้นได้มากกว่า 1 เช่น เมื่อกำหนดค่า N = 5 จะส่งผลให้แต่ละ Bit สามารถ

รองรับค่าได้ตั้งแต่ 0-4 ทำให้สามารถลดจำนวนของ Dummy Feature ในการรองรับค่าที่เป็นไปได้ทั้งหมดในคุณลักษณะนั้นๆ ได้ ซึ่งในงานวิจัยครั้งนี้ได้ใช้งานค่า N = 5

3.4.3 ข้อมูลชนิดประเภทแบบมีลำดับ

จะใช้วิธีการแบบ Ordinal Encoding ซึ่งจะเป็นการเปลี่ยนแปลงค่าของข้อมูลจากชนิดประเภทให้กลายเป็นตัวเลขแบบเรียงลำดับตามลำดับของข้อมูลที่กำหนดไว้ล่วงหน้า ทำให้ลำดับหรือความสำคัญของข้อมูลไม่สูญหายไป และมีการกำหนดให้ทำการแทนที่ข้อมูลที่เป็นค่าว่างด้วยค่าของข้อมูลที่ปรากฏมากที่สุดในคุณลักษณะ

3.4.4 ข้อมูลชนิดประเภทแบบไบนารี

จะใช้วิธีการแบบ Label Encoding ซึ่งจะเปลี่ยนแปลงค่าของข้อมูลผลลัพธ์จากชนิดประเภท ‘yes’ และ ‘no’ ให้กลายเป็น ‘1’ และ ‘0’ ซึ่งเหมาะกับการใช้งานกับข้อมูลคุณลักษณะที่เป็นผลลัพธ์ของการจำแนกประเภทซึ่งไม่มีการใช้งานลำดับของตัวเลข

โดยการทำวิศวกรรมคุณลักษณะและการจัดเตรียมข้อมูลเบื้องต้นจะมีการใช้งานการทำงานแบบสายท่อ หรือ Pipeline เพื่อให้การทำงานของคุณลักษณะแต่ละชนิดเป็นไปตามลำดับขั้นตอนที่มีการวางแผนไว้และเพื่อเป็นการป้องกันการรั่วไหลของข้อมูล จากนั้นมีการใช้งานวิธีการแบบ ColumnTransformer เพื่อรวบรวมการทำงานแบบสายท่อหลายๆ อันเข้าไว้ด้วยกัน

3.5 การสร้างแบบจำลองพร้อมจัดการความไม่สมดุลกันของข้อมูลและการประเมินผลประสิทธิภาพ

ในขั้นตอนการดำเนินงานนี้จะทำการสร้างแบบจำลองโดยใช้งานคุณลักษณะของข้อมูลทั้งหมดหลังจากการทำวิศวกรรมคุณลักษณะและการจัดเตรียมข้อมูลเบื้องต้น โดยไม่ได้มีการคัดเลือกคุณลักษณะที่จะนำไปใช้ในการสร้างแบบจำลอง เนื่องจากต้องการให้แต่ละแบบจำลองที่เลือกนำมาใช้งานทำงานกับคุณลักษณะของข้อมูลทั้งหมดก่อน จากนั้นจะนำแบบจำลองที่มีประสิทธิภาพสูงมาตรวจสอบและวิเคราะห์เพื่อค้นหาความสำคัญของแต่ละคุณลักษณะในแบบจำลองนั้นๆ เพื่อนำไปดำเนินการคัดเลือกคุณลักษณะต่อไป

3.5.1 การกำหนดและการปรับแต่งแบบจำลอง

ในการวิจัยได้นำแบบจำลองที่เป็นที่นิยมจำนวน 4 แบบจำลองมาใช้การจำแนกประเภทซึ่งประกอบด้วย Logistic Regression, Random Forest, LightGBM และ XGBoost โดยแบบจำลองที่ทำการกำหนดขึ้นจะถูกรวมเข้าไปเป็นส่วนหนึ่งของการทำงานแบบสายท่อ

นอกจากนี้ในแต่ละแบบจำลองจะมีการปรับแต่งตัวแปรปัจจัย (Hyperparameter) เพื่อค้นหาแบบจำลองและชุดของตัวแปรปัจจัยที่ดีที่สุดในการทำให้ประสิทธิภาพของแบบจำลอง

ออกมาสูงที่สุด โดยการปรับแต่งตัวแปรปัจจัยของแบบจำลองจะมีการใช้งาน Grid Search สำหรับการวนปรับแต่งตัวแปรปัจจัยที่กำหนดไว้ล่วงหน้าให้ครบทุกความเป็นไปได้ของชุดตัวแปรปัจจัย

3.5.2 การจัดการความไม่สมดุลกันของข้อมูล

ระหว่างกระบวนการสร้างแบบจำลองจะมีการดำเนินการจัดการกับความไม่สมดุลกันของข้อมูลโดยวิธีที่นำมาใช้ประกอบด้วย การปรับค่าน้ำหนักของข้อมูล (Class Weight), Random Undersampling และ SMOTE โดยการจัดการความไม่สมดุลกันของข้อมูลแบบการปรับค่าน้ำหนักของข้อมูลจะถูกกำหนดไปพร้อมกับการปรับแต่งตัวแปรปัจจัยของแบบจำลอง ส่วนการจัดการความไม่สมดุลแบบ Random Undersampling และ SMOTE จะถูกรวมเข้าไปอยู่ในการทำงานแบบสายท่อร่วมกับการทำวิศวกรรมคุณลักษณะ, การจัดเตรียมข้อมูลเบื้องต้น และการสร้างแบบจำลอง

3.5.3 การเรียนรู้ด้วยเครื่องเพื่อสร้างแบบจำลอง

ในการเรียนรู้ด้วยเครื่องเพื่อสร้างแบบจำลองที่มีประสิทธิภาพสูงที่สุดออกมาได้มีการใช้งานเทคนิค Cross Validation ซึ่งจะเป็นการแบ่งชุดข้อมูลสำหรับการเรียนรู้ออกเป็นส่วนย่อยๆ ตามที่กำหนด จากนั้นจะทำการสร้างแบบจำลองเป็นจำนวนรอบเท่ากับจำนวนส่วนย่อยของข้อมูลที่กำหนดไว้ โดยในแต่ละรอบจะมีข้อมูลหนึ่งส่วนย่อยถูกแยกออกมาสำหรับเป็นข้อมูลในการทดสอบแบบจำลองของรอบนั้นๆ ส่วนข้อมูลส่วนย่อยที่เหลือจะถูกนำไปใช้ในการสร้างแบบจำลองของรอบนั้นๆ โดยในการวิจัยครั้งนี้จะมีการกำหนด Cross Validation ของแบบจำลองไว้ที่ 5 ครั้ง

ซึ่งในการวิจัยจะมีการใช้งานวิธีแบบ GridSearchCV สำหรับการใช้งาน Cross Validation ร่วมกับการใช้งาน Grid Search ซึ่งจะส่งผลให้มีการสร้างแบบจำลองในรูปแบบที่หลากหลายทั้งในด้านของตัวแปรปัจจัยของแบบจำลองและในด้านของชุดข้อมูลที่ใช้ในการสร้างและการทดสอบแบบจำลอง โดยเมื่อพบชุดของตัวแปรปัจจัยที่ทำให้แบบจำลองมีประสิทธิภาพสูงที่สุดแล้ว จะมีการนำแบบจำลองพร้อมทั้งชุดตัวแปรปัจจัยมาใช้ในการเรียนรู้กับชุดข้อมูลสำหรับการเรียนรู้ทั้งหมดอีกครั้ง

ในการกำหนดค่าสำหรับการประเมินประสิทธิภาพของแบบจำลอง มีการเลือกใช้งานค่า Recall ซึ่งสามารถกำหนดเข้าไปเป็นส่วนหนึ่งของการทำ GridSearchCV ได้ เพื่อให้แบบจำลองมีการเรียนรู้โดยเน้นประสิทธิภาพของค่า Recall ให้ออกมาสูงที่สุด โดยสาเหตุของการมุ่งเน้นไปที่ค่าของ Recall เนื่องจากว่าในการนำเสนอผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์เราต้องการที่จะค้นหากลุ่มของลูกค้าที่มีแนวโน้มที่จะสมัครผลิตภัณฑ์ให้ได้มากที่สุด โดยยอมรับกับความผิดพลาดได้เล็กน้อยในการจำแนกลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ว่าจะทำการสมัคร

ผลิตภัณฑ์ ซึ่งหากแบบจำลองไม่สามารถค้นหาลูกค้าที่ทำการสมัครผลิตภัณฑ์ได้ครบถ้วนเพียงพอ ก็จะทำให้ธนาคารเสียโอกาสและรายได้ในส่วนนั้นๆ ไป

3.5.4 การประเมินผลประสิทธิภาพของแบบจำลอง

ในการประเมินผลประสิทธิภาพของแบบจำลองเบื้องต้นมีการใช้งาน Classification Report เพื่อดูผลค่าประสิทธิภาพต่างๆ เช่น ค่าความแม่นยำ, Precision, Recall และ F1-Score ผ่านทางการใช้งานวิธีการแบบ classification_report โดยมุ่งเน้นไปยังแบบจำลองที่มีค่าของ Recall และ F1-Score ที่มีประสิทธิภาพสูง โดยมีการใช้งานควบคู่กับ Confusion Matrix เพื่อดูจำนวนของตัวอย่างข้อมูลที่มีการจำแนกถูกและผิดตามจริง

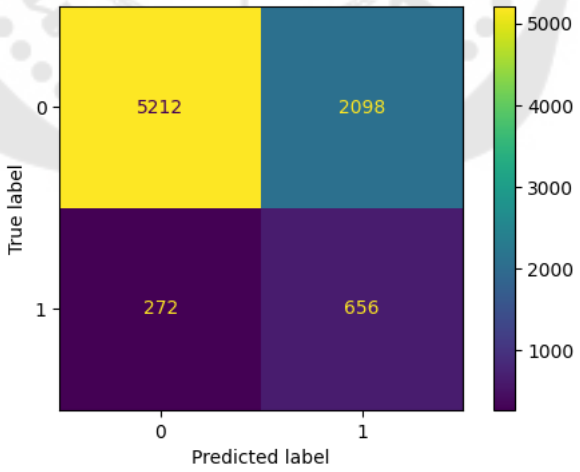
```
Best Param : {'logisticregression__C': 1e-05, 'logisticregression__solver': 'liblinear'}
Train Accuracy : 0.7176682278081501
Test Accuracy : 0.7123088128186453

Classification Report :
              precision    recall  f1-score   support

           0       0.95      0.71      0.81      7310
           1       0.24      0.71      0.36       928

    accuracy                           0.71      8238
   macro avg       0.59      0.71      0.59      8238
weighted avg       0.87      0.71      0.76      8238
```

ภาพประกอบ 39 แสดงการประเมินผลประสิทธิภาพของแบบจำลองด้วยวิธีการแบบ classification_report พร้อมตัวแปรปัจจัยที่ส่งผลให้แบบจำลองมีประสิทธิภาพสูงที่สุด



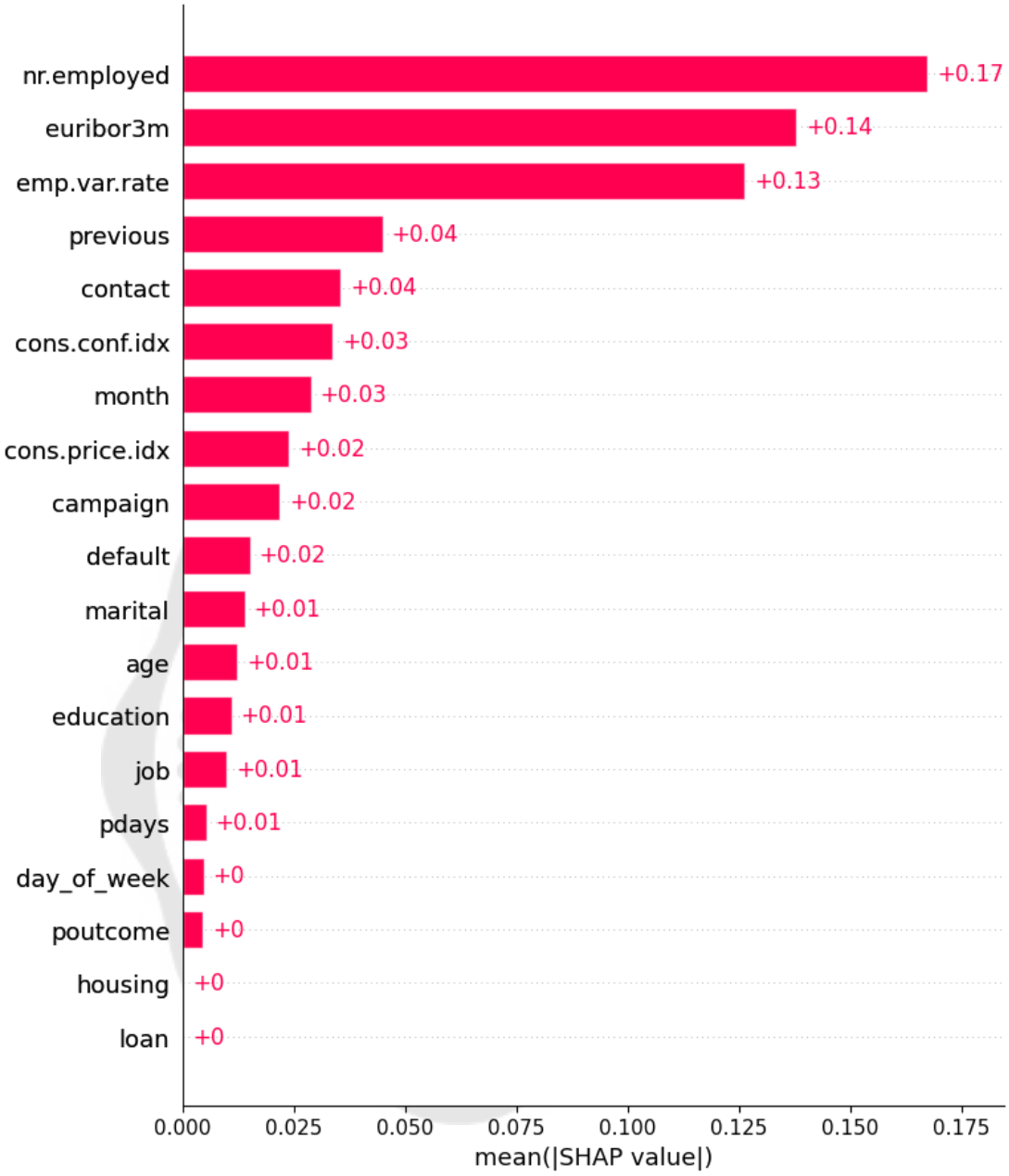
ภาพประกอบ 40 แสดง Confusion Matrix ของแบบจำลองในงานวิจัย

3.6 การอธิบายแบบจำลองโดยใช้การเรียนรู้ด้วยเครื่องแบบอธิบายได้สำหรับการคัดเลือกคุณลักษณะ

ในการอธิบายแบบจำลองในขั้นตอนนี้จะมีการใช้งานวิธีการแบบ SHAP เพื่อนำมาอธิบายแบบจำลองที่มีประสิทธิภาพสูงที่สุด เพื่อค้นหาว่าแต่ละคุณลักษณะมีค่าความสำคัญหรือมีอิทธิพลต่อการจำแนกประเภทด้วยสัดส่วนเท่าใดบ้าง และความสัมพันธ์ของความมีอิทธิพลและค่าที่ปรากฏในแต่ละคุณลักษณะ

โดยในการอธิบายแบบจำลองนั้นข้อมูลคุณลักษณะชนิดประเภทที่มีการทำ One-Hot Encoding นั้น จะมีการดำเนินการเพื่อรวบรวมคุณลักษณะใหม่ที่เกิดขึ้นจากคุณลักษณะเดียวกันกลับมาไว้รวมกันเป็นคุณลักษณะเดียว ซึ่งค่าของความสำคัญในทุกคุณลักษณะใหม่จะถูกคำนวณเพื่อหาผลรวมของความสำคัญออกมา

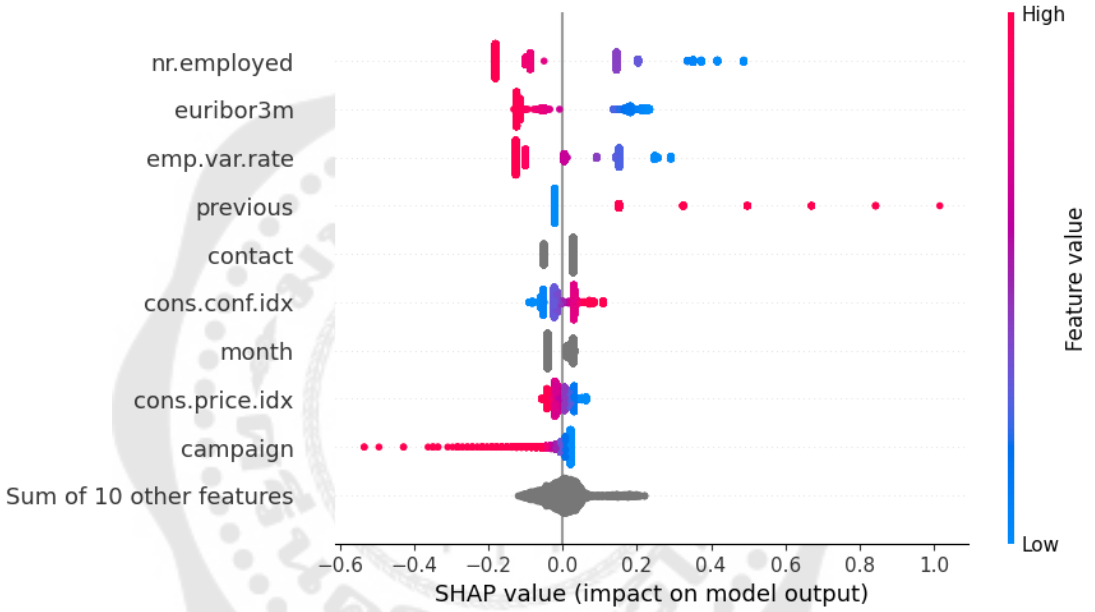
สามารถใช้งาน Bar Plot เพื่อแสดงค่าความสำคัญของแต่ละคุณลักษณะได้ โดยกราฟที่แสดงจะเรียงลำดับจากคุณลักษณะที่มีความสำคัญสูงที่สุดไล่ลงมายังคุณลักษณะที่มีความสำคัญต่ำที่สุด



ภาพประกอบ 41 แสดงค่าความสำคัญของคุณลักษณะโดยใช้วิธีการ SHAP Bar Plot

นอกจากนี้ยังสามารถใช้งานวิธีการวาดกราฟอื่นๆ เพื่อใช้ในการแสดงค่าความสำคัญของคุณลักษณะในรูปแบบต่างๆ ได้ เช่น Beeswarm Plot ซึ่งจะเป็นการวาดกราฟโดยแสดงถึงความสัมพันธ์ของความสำคัญของคุณลักษณะทั้งในทางเชิงบวกและเชิงลบเมื่อเปรียบเทียบกับค่าของข้อมูลดังรูปด้านล่าง ซึ่งเหมาะกับการแสดงค่าความสำคัญของคุณลักษณะแบบตัวเลข ตัวอย่างเช่น คุณลักษณะอัตราดอกเบี้ยการกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)

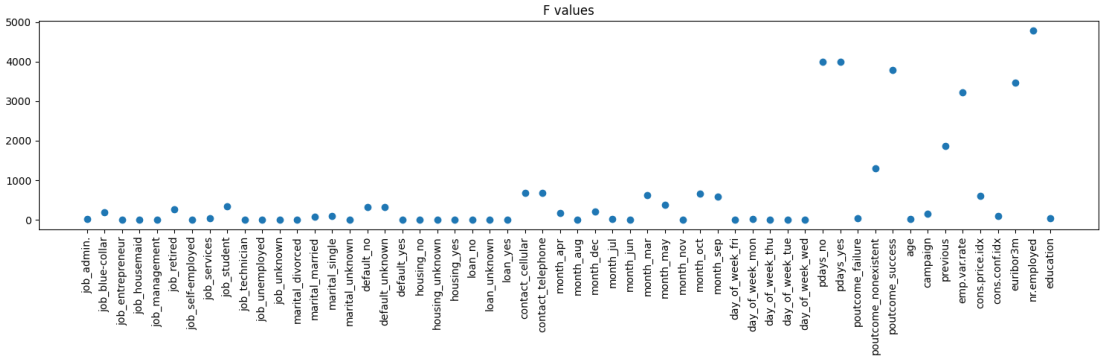
เมื่อมีค่าของข้อมูลที่สูง (สีชมพู) จะมีค่าของ SHAP Value หรือค่าความสำคัญที่สูงขึ้นในเชิงลบต่อการจำแนกประเภท และเมื่อมีค่าของข้อมูลที่ต่ำ (สีฟ้า) จะมีค่าของ SHAP Value ที่สูงขึ้นในเชิงบวก ซึ่งสามารถบ่งบอกได้ว่าเมื่ออัตราดอกเบี้ยการกู้ยืมระหว่างธนาคารภายในยุโรปรายวันมีค่าที่ต่ำลงจะทำให้มีโอกาสที่ลูกค้าจะทำการสมัครผลิตภัณฑ์เงินฝากประจำผ่านการนำเสนอทางโทรศัพท์สูงมากขึ้น ส่วนความหนาหรือสูงของเส้นกราฟจะบ่งบอกถึงความหนาแน่นของประชากรในค่าของข้อมูลบริเวณนั้น ในส่วนของคุณลักษณะชนิดประเภทสามารถใช้งานการวาดกราฟด้วยวิธีการอื่นๆ เพื่อแสดงค่าความสำคัญได้



ภาพประกอบ 42 แสดงความสัมพันธ์ระหว่างค่าในคุณลักษณะนั้นๆ และ SHAP Value

3.7 การคัดเลือกคุณลักษณะด้วยวิธีการต่างๆ รวมถึงการเรียนรู้ด้วยเครื่องแบบอธิบายได้

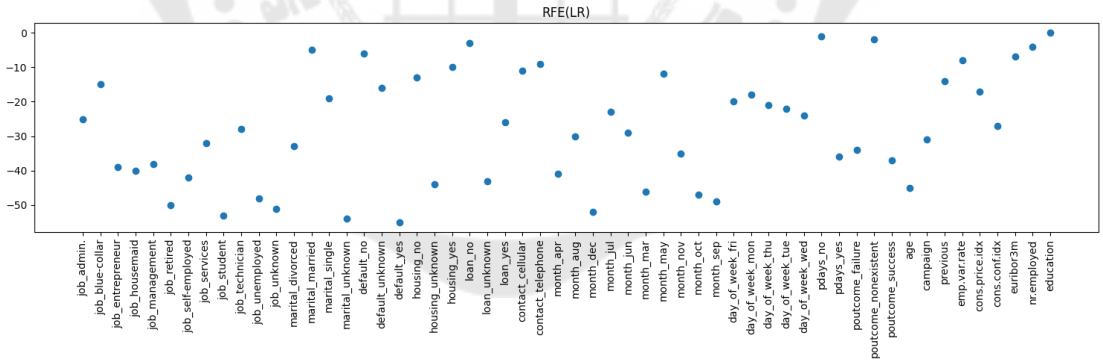
ในกระบวนการคัดเลือกคุณลักษณะมีการใช้งานวิธีการต่างๆ ประกอบด้วย F-Value, RFE และ SHAP โดย F-Value จะเป็นวิธีการที่ค้นหาความสำคัญของคุณลักษณะโดยใช้งานกระบวนการทางสถิติ ซึ่งไม่ต้องมีการสร้างแบบจำลองใดๆ ประกอบการดำเนินการ ซึ่งคุณลักษณะใดที่มีค่าของ F-Value ที่สูงจะหมายถึงคุณลักษณะนั้นมีค่าความสำคัญและมีความสัมพันธ์ต่อการจำแนกประเภทที่สูง โดยสามารถใช้งานวิธีการแบบ f_classif เพื่อหาค่าของ F-Value ได้ จากภาพประกอบที่ 43 จะพบว่าคุณลักษณะจำนวนพนักงานรายไตรมาส (nr.employ) มีค่าของความสำคัญที่สูงที่สุด



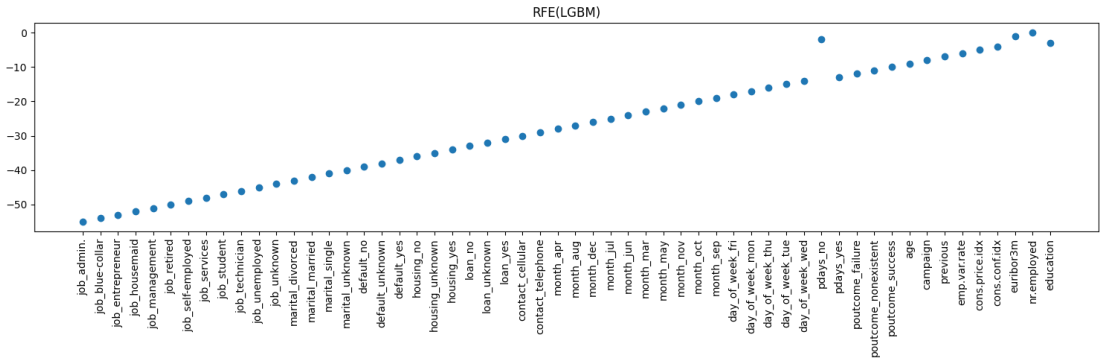
ภาพประกอบ 43 แสดงค่าความสำคัญของคุณลักษณะโดยการใช้วิธีการแบบ F-Value

ส่วนวิธีการแบบ RFE ซึ่งต้องอาศัยแบบจำลองเพื่อนำมาใช้ประเมินค่าความสำคัญของคุณลักษณะ ในการดำเนินงานวิจัยนี้มีการใช้งานวิธีการแบบ RFE ร่วมกับแบบจำลองที่มีประสิทธิภาพสูงที่สุดจากการสร้างแบบจำลองในขั้นตอนก่อนหน้าประกอบด้วยแบบจำลอง Logistic Regression, LightGBM และ XGBoost

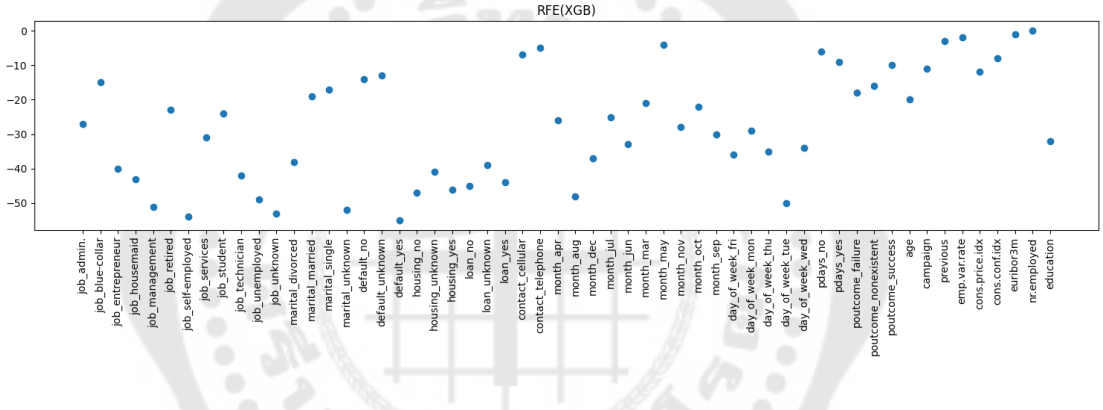
จากภาพประกอบที่ 44, 45 และ 46 ซึ่งเป็นค่าความสำคัญของคุณลักษณะจากการใช้งานวิธีแบบ RFE(LR), RFE(LGBM) และ RFE(XGB) ตามลำดับ จะเห็นว่าค่าความสำคัญของคุณลักษณะจะมีความแตกต่างกันอยู่ระหว่างทั้งสามแบบจำลอง



ภาพประกอบ 44 แสดงค่าความสำคัญของคุณลักษณะโดยการใช้วิธีการแบบ RFE ร่วมกับแบบจำลอง Logistic Regression

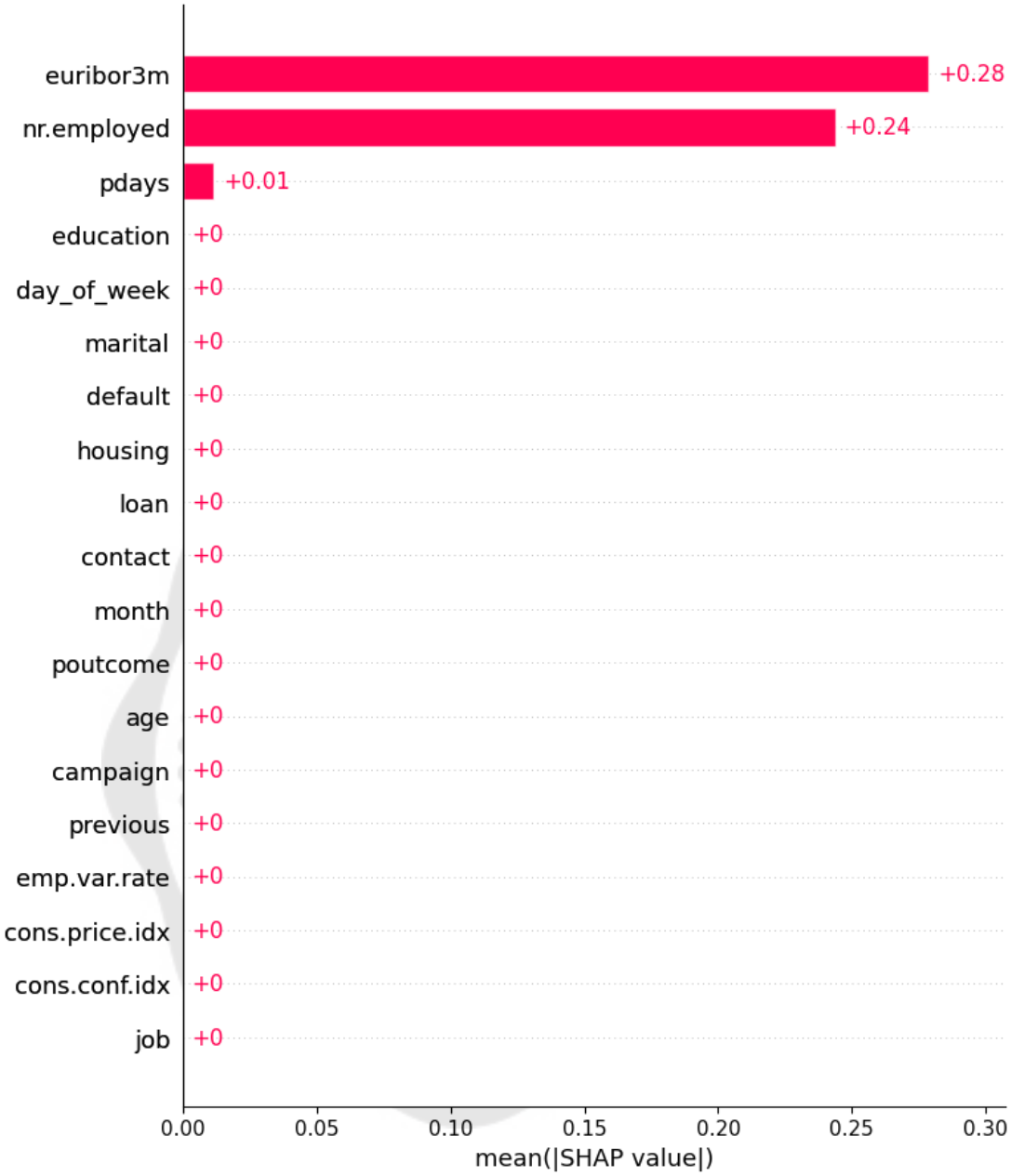


ภาพประกอบ 45 แสดงค่าความสำคัญของคุณลักษณะโดยการใช้วิธีการแบบ RFE ร่วมกับแบบจำลอง LightGBM

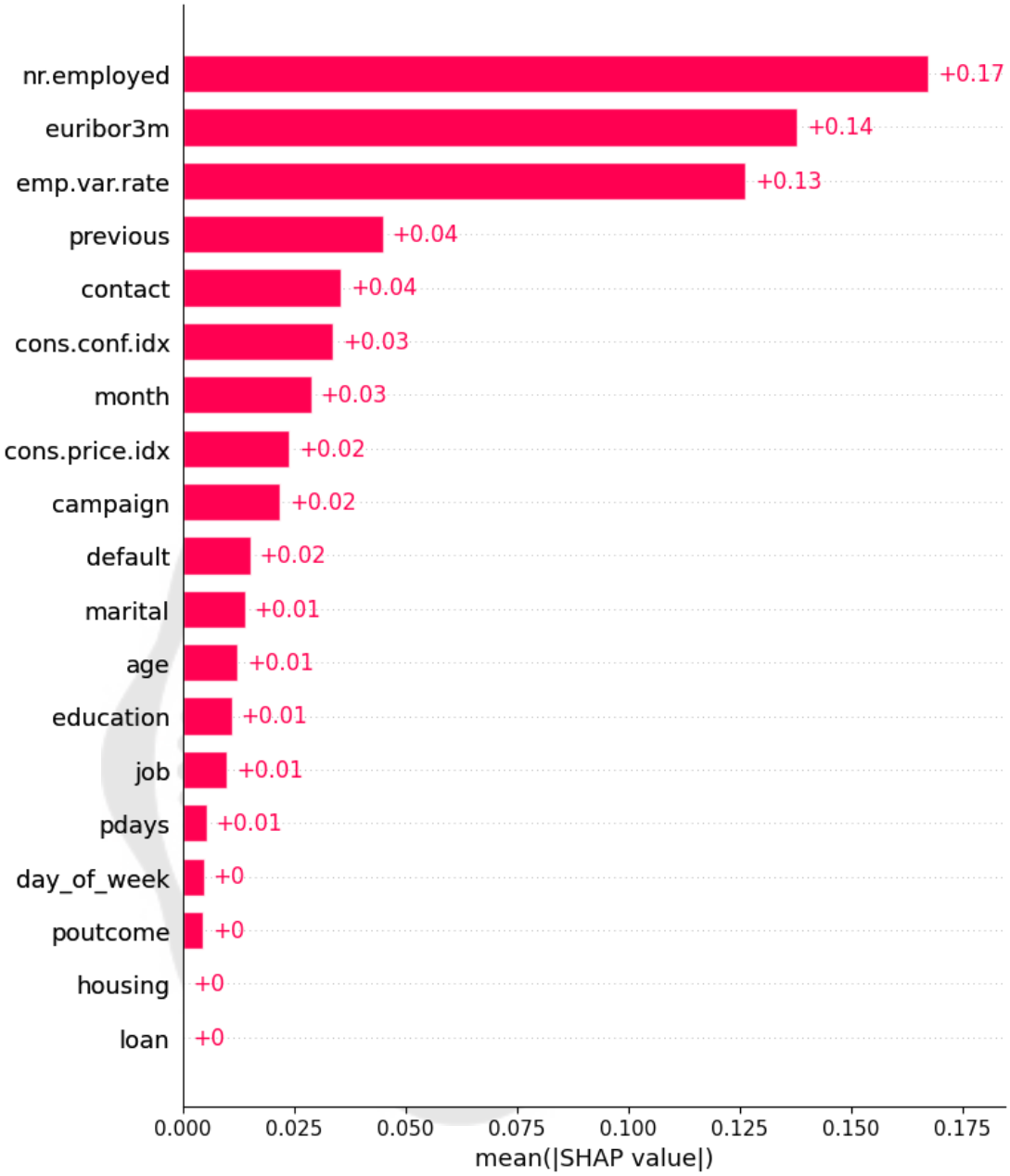


ภาพประกอบ 46 แสดงค่าความสำคัญของคุณลักษณะโดยการใช้วิธีการแบบ RFE ร่วมกับแบบจำลอง XGBoost

จากนั้นจะดำเนินการคัดเลือกคุณลักษณะที่มีค่าความสำคัญสูงที่สุดซึ่งได้มาจากการอธิบายแบบจำลองของการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธีการแบบ SHAP เพื่อใช้นำไปอธิบายแบบจำลอง LightGBM และ XGBoost ดังภาพประกอบที่ 47 และ 48 ตามลำดับ ซึ่งจะได้ค่าความสำคัญของคุณลักษณะดังรูปด้านล่าง จากนั้นจะมีการนำไปใช้ในการสร้างแบบจำลองในขั้นตอนถัดไป



ภาพประกอบ 47 แสดงค่าความสำคัญของคุณลักษณะโดยการใช้วิธีการแบบ SHAP ร่วมกับแบบจำลอง LightGBM



ภาพประกอบ 48 แสดงค่าความสำคัญของคุณลักษณะโดยการใช้วิธีการแบบ SHAP ร่วมกับแบบจำลอง XGBoost

ในการคัดเลือกคุณลักษณะทั้ง 6 วิธีการที่นำมาใช้งานจะทำการคัดเลือกคุณลักษณะที่มีค่าของความสำคัญสูงที่สุดเพียงแค่ 2 คุณลักษณะแรก เพื่อนำไปใช้ในการสร้างแบบจำลองในขั้นตอนถัดไป ซึ่งแต่ละวิธีการจะได้คุณลักษณะที่มีค่าความสำคัญสูงสุดสองอันดับแรกออกมาดังตารางที่ 7 ซึ่งจะพบว่าวิธีการคัดเลือกคุณลักษณะแบบ RFE(LGBM), RFE(XGB), SHAP(LGBM)

และ SHAP(XGB) มีคุณลักษณะที่มีค่าความสำคัญสูงสุดสองอันดับแรกที่เหมือนกัน ดังนั้นจะมีการใช้งานชุดของคุณลักษณะทั้งหมด 3 ชุดดังตารางที่ 7 ในการสร้างแบบจำลองในขั้นตอนถัดไป

ตาราง 7 ชุดของคุณลักษณะที่ดีที่สุดจำนวนสองคุณลักษณะจากการคัดเลือกด้วยวิธีการต่างๆ

วิธีการคัดเลือกคุณลักษณะ	**คุณลักษณะที่ 1**	**คุณลักษณะที่ 2**
F-Value	จำนวนวันจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า (pdays)	จำนวนพนักงานรายไตรมาส (nr.employed)
RFE(LR)	จำนวนวันจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า (pdays)	ระดับการศึกษา (education)
RFE(LGBM)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)
RFE(XGB)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)
SHAP(LGBM)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)
SHAP(XGB)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)

โดยในขั้นตอนนี้ยังรวมไปถึงการดำเนินการสร้างการทำงานแบบสายท่อเพื่อรองรับการทำวิศวกรรมคุณลักษณะและการจัดเตรียมข้อมูลเบื้องต้นสำหรับแต่ละชุดคุณลักษณะย่อยที่ดำเนินการคัดเลือกมา

3.8 การสร้างแบบจำลองจากการคัดเลือกคุณลักษณะและการประเมินผลประสิทธิภาพ

ในการดำเนินการของขั้นตอนนี้จะคล้ายคลึงกับการดำเนินการในขั้นตอนการสร้างแบบจำลองพร้อมจัดการความไม่สมดุลกันของข้อมูลและการประเมินผลประสิทธิภาพ โดยจะมีความแตกต่างคือในการดำเนินการสร้างแบบจำลองของขั้นตอนนี้จะมีการใช้งานชุดคุณลักษณะ

ย่อยที่ทำการคัดเลือกมาแล้ว ซึ่งจะส่งผลต่อการสร้างแบบจำลองในแง่ดีในเรื่องของประสิทธิภาพทางด้านเวลา และทรัพยากรที่จะมีการใช้งานที่ลดลงเมื่อเปรียบเทียบกับการสร้างแบบจำลองที่ใช้งานทุกคุณลักษณะ

ในการประเมินผลของประสิทธิภาพของแบบจำลองในขั้นตอนนี้จะมีการใช้งาน Classification Report และ Confusion Matrix เช่นเดียวกับการประเมินผลในขั้นตอนก่อนหน้า

นอกจากนี้จะดำเนินการเปรียบเทียบประสิทธิภาพของแบบจำลองระหว่างแบบจำลองที่มีการคัดเลือกคุณลักษณะและแบบจำลองไม่มีการคัดเลือกคุณลักษณะในด้านของประสิทธิภาพความแม่นยำ และเวลาที่ใช้ในการเรียนรู้ของแบบจำลอง

3.9 การอธิบายแบบจำลองโดยใช้การเรียนรู้ด้วยเครื่องแบบอธิบายได้

ในการอธิบายแบบจำลองในขั้นตอนนี้จะมีการใช้งานวิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธีการแบบ SHAP ในการอธิบายทั้งในระดับแบบจำลองและระดับรายตัวอย่างข้อมูล ซึ่งจะทำให้สามารถเปรียบเทียบความสำคัญของคุณลักษณะทั้งในระดับแบบจำลองและความสำคัญของคุณลักษณะในระดับรายตัวอย่างข้อมูลได้ ซึ่งอาจจะมีผลลัพธ์ที่แตกต่างกันออกไป

โดยการอธิบายในระดับแบบจำลองจะเป็นการหาค่าความสำคัญของคุณลักษณะโดยเฉลี่ยจากทุกตัวอย่างข้อมูลในชุดข้อมูล ซึ่งจะสามารถช่วยอธิบายได้ว่าคุณลักษณะได้ที่ส่งผลต่อการทำนายของแบบจำลองและส่งผลมากน้อยเพียงใด ส่วนการอธิบายในระดับรายตัวอย่างข้อมูลจะเป็นการอธิบายว่าด้วยค่าของคุณลักษณะนั้นๆ ของตัวอย่างข้อมูล จะมีความสำคัญมากหรือน้อยเพียงใดต่อการทำให้แบบจำลองเลือกตัดสินใจว่าตัวอย่างข้อมูลนั้นๆ จะถูกทำนายไปยังกลุ่มผลลัพธ์ใด

3.10 การวิเคราะห์ความผิดพลาดของแบบจำลอง

ในส่วนของการวิเคราะห์ความผิดพลาดของแบบจำลองจะเป็นขั้นตอนในการคัดเลือกแบบจำลองที่มีประสิทธิภาพเป็นที่น่าพึงพอใจมาวิเคราะห์เพื่อค้นหาสาเหตุว่าเหตุใดจึงมีการทำนายบางตัวอย่างข้อมูลผิดพลาดไป โดยการวิเคราะห์ความผิดพลาดในขั้นตอนนี้จะมีลักษณะเป็นการอธิบายด้วยข้อมูลจริง เพื่อวิเคราะห์ว่าด้วยคุณลักษณะและค่าของตัวอย่างข้อมูลนั้นๆ มีลักษณะอย่างไร และส่งผลให้แบบจำลองทำงานผิดพลาดได้หรือไม่

ซึ่งจากการตั้งข้อสันนิษฐานเบื้องต้นได้มีการตั้งประเด็นหลักไว้ว่าอาจมีสาเหตุมาจากการที่คุณลักษณะของตัวอย่างข้อมูลที่มีการทำนายผิดพลาดมีลักษณะที่ค่อนข้างใกล้เคียงกับข้อมูล

ส่วนใหญ่ในบริเวณเดียวกันซึ่งมีข้อมูลผลลัพธ์คนละกลุ่มกัน ส่งผลให้แบบจำลองไม่สามารถจำแนกตัวอย่างข้อมูลได้

3.11 การสรุปผลการวิจัยและการอภิปรายการวิจัย

ในส่วนของผลการวิจัย การสรุปผลการวิจัย และการอภิปรายผลการวิจัย จะมีการดำเนินการหลังจากการทดลองได้เสร็จสิ้นลงแล้ว

บทที่ 4
ผลการดำเนินงานวิจัย

ในการวิจัยเกี่ยวกับการใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคารครั้งนี้ ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามกระบวนการและขั้นตอนต่างๆ ตลอดจนการประเมินประสิทธิภาพของแบบจำลอง เพื่อให้บรรลุวัตถุประสงค์ของการวิจัยที่ได้ตั้งเป้าหมายไว้ ผู้วิจัยได้มีการแบ่งหัวข้อต่างๆ สำหรับการนำเสนอผลการดำเนินงานวิจัย โดยมีรายละเอียดดังต่อไปนี้

1. มาตรวัดประสิทธิภาพของแบบจำลอง

2. การเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะ

3. การเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งมีการใช้งานกลุ่มคุณลักษณะย่อยจากชุดข้อมูล

4. ชุดของคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับจากการคัดเลือกคุณลักษณะแบบต่างๆ

5. การเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับ

โดยในขั้นตอนการดำเนินการสร้างแบบจำลอง มีการใช้งานหลักการ GridSearch ร่วมกับ Cross Validation สำหรับการค้นหาค่าตัวแปรของแบบจำลองซึ่งจะส่งผลให้แบบจำลองมีประสิทธิภาพสูงที่สุด ร่วมกับการจัดการความไม่สมดุลกันของข้อมูลด้วยหลักการ Random Undersampling, SMOTE และ Class Weight เพื่อทำให้แบบจำลองไม่เกิดความโน้มเอียงไปยังข้อมูลกลุ่มใหญ่ ซึ่งแบบจำลองที่นำมาใช้ในการดำเนินการวิจัยประกอบด้วย Logistic Regression, Random Forest, LightGBM และ XGBoost

ในส่วนของการทำวิศวกรรมคุณลักษณะและการจัดเตรียมข้อมูลเบื้องต้น ได้ใช้งานหลักการ One-Hot Encoding, CatBoost Encoding, BaseN Encoding ในการจัดการกับคุณลักษณะชนิดประเภทแบบไม่มีลำดับ มีการใช้งานหลักการ Ordinal Encoding สำหรับจัดการคุณลักษณะชนิดประเภทแบบมีลำดับ ใช้งานหลักการ Standard Scaling สำหรับจัดการคุณลักษณะชนิดตัวเลข และใช้งานหลักการ Label Encoding สำหรับจัดการข้อมูลผลลัพธ์ของการทำนาย

นอกจากนี้ในการสร้างแบบจำลองได้มีการกำหนดค่าเพื่อให้แบบจำลองมุ่งเน้นประสิทธิภาพไปยังค่า Recall สำหรับการตรวจจับลูกค้าที่มีแนวโน้มสมัครผลิตภัณฑ์เงินฝากประจำ โดยการประเมินประสิทธิภาพของแบบจำลองจะมีการใช้งาน Classification Report ร่วมกับ Confusion Matrix โดยมุ่งเน้นไปยังค่า Recall เป็นหลัก โดยที่ค่าประสิทธิภาพของมาตรวัดอื่นๆ โดยรวมของแบบจำลองยังอยู่ในเกณฑ์ที่น่าพึงพอใจ

4.1 มาตรวัดประสิทธิภาพของแบบจำลอง

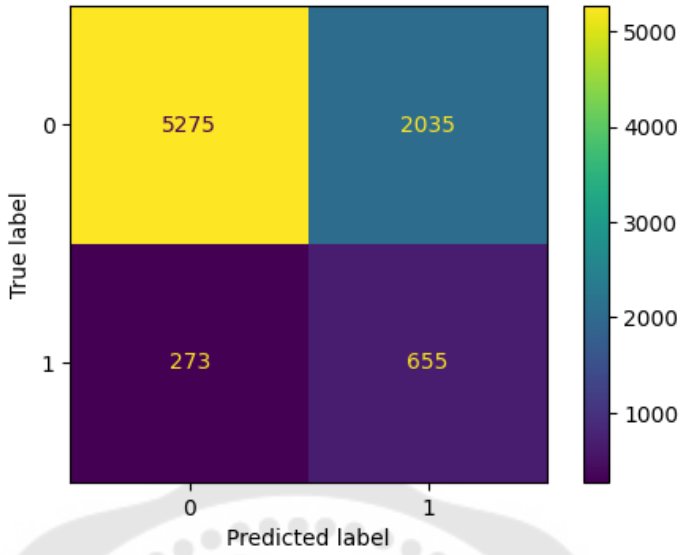
ในการประเมินประสิทธิภาพของแบบจำลองของงานในการจำแนกประเภทแบบไบนารีสำหรับการเรียนรู้ด้วยเครื่องแบบมีผู้สอนจะสามารถแบ่งผลของการทำนายออกได้เป็น 4 กลุ่มตามความถูกต้องของการทำนายและป้ายกำกับของข้อมูล ซึ่งผลของการทำนายทั้ง 4 กลุ่มประกอบด้วย

- จำนวนของการทำนายได้ถูกต้องของกลุ่มตัวอย่างที่สนใจ (TP)
- จำนวนของการทำนายผิดของกลุ่มตัวอย่างที่สนใจ (FP)
- จำนวนของการทำนายได้ถูกต้องของกลุ่มตัวอย่างที่ไม่ได้สนใจ (TN)
- จำนวนของการทำนายผิดของกลุ่มตัวอย่างที่ไม่ได้สนใจ (FN)

โดยในการประเมินประสิทธิภาพของแบบจำลองในการวิจัยครั้งนี้มีการใช้งานมาตรวัดต่างๆ ซึ่งเหมาะสมกับงาน ดังนี้

4.1.1 Confusion Matrix

Confusion Matrix เป็นตารางที่สามารถแสดงถึงจำนวนของผลลัพธ์การทำนายตัวอย่างข้อมูล ว่าสามารถทำนายได้ถูกต้องหรือทำนายผิดมากน้อยเพียงใด โดยค่าที่แสดงใน Confusion Matrix จะประกอบด้วยค่าของ TP, FP, TN และ FN



ภาพประกอบ 49 ตัวอย่างของ Confusion Matrix ในการแสดงผลลัพธ์จากการทำนายของแบบจำลอง

4.1.2 Accuracy หรือ ค่าความแม่นยำ

เป็นการประเมินประสิทธิภาพโดยรวมของแบบจำลองว่ามีความแม่นยำมากน้อยเพียงใด โดยเป็นการคำนวณรวมจากทุกการทำนายถูกและผิดของทุกตัวอย่างข้อมูล

4.1.3 Recall หรือ ค่าความครบถ้วนของกลุ่มข้อมูลที่สนใจ

การประเมินประสิทธิภาพด้วยค่า Recall จะเป็นการประเมินประสิทธิภาพของแบบจำลองว่าสามารถทำนายตัวอย่างของกลุ่มข้อมูลที่สนใจ หรือกลุ่มบวก (Positive Class หรือ Class 1) ได้ครบถ้วนมากน้อยเพียงใด

4.1.4 Precision หรือ ค่าความถูกต้องของกลุ่มข้อมูลที่สนใจ

การประเมินประสิทธิภาพด้วยค่า Precision จะเป็นการเป็นการประเมินประสิทธิภาพของแบบจำลองว่าสามารถทำนายตัวอย่างของกลุ่มข้อมูลที่สนใจ หรือกลุ่มบวก (Positive Class หรือ Class 1) ได้ถูกต้องมากน้อยเพียงใด

4.1.5 F1-Score หรือ ค่าเฉลี่ยระหว่าง Recall และ Precision.

การประเมินประสิทธิภาพด้วยค่า F1-Score จะเป็นการเป็นการประเมินประสิทธิภาพของแบบจำลองซึ่งบ่งบอกถึงค่าเฉลี่ยแบบ Harmonic ระหว่างค่าของ Recall และ Precision

4.2 การเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะ

ผลของการดำเนินงานวิจัยในหัวข้อนี้จะประกอบด้วยแบบจำลองจำนวนมากซึ่งเกิดจากการประยุกต์ใช้หลักการต่างๆ ที่ทำการศึกษาเข้าด้วยกัน ซึ่งแบบจำลองจะมีการใช้งานคุณลักษณะทั้งหมดจากชุดข้อมูลสำหรับการทดสอบ โดยไม่ได้มีการดำเนินการคัดเลือกคุณลักษณะ ซึ่งจำนวนคุณลักษณะก่อนการทำวิศวกรรมคุณลักษณะมีประกอบด้วย 19 คุณลักษณะ

4.2.1 แบบจำลองซึ่งใช้งาน One-Hot Encoding จัดการคุณลักษณะชนิดประเภท

เป็นแบบจำลองซึ่งมีการทำวิศวกรรมคุณลักษณะกับตัวแปรชนิดประเภทด้วยวิธีการ One-Hot Encoding ซึ่งส่งผลให้มีจำนวนคุณลักษณะเพิ่มขึ้นจาก 19 เป็น 56 คุณลักษณะ โดยมีการสร้างแบบจำลองทั้งสิ้น 12 แบบจำลองประกอบไปด้วย

1. Logistic Regression ที่ใช้งาน Class Weight เพื่อจัดการความไม่สมดุลกันของข้อมูล (LR-ClassW)

2. Logistic Regression ที่ใช้งาน Random Undersampling เพื่อจัดการความไม่สมดุลกันของข้อมูล (LR-Under)

3. Logistic Regression ที่ใช้งาน SMOTE เพื่อจัดการความไม่สมดุลกันของข้อมูล (LR-Smote)

4. Random Forest ที่ใช้งาน Class Weight เพื่อจัดการความไม่สมดุลกันของข้อมูล (RF-ClassW)

5. Random Forest ที่ใช้งาน Random Undersampling เพื่อจัดการความไม่สมดุลกันของข้อมูล (RF-Under)

6. Random Forest ที่ใช้งาน SMOTE เพื่อจัดการความไม่สมดุลกันของข้อมูล (RF-Smote)

7. LightGBM ที่ใช้งาน Class Weight เพื่อจัดการความไม่สมดุลกันของข้อมูล (LGBM-ClassW)

8. LightGBM ที่ใช้งาน Random Undersampling เพื่อจัดการความไม่สมดุลกันของข้อมูล (LGBM-Under)

9. LightGBM ที่ใช้งาน SMOTE เพื่อจัดการความไม่สมดุลกันของข้อมูล (LGBM-Smote)

10. XGBoost ที่ใช้งาน Class Weight เพื่อจัดการความไม่สมดุลกันของข้อมูล (XGB-ClassW)

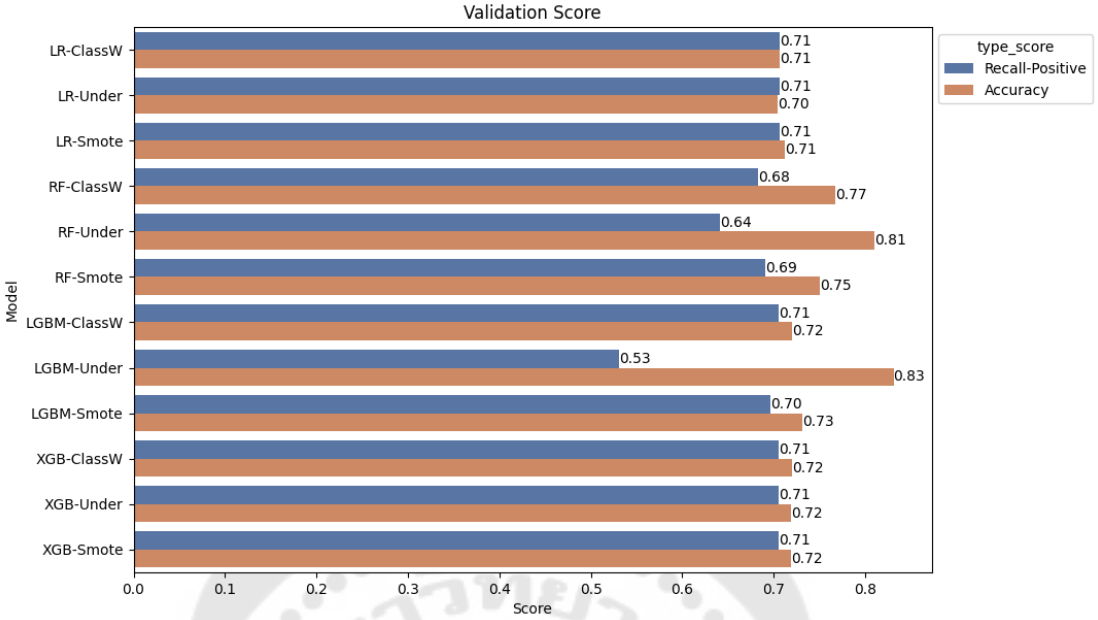
11. XGBoost ที่ใช้งาน Random Undersampling เพื่อจัดการความไม่สมดุลกันของข้อมูล (XGB-Under)

12. XGBoost ที่ใช้งาน SMOTE เพื่อจัดการความไม่สมดุลกันของข้อมูล (XGB-Smote)

ในการประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลอง ได้มุ่งเน้นไปยังค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และ Accuracy ของแบบจำลอง เพื่อค้นหาแบบจำลองที่มีประสิทธิภาพในการจำแนกลูกค้าที่มีแนวโน้มในการสมัครผลิตภัณฑ์ได้อย่างครบถ้วน โดยที่ความแม่นยำของแบบจำลองยังอยู่ในเกณฑ์ที่น่าพอใจ

ภาพประกอบที่ 50 แสดงการเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และค่า Accuracy ของแบบจำลองทั้ง 12 แบบจำลอง จากกราฟจะเห็นได้ว่าแบบจำลองส่วนใหญ่มีประสิทธิภาพของค่า Recall และ Accuracy อยู่ที่ประมาณ 0.70 (70%) โดยแบบจำลอง RF-ClassW, RF-Under และ LGBM-Under ซึ่งแม้จะมีค่าของ Recall ที่ลดลงแต่จะส่งผลให้มีค่าของ Accuracy ที่เพิ่มสูงมากขึ้น โดยสามารถสังเกตได้ว่าเมื่อแบบจำลองมีค่าของ Recall ที่ลดลงมากขึ้น จะยิ่งทำให้แบบจำลองมีค่าของ Accuracy ที่สูงเพิ่มขึ้นตามไปด้วย

แบบจำลองที่มีความแม่นยำสูงที่สุดได้แก่ LGBM-Under โดยมีค่า Accuracy อยู่ที่ 0.83 (83%) และมีค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) อยู่ที่ 0.53 (53%) โดยค่า Accuracy ที่เพิ่มสูงขึ้นเกิดจากการแปรผกผันกับค่า Recall ที่สนใจ เนื่องจากชุดข้อมูลมีความไม่สมดุลกันอย่างมากโดยกลุ่มข้อมูลส่วนใหญ่เป็นลูกค้าที่ไม่ได้ทำการสมัครผลิตภัณฑ์ ซึ่งการทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ครบถ้วนลดลงจึงส่งผลให้การทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ซึ่งเป็นกลุ่มข้อมูลส่วนใหญ่มีจำนวนที่ถูกต้องมากยิ่งขึ้น ดังนั้นจึงส่งผลให้ค่า Accuracy ของแบบจำลองเพิ่มสูงขึ้น ส่วนแบบจำลองที่มีความแม่นยำรองลงมา ได้แก่ RF-Under และ RF-ClassW โดยมีค่า Accuracy เป็น 0.81 (81%) และ 0.77 (77%) ตามลำดับ และในส่วนของค่า Recall เป็น 0.64 (64%) และ 0.68 (68%) ตามลำดับ



ภาพประกอบ 50 การเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ และค่า Accuracy ของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะและจัดการคุณลักษณะชนิดประเภทด้วยวิธีการ One-Hot Encoding

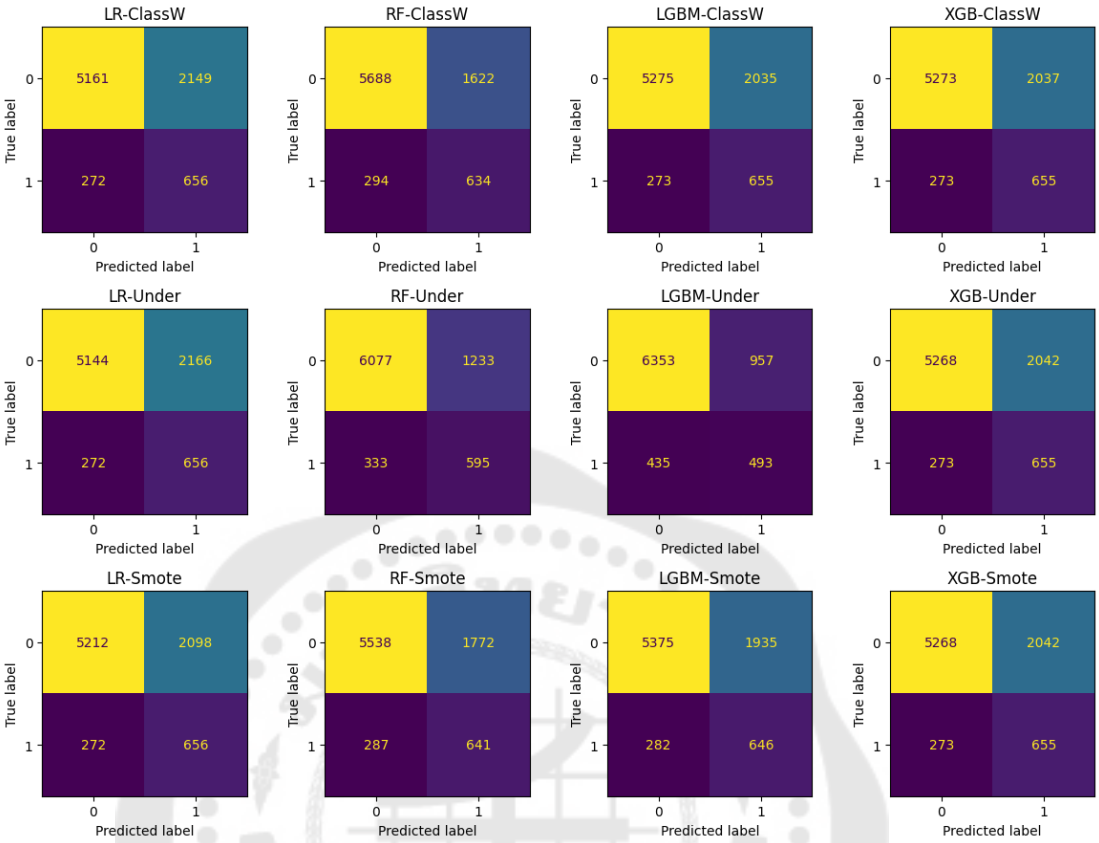
จากภาพประกอบที่ 51 แสดง Confusion Matrix ของจำนวนการทำนายทั้งถูกและผิดของแบบจำลองซึ่งดำเนินการกับชุดข้อมูลสำหรับการทดสอบ โดยประกอบด้วยจำนวนข้อมูลทั้งหมด 8,238 ตัวอย่าง แบ่งเป็นตัวอย่างข้อมูลของลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) จำนวน 928 ตัวอย่าง และตัวอย่างข้อมูลของลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) จำนวน 7,310 ตัวอย่าง

โดยแบบจำลอง LGBM-Under ซึ่งมีค่า Accuracy สูงสุดอยู่ที่ 0.83 (83%) สามารถทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้องสูงที่สุดเป็นจำนวน 6,353 ตัวอย่าง แต่สามารถทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้เพียง 493 ตัวอย่าง ซึ่งถือว่าต่ำที่สุดจากทั้ง 12 แบบจำลอง โดยจำนวนตัวอย่างข้อมูลที่ทำนายได้ถูกต้อง (TP + TN) มีผลรวมเป็น 6,846 ตัวอย่าง ซึ่งเมื่อนำมาคำนวณร่วมกับจำนวนตัวอย่างข้อมูลในชุดข้อมูลสำหรับทดสอบทั้งหมด (TP + FP + TN + FN) ที่มีจำนวน 8,233 ตัวอย่าง เพื่อหาความแม่นยำ จะส่งผลให้ได้ค่า Accuracy ออกมาที่ 0.83 (83%) และเมื่อนำจำนวนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ที่ทำนายได้ถูกต้อง (TP) จำนวน 493 ตัวอย่าง มาคำนวณร่วมกับจำนวน

ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ทั้งหมดของชุดข้อมูลสำหรับทดสอบ (TP + FN) ที่มีจำนวน 928 ตัวอย่าง ส่งผลให้ได้ค่า Recall ออกมาที่ 0.53 (53%)

ในส่วนของแบบจำลองอื่นๆ ซึ่งมีค่า Accuracy และค่า Recall ที่ประมาณ 0.70 (70%) พบว่าจำนวนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ที่ทำนายได้ถูกต้อง (TP) อยู่ระหว่าง 641 ถึง 656 ตัวอย่างข้อมูล จำนวนของกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ที่ทำนายได้ถูกต้อง (TN) อยู่ระหว่าง 5,144 ถึง 5,538 ตัวอย่างข้อมูล และจำนวนที่ทำนายได้ถูกต้องของทั้งสองกลุ่มข้อมูล (TP + TN) อยู่ระหว่าง 5,800 ถึง 6,179 ตัวอย่างข้อมูล

จากข้อมูลใน Confusion Matrix แสดงให้เห็นว่าในแบบจำลองซึ่งใช้งานอัลกอริทึม Logistic Regression และ XGBoost ให้ผลของจำนวนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ที่ทำนายได้ถูกต้อง (TP) ออกมาเท่ากันในแต่ละอัลกอริทึมนั้นๆ โดยทำนายได้ถูกต้อง 656 และ 655 ตัวอย่างข้อมูลตามลำดับ และเมื่อลองเปรียบเทียบภายในแต่ละอัลกอริทึมด้วยจำนวนของกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ที่ทำนายได้ถูกต้อง (TN) จะพบว่าแบบจำลอง LR-Smote ซึ่งดำเนินการจัดการความไม่สมดุลกันของข้อมูลด้วยวิธีการแบบ SMOTE ให้ประสิทธิภาพที่สูงกว่าแบบจำลองอื่นๆ ภายใต้อัลกอริทึมเดียวกัน โดยทำนายได้ถูกต้องถึง 5,212 ตัวอย่าง ส่วนแบบจำลอง XGB-ClassW ซึ่งเป็นการให้อัลกอริทึม XGBoost ปรับค่าน้ำหนักของตัวอย่างข้อมูลเพื่อจัดการความไม่สมดุลกันของข้อมูล ให้ประสิทธิภาพที่สูงกว่าแบบจำลองอื่นๆ ภายใต้อัลกอริทึมเดียวกัน โดยทำนายได้ถูกต้องถึง 5,273 ตัวอย่าง



ภาพประกอบ 51 การเปรียบเทียบ Confusion Matrix ของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะและจัดการคุณลักษณะชนิดประเภทด้วยวิธีการ One-Hot Encoding

```
Classification Report :
              precision    recall  f1-score   support

           0       0.95      0.71      0.81      7310
           1       0.23      0.71      0.35       928

    accuracy                           0.71      8238
   macro avg       0.59      0.71      0.58      8238
weighted avg       0.87      0.71      0.76      8238
```

ภาพประกอบ 52 ผลลัพธ์จากการเรียกใช้งานฟังก์ชัน Classification Report ของไลบรารี Scikit-Learn เพื่อแสดงผลมาตรวัดประสิทธิภาพต่างๆ ของแบบจำลอง

ตารางที่ 8 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลอง ประกอบไปด้วย Accuracy, Recall, Precision และ F1-Score ของทั้งสองกลุ่มของตัวอย่างข้อมูล นอกจากนี้ยังแสดงค่า

ประสิทธิภาพของ Recall, Precision และ F1-Score โดยรวมของแบบจำลองแบบ Weighted Average ซึ่งจะมีการคำนวณโดยคำนึงถึงสัดส่วนของตัวอย่างข้อมูลในแต่ละกลุ่ม

ตาราง 8 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะและจัดการคุณลักษณะชนิดประเภทด้วยวิธีการ One-Hot Encoding

แบบจำลอง	Acc	Precision			Recall			F1-Score		
		C-0	C-1	W-avg	C-0	C-1	W-avg	C-0	C-1	W-avg
LR-ClassW	0.71	0.95	0.23	0.87	0.71	0.71	0.71	0.81	0.35	0.76
LR-Under	0.70	0.95	0.23	0.87	0.70	0.71	0.70	0.81	0.35	0.76
LR-Smote	0.71	0.95	0.24	0.87	0.71	0.71	0.71	0.81	0.36	0.76
RF-ClassW	0.77	0.95	0.28	0.88	0.78	0.68	0.77	0.86	0.40	0.80
RF-Under	0.81	0.95	0.33	0.88	0.83	0.64	0.81	0.89	0.43	0.83
RF-Smote	0.75	0.95	0.27	0.87	0.76	0.69	0.75	0.84	0.38	0.79
LGBM-ClassW	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
LGBM-Under	0.83	0.94	0.34	0.87	0.87	0.53	0.83	0.90	0.41	0.85
LGBM-Smote	0.73	0.95	0.25	0.87	0.74	0.70	0.73	0.83	0.37	0.78
XGB-ClassW	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB -Under	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB -Smote	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77

4.2.2 แบบจำลองซึ่งใช้งานหลักการอื่น ๆ จัดการคุณลักษณะชนิดประเภท

เป็นแบบจำลองซึ่งมีการทำวิศวกรรมคุณลักษณะกับตัวแปรชนิดประเภทด้วยวิธีการอื่นๆ ที่นอกเหนือจาก One-Hot Encoding ซึ่งประกอบด้วยวิธีการแบบ CatBoost Encoding และ BaseN Encoding โดยมีการใช้งานร่วมกับการจัดการความไม่สมดุลกันของข้อมูลด้วยวิธีการ

ต่างๆ ซึ่งให้ประสิทธิภาพที่ดีแตกต่างกันไปตามแต่ละอัลกอริทึมของแบบจำลองจากหัวข้อก่อนหน้า

CatBoost Encoding เป็นวิธีการในการจัดการคุณลักษณะชนิดประเภทซึ่งพัฒนาต่อยอดจากหลักการของ Target Encoding ซึ่งจะดำเนินการแทนที่คุณลักษณะชนิดประเภทด้วยค่าที่คำนวณจากความน่าจะเป็นของผลลัพธ์ที่สอดคล้องกับค่าของข้อมูลนั้นๆ ในแต่ละคุณลักษณะร่วมกับค่าความน่าจะเป็นของผลลัพธ์ในชุดข้อมูล ซึ่งคำนวณได้จากสมการดังนี้

$$Value = \frac{TargetSum + prior}{FeatureCount + 1} \qquad (7)$$

โดย TargetSum คือ ค่าผลรวมของผลลัพธ์ (เช่น 0 หรือ 1 ของผลลัพธ์แบบไบนารี) ที่สอดคล้องกับค่าของข้อมูลนั้นๆ ภายใต้คุณลักษณะนั้นๆ ที่กำลังดำเนินการของชุดข้อมูลสำหรับการเรียนรู้ ส่วน Prior คือ ค่าคงที่ซึ่งคำนวณได้จากการหาค่าผลรวมของผลลัพธ์ของทั้งชุดข้อมูลสำหรับการเรียนรู้ ด้วยจำนวนตัวอย่างข้อมูลทั้งหมดของชุดข้อมูลสำหรับการเรียนรู้ และ FeatureCount คือ จำนวนของตัวอย่างข้อมูลที่ปรากฏค่าที่สนใจภายใต้คุณลักษณะนั้นๆ ที่กำลังดำเนินการ

ซึ่ง CatBoost Encoding มีการปรับปรุงในส่วนของการใช้งานเพียงแค่ตัวอย่างข้อมูลก่อนหน้าที่เคยเห็นมาแล้วมาใช้สำหรับการคำนวณแทนที่การใช้งานทั้งชุดข้อมูล เพื่อเป็นการลดการเกิดการรั่วไหลของข้อมูลของ Target Encoding

BaseN Encoding เป็นวิธีการในการจัดการคุณลักษณะชนิดประเภทซึ่งมีความคล้ายคลึงกับการใช้งานวิธีการแบบ Binary Encoding ซึ่งจะทำการแปลงค่าของข้อมูลที่แตกต่างกันในแต่ละคุณลักษณะให้เป็น Bit String (0 และ 1) โดยจะทำการเพิ่ม Dummy Feature เพื่อขยายความยาวของ bit string เมื่อมีขนาดไม่เพียงพอสำหรับการรองรับค่าที่เป็นไปได้ทั้งหมดในคุณลักษณะนั้นๆ ส่วนวิธีการแบบ BaseN จะสามารถกำหนดจำนวนของค่าที่เพิ่มขึ้นได้มากกว่า 1 เช่น เมื่อกำหนดค่า N = 5 จะส่งผลให้แต่ละ bit สามารถรองรับค่าได้ตั้งแต่ 0-4 ทำให้สามารถลดจำนวนของ Dummy Feature ในการรองรับค่าที่เป็นไปได้ทั้งหมดในคุณลักษณะนั้นๆ ได้ ซึ่งในงานวิจัยครั้งนี้ได้ใช้งานค่า N = 5

โดยในขั้นตอนการวิจัยนี้มีการสร้างแบบจำลองทั้งสิ้น 12 แบบจำลองประกอบด้วย

1. Logistic Regression ที่ใช้งาน SMOTE และ CatBoost Encoding (LR-Smote-Cb)

2. Logistic Regression ที่ใช้งาน Random Undersampling และ CatBoost Encoding (LR-Under-Cb)

3. Logistic Regression ที่ใช้งาน SMOTE และ BaseN Encoding (LR-Smote-Bn)

4. Random Forest ที่ใช้งาน SMOTE และ CatBoost Encoding (RF-Smote-Cb)

5. Random Forest ที่ใช้งาน Random Undersampling และ CatBoost Encoding (RF-Under-Cb)

6. Random Forest ที่ใช้งาน SMOTE และ BaseN Encoding (RF-Smote-Bn)

7. LightGBM ที่ใช้งาน Class Weight และ CatBoost Encoding (LGBM-ClassW-Cb)

8. LightGBM ที่ใช้งาน Random Undersampling และ CatBoost Encoding (LGBM-Under-Cb)

9. LightGBM ที่ใช้งาน Class Weight และ BaseN Encoding (LGBM-ClassW-Bn)

10. XGBoost ที่ใช้งาน Class Weight และ CatBoost Encoding (XGB-ClassW-Cb)

11. XGBoost ที่ใช้งาน Random Undersampling และ CatBoost Encoding (XGB-Under-Cb)

12. XGBoost ที่ใช้งาน Class Weight และ BaseN Encoding (XGB-ClassW-Bn)

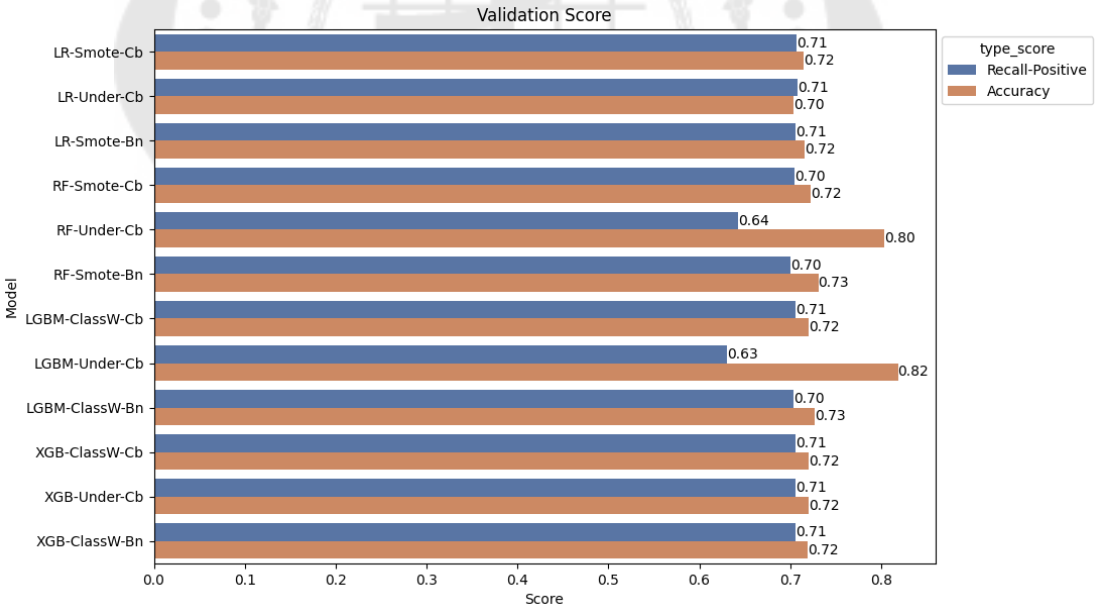
เช่นเดียวกับการประเมินประสิทธิภาพในหัวข้อก่อนหน้า การประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลองได้มุ่งเน้นไปยังค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และ Accuracy ของแบบจำลอง เพื่อค้นหาแบบจำลองที่มีประสิทธิภาพในการจำแนกลูกค้าที่มีแนวโน้มในการสมัครผลิตภัณฑ์ได้อย่างครบถ้วน โดยที่ความแม่นยำของแบบจำลองยังอยู่ในเกณฑ์ที่น่าพอใจ

ภาพประกอบที่ 53 แสดงการเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และค่า Accuracy ของแบบจำลองทั้ง 12 แบบจำลอง จากกราฟจะเห็นได้ว่าการเปลี่ยนแปลงการใช้งานวิธีการในการทำวิศกรรมคุณลักษณะจาก One-Hot Encoding เป็น

วิธีการอื่นๆ ไม่ได้ส่งผลที่ชัดเจนในการเปลี่ยนแปลงของประสิทธิภาพทั้งค่า Recall และ Accuracy โดยแบบจำลองส่วนใหญ่ยังคงได้ประสิทธิภาพของค่า Recall และ Accuracy อยู่ที่ประมาณ 0.70 (70%) และประสิทธิภาพของแบบจำลอง RF-Under-Cb ก็มีค่าที่ใกล้เคียงกับการทำ One-Hot Encoding โดยมีค่า Recall อยู่ที่ 0.64 (64%) และมีค่า Accuracy อยู่ที่ 0.80 (80%)

แบบจำลองที่แสดงผลการเปลี่ยนแปลงของประสิทธิภาพมากที่สุดเมื่อเปลี่ยนวิธีการทำวิศวกรรมคุณลักษณะ คือแบบจำลอง LGBM-Under-Cb ซึ่งเป็นการเปลี่ยนวิธีจัดการกับคุณลักษณะชนิดประเภทจาก One-Hot Encoding มาใช้งาน CatBoost Encoding โดยแม้ว่าค่า Accuracy จะลดลงเล็กน้อยจาก 0.83 (83%) เหลือ 0.82 (82%) แต่ยังคงเป็นแบบจำลองที่มีค่า Accuracy ที่สูงที่สุด และยังสามารถเพิ่มค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้อย่างมากจาก 0.53 (53%) ขึ้นมาเป็น 0.63 (63%)

โดยแบบจำลองซึ่งใช้อัลกอริทึม XGBoost ทั้ง 3 แบบจำลอง ยังคงค่าของให้ประสิทธิภาพเท่าเดิมเมื่อเปรียบเทียบกับการใช้งานหลักการ One-Hot Encoding ในการจัดการคุณลักษณะชนิดประเภท



ภาพประกอบ 53 การเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ และค่า Accuracy ของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะและจัดการคุณลักษณะชนิดประเภทด้วยวิธีการ CatBoost Encoding หรือ BaseN Encoding

จากภาพประกอบที่ 54 แสดง Confusion Matrix ของจำนวนการทำนายทั้งถูกและผิดของแบบจำลองซึ่งดำเนินการกับชุดข้อมูลสำหรับการทดสอบ โดยประกอบด้วยจำนวนข้อมูลทั้งหมด 8,238 ตัวอย่าง แบ่งเป็นตัวอย่างข้อมูลของลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) จำนวน 928 ตัวอย่าง และตัวอย่างข้อมูลของลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) จำนวน 7,310 ตัวอย่าง

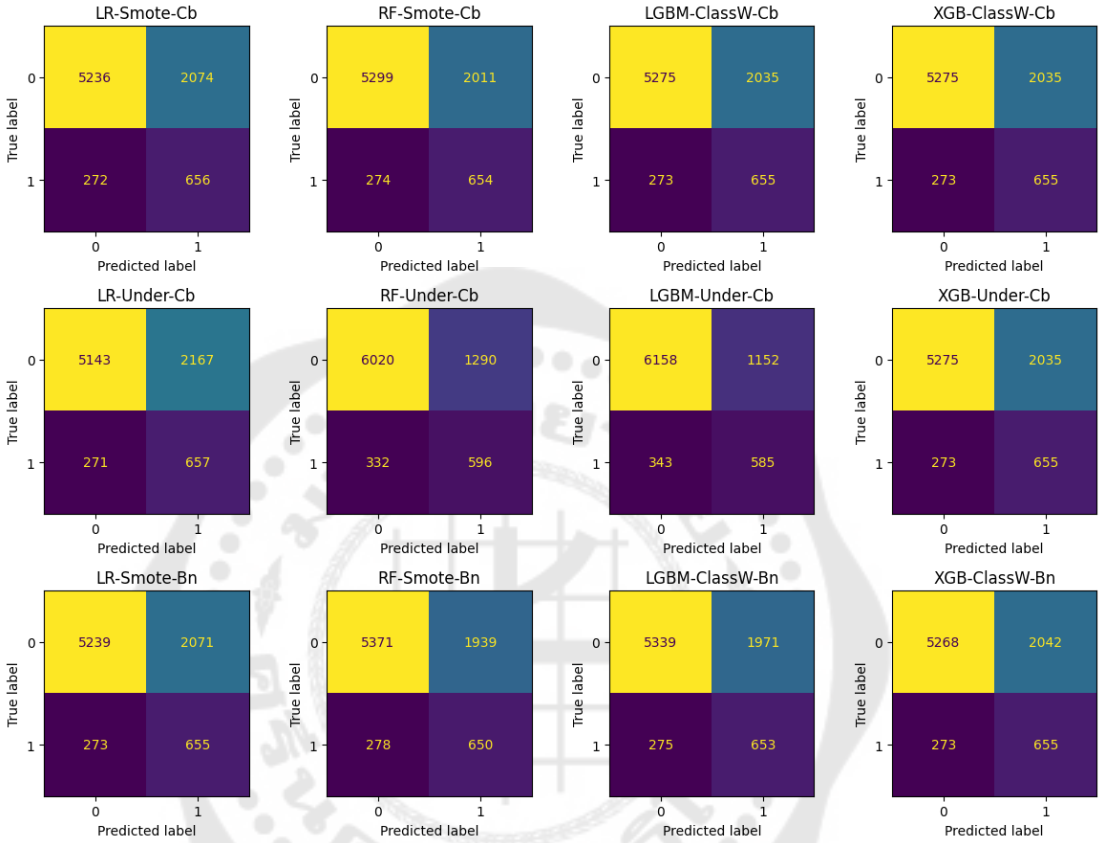
โดยแบบจำลอง LGBM-Under-Cb ซึ่งมีค่า Accuracy สูงสุดที่ 0.82 (82%) สามารถทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้องสูงที่สุดเป็นจำนวน 6,158 ตัวอย่าง ซึ่งลดลงเล็กน้อยจากแบบจำลองที่ใช้งาน One-Hot Encoding ที่ทำนายถูกที่ 6,353 ตัวอย่าง แต่สามารถทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้สูงขึ้นเป็น 585 ตัวอย่าง จากเดิมที่ 493 ตัวอย่าง ซึ่งทำนายถูกต้องสูงขึ้น 92 ตัวอย่าง โดยจำนวนตัวอย่างข้อมูลที่ทำนายได้ถูกต้อง (TP + TN) มีผลรวมเป็น 6,743 ตัวอย่าง ซึ่งลดลงเล็กน้อยจาก 6,846 ตัวอย่าง ซึ่งเมื่อนำมาคำนวณร่วมกับจำนวนตัวอย่างข้อมูลในชุดข้อมูลสำหรับทดสอบทั้งหมด (TP + FP + TN + FN) ที่มีจำนวน 8,233 ตัวอย่าง เพื่อหาความแม่นยำ จะส่งผลให้ได้ค่า Accuracy ออกมาที่ 0.82 (82%) และเมื่อนำจำนวนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ที่ทำนายได้ถูกต้อง (TP) จำนวน 585 ตัวอย่าง มาคำนวณร่วมกับจำนวนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ทั้งหมดของชุดข้อมูลสำหรับทดสอบ (TP + FN) ที่มีจำนวน 928 ตัวอย่าง ส่งผลให้ได้ค่า Recall ออกมาที่ 0.63 (63%)

จากข้อมูลใน Confusion Matrix แสดงให้เห็นว่าในแบบจำลองส่วนใหญ่ซึ่งใช้งานอัลกอริทึม Logistic Regression และ XGBoost ให้ผลของจำนวนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ที่ทำนายได้ถูกต้อง (TP) ออกมาใกล้เคียงกับการใช้งาน One-Hot Encoding แต่สามารถเพิ่มประสิทธิภาพของการทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) สูงมากขึ้น โดยสังเกตได้จากแบบจำลอง LR-Smote-Cb, LR-Smote-Bn, XGB-ClassW-Cb และ XGB-Under-Cb

ในทางกลับกันแบบจำลอง RF-Smote-Cb และ RF-Smote-Bn สามารถทำนายกลุ่มของลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ถูกต้อง (TP) สูงขึ้น แต่ก็ส่งผลให้การทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) ลดลงเล็กน้อย

จากผลการเปรียบเทียบประสิทธิภาพของแบบจำลองแสดงให้เห็นว่าการทำวิศวกรรมคุณลักษณะด้วยหลักการ CatBoost Encoding ให้ประสิทธิภาพในการทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ดีกว่าการทำวิศวกรรมคุณลักษณะด้วยหลักการ BaseN Encoding

และให้ประสิทธิภาพในการทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ดีกว่าการทำวิศวกรรมคุณลักษณะด้วยหลักการ BaseN Encoding เมื่อให้ผลประสิทธิภาพของการทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ที่ใกล้เคียงกัน



ภาพประกอบ 54 การเปรียบเทียบ Confusion Matrix ของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะและจัดการคุณลักษณะชนิดประเภทด้วยวิธีการ CatBoost Encoding หรือ BaseN Encoding

ตารางที่ 9 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลอง ประกอบไปด้วย Accuracy, Recall, Precision และ F1-Score ของทั้งสองกลุ่มของตัวอย่างข้อมูล นอกจากนี้ยังแสดงค่าประสิทธิภาพของ Recall, Precision และ F1-Score โดยรวมของแบบจำลองแบบ weighted average ซึ่งจะมีการคำนวณโดยคำนึงถึงสัดส่วนของตัวอย่างข้อมูลในแต่ละกลุ่ม

ตาราง 9 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลองซึ่งไม่ได้มีการใช้งานคัดเลือกคุณลักษณะและจัดการคุณลักษณะชนิดประเภทด้วยวิธีการ CatBoost Encoding หรือ BaseN Encoding

แบบจำลอง	Acc	Precision			Recall			F1-Score		
		C-0	C-1	W-avg	C-0	C-1	W-avg	C-0	C-1	W-avg
LR-Smote-Cb	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
LR-Under-Cb	0.70	0.95	0.23	0.87	0.70	0.71	0.70	0.81	0.35	0.76
LR-Smote-Bn	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
RF-Smote-Cb	0.72	0.95	0.25	0.87	0.72	0.70	0.72	0.82	0.36	0.77
RF-Under-Cb	0.80	0.95	0.32	0.88	0.82	0.64	0.80	0.88	0.42	0.83
RF-Smote-Bn	0.73	0.95	0.25	0.87	0.73	0.70	0.73	0.83	0.37	0.78
LGBM-ClassW-Cb	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
LGBM-Under-Cb	0.82	0.95	0.34	0.88	0.84	0.63	0.82	0.89	0.44	0.84
LGBM-ClassW-Bn	0.73	0.95	0.25	0.87	0.73	0.70	0.73	0.83	0.37	0.77
XGB-ClassW-Cb	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB-Under-Cb	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB-ClassW-Bn	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77

4.3 การเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งมีการใช้งานกลุ่มคุณลักษณะย่อยจากชุดข้อมูล

ภายใต้ชุดข้อมูลที่มีการนำมาใช้งานในการวิจัยนั้นประกอบด้วยคุณลักษณะทั้งหมดจำนวน 21 คุณลักษณะ แบ่งเป็นคุณลักษณะของตัวอย่างข้อมูลหรือตัวแปรต้นจำนวน 20 คุณลักษณะ และผลลัพธ์ของตัวอย่างข้อมูลหรือตัวแปรตามจำนวน 1 คุณลักษณะ โดยในการใช้งานชุดข้อมูลจะมีการตัดออกของหนึ่งคุณลักษณะซึ่งค่อนข้างส่งผลกับผลลัพธ์ของการทำนาย เนื่องจากเป็นข้อมูลที่ได้รับมาหลังจากการทราบผลลัพธ์ของการสมัครหรือไม่สมัครผลิตภัณฑ์ จึงทำให้เหลือคุณลักษณะจำนวน 19 คุณลักษณะ เพื่อนำมาใช้งานในการวิจัย

คุณลักษณะทั้ง 19 คุณลักษณะ สามารถแบ่งออกได้เป็น 3 กลุ่มข้อมูลย่อย ได้แก่ กลุ่มข้อมูลส่วนตัวของลูกค้าธนาคาร กลุ่มข้อมูลการติดต่อระหว่างธนาคารและลูกค้ารายนั้นๆ และสุดท้าย กลุ่มข้อมูลทางเศรษฐศาสตร์ในช่วงเวลานั้นๆ ที่ทำการนำเสนอผลิตภัณฑ์แก่ลูกค้า ซึ่งในแต่ละกลุ่มข้อมูลย่อยประกอบไปด้วยคุณลักษณะต่างๆ ดังนี้

1. กลุ่มข้อมูลส่วนตัวของลูกค้าธนาคาร (Personal) ประกอบด้วย อายุ (age), อาชีพ (job), สถานะสมรส (marital), ระดับการศึกษา (education), การเคยผิดนัดชำระหนี้ (default), การมีสินเชื่อเคหะสถาน (housing), การมีสินเชื่อส่วนบุคคล (loan)

2. กลุ่มข้อมูลการติดต่อระหว่างธนาคารและลูกค้ารายนั้นๆ (Contact) ประกอบด้วย ประเภทของการติดต่อ (contact), เดือนที่ติดต่อล่าสุด (month), วันในสัปดาห์ที่ติดต่อล่าสุด (day_of_week), จำนวนครั้งการติดต่อเพื่อนำเสนอผลิตภัณฑ์ (campaign), จำนวนวันที่ห่างจากการนำเสนอผลิตภัณฑ์ก่อน (pdays), จำนวนครั้งการติดต่อก่อนนำเสนอผลิตภัณฑ์ (previous), ผลลัพธ์จากการนำเสนอผลิตภัณฑ์ก่อน (poutcome)

3. กลุ่มข้อมูลทางเศรษฐศาสตร์ในช่วงเวลานั้นๆ ที่ทำการนำเสนอผลิตภัณฑ์แก่ลูกค้า (Economics) ประกอบด้วย อัตราการจ้างงานรายไตรมาส (emp.var.rate), ดัชนีราคาผู้บริโภครายเดือน (cons.price.idx), ดัชนีความเชื่อมั่นผู้บริโภครายเดือน (cons.conf.idx), อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m), จำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed)

ซึ่งในหัวข้อนี้จะมีการสร้างแบบจำลองสำหรับทำงานกับแต่ละชุดคุณลักษณะย่อยทั้งสามชุดรวมทั้งหมด 12 แบบจำลอง โดยมีการใช้งานหลักการวิศวกรรมคุณลักษณะและการจัดการความไม่สมดุลกันของข้อมูลแตกต่างกันไปในแต่ละอัลกอริทึมของแบบจำลอง โดยเลือกประยุกต์ใช้งานจากข้อมูลประสิทธิภาพและความเหมาะสมจากแบบจำลองในหัวข้อก่อนหน้า โดยทั้ง 12 แบบจำลองประกอบด้วย

1. Logistic Regression ที่ใช้งาน SMOTE และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Personal (LR-Smote-Cb-Per)

2. Logistic Regression ที่ใช้งาน SMOTE และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Contact (LR-Smote-Cb-Con)

3. Logistic Regression ที่ใช้งาน SMOTE และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Economics (LR-Smote-Cb-Eco)

4. Random Forest ที่ใช้งาน SMOTE และ BaseN Encoding ร่วมกับชุดคุณลักษณะ Personal (RF-Smote-Bn-Per)

5. Random Forest ที่ใช้งาน SMOTE และ BaseN Encoding ร่วมกับชุดคุณลักษณะ Contact (RF-Smote-Bn-Con)

6. Random Forest ที่ใช้งาน SMOTE และ BaseN Encoding ร่วมกับชุดคุณลักษณะ Economics (RF-Smote-Bn-Eco)

7. LightGBM ที่ใช้งาน Class Weight และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Personal (LGBM-ClassW-Cb-Per)

8. LightGBM ที่ใช้งาน Class Weight และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Contact (LGBM-ClassW-Cb-Con)

9. LightGBM ที่ใช้งาน Class Weight และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Economics (LGBM-ClassW-Cb-Eco)

10. XGBoost ที่ใช้งาน Random Undersampling และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Personal (XGB-Under-Cb-Per)

11. XGBoost ที่ใช้งาน Random Undersampling และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Contact (XGB-Under-Cb-Con)

12. XGBoost ที่ใช้งาน Random Undersampling และ CatBoost Encoding ร่วมกับชุดคุณลักษณะ Economics (XGB-Under-Cb-Eco)

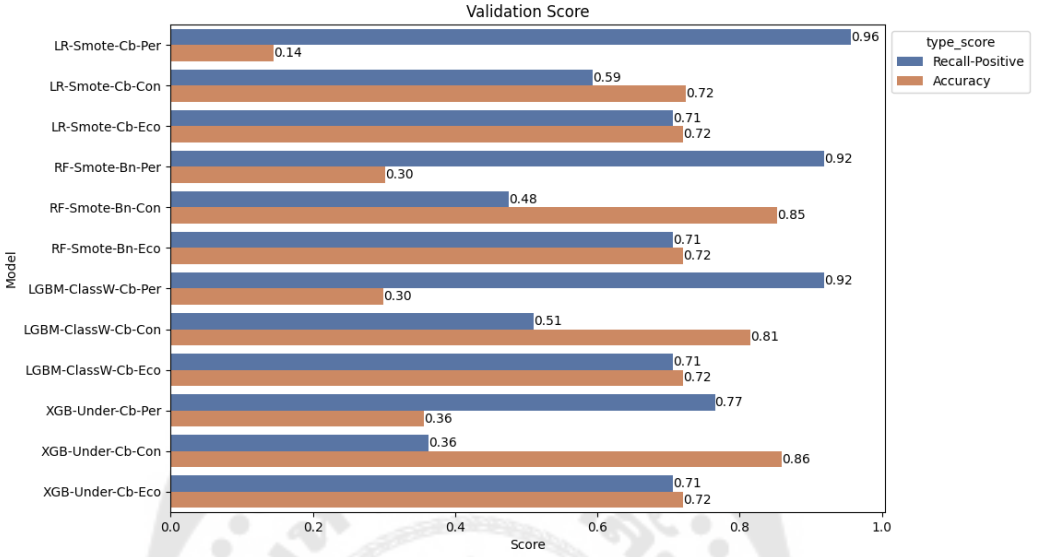
การประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลองยังคงมุ่งเน้นไปยังค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และ Accuracy ของแบบจำลอง เพื่อค้นหาแบบจำลองที่มีประสิทธิภาพในการจำแนกลูกค้าที่มีแนวโน้มในการสมัครผลิตภัณฑ์ได้อย่างครบถ้วน โดยที่ความแม่นยำของแบบจำลองยังอยู่ในเกณฑ์ที่น่าพอใจ

ภาพประกอบที่ 55 แสดงการเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และค่า Accuracy ของแบบจำลองทั้ง 12 แบบจำลอง จากกราฟจะเห็นได้ว่าการใช้งานชุดของคุณลักษณะ Personal ส่งผลให้แบบจำลองมีค่าของ Recall ที่สูงขึ้นเมื่อเปรียบเทียบกับแบบจำลองในหัวข้อก่อนหน้าอย่างเห็นได้ชัด โดยแบบจำลอง LR-Smote-Cb-Per ให้ประสิทธิภาพของ Recall สูงสุดอยู่ที่ 0.96 (96%) แต่จะมีประสิทธิภาพของความแม่นยำโดยรวมลดลงอย่างมาก โดยค่า Accuracy ลดลงเหลือเพียง 0.14 (14%) ส่วนแบบจำลอง RF-Smote-Bn-Per และ LGBM-ClassW-Cb-Per ให้ค่าของ Recall และ Accuracy ที่เท่ากันที่ 0.92 (92%) และ 0.30 (30%) ตามลำดับ ในขณะที่แบบจำลอง XGB-Under-Cb-Per มีการเพิ่มขึ้นของประสิทธิภาพค่า Recall ที่ไม่ได้สูงมากนักอยู่ที่ 0.77 (77%) แต่ค่าของ Accuracy กลับลดลงเหลือเพียง 0.36 (36%)

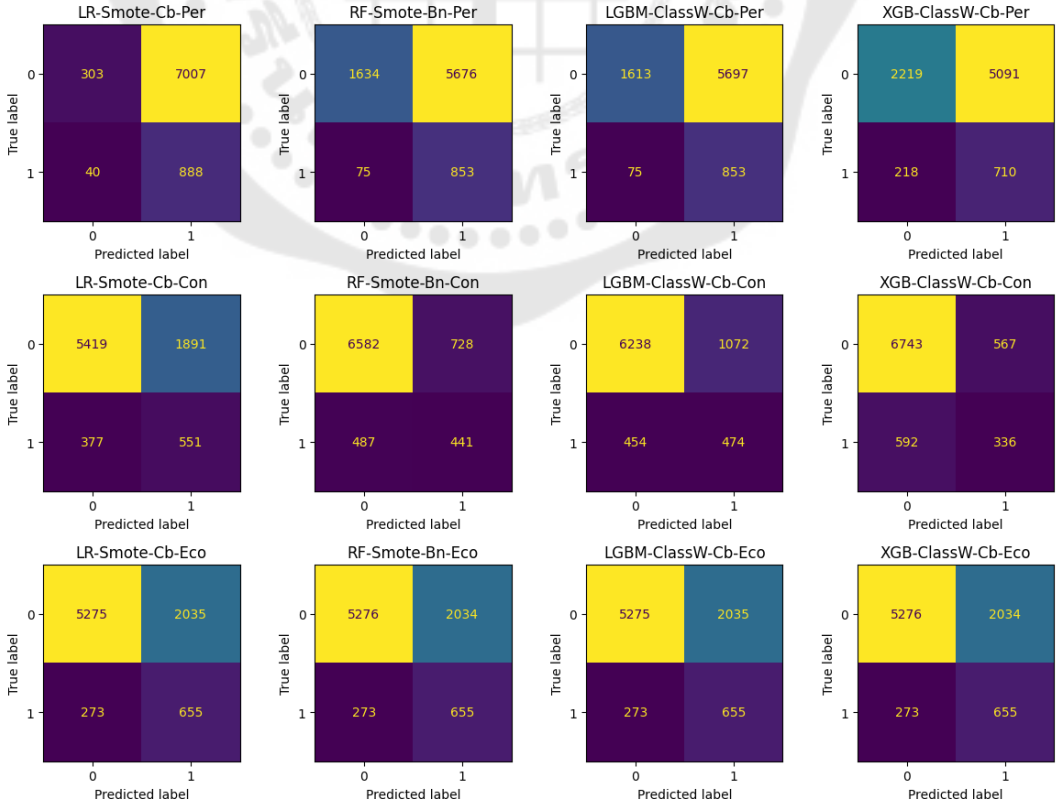
ในทางกลับกันการใช้งานชุดของคุณลักษณะ Contact จะสังเกตได้ว่าแบบจำลองมีประสิทธิภาพของความแม่นยำโดยรวมที่สูงขึ้น โดยมีค่าของ Accuracy ที่ใกล้เคียงหรือเพิ่มสูงขึ้นเมื่อเปรียบเทียบแบบจำลองในหัวข้อก่อนหน้าแต่จะมีค่าของ Recall ที่ลดลงอย่างเห็นได้ชัดเช่นกัน โดยแบบจำลองที่มีค่า Accuracy สูงที่สุดอยู่ที่ 0.86 (86%) ได้แก่ XGB-Under-Cb-Con ส่วนค่าของ Recall อยู่ที่ 0.36 (36%) โดยแบบจำลองที่มีค่า Accuracy ในอันดับถัดๆ มา ได้แก่ แบบจำลอง RF-Smote-Bn-Con, LGBM-ClassW-Cb-Con และ LR-Smote-Cb-Con ตามลำดับ ซึ่งมีค่าของ Accuracy เป็น 0.85 (85%), 0.81 (81%) และ 0.72 (72%) ตามลำดับ ส่วนค่าของ Recall มีค่าเป็น 0.48 (48%), .0.51 (51%) และ 0.59 (59%) ตามลำดับ

ในส่วนของแบบจำลองซึ่งมีการใช้งานชุดของคุณลักษณะ Economics พบว่าทุกแบบจำลองซึ่งประกอบด้วย LR-Smote-Cb-Eco, RF-Smote-Bn-Eco, LGBM-ClassW-Cb-Eco และ XGB-Under-Cb-Eco ให้ผลประสิทธิภาพของ Recall และ Accuracy ที่เท่ากันในทุกๆ แบบจำลอง โดยมีค่าเป็น 0.71 (71%) และ 0.72 (72%) ตามลำดับ ซึ่งค่าประสิทธิภาพมีความใกล้เคียงกับแบบจำลองส่วนใหญ่ในหัวข้อก่อนๆ หน้า โดยเมื่อผลที่แสดงใน Confusion Matrix ดังรูปที่ 56 จะพบว่าแบบจำลองทั้งสี่ที่ใช้งานชุดของคุณลักษณะ Economics จะทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ถูกต้อง (TP) จำนวนเท่ากันทั้งหมดที่ 655 ตัวอย่างข้อมูล และมีจำนวนตัวอย่างข้อมูลที่มีความแตกต่างเกิดขึ้นเพียง 1 ตัวอย่างข้อมูล โดยแบบจำลอง LR-Smote-Cb-Eco และ LGBM-ClassW-Cb-Eco สามารถทำนายกลุ่มลูกค้าทีไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) เป็นจำนวนเท่ากันที่ 5275 ตัวอย่างข้อมูล ส่วนแบบจำลองที่

เหลือสามารถทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) เป็นจำนวน 5276 ตัวอย่างข้อมูล



ภาพประกอบ 55 การเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์และค่า Accuracy ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะ Personal, Contact และ Economics



ภาพประกอบ 56 การเปรียบเทียบ Confusion Matrix ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะ Personal, Contact และ Economics

ตารางที่ 10 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะ Personal, Contact และ Economics ประกอบไปด้วย Accuracy, Recall, Precision และ F1-Score ของทั้งสองกลุ่มของตัวอย่างข้อมูล นอกจากนี้ยังแสดงค่าประสิทธิภาพของ Recall, Precision และ F1-Score โดยรวมของแบบจำลองแบบ weighted average ซึ่งจะมีการคำนวณโดยคำนึงถึงสัดส่วนของตัวอย่างข้อมูลในแต่ละกลุ่ม

ตาราง 10 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะ Personal, Contact และ Economics

แบบจำลอง	Acc	Precision			Recall			F1-Score		
		C-0	C-1	W-avg	C-0	C-1	W-avg	C-0	C-1	W-avg
LR-Smote-Cb-Per	0.14	0.88	0.11	0.80	0.04	0.96	0.14	0.08	0.20	0.09
LR-Smote-Cb-Con	0.72	0.93	0.23	0.86	0.74	0.59	0.72	0.83	0.33	0.77
LR-Smote-Cb-Eco	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
RF-Smote-Bn-Per	0.30	0.96	0.13	0.86	0.22	0.92	0.30	0.36	0.23	0.35
RF-Smote-Bn-Con	0.85	0.93	0.38	0.87	0.90	0.48	0.85	0.92	0.42	0.86
RF-Smote-Bn-Eco	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
LGBM-ClassW-Cb-Per	0.30	0.96	0.13	0.86	0.22	0.92	0.30	0.36	0.23	0.34
LGBM-	0.81	0.93	0.31	0.86	0.85	0.51	0.81	0.89	0.38	0.83

ตาราง 10 (ต่อ)

แบบจำลอง	Acc	Precision			Recall			F1-Score		
		C-0	C-1	W-avg	C-0	C-1	W-avg	C-0	C-1	W-avg
ClassW-Cb-Con										
LGBM-ClassW-Cb-Eco	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB-Under-Cb-Per	0.36	0.91	0.12	0.82	0.30	0.77	0.36	0.46	0.21	0.43
XGB-Under-Cb-Con	0.86	0.92	0.37	0.86	0.92	0.36	0.86	0.92	0.37	0.86
XGB-Under-Cb-Eco	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77

4.4 ชุดของคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับจากการคัดเลือกคุณลักษณะ

ในการวิจัยครั้งนี้มีการใช้งานหลักการการคัดเลือกคุณลักษณะที่สำคัญด้วยวิธีการต่างๆ เพื่อค้นหาคุณลักษณะที่ส่งผลต่อการทำนายสูงที่สุดสองอันดับแรก เพื่อเป็นการลดจำนวนของข้อมูลสำหรับการนำไปใช้งานกับแบบจำลอง เพื่อลดความซับซ้อนของแบบจำลองและเป็นการเพิ่มประสิทธิภาพในแง่ของทรัพยากรและเวลา โดยหลักการการคัดเลือกคุณลักษณะที่นำมาใช้งาน ได้แก่

1. F-Value

2. Recursive Feature Elimination (RFE) กับแบบจำลอง Logistic Regression ที่ใช้งานร่วมกับ SMOTE และ One-Hot Encoding (RFE(LR))

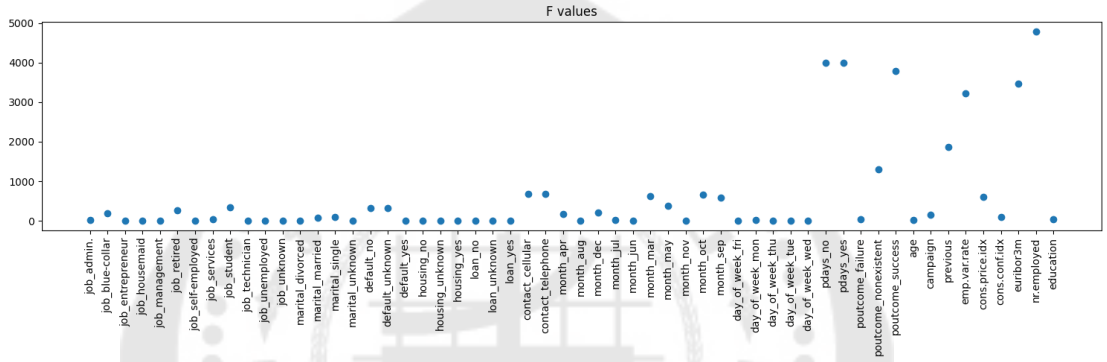
3. Recursive Feature Elimination (RFE) กับแบบจำลอง LightGBM ที่ใช้งานร่วมกับ Class Weight และ One-Hot Encoding (RFE(LGBM))

4. Recursive Feature Elimination (RFE) กับแบบจำลอง XGBoost ที่ใช้งานร่วมกับ Class Weight และ One-Hot Encoding (RFE(XGB))

5. SHAP กับแบบจำลอง LightGBM ที่ใช้งานร่วมกับ Class Weight และ One-Hot Encoding (SHAP(LGBM))

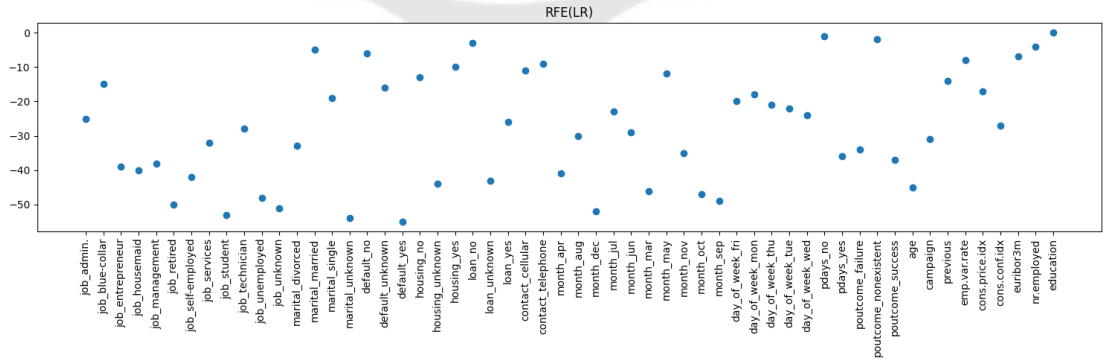
6. SHAP กับแบบจำลอง XGBoost ที่ใช้งานร่วมกับ Class Weight และ One-Hot Encoding (SHAP(XGB))

จากผลการทดลองเพื่อหาความสำคัญของคุณลักษณะของชุดข้อมูลพบว่าชุดของคุณลักษณะมีความสำคัญสูงสุดสองอันดับที่ได้จากการใช้งานหลักการ F-Value ประกอบด้วย จำนวนวันที่ห่างจากการนำเสนอผลิตภัณฑ์ก่อน (pdays) และจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) ดังภาพประกอบที่ 57



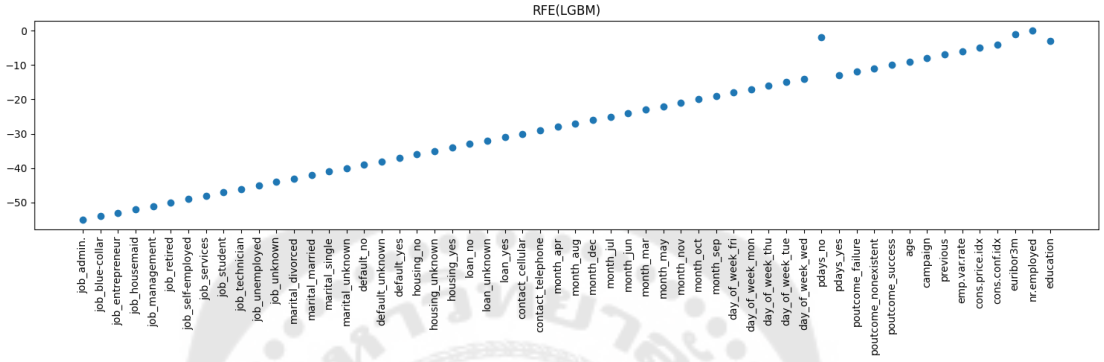
ภาพประกอบ 57 ผลลัพธ์การหาค่าความสำคัญของคุณลักษณะด้วยวิธีการ F-Value

ส่วนการใช้งานหลักการ RFE(LR) จะได้ชุดของคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับประกอบด้วย จำนวนวันที่ห่างจากการนำเสนอผลิตภัณฑ์ก่อน (pdays) และระดับการศึกษา (education) ดังภาพประกอบที่ 58

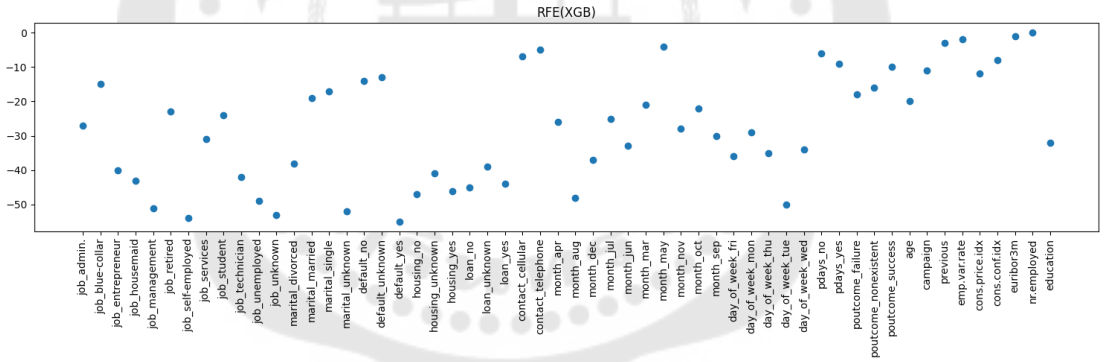


ภาพประกอบ 58 ผลลัพธ์การหาค่าความสำคัญของคุณลักษณะด้วยวิธีการ RFE(LR)

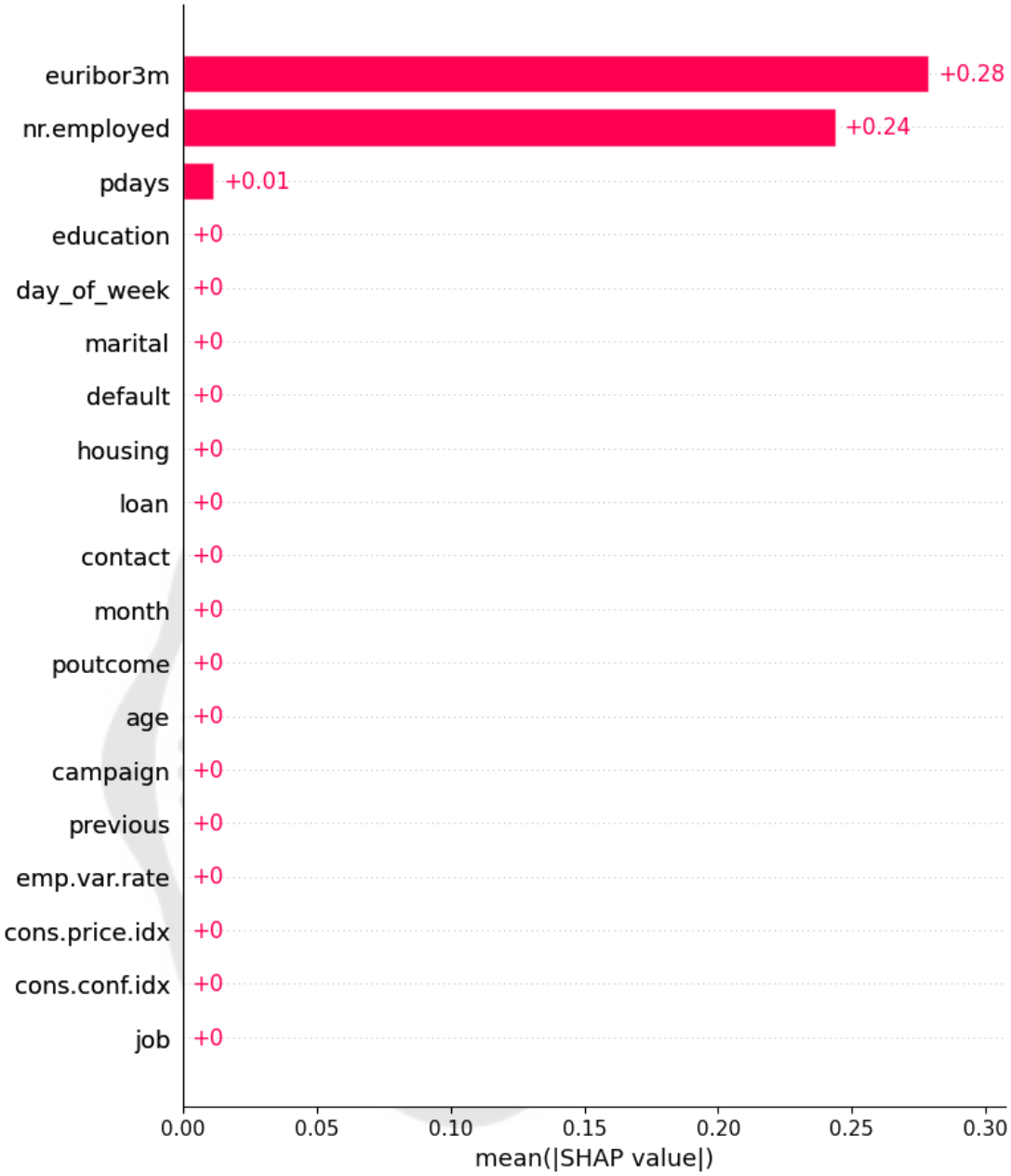
สำหรับชุดของคุณลักษณะที่ได้จากการใช้งานหลักการ RFE(LGBM), RFE(XGB), SHAP(LGBM) และ SHAP(XGB) ให้ผลลัพธ์ของชุดของคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับที่เหมือนกันซึ่งประกอบด้วย อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) และจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) ดังภาพประกอบที่ 59, 60, 61 และ 62 ตามลำดับ



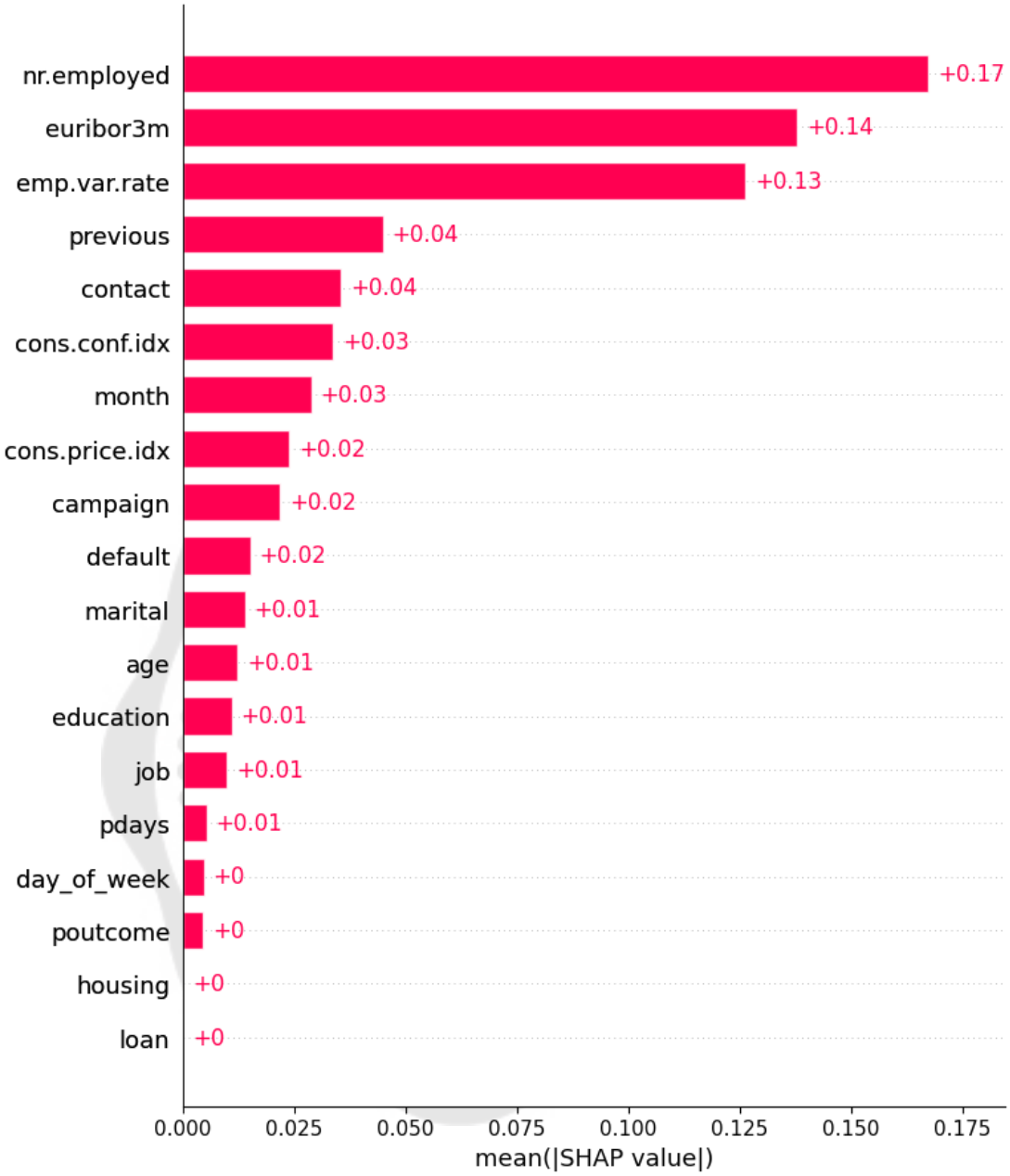
ภาพประกอบ 59 ผลลัพธ์การหาค่าความสำคัญของคุณลักษณะด้วยวิธีการ RFE(LGBM)



ภาพประกอบ 60 ผลลัพธ์การหาค่าความสำคัญของคุณลักษณะด้วยวิธีการ RFE(XGB)



ภาพประกอบ 61 ผลลัพธ์การหาค่าความสำคัญของคุณลักษณะด้วยวิธีการ SHAP(LGBM)



ภาพประกอบ 62 ผลลัพธ์การหาค่าความสำคัญของคุณลักษณะด้วยวิธีการ SHAP(XGB)

โดยจะสังเกตได้ว่าคุณลักษณะ จำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) ปรากฏอยู่ในชุดของคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับอยู่ถึง 5 ชุด จาก 6 ชุด คุณลักษณะย่อย

ในส่วนของการใช้งานการเรียนรู้ด้วยเครื่องแบบอธิบายได้ หรือ SHAP เพื่อค้นหาระดับความสำคัญ หรือ SHAP Value ของแต่ละคุณลักษณะ จะพบว่าหลักการ SHAP(LGBM) จะมี

เพียง 2 คุณลักษณะ ที่มีค่าของ SHAP Value ซึ่งโดดเด่นออกมา โดยคุณลักษณะ อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) และจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) มีค่าของ SHAP Value อยู่ที่ 0.28 และ 0.24 ตามลำดับ ส่วนการใช้งานหลักการ SHAP(XGB) จะมี 3 คุณลักษณะ ที่มีค่าของ SHAP Value โดดเด่นออกมา ซึ่งคุณลักษณะที่มีค่า SHAP Value โดดเด่นเพิ่มเติมขึ้นมาคือ อัตราการจ้างงานรายไตรมาส (emp.var.rate) โดยมีค่าอยู่ที่ 0.13 ส่วนคุณลักษณะสองดับดับแรก ได้แก่ จำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) โดยมีค่า SHAP Value เป็น 0.17 และ 0.14 ตามลำดับ

ตารางที่ 11 แสดงผลสรุปจากการคัดเลือกคุณลักษณะที่มีความสูงสุดสองอันดับแรกจากวิธีการต่างๆ

ตาราง 11 คุณลักษณะที่มีความสูงสุดสองอันดับจากการคัดเลือกคุณลักษณะด้วยวิธีการต่างๆ

วิธีการคัดเลือกคุณลักษณะ	**คุณลักษณะที่ 1**	**คุณลักษณะที่ 2**
F-Value	จำนวนวันจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า (pdays)	จำนวนพนักงานรายไตรมาส (nr.employed)
RFE(LR)	จำนวนวันจากการติดต่อเพื่อนำเสนอผลิตภัณฑ์ก่อนหน้า (pdays)	ระดับการศึกษา (education)
RFE(LGBM)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)
RFE(XGB)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)
SHAP(LGBM)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)
SHAP(XGB)	จำนวนพนักงานรายไตรมาส (nr.employed)	อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารภายในยุโรปรายวัน (euribor3m)

4.5 การเปรียบเทียบประสิทธิภาพของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับ

จากผลลัพธ์ของการค้นหาชุดของคุณลักษณะย่อยที่มีความสำคัญสูงที่สุดสองอันดับจากการใช้งานวิธีการคัดเลือกคุณลักษณะทั้ง 6 แบบในหัวข้อก่อนหน้า สามารถเลือกนำชุดคุณลักษณะย่อยมาใช้งานได้ทั้งหมดจำนวน 3 ชุดที่ไม่ซ้ำกัน ได้แก่

- F-Value ประกอบด้วยคุณลักษณะ จำนวนวันที่ห่างจากการนำเสนอผลิตภัณฑ์ก่อน (pdays) และจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed)

- RFE(LR) ประกอบด้วยคุณลักษณะ จำนวนวันที่ห่างจากการนำเสนอผลิตภัณฑ์ก่อน (pdays) และระดับการศึกษา (education)

- RFE(LGBM) ประกอบด้วยคุณลักษณะ จำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m)

ซึ่งในหัวข้อนี้จะมีการสร้างแบบจำลองสำหรับทำงานกับแต่ละชุดคุณลักษณะย่อยทั้งสามชุดรวมทั้งหมด 12 แบบจำลอง โดยคุณลักษณะ จำนวนวันที่ห่างจากการนำเสนอผลิตภัณฑ์ก่อน (pdays) มีการทำวิศวกรรมคุณลักษณะเพื่อเปลี่ยนจากคุณลักษณะชนิดตัวเลขมาเป็นคุณลักษณะชนิดประเภทแบบไม่เรียงลำดับ ดังนั้นจึงมีการใช้งานหลักการวิศวกรรมคุณลักษณะ Standard Scaling และ Ordinal Encoding เพื่อจัดการกับคุณลักษณะชนิดตัวเลข และคุณลักษณะชนิดประเภทแบบมีลำดับ และใช้งาน One-Hot Encoding สำหรับจัดการกับคุณลักษณะชนิดประเภทแบบไม่มีลำดับ ส่วนการจัดการความไม่สมดุลกันของข้อมูลจะมีการใช้งานแตกต่างกันไปในแต่ละอัลกอริทึมของแบบจำลอง โดยอัลกอริทึมแบบ Logistic Regression และ Random Forest จะใช้งานหลักการ SMOTE ส่วนอัลกอริทึมแบบ LightGBM และ XGBoost จะใช้งานหลักการ Class Weight ซึ่งทั้ง 12 แบบจำลองประกอบด้วย

1. Logistic Regression ที่ใช้งาน SMOTE ร่วมกับชุดคุณลักษณะ F-Value (LR-Smote-Fval)

2. Logistic Regression ที่ใช้งาน SMOTE ร่วมกับชุดคุณลักษณะ RFE(LR) (LR-Smote-rfeLR)

3. Logistic Regression ที่ใช้งาน SMOTE ร่วมกับชุดคุณลักษณะ RFE(LGBM) (LR-Smote-rfeLGBM)

4. Random Forest ที่ใช้งาน SMOTE ร่วมกับชุดคุณลักษณะ F-Value (RF-Smote-Fval)

5. Random Forest ที่ใช้งาน SMOTE ร่วมกับชุดคุณลักษณะ RFE(LR) (RF-Smote-rfeLR)

6. Random Forest ที่ใช้งาน SMOTE ร่วมกับชุดคุณลักษณะ RFE(LGBM) (RF-Smote-rfeLGBM)

7. LightGBM ที่ใช้งาน Class Weight ร่วมกับชุดคุณลักษณะ F-Value (LGBM-ClassW-Fval)

8. LightGBM ที่ใช้งาน Class Weight ร่วมกับชุดคุณลักษณะ RFE(LR) (LGBM-ClassW-rfeLR)

9. LightGBM ที่ใช้งาน Class Weight ร่วมกับชุดคุณลักษณะ RFE(LGBM) (LGBM-ClassW-rfeLGBM)

10. XGBoost ที่ใช้งาน Class Weight ร่วมกับชุดคุณลักษณะ F-Value (XGB-ClassW-Fval)

11. XGBoost ที่ใช้งาน Class Weight ร่วมกับชุดคุณลักษณะ RFE(LR) (XGB-ClassW-rfeLR)

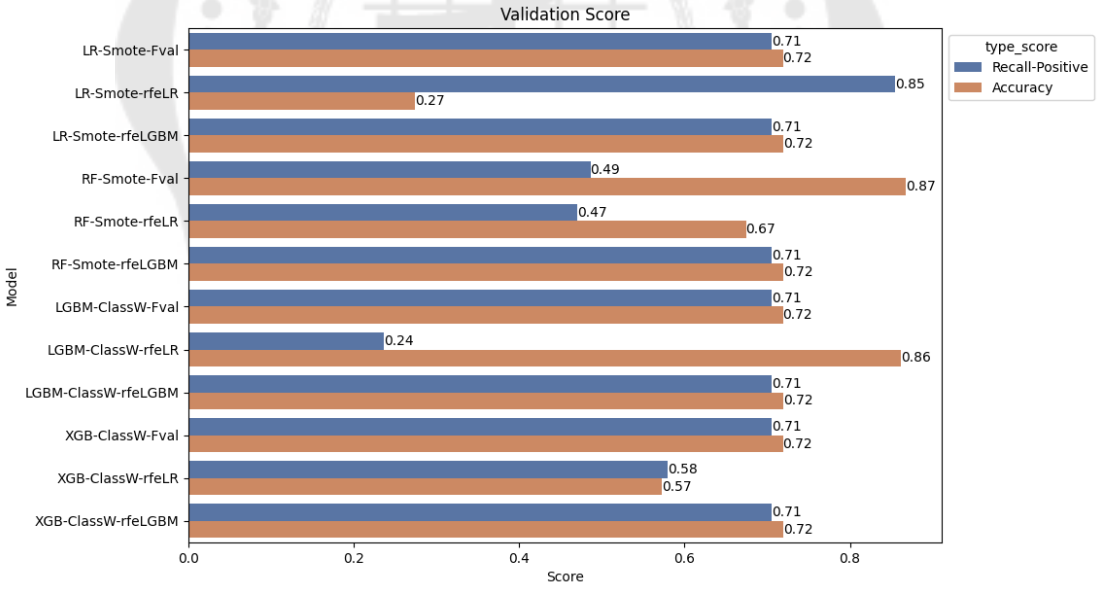
12. XGBoost ที่ใช้งาน Class Weight ร่วมกับชุดคุณลักษณะ RFE(LGBM) (XGB-ClassW-rfeLGBM)

การประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลองมุ่งเน้นไปยังค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และ Accuracy ของแบบจำลอง เพื่อค้นหาแบบจำลองที่มีประสิทธิภาพในการจำแนกลูกค้าที่มีแนวโน้มในการสมัครผลิตภัณฑ์ได้อย่างครบถ้วน โดยที่ความแม่นยำของแบบจำลองยังอยู่ในเกณฑ์ที่น่าพอใจ

ภาพประกอบที่ 63 แสดงการเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และค่า Accuracy ของแบบจำลองทั้ง 12 แบบจำลอง จากกราฟจะเห็นว่าการใช้งานชุดของคุณลักษณะ F-Value จะให้ประสิทธิภาพของทั้งค่า Recall และ Accuracy ที่ใกล้เคียงกันในการใช้งานกับ 3 แบบจำลอง ประกอบด้วย LR-Smote-Fval, LGBM-ClassW-Fval และ XGB-ClassW-Fval โดยมีค่าประสิทธิภาพของ Recall ที่ 0.71 (71%) และค่า Accuracy ที่ 0.72 (72%) ส่วนการใช้งานกับแบบจำลอง RF-Smote-Fval พบว่าให้ค่า Accuracy ที่ 0.87 (87%) ซึ่งสูงที่สุดในการดำเนินการวิจัยในครั้งนี้ แต่จะมีค่าของ Recall ที่ค่อนข้างต่ำที่ 0.49 (49%)

การใช้งานชุดของคุณลักษณะ RFE(LR) จะให้ผลลัพธ์ประสิทธิภาพของแบบจำลองที่ค่อนข้างแปรปรวน การใช้งานชุดคุณลักษณะ RFE(LR) กับแบบจำลอง LR-Smote-rfeLR ให้ค่าประสิทธิภาพของ Recall ที่ค่อนข้างสูงที่ 0.85 (85%) แต่มีค่าของ Accuracy ที่ 0.27 (27%) ซึ่งค่อนข้างต่ำ ส่วนการใช้งานกับแบบจำลอง LGBM-ClassW-rfeLR จะให้ผลของประสิทธิภาพที่สวนทางกัน โดยจะมีค่า Recall ที่ค่อนข้างต่ำแต่จะได้ค่าของ Accuracy ที่ค่อนข้างสูง โดยมีค่าอยู่ที่ 0.24 (24%) และ 0.86 (86%) ตามลำดับ การใช้งานกับแบบจำลอง RF-Smote-rfeLR จะให้ค่า Recall ที่ 0.47 (47%) และค่า Accuracy ที่ 0.67 (67%) และการใช้งานกันแบบจำลอง XGB-ClassW-rfeLR จะได้ผลประสิทธิภาพที่ค่อนข้างต่ำทั้งในส่วนของค่า Recall และ Accuracy โดยมีค่าที่ 0.58 (58%) และ 0.57 (57%) ตามลำดับ

ส่วนของชุดของคุณลักษณะ RFE(LGBM) เมื่อนำไปใช้งานกับแบบจำลองทั้งสี่แล้วส่งผลให้ผลลัพธ์ประสิทธิภาพของทุกแบบจำลองมีสัดส่วนออกมาที่ใกล้เคียงกันอย่างมาก โดยประสิทธิภาพของทั้งสี่แบบจำลองในส่วนของค่า Recall และ Accuracy มีสัดส่วนที่คำนวณออกมาได้เท่ากันที่ 0.71 (71%) และ 0.72 (72%) ตามลำดับ

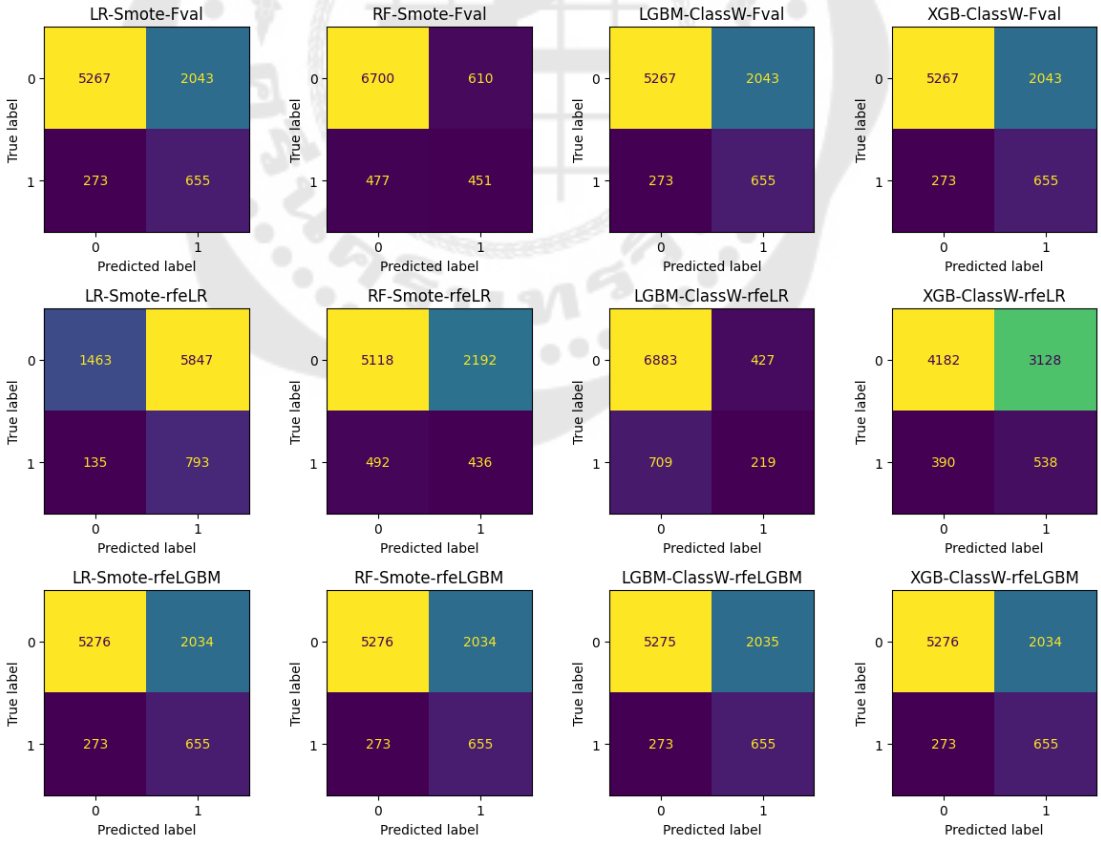


ภาพประกอบ 63 การเปรียบเทียบค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ และค่า Accuracy ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับด้วยวิธีการ F-Value, RFE(LR) และ RFE(LGBM)

ผลลัพธ์ที่แสดงใน Confusion Matrix ดังภาพประกอบที่ 64 จะพบว่าแบบจำลอง LR-Smote-rfeLR สามารถทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ถูกต้อง (TP) ใน

จำนวนที่สูงมากถึง 793 จาก 928 ตัวอย่างข้อมูล แต่ก็จะไปลดการทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) เหลือเพียงแค่ 1,463 ตัวอย่างข้อมูล ซึ่งค่อนข้างต่ำมากเมื่อเทียบกับแบบจำลองอื่นๆ ส่วนแบบจำลอง LGBM-ClassW-rfeLR สามารถทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) ได้สูงถึง 6,883 จาก 7,310 ตัวอย่างข้อมูล แต่จะสามารถทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ถูกต้อง (TP) ได้ค่อนข้างน้อยเพียง 219 ตัวอย่างข้อมูล

ส่วน 7 แบบจำลองซึ่งประกอบด้วย LR-Smote-Fval, LR-Smote-rfeLGBM, RF-Smote-rfeLGBM, LGBM-ClassW-Fval, LGBM-ClassW-rfeLR, XGB-ClassW-Fval และ XGB-ClassW-rfeLR ให้ผลลัพธ์จำนวนของการทำนายที่ถูกต้องอยู่ในระดับที่ใกล้เคียงกันมาก โดยทั้ง 7 แบบจำลองสามารถทำนายกลุ่มของลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ถูกต้อง (TP) เท่ากันที่ 655 จากทั้งหมด 928 ตัวอย่างข้อมูล และมีความแตกต่างกันเล็กน้อยในการทำนายกลุ่มของลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) อยู่ระหว่าง 5,267 ถึง 5,276 ตัวอย่างข้อมูล



ภาพประกอบ 64 การเปรียบเทียบ Confusion Matrix ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับด้วยวิธีการ F-Value, RFE(LR) และ RFE(LGBM)

ตารางที่ 12 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลองซึ่งมีการใช้งานคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับด้วยวิธีการ F-Value, RFE(LR) และ RFE(LGBM) ประกอบไปด้วย Accuracy, Recall, Precision และ F1-Score ของทั้งสองกลุ่มของตัวอย่างข้อมูล นอกจากนี้ยังแสดงค่าประสิทธิภาพของ Recall, Precision และ F1-Score โดยรวมของแบบจำลองแบบ weighted average ซึ่งจะมีการคำนวณโดยคำนึงถึงสัดส่วนของตัวอย่างข้อมูลในแต่ละกลุ่ม

ตาราง 12 แสดงค่าประสิทธิภาพต่างๆ ของแบบจำลองซึ่งมีการใช้งานชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับด้วยวิธีการ F-Value, RFE(LR) และ RFE(LGBM)

แบบจำลอง	Acc	Precision			Recall			F1-Score		
		C-0	C-1	W-avg	C-0	C-1	W-avg	C-0	C-1	W-avg
LR-Smote-Fval	0.72	0.95	0.24	0.87	0.71	0.71	0.72	0.82	0.36	0.77
LR-Smote-rfeLR	0.27	0.92	0.12	0.83	0.20	0.85	0.27	0.33	0.21	0.32
LR-Smote-rfeLGBM	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
RF-Smote-Fval	0.87	0.93	0.43	0.88	0.92	0.49	0.87	0.92	0.45	0.87
RF-Smote-rfeLR	0.67	0.91	0.17	0.83	0.70	0.47	0.67	0.79	0.25	0.73
RF-Smote-rfeLGBM	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
LGBM-ClassW-Fval	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
LGBM-	0.86	0.91	0.34	0.84	0.94	0.24	0.86	0.92	0.28	0.85

ตาราง 12 (ต่อ)

แบบจำลอง	Acc	Precision			Recall			F1-Score		
		C-0	C-1	W-avg	C-0	C-1	W-avg	C-0	C-1	W-avg
ClassW-rfeLR										
LGBM-ClassW-rfeLGBM	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB-ClassW-Fval	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77
XGB-ClassW-rfeLR	0.57	0.91	0.15	0.83	0.57	0.58	0.57	0.70	0.23	0.65
XGB-ClassW-rfeLGBM	0.72	0.95	0.24	0.87	0.72	0.71	0.72	0.82	0.36	0.77

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยเพื่อศึกษาเกี่ยวกับการใช้วิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้เพื่อวิเคราะห์ข้อมูลการนำเสนอผลิตภัณฑ์ทางโทรศัพท์ของธนาคาร ผู้วิจัยได้มีการสร้างแบบจำลองสำหรับทำนายผลของการสมัครผลิตภัณฑ์เงินฝากประจำของลูกค้าธนาคารโดยมีการใช้วิธีการและหลักการที่หลากหลายประกอบไปด้วย อัลกอริทึมในการสร้างแบบจำลอง เช่น XGBoost การจัดการความไม่สมดุลกันของข้อมูล เช่น SMOTE การทำวิศวะกรรมคุณลักษณะ เช่น One-Hot Encoding การคัดเลือกคุณลักษณะ เช่น F-Value การเรียนรู้ด้วยเครื่องแบบอธิบายได้ เช่น SHAP รวมไปถึงวิธีการและหลักการอื่นๆ จากนั้นได้มีการประเมินผลและเปรียบเทียบประสิทธิภาพของแบบจำลองต่างๆ รวมไปถึงการนำเสนอผลการวิจัย เพื่อค้นหาว่าแบบจำลองใดที่มีประสิทธิภาพที่เหมาะสมที่สุดสำหรับการนำไปประยุกต์ใช้งาน หรือปัจจัยใดที่ส่งผลต่อการทำนายผลลัพธ์ที่มีประสิทธิภาพ และนอกจากนี้ยังรวมไปถึงการนำการเรียนรู้ด้วยเครื่องแบบอธิบายได้มาเพื่ออธิบายแบบจำลองและการทำนายในระดับรายตัวอย่าง โดยสามารถแบ่งหัวข้อของการสรุปผลได้ดังต่อไปนี้

1. สรุปผลการวิจัย
2. การวิเคราะห์ความผิดพลาดของแบบจำลอง (Error Analysis)
3. อภิปรายผลการวิจัย
4. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในงานวิจัยนี้มีการสร้างแบบจำลองสำหรับการทำนายการสมัครหรือไม่สมัครผลิตภัณฑ์เงินฝากประจำของลูกค้าธนาคาร โดยมีการนำวิธีการเรียนรู้ด้วยเครื่องแบบอธิบายได้มาช่วยสำหรับการอธิบายแบบจำลอง วิเคราะห์ข้อมูลคุณลักษณะที่มีความสำคัญ และอธิบายการทำนายในระดับรายบุคคล โดยแบบจำลองที่สร้างขึ้นเกิดจากการประยุกต์ใช้งานวิธีการและหลักการต่างๆ เข้าด้วยกัน

5.1.1 การสร้างแบบจำลอง

ในส่วนของอัลกอริทึมสำหรับสร้างแบบจำลองมีการใช้งานทั้งหมด 4 อัลกอริทึม ได้แก่ Logistic Regression, Random Forest, LightGBM และ XGBoost ในส่วนของการจัดการความไม่สมดุลกันของข้อมูลมีการใช้งานทั้งหมด 3 วิธีการ ได้แก่ Class Weight, Random

Undersampling และ SMOTE ในส่วนของการทำวิศวะกรรมคุณลักษณะกับคุณลักษณะชนิดประเภทแบบมีลำดับและคุณลักษณะชนิดตัวเลข มีการใช้งานวิธีการ Ordinal Encoding และ Standard Scaling ตามลำดับ ในส่วนของการทำวิศวกรรมคุณลักษณะกับคุณลักษณะชนิดประเภทแบบไม่มีลำดับมีการใช้งานวิธีการทั้งหมด 3 แบบ ได้แก่ One-Hot Encoding, CatBoost Encoding และ BaseN Encoding ในส่วนของการคัดเลือกคุณลักษณะมีการใช้งานวิธีการต่างๆ ได้แก่ F-Value, Recursive Feature Elimination (RFE) และ SHAP โดยการใช้งาน RFE และ SHAP มีการใช้งานร่วมกับแบบจำลอง Logistic Regression, LightGBM และ XGBoost

5.1.2 การคัดเลือกคุณลักษณะ

มีการแบ่งคุณลักษณะออกเป็นชุดย่อยๆ เพื่อค้นหาชุดของคุณลักษณะที่ส่งผลให้แบบจำลองมีประสิทธิภาพที่ดีโดยมีการแบ่งการทดลองกับชุดข้อมูลย่อยออกเป็น 3 ขั้นตอน ได้แก่ การใช้งานทุกคุณลักษณะในชุดข้อมูล การใช้งานชุดคุณลักษณะย่อยตามประเภท และการใช้งานชุดคุณลักษณะที่มีความสำคัญสูงที่สุดสองอันดับ ซึ่งได้จากขั้นตอนของการคัดเลือกคุณลักษณะ

ในส่วนของชุดคุณลักษณะย่อยตามประเภท สามารถแบ่งย่อยลงไปได้ 3 ชุด ได้แก่ ชุดคุณลักษณะเกี่ยวกับข้อมูลส่วนตัวของลูกค้าธนาคาร ชุดคุณลักษณะเกี่ยวกับการติดต่อระหว่างธนาคารและลูกค้าธนาคาร และชุดคุณลักษณะข้อมูลทางเศรษฐศาสตร์ในช่วงเวลานั้นๆ ในส่วนของชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับมีการซ้ำกันของชุดคุณลักษณะย่อยจากการใช้งานวิธีการ RFE(LGBM), RFE(XGB), SHAP(LGBM) และ SHAP(XGB) จึงมีการเลือกมาเพียงชุดคุณลักษณะเดียว ดังนั้นชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับจึงประกอบด้วยชุดของคุณลักษณะที่มาจากวิธีการ F-Value, RFE(LR) และ RFE(LGBM)

การใช้งานชุดของคุณลักษณะทั้งหมดของชุดข้อมูลค่อนข้างมีประสิทธิภาพในแง่ของความเสถียรในการใช้งานกับแบบจำลองต่างๆ ซึ่งเกิดจากการประยุกต์ใช้อัลกอริทึมและหลักการต่างๆ เข้าด้วยกัน โดยแบบจำลองส่วนใหญ่ที่ใช้ครบทุกคุณลักษณะจะให้ค่าของ Accuracy และค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) อยู่ที่ประมาณ 0.70 (70%) ถึง 0.72 (72%) ส่วนการใช้งานชุดของคุณลักษณะที่มีจำนวนลดลงส่งผลให้ประสิทธิภาพของการสร้าง การเรียนรู้ และการทำนายผลของแบบจำลองเพิ่มสูงขึ้นในแง่ของเวลาที่ใช้งาน ซึ่งการใช้งานเพียงสองคุณลักษณะที่สำคัญที่สุดยังคงสามารถให้ประสิทธิภาพของแบบจำลองในส่วนของมาตรวัดที่สนใจได้ใกล้เคียงหรือเทียบเท่ากับการใช้งานทุกคุณลักษณะในชุดข้อมูล เมื่อมีการใช้งานกับแบบจำลองที่เหมาะสม

5.1.3 ผลการทดลอง

จากการสร้างแบบจำลองทั้งหมด 48 แบบจำลอง จะพบว่าแบบจำลองส่วนมากจะให้ค่าประสิทธิภาพซึ่งสนใจในการวิจัยครั้งนี้ ได้แก่ ค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และค่า Accuracy ของแบบจำลองอยู่ที่ประมาณ 0.70 (70%) ถึง 0.72 (72%) โดยแบบจำลองอื่นๆส่วนใหญ่ ที่นอกเหนือจากนี้จะมีสัดส่วนของค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และค่า Accuracy ที่ค่อนข้างแปรผกผันกัน โดยเมื่อค่า Recall มีค่าสูงมากขึ้นจะยิ่งทำให้ Accuracy มีค่าลดลง และในทางกลับกันค่าของ Accuracy ที่ยิ่งสูงมากขึ้นจะยิ่งทำให้ค่าของ Recall ลดต่ำลง ยกเว้นแบบจำลอง XGB-ClassW-rfeLR ซึ่งมีค่าของ Recall และ Accuracy ที่ต่ำลงทั้งคู่อยู่ที่ 0.58 (58%) และ 0.57 (57%) ตามลำดับ

จากผลการวิจัย แบบจำลองที่มีค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) สูงที่สุดคือ LR-Smote-Cb-Per โดยเป็นแบบจำลองที่เกิดจากการใช้งาน Logistic Regression ร่วมกับ SMOTE และ CatBoost Encoding โดยใช้ชุดคุณลักษณะข้อมูลส่วนตัวของลูกค้าธนาคารในการสร้างแบบจำลอง ซึ่งมีค่าของ Recall สูงถึง 0.96 (96%) แต่จะให้ค่า Accuracy ที่ค่อนข้างต่ำมากเพียง 0.14 (14%)

ส่วนแบบจำลองที่มีค่าของ Accuracy สูงที่สุดได้แก่ RF-Smote-Fval โดยเป็นแบบจำลองที่เกิดจากการใช้งาน Random Forest ร่วมกับ SMOTE และ One-Hot Encoding โดยใช้ชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับจากวิธี F-Value ซึ่งมีค่าของ Accuracy อยู่ที่ 0.87 (87%) แต่จะมีค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) เพียง 0.49 (49%)

5.1.4 การอธิบายแบบจำลองด้วยเรียนรู้ด้วยเครื่องแบบอธิบายได้

การอธิบายในระดับแบบจำลอง

มีการนำการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธีการแบบ SHAP มาใช้งานสำหรับการอธิบายแบบจำลองโดยการวาดกราฟต่างๆ เช่น Bar Plot หรือ Beeswamp Plot เป็นต้น เพื่อให้สามารถเห็นได้ว่าแบบจำลองมีวิธีการทำนายผลลัพธ์โดยอ้างอิงจากคุณลักษณะใดด้วยค่าความสำคัญมากน้อยเพียงใด โดยเป็นการอธิบายการทำงานของแบบจำลองที่มีต่อแต่ละคุณลักษณะ โดยการอธิบายแบบจำลองสามารถเพิ่มความน่าเชื่อถือของการทำนายและยังสามารถนำไปประยุกต์ใช้กับกระบวนการในการช่วยตัดสินใจ หรือ Support Decision Making Process ได้ นอกจากนี้ SHAP ยังสามารถอธิบายการทำนายในระดับย่อยของแต่ละตัวอย่างข้อมูลได้ โดยจะเป็นการอธิบายว่าค่าของคุณลักษณะใดที่ส่งผลต่อการทำนายและส่งผลมากน้อยเพียงใดจึงทำให้ผลลัพธ์ออกมาในรูปแบบนั้น

นอกจากนี้ในการวิจัยยังมีการนำการเรียนรู้ด้วยเครื่องแบบอธิบายได้มาประยุกต์ใช้สำหรับขั้นตอนการคัดเลือกคุณลักษณะ โดยมีการใช้งานร่วมกับแบบจำลองที่มาจากอัลกอริทึมแบบ LightGBM และ XGBoost โดยผลของการใช้งานการคัดเลือกคุณลักษณะด้วย SHAP กับทั้งสองแบบจำลอง ให้ผลลัพธ์ของชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับออกมาเหมือนกัน และยังเหมือนกับชุดของคุณลักษณะที่ได้จากวิธีการ RFE ที่ใช้งานร่วมกับแบบจำลองที่มาจากอัลกอริทึมแบบ LightGBM และ XGBoost โดยคุณลักษณะทั้งสอง ได้แก่ จานวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m)

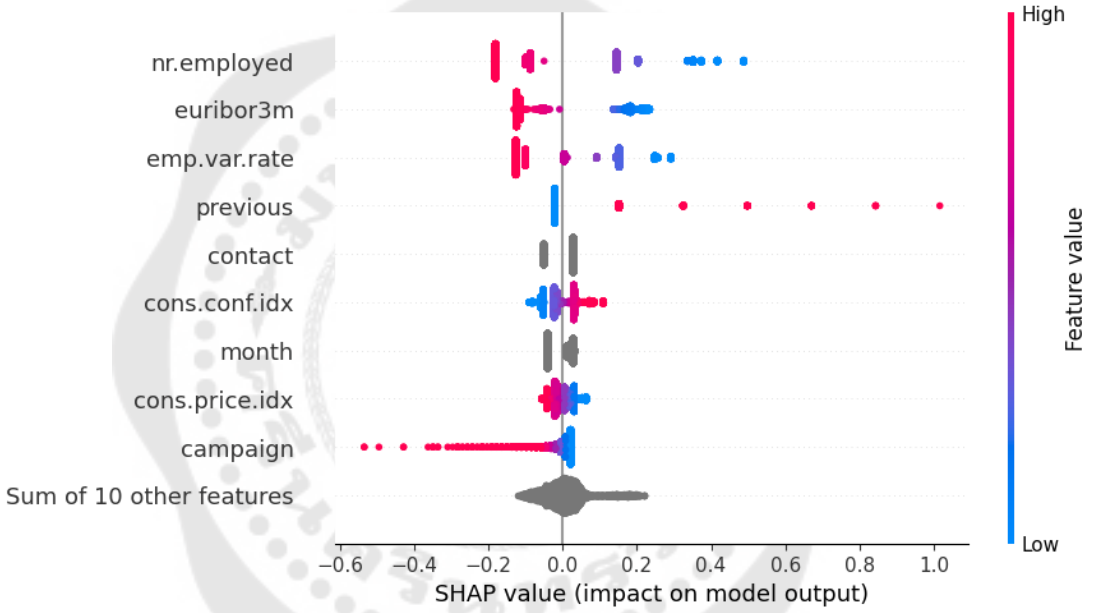
ภาพประกอบที่ 65 เป็นการวาดกราฟ Beeswamp Plot ของ SHAP Value ที่ได้จากแบบจำลอง XGB-ClassW ซึ่งเป็นแบบจำลองที่มีการใช้งาน XGBoost และ One-Hot Encoding ร่วมกับการจัดการความไม่สมดุลกันของข้อมูลด้วย Class Weight โดยใช้ชุดคุณลักษณะทั้งหมด ซึ่งให้ค่า Accuracy และ Recall ที่ 0.72 (72%) และ 0.71 (71%) ตามลำดับ โดยจะเป็นกราฟ Beeswamp Plot ซึ่งเกี่ยวข้องกับกราฟ Bar Plot ในภาพประกอบที่ 61 เนื่องจากใช้งาน SHAP Value จากแบบจำลองเดียวกัน

โดยกราฟ Beeswamp Plot จะเป็นการวาดค่า SHAP Value ของแต่ละจุดข้อมูลลงไปยังแกนในแนวนอน ซึ่งเป็นค่าของ SHAP Value ทั้งในทางบวกและในทางลบ โดยฝั่งขวาของแกนจะเป็นค่าในทางบวกซึ่งจะส่งผลให้แบบจำลองมีแนวโน้มทำนายเป็น Positive Class ส่วนฝั่งซ้ายของแกนจะเป็นค่าในทางลบซึ่งจะส่งผลให้แบบจำลองมีแนวโน้มทำนายเป็น Negative Class โดยสีของตัวอย่างข้อมูลจะหมายถึงค่าจริงของคุณลักษณะของตัวอย่างข้อมูลนั้นๆ สีชมพูจะหมายถึงค่าของคุณลักษณะนั้นๆ มีค่าที่สูง ส่วนสีฟ้าจะหมายถึงค่าของคุณลักษณะนั้นๆ มีค่าที่ต่ำ นอกจากนี้ยังสามารถดูความหนาแน่นและของข้อมูลได้โดยการดูจำนวนจุดที่หนาแน่นและเพิ่มสูงขึ้นใน SHAP Value นั้นๆ โดยคุณลักษณะจะถูกวาดไล่เรียงลงมาตามลำดับความสำคัญของคุณลักษณะ โดยบนสุดหมายถึงมีความสำคัญมากที่สุด

โดยจากกราฟจะสามารถแปลผลได้ว่าคุณลักษณะ nr.employed มีความสำคัญและส่งผลสูงที่สุดต่อการทำนายของแบบจำลอง โดยเมื่อ nr.employed มีค่าที่ต่ำ (สีฟ้า) จะส่งผลให้แบบจำลองมีแนวโน้มที่จะทำนายตัวอย่างข้อมูลเป็น Positive Class เนื่องจากจะสังเกตได้ว่าจุดข้อมูลสีฟ้าจะอยู่เลยไปในทางขวาของแกน หรือมีค่า SHAP Value ในทางบวก ส่วนเมื่อ nr.employed มีค่าที่สูง (สีชมพู) ซึ่งมีความหนาแน่นของข้อมูลที่ค่อนข้างมาก สังเกตได้จากมีจุดข้อมูลเรียงตัวกันหนาแน่นและมีความสูงเพิ่มขึ้นที่ค่า SHAP Value นั้นๆ จะส่งผลให้แบบจำลองมี

แนวโน้มที่จะทำนายตัวอย่างข้อมูลเป็น Negative Class เป็นต้น โดยยิ่งจุดข้อมูลอยู่ห่างจากแกนกลางของ SHAP Value มากขึ้นเท่าไหร่ ก็จะยิ่งแสดงถึงความส่งผลต่อการทำนายของแบบจำลองที่ยิ่งสูงขึ้นด้วยทั้งในทางบวกและทางลบ

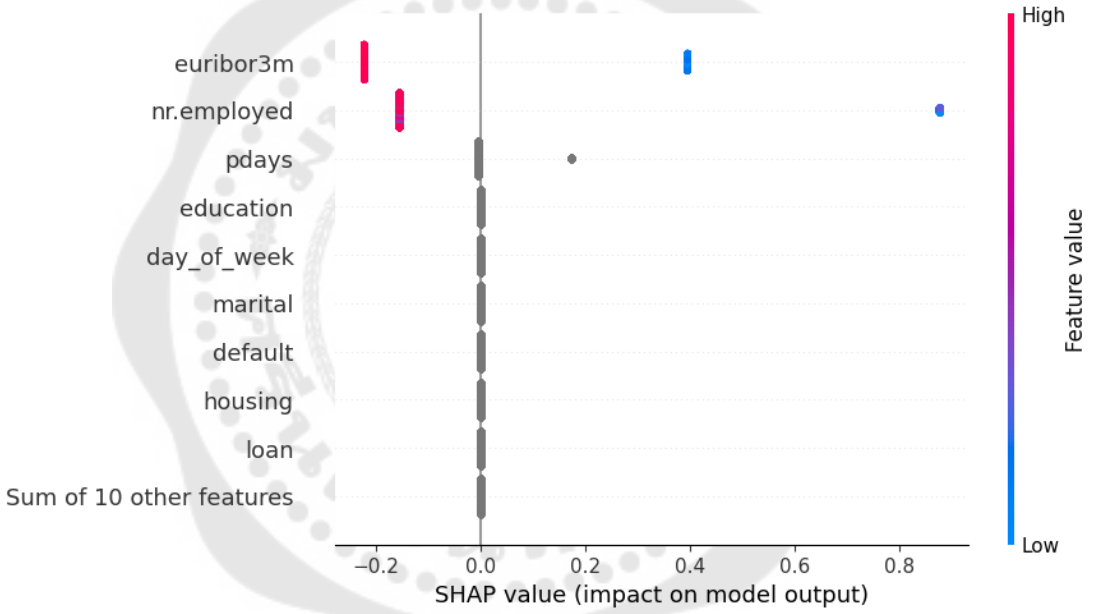
โดย Beeswamp Plot จะเหมาะสำหรับการทำงานกับคุณลักษณะที่เป็นชนิดตัวเลข โดยหากมีคุณลักษณะที่เป็นชนิดประเภทจะไม่สามารถแสดงผลได้อย่างเหมาะสม เช่น คุณลักษณะ contact ซึ่งประกอบด้วยค่า ‘mobile’ และ ‘cellular’ จะถูกวาดกราฟออกมาเป็นจุดสีเทาเนื่องจากไม่สามารถวัดค่าความสูงต่ำของข้อมูลได้ ส่วนคุณลักษณะอื่นๆ ก็จะมีความมีอิทธิพลในการทำนายของแบบจำลองลดหลั่นลงไปตามลำดับ



ภาพประกอบ 65 กราฟ Beeswamp Plot แสดงความสัมพันธ์ระหว่างค่าของคุณลักษณะและค่าความสำคัญของคุณลักษณะ หรือ SHAP Value ที่ได้จากแบบจำลอง XGB-ClassW

ภาพประกอบที่ 66 เป็นการวาดกราฟ Beeswamp Plot ของ SHAP Value ที่ได้จากแบบจำลอง LGBM-ClassW ซึ่งเป็นแบบจำลองที่มีการใช้งาน LightGBM และ One-Hot Encoding ร่วมกับการจัดการความไม่สมดุลกันของข้อมูลด้วย Class Weight โดยใช้ชุดคุณลักษณะทั้งหมด ซึ่งให้ค่า Accuracy และ Recall ที่ 0.72 (72%) และ 0.71 (71%) ตามลำดับ ซึ่งเท่ากับแบบจำลอง XGB-ClassW แต่จะสามารถสังเกตได้ว่าค่า SHAP Value ของแบบจำลองนี้จะให้ผลลัพธ์ที่แตกต่างอย่างมากจากกราฟ Beeswamp Plot ในภาพประกอบที่ 65 โดยคุณลักษณะที่มีความสำคัญสูงที่สุดสองอันดับของแบบจำลองนี้จากการใช้งาน SHAP จะพบว่าได้

เปลี่ยนแปลงเล็กน้อยโดยสลับตำแหน่งกันระหว่าง nr.employed และ euribor3m ทำให้คุณลักษณะที่สำคัญที่สุดจะกลายเป็น euribor3m แทน นอกจากนี้ความมีอิทธิพลต่อแบบจำลองหรือค่าของ SHAP Value ของคุณลักษณะ nr.employed ที่อยู่ในช่วงสูงสุดขยายจากค่าที่ประมาณ 0.5 ออกไปเป็นค่าที่ประมาณ 0.9 ซึ่งจะแสดงให้เห็นเมื่อตัวอย่างข้อมูลมีค่าของ nr.employed ที่อยู่ในระดับกลางลงไปถึงค่าต่ำ จะส่งผลให้แบบจำลองมีแนวโน้มในการทำนายผลลัพธ์เป็น Positive Class หรือ ลูกค้าทำการสมัครผลิตภัณฑ์ เพิ่มสูงขึ้น ส่วนคุณลักษณะอื่นๆ ที่เหลือจะเห็นได้ว่านอกจาก pdays แล้วคุณลักษณะอื่นๆ ไม่มีอิทธิพลหรือส่งผลต่อแนวโน้มในการทำนายของแบบจำลอง โดยมีค่าของ SHAP Value เป็น 0 ซึ่งสอดคล้องกับภาพประกอบที่ 62 ซึ่งเป็นกราฟ Bar Plot ที่ได้จากการใช้งาน SHAP Value ของแบบจำลองเดียวกัน



ภาพประกอบ 66 กราฟ Beeswamp Plot แสดงความสัมพันธ์ระหว่างค่าของคุณลักษณะและค่าความสำคัญของคุณลักษณะ หรือ SHAP Value ที่ได้จากแบบจำลอง LGBM-ClassW

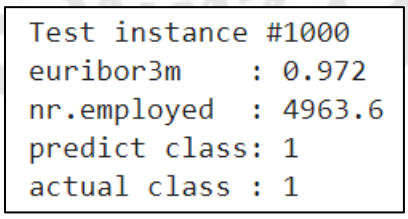
การอธิบายในระดับตัวอย่างข้อมูล

การอธิบายในระดับรายตัวอย่างข้อมูลจะสามารถทำให้อธิบายพฤติกรรมของแบบจำลองในการดำเนินการกับตัวอย่างข้อมูลนั้นๆ ได้อย่างแม่นยำและมีประสิทธิภาพ โดยสามารถวิเคราะห์ได้ว่าในแต่ละตัวอย่างข้อมูลแบบจำลองใช้งานคุณลักษณะใด และค่าจริงของคุณลักษณะนั้นๆ ส่งผลอย่างไรในการทำนายผลของตัวอย่างข้อมูลนั้นๆ ซึ่งเหมาะสมอย่างยิ่ง

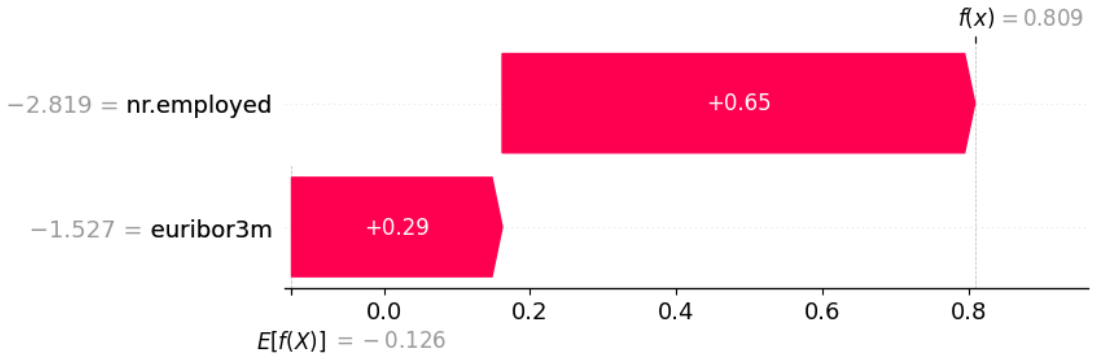
สำหรับการนำไปประยุกต์ใช้เพื่อประกอบการรองรับกระบวนการการช่วยตัดสินใจที่แม่นยำและมีประสิทธิภาพมากขึ้นในระดับรายตัวอย่างข้อมูล

จากภาพประกอบที่ 68 เป็นการใช้งานกราฟ Waterfall Plot เพื่ออธิบายผลของการทำนายตัวอย่างข้อมูลหมายเลข 1000 จากชุดข้อมูลสำหรับการทดสอบ โดยเป็นการอธิบายการทำนายตัวอย่างข้อมูลของแบบจำลอง XGB-ClassW-rfeLGBM ซึ่งใช้ชุดคุณลักษณะที่ประกอบด้วย 2 คุณลักษณะ คือ nr.employed และ euribor3m โดยผลลัพธ์ของตัวอย่างข้อมูลนี้ คือ ลูกค้าทำการสมัครผลิตภัณฑ์ (Positive Class หรือ Class 1) โดยจะเห็นได้ว่าคุณลักษณะ nr.employed ซึ่งมีค่าจริงของข้อมูลหลังทำการ Standard Scaling อยู่ที่ -2.819 จะมีค่า SHAP Value เป็น 0.65 ซึ่งส่งผลอย่างมากในการทำให้แบบจำลองทำนายเป็น Positive Class ส่วนคุณลักษณะ euribor3m ซึ่งมีค่าจริงของข้อมูลหลังทำการ Standard Scaling อยู่ที่ -1.527 จะมีค่า SHAP Value เป็น 0.29 ซึ่งส่งผลน้อยกว่าในการทำให้แบบจำลองทำนายเป็น Positive Class และเนื่องจากว่าแบบจำลองมีค่าฐาน หรือ Base Value ของ SHAP ตั้งต้นที่ -0.126 ดังนั้นเมื่อรวมผลของ SHAP Value ของทั้งสองคุณลักษณะ จะทำให้ได้ค่า SHAP Value รวมของตัวอย่างข้อมูลนี้ที่ 0.809 ซึ่งยังคงมากกว่า 0 ดังนั้นจึงทำให้แบบจำลองทำนายตัวอย่างข้อมูลเป็น Positive Class ซึ่งทำนายได้ถูกต้องดังรายละเอียดของตัวอย่างข้อมูลในภาพประกอบที่ 67

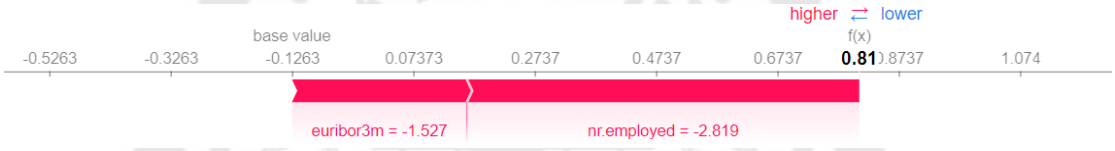
ภาพประกอบที่ 69 เป็นการวาดกราฟ Additive Force Plot ซึ่งสามารถใช้งานในการอธิบายแบบจำลองในการทำนายในระดับตัวอย่างข้อมูลได้เช่นกัน โดยกราฟจะต่างกับ Waterfall Plot เพียงเล็กน้อยในด้านของการแสดงผลและการจัดการข้อมูลจุดทศนิยม



ภาพประกอบ 67 รายละเอียดข้อมูลของตัวอย่างข้อมูลหมายเลข 1000 ประกอบด้วยค่าของคุณลักษณะ euribor3m และ nr.employed, ผลลัพธ์จากการทำนายของแบบจำลอง และผลลัพธ์จริงของตัวอย่างข้อมูล



ภาพประกอบ 68 กราฟ Waterfall Plot อธิบายความสำคัญของข้อมูลในแต่ละคุณลักษณะว่าส่งผลมากน้อยเพียงใดและในทางบวกหรือลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 1000

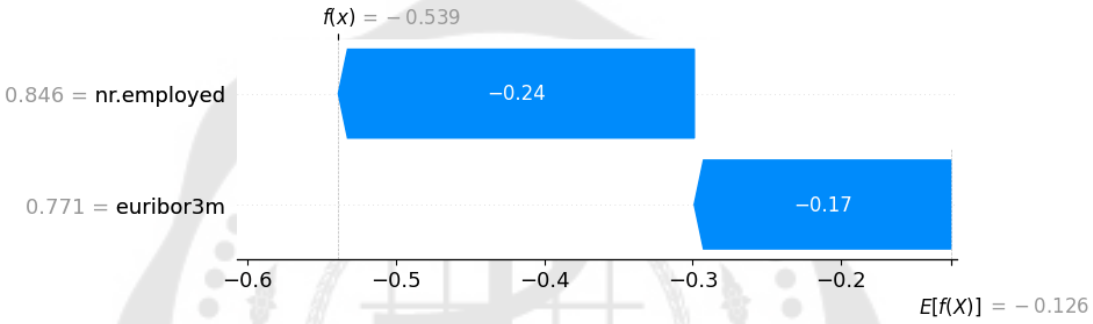


ภาพประกอบ 69 กราฟ Additive Force Plot อธิบายความสำคัญของข้อมูลในแต่ละคุณลักษณะว่าส่งผลมากน้อยเพียงใดและในทางบวกหรือลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 1000

ภาพประกอบที่ 71 และ 72 เป็นการใช้งานกราฟ Waterfall Plot และ Additive Force Plot เพื่ออธิบายผลของการทำนายตัวอย่างข้อมูลหมายเลข 1002 จากชุดข้อมูลสำหรับการทดสอบ ซึ่งใช้งานแบบจำลองและชุดข้อมูลสำหรับทดสอบเดียวกัน โดยผลลัพธ์ของตัวอย่างข้อมูลนี้คือ ลูกค้าไม่ทำการสมัครผลิตภัณฑ์ (Negative Class หรือ Class 0) โดยจะเห็นได้ว่าคุณลักษณะ nr.employed ซึ่งมีค่าจริงของข้อมูลหลังทำการ Standard Scaling อยู่ที่ 0.846 จะมีค่า SHAP Value เป็น -0.24 ซึ่งส่งผลในการทำให้แบบจำลองทำนายเป็น Negative Class ส่วนคุณลักษณะ euribor3m ซึ่งมีค่าจริงของข้อมูลหลังทำการ Standard Scaling อยู่ที่ 0.771 จะมีค่า SHAP Value เป็น -0.17 ซึ่งส่งผลน้อยกว่าในการทำให้แบบจำลองทำนายเป็น Negative Class และเนื่องจากว่าแบบจำลองมีค่าฐาน หรือ Base Value ของ SHAP ตั้งต้นที่ -0.126 ดังนั้นเมื่อรวมผลของ SHAP Value ของทั้งสองคุณลักษณะ จะทำให้ได้ค่า SHAP Value รวมของตัวอย่างข้อมูลนี้ที่ -0.539 ซึ่งน้อยกว่า 0 ดังนั้นจึงทำให้แบบจำลองทำนายตัวอย่างข้อมูลเป็น Negative Class ซึ่งทำนายได้ถูกต้องดังรายละเอียดของตัวอย่างข้อมูลในภาพประกอบที่ 70

```
Test instance #1002
euribor3m     : 0.972
nr.employed   : 4963.6
predict class: 1
actual class : 1
```

ภาพประกอบ 70 รายละเอียดข้อมูลของตัวอย่างข้อมูลหมายเลข 1002 ประกอบด้วยค่าของคุณลักษณะ euribor3m และ nr.employed, ผลลัพธ์จากการทำนายของแบบจำลอง และผลลัพธ์จริงของตัวอย่างข้อมูล



ภาพประกอบ 71 กราฟ Waterfall Plot อธิบายความสำคัญของข้อมูลในแต่ละคุณลักษณะว่าส่งผลมากน้อยเพียงใดและในทางบวกหรือลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 1002

ภาพประกอบ 72 กราฟ Additive Force Plot อธิบายความสำคัญของข้อมูลในแต่ละคุณลักษณะว่าส่งผลมากน้อยเพียงใดและในทางบวกหรือลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 1002

5.2 **การวิเคราะห์ความผิดพลาดของแบบจำลอง** (Error Analysis)

ในหัวข้อของการวิเคราะห์ความผิดพลาดจะเป็นส่วนของวิเคราะห์เพื่อค้นหาสาเหตุซึ่งส่งผลให้แบบจำลองมีการทำงานที่ผิดพลาด โดยจะมีการคัดเลือกแบบจำลองขึ้นมาหนึ่งแบบจำลองสำหรับการวิเคราะห์และอธิบายความผิดพลาดด้วยหลักการต่างๆ

แบบจำลองที่เลือกนำมาใช้สำหรับการวิเคราะห์ความผิดพลาดในงานวิจัยครั้งนี้ คือ แบบจำลอง XGB-ClassW-rfeLGBM ซึ่งเป็นการประยุกต์ใช้งาน XGBoost ร่วมกับ One-Hot Encoding และ Class Weight ในการจัดการความไม่สมดุลกันของข้อมูล โดยใช้งานชุดคุณลักษณะที่มีความสำคัญสูงสุดสองอันดับที่ได้จากการทำ RFE(LightGBM) ประกอบด้วย จำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m)

โดยค่าประสิทธิภาพของแบบจำลอง XGB-ClassW-rfeLGBM มีค่า Accuracy เป็น 0.72 (72%) และมีค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ที่ 0.71 (71%) โดยมีค่าประสิทธิภาพอื่นๆ ที่ได้จากการเรียกใช้งานฟังก์ชัน Classification Report ดังภาพที่ 73 และในส่วนจำนวนของการทำนายทั้งถูกและผิดของแบบจำลองเป็นไปตาม Confusion Matrix ดังภาพประกอบที่ 74

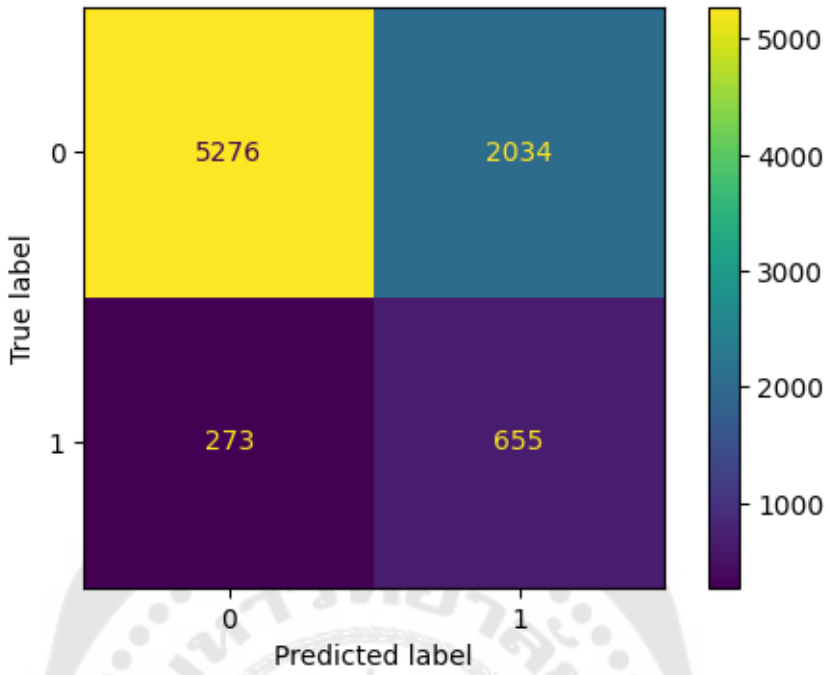
```
Train Accuracy : 0.7128204663109053
Test Accuracy : 0.7199563000728332

Classification Report :
              precision    recall  f1-score   support

           0       0.95      0.72      0.82      7310
           1       0.24      0.71      0.36       928

    accuracy                           0.72      8238
   macro avg       0.60      0.71      0.59      8238
weighted avg       0.87      0.72      0.77      8238
```

ภาพประกอบ 73 ค่าประสิทธิภาพต่างๆ ของแบบจำลอง XGB-ClassW-rfeLGBM จากการเรียกใช้งานฟังก์ชัน Classification Report



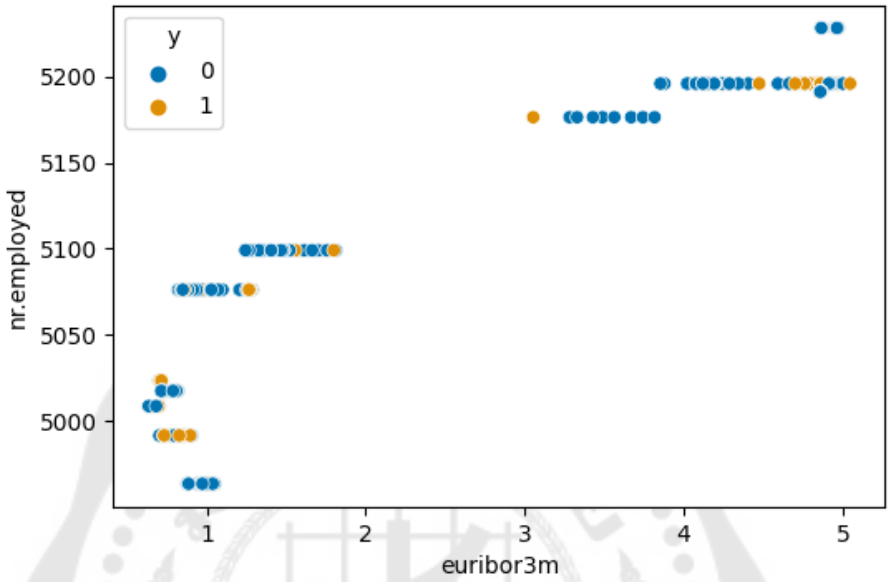
ภาพประกอบ 74 Confusion Matrix แสดงผลลัพธ์จากการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM

จาก Confusion Matrix ในภาพประกอบที่ 74 จะเห็นได้ว่าจากตัวอย่างข้อมูลทั้งหมดในชุดข้อมูลสำหรับการทดสอบจำนวน 8,238 ตัวอย่างข้อมูล มีจำนวนของการทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ได้ถูกต้อง (TP) จำนวน 655 ตัวอย่างข้อมูล จำนวนของการทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ได้ถูกต้อง (TN) จำนวน 5,276 ตัวอย่างข้อมูล จำนวนของการทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ผิดพลาด (FN) หรือทำนายเป็นไม่สมัครผลิตภัณฑ์จำนวน 273 ตัวอย่างข้อมูล และจำนวนของการทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) ผิดพลาด (FP) หรือทำนายเป็นสมัครผลิตภัณฑ์จำนวน 2,034 ตัวอย่างข้อมูล

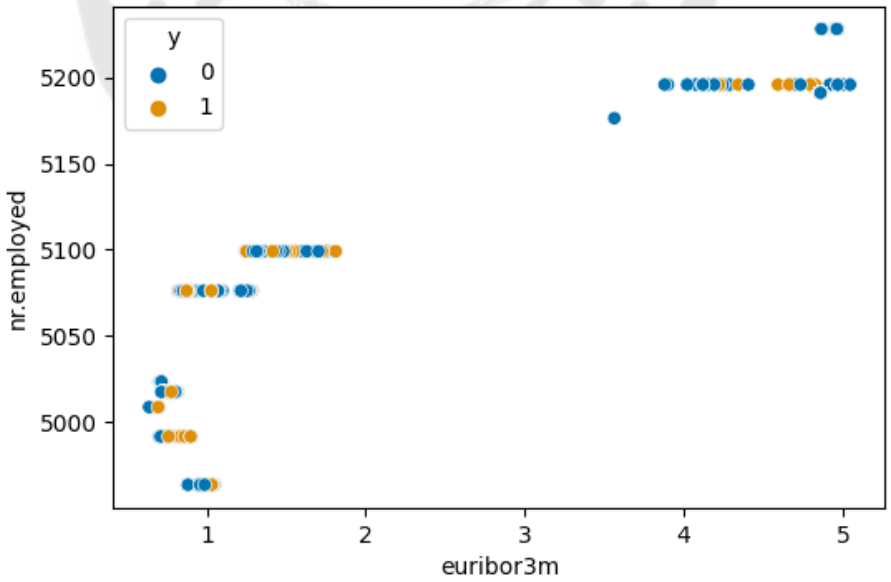
5.2.1 ความหนาแน่นและการกระจายตัวของข้อมูล

ในขั้นต้นของการวิเคราะห์ความผิดพลาดได้มีการวาดกราฟ Scatter Plot เพื่อแสดงความสัมพันธ์ระหว่างคุณลักษณะ euribor3m และ nr.employed ของชุดข้อมูลสำหรับการเรียนรู้และชุดข้อมูลสำหรับการทดสอบ ดังแสดงในภาพประกอบที่ 75 และ 76 ตามลำดับ โดยมีการแบ่งแยกผลลัพธ์ หรือ Class ของแต่ละตัวอย่างข้อมูลด้วยสี จากการสังเกตเบื้องต้นจะพบว่าตัวอย่างข้อมูลสามารถแบ่งแยกออกจากกันได้ชัดเจนเป็น 2 กลุ่ม คือ กลุ่มแรกจะมีค่า euribor3m

และ nr.employed ต่ำ และกลุ่มที่สองจะมีค่า euribor3m และ nr.employed สูง ซึ่งภายในแต่ละกลุ่มมีตัวอย่างข้อมูลทั้ง กลุ่มลูกค้าที่ทำการสมัคร (สีเหลือง) และ กลุ่มลูกค้าที่ไม่ทำการสมัคร (สีฟ้า) อยู่ปะปนกัน และไม่สามารถแยกจากกันได้อย่างชัดเจนนัก



ภาพประกอบ 75 กราฟ Scatter Plot แสดงความสัมพันธ์ระหว่างคุณลักษณะ euribor3m และ nr.employed ของชุดข้อมูลการเรียนรู้โดยใช้งานผลลัพธ์ของตัวอย่างข้อมูลในการจัดกลุ่ม

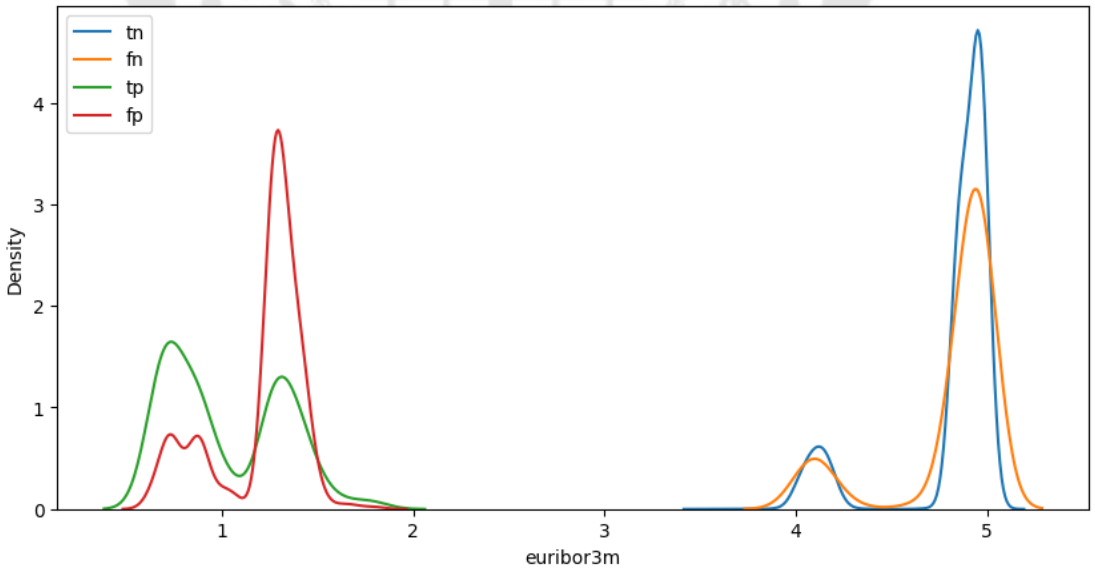


ภาพประกอบ 76 กราฟ Scatter Plot แสดงความสัมพันธ์ระหว่างคุณลักษณะ euribor3m และ nr.employed ของชุดข้อมูลการทดสอบโดยใช้งานผลลัพธ์ของตัวอย่างข้อมูลในการจัดกลุ่ม

ภาพประกอบที่ 77 เป็นการวาดกราฟเพื่อแสดงความหนาแน่นและการกระจายตัวของข้อมูลสำหรับคุณลักษณะ euribor3m จากชุดข้อมูลสำหรับทดสอบ โดยแบ่งข้อมูลความหนาแน่นและการกระจายตัวออกเป็น 4 กลุ่ม โดยใช้ผลลัพธ์จากการทำนายทั้ง 4 ประเภท จากการสังเกตพบว่าค่าของคุณลักษณะมีการทับซ้อนกันค่อนข้างสูงทั้งในส่วนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และไม่ทำการสมัครผลิตภัณฑ์ (Class 0)

กราฟเส้นสีเขียวจะแสดงความหนาแน่นและการกระจายตัวของค่า euribor3m ซึ่งเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ได้ถูกต้อง (TP) ส่วนกราฟเส้นสีแดงจะเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ผิดพลาด (FP) โดยจะเห็นได้ว่าการกระจายตัวของข้อมูลมีความทับซ้อนกันอยู่เป็นอย่างมาก โดยมีค่าที่สูงที่สุดอยู่ที่ไม่เกิน 3

กราฟเส้นสีฟ้าจะแสดงความหนาแน่นและการกระจายตัวของค่า euribor3m ซึ่งเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ได้ถูกต้อง (TN) ส่วนกราฟเส้นสีส้มจะเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ผิดพลาด (FN) ซึ่งจะเห็นได้ว่าการกระจายตัวของข้อมูลมีความทับซ้อนกันอยู่เป็นอย่างมากเช่นเดียวกัน โดยมีค่าตั้งแต่ประมาณ 3 ขึ้นไป

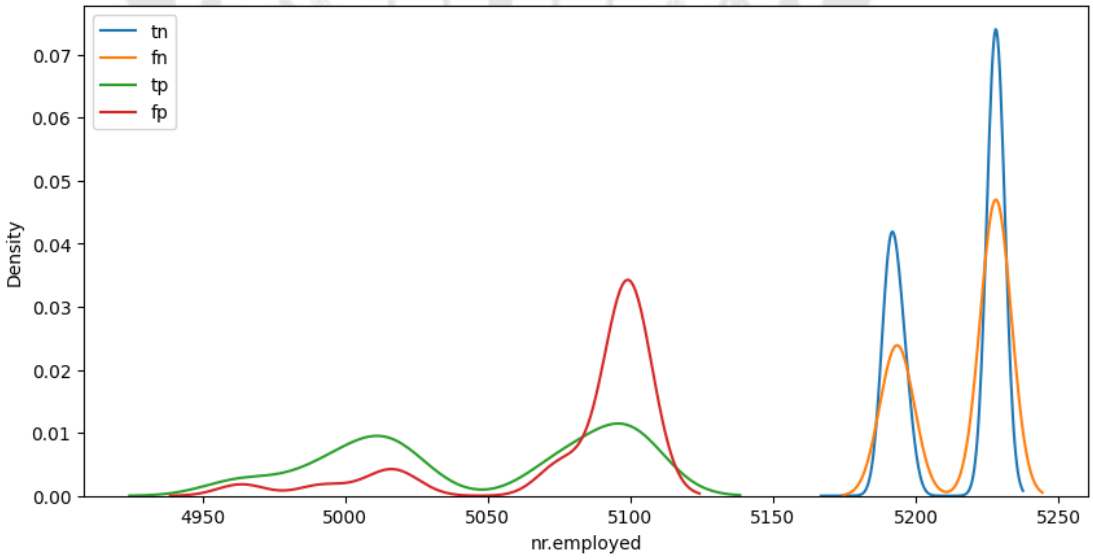


ภาพประกอบ 77 กราฟแสดงความหนาแน่นและการกระจายตัวของข้อมูลสำหรับคุณลักษณะ euribor3m จากชุดข้อมูลสำหรับการทดสอบโดยใช้งานผลลัพธ์จากการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM ในการจัดกลุ่ม

ภาพประกอบที่ 78 เป็นการวาดกราฟเพื่อแสดงความหนาแน่นและการกระจายตัวของข้อมูลสำหรับคุณลักษณะ nr.employed จากชุดข้อมูลสำหรับทดสอบ โดยแบ่งข้อมูลความหนาแน่นและการกระจายตัวออกเป็น 4 กลุ่ม โดยใช้ผลลัพธ์จากการทำนายทั้ง 4 ประเภท จากการสังเกตพบว่าค่าของคุณลักษณะมีการทับซ้อนกันค่อนข้างสูงทั้งในส่วนของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) และไม่ทำการสมัครผลิตภัณฑ์ (Class 0)

กราฟเส้นสีเขียวจะแสดงความหนาแน่นและการกระจายตัวของค่า nr.employed ซึ่งเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ได้ถูกต้อง (TP) ส่วนกราฟเส้นสีแดงจะเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ผิดพลาด (FP) โดยจะเห็นได้ว่าการกระจายตัวของข้อมูลมีความทับซ้อนกันอยู่เป็นอย่างมาก โดยมีค่าที่สูงที่สุดอยู่ที่ไม่เกิน 5,150

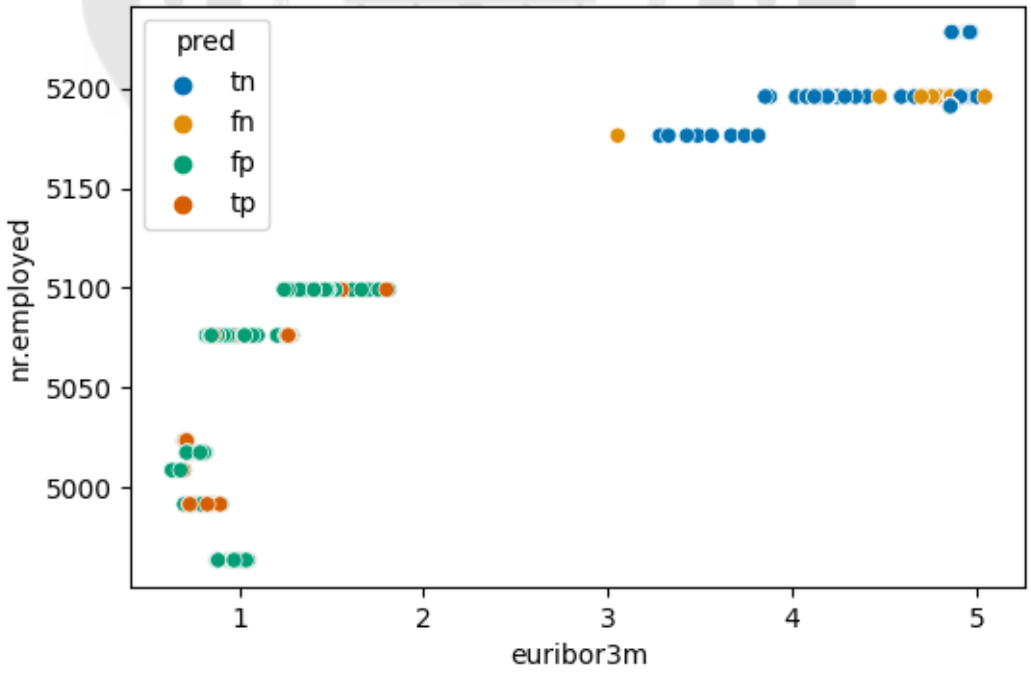
กราฟเส้นสีฟ้าจะแสดงความหนาแน่นและการกระจายตัวของค่า nr.employed ซึ่งเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ได้ถูกต้อง (TN) ส่วนกราฟเส้นสีส้มจะเป็นกลุ่มข้อมูลที่แบบจำลองทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ผิดพลาด (FN) ซึ่งจะเห็นได้ว่าการกระจายตัวของข้อมูลมีความทับซ้อนกันอยู่เป็นอย่างมากเช่นเดียวกัน โดยมีค่าตั้งแต่ประมาณ 5,150 ขึ้นไป



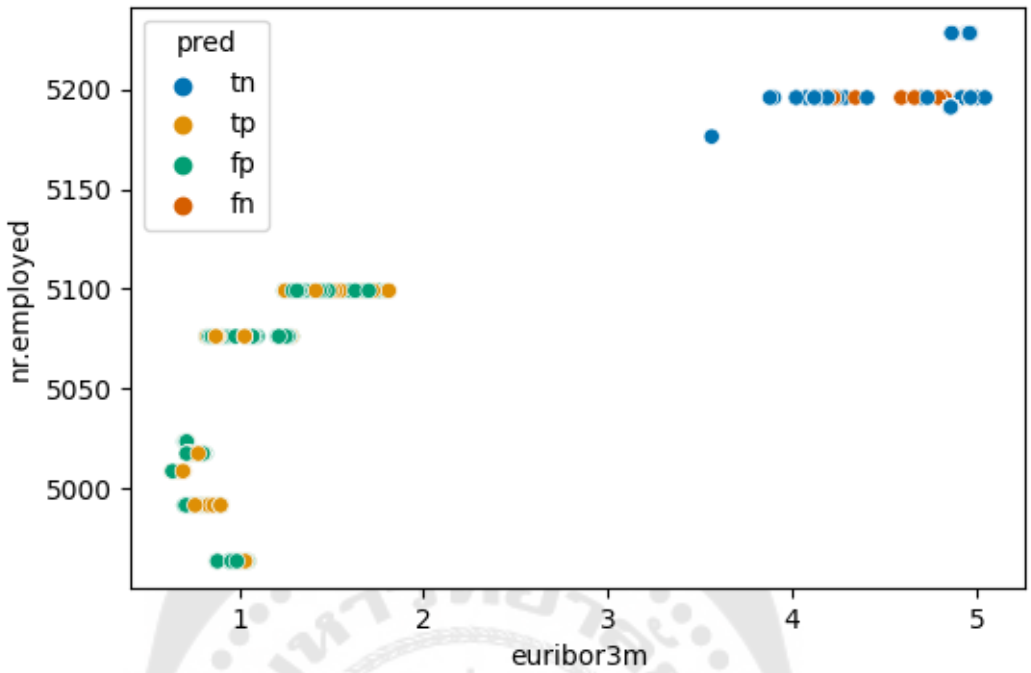
ภาพประกอบ 78 กราฟแสดงความหนาแน่นและการกระจายตัวของข้อมูลสำหรับคุณลักษณะ nr.employed จากชุดข้อมูลสำหรับการทดสอบโดยใช้งานผลลัพธ์จากการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM ในการจัดกลุ่ม

เบื้องต้นจากการสังเกตกราฟความหนาแน่นและการกระจายตัวของข้อมูลทำให้สามารถสรุปได้ว่าเนื่องจากค่าของสองคุณลักษณะที่นำมาใช้ในการเรียนรู้ของแบบจำลองค่อนข้างมีความทับซ้อนกันอยู่ โดยไม่สามารถแบ่งแยกกลุ่มของผลลัพธ์ได้อย่างชัดเจน ดังนั้นจึงทำให้แบบจำลองไม่สามารถแยกแยะบางตัวอย่างข้อมูลที่ค่าของคุณลักษณะไปตกภายในบริเวณของอีกลุ่มได้อย่างถูกต้อง

เมื่อนำผลลัพธ์จากการทำนายของแบบจำลองมาใช้สำหรับแบ่งแยกข้อมูลในการวาดกราฟ Scatter Plot เพื่อแสดงความสัมพันธ์ระหว่างคุณลักษณะ euribor3m และ nr.employed ของชุดข้อมูลสำหรับการเรียนรู้และชุดข้อมูลสำหรับการทดสอบ ดังแสดงในภาพประกอบที่ 79 และ 80 ตามลำดับ โดยจะพบว่าแม้ใช้ชุดข้อมูลสำหรับการเรียนรู้ในการทำนายก็ยังสามารถเกิดความผิดพลาดในการทำนายได้ (สีเหลืองและเขียว) โดยกลุ่มข้อมูลที่มีค่า euribor3m และ nr.employed ต่ำ แบบจำลองจะทำนายเป็น กลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (Class 1) ส่วนกลุ่มข้อมูลที่มีค่า euribor3m และ nr.employed สูง จะถูกทำนายเป็น กลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (Class 0) โดยในการทำนายชุดข้อมูลสำหรับการทดสอบก็มีแนวโน้มในการทำนายเป็นไปในทางเดียวกัน



ภาพประกอบ 79 กราฟ Scatter Plot แสดงความสัมพันธ์ระหว่างคุณลักษณะ euribor3m และ nr.employed ของชุดข้อมูลสำหรับการเรียนรู้โดยใช้งานผลลัพธ์จากการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM ในการจัดกลุ่ม



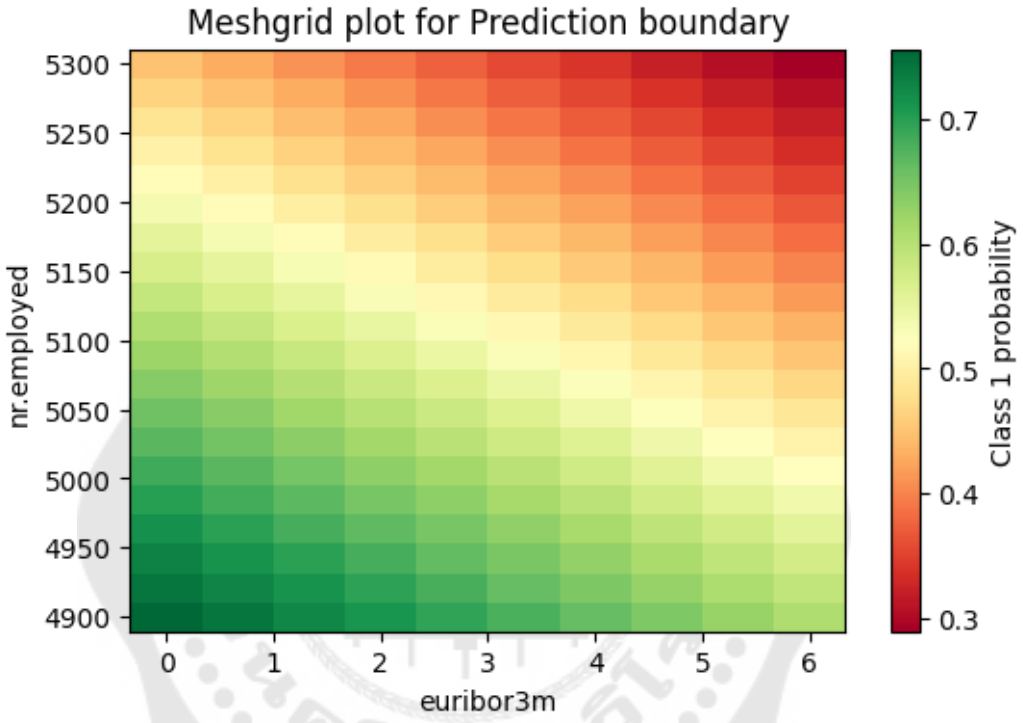
ภาพประกอบ 80 กราฟ Scatter Plot แสดงความสัมพันธ์ระหว่างคุณลักษณะ euribor3m และ nr.employed ของชุดข้อมูลสำหรับการทดสอบโดยใช้งานผลลัพธ์จากการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM ในการจัดกลุ่ม

ภาพประกอบที่ 81 เป็นกราฟ Meshgrid เพื่อแสดงขอบเขตการตัดสินใจ หรือ Decision Boundary ของแบบจำลอง XGB-ClassW-rfeLGBM ว่าค่าของคุณลักษณะ euribor3m และ nr.employed ในแต่ละคู่จะส่งผลต่อการทำนายของแบบจำลองในลักษณะใด การวาดกราฟเลือกใช้งานสีเพื่อแสดงค่าความน่าจะเป็นของผลลัพธ์การสมัครผลิตภัณฑ์ หรือ Positive Class หรือ Class 1 ของแต่ละคู่ข้อมูลภายในกราฟ โดยสีเขียวจะหมายถึงการมีค่าความน่าจะเป็นของการสมัครผลิตภัณฑ์ที่สูง (มากกว่า 0.5) ซึ่งส่งผลให้แบบจำลองทำนายข้อมูลในบริเวณสีเขียวเป็นลูกค้าที่ทำการสมัครผลิตภัณฑ์ ในส่วนของสีแดงจะหมายถึงการมีค่าความน่าจะเป็นของการสมัครผลิตภัณฑ์ที่ต่ำ (น้อยกว่า 0.5) ซึ่งจะส่งผลให้แบบจำลองทำนายข้อมูลในบริเวณสีแดงเป็นลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์

โดยเมื่อค่าของคุณลักษณะ euribor3m และ nr.employed มีค่าที่ค่อนไปในทางสูงทั้งคู่ แบบจำลองจะทำนายความน่าจะเป็นของการสมัครผลิตภัณฑ์ออกมาต่ำ ส่วนเมื่อค่าของทั้งสองคุณลักษณะมีค่าที่ค่อนไปในทางต่ำ แบบจำลองจะทำนายความน่าจะเป็นของการสมัคร

ผลิตภัณฑ์ออกมาค่อนข้างสูง โดยค่าความน่าจะเป็นของการสมัครผลิตภัณฑ์ที่แบบจำลองทำนายออกมาอยู่ในช่วงระหว่าง 0.2 ถึง 0.8

เมื่อเปรียบเทียบกราฟ Meshgrid กับกราฟผลลัพธ์จากการทำนายในภาพประกอบที่ 79 และ 80 จะเห็นได้ว่าแบบจำลองทำการทำนายตัวอย่างข้อมูลโดยเป็นไปในลักษณะเดียวกับขอบเขตการตัดสินใจที่สังเกตได้จากกราฟ Meshgrid



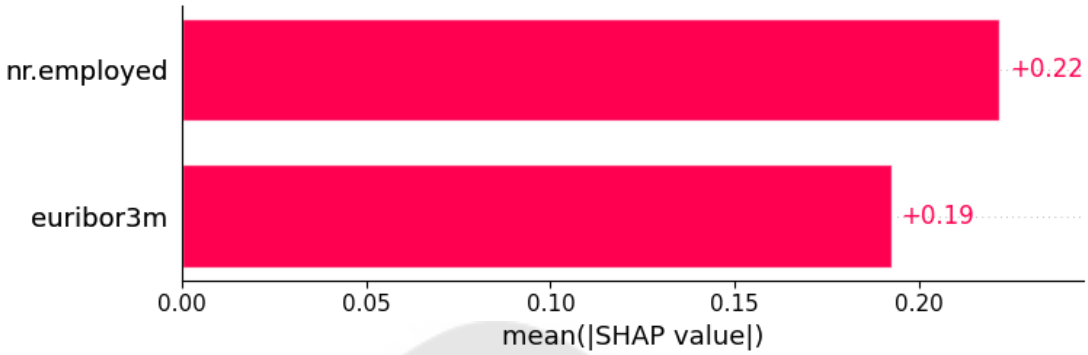
ภาพประกอบ 81 กราฟ Meshgrid เพื่อแสดงขอบเขตการตัดสินใจ หรือ Decision Boundary ของแบบจำลอง XGB-ClassW-rfeLGBM

5.2.2 การอธิบายแบบจำลองและผลการทำนายด้วยเรียนรู้ด้วยเครื่องแบบอธิบายได้

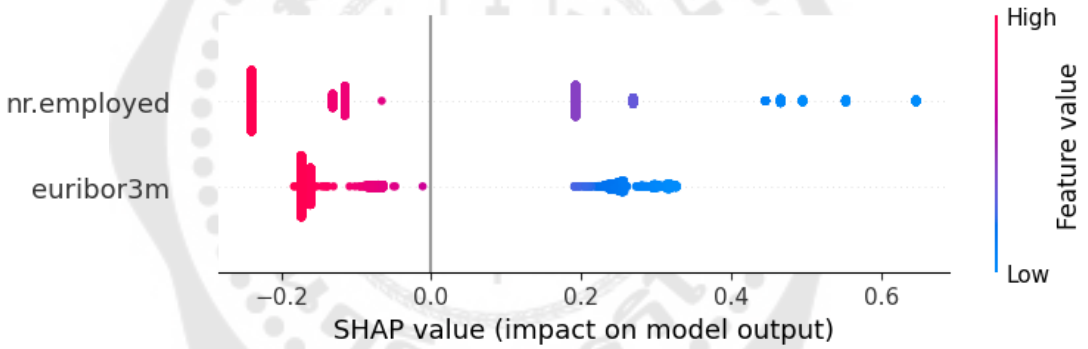
การอธิบายในระดับแบบจำลอง

จากการใช้งาน SHAP เพื่ออธิบายแบบจำลองดังภาพประกอบที่ 82 จะพบว่าคุณลักษณะ nr.employed มีอิทธิพลและความสำคัญต่อการทำนายของแบบจำลองที่สูงกว่าคุณลักษณะ euribor3m อยู่เล็กน้อย โดยมีค่า SHAP Value อยู่ที่ 0.22 และ 0.19 ตามลำดับ โดยในส่วนของกราฟในภาพประกอบที่ 83 สามารถบ่งบอกได้ว่าตัวอย่างข้อมูลที่มีค่าของ nr.employed ที่สูงจะมีค่าของ SHAP Value ที่ต่ำ ซึ่งจะส่งผลให้แบบจำลองมีโอกาสทำนายผลลัพธ์เป็น Negative Class ในทางกลับกันตัวอย่างข้อมูลที่มีค่าของ nr.employed ที่ต่ำจะมีค่า

ของ SHAP Value ที่สูง ซึ่งจะส่งผลให้แบบจำลองมีโอกาสทำนายผลลัพธ์เป็น Positive Class มากขึ้น ซึ่งในส่วนของคุณลักษณะ euribor3m ก็สามารถแปลผลได้เป็นไปในทิศทางเดียวกัน

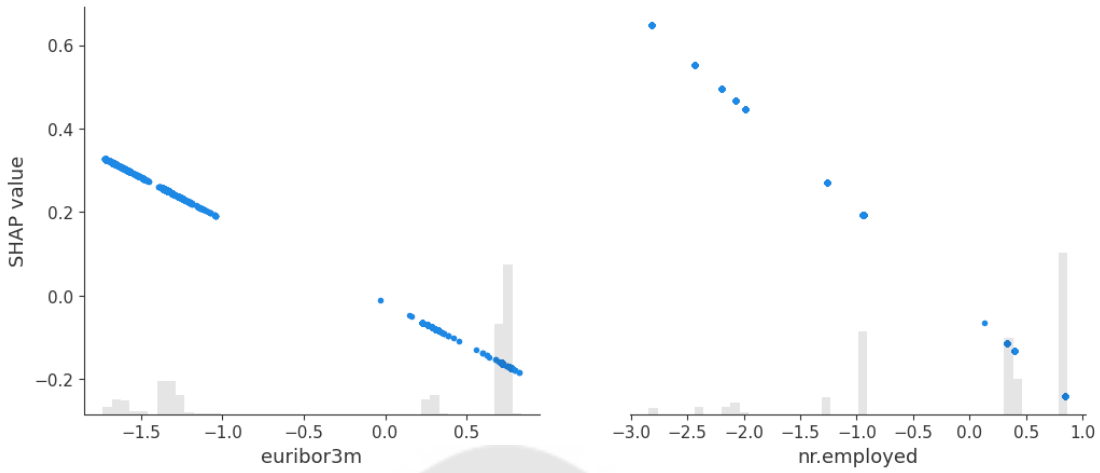


ภาพประกอบ 82 กราฟ Bar Plot แสดงความสำคัญของคุณลักษณะ หรือ ค่า SHAP Value ที่ส่งผลต่อการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM



ภาพประกอบ 83 กราฟ Beeswamp Plot แสดงความสัมพันธ์ระหว่างค่าของคุณลักษณะและค่าความสำคัญของคุณลักษณะ หรือ SHAP Value ที่ส่งผลต่อการทำนายของแบบจำลอง XGB-ClassW-rfeLGBM

ภาพประกอบที่ 84 เป็นกราฟ Scatter Plot ของทั้งสองคุณลักษณะเพื่อแสดงความสัมพันธ์ระหว่างค่าของคุณลักษณะ และค่าของ SHAP Value โดยแกนตั้งจะเป็นค่าของ SHAP Value และแกนนอนจะเป็นค่าของคุณลักษณะหลังการทำ Standard Scaling แล้ว ซึ่งจะเห็นได้ว่ามีแนวโน้มความสัมพันธ์ไปในทิศทางเดียวกันกับการแปลผลจากกราฟ Bar Plot และ Beeswamp Plot



ภาพประกอบ 84 กราฟ Scatter Plot แสดงความสัมพันธ์ระหว่างค่าของคุณลักษณะและค่าความสำคัญของคุณลักษณะ หรือ SHAP Value ของคุณลักษณะ euribor3m และ nr.employed

การอธิบายในระดับตัวอย่างข้อมูล

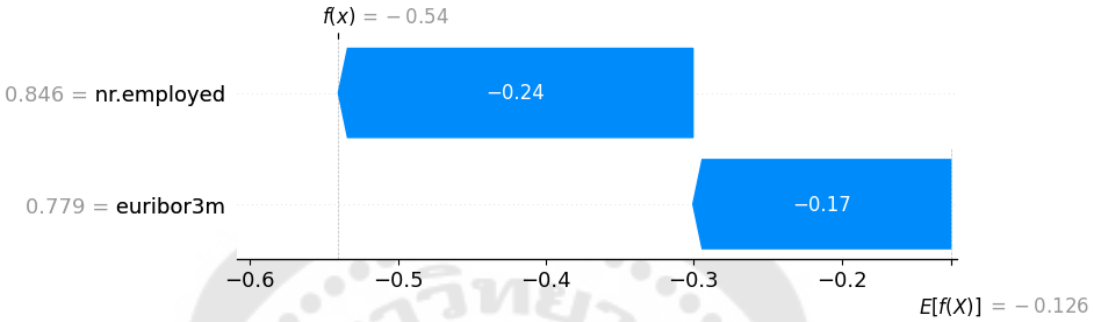
การทำนายผิดพลาดของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ (FN)

จากแบบจำลองที่คัดเลือกมาเพื่อวิเคราะห์ความผิดพลาด มีการดึงข้อมูลของตัวอย่างข้อมูลหมายเลข 211 จากชุดข้อมูลสำหรับทดสอบขึ้นมาเพื่อทำการวิเคราะห์และเปรียบเทียบการทำงานของแบบจำลอง โดยข้อมูลค่าจริงของคุณลักษณะ ผลลัพธ์จากการทำนายและผลลัพธ์จริงของตัวอย่างข้อมูลเป็นไปตามภาพประกอบที่ 85 โดยมีค่าของคุณลักษณะ euribor3m ที่ 4.97 และค่าของคุณลักษณะ nr.employed เป็น 5228.1 โดยผลลัพธ์จริงของตัวอย่างข้อมูลคือ Class 0 หรือ ไม่ทำการสมัครผลิตภัณฑ์ โดยแบบจำลองสามารถทำนายตัวอย่างข้อมูลหมายเลข 211 ได้ถูกต้อง

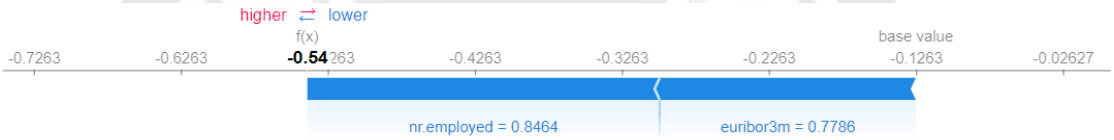
```
Test instance #211
euribor3m    : 4.97
nr.employed  : 5228.1
predict class: 0
actual class : 0
```

ภาพประกอบ 85 รายละเอียดข้อมูลของตัวอย่างข้อมูลหมายเลข 211 ประกอบด้วยค่าของคุณลักษณะ euribor3m และ nr.employed, ผลลัพธ์จากการทำนายของแบบจำลอง และผลลัพธ์จริงของตัวอย่างข้อมูล

ภาพประกอบที่ 86 และ 87 เป็นการวาดกราฟ Waterfall Plot และ Additive Force Plot จาก SHAP Value ของตัวอย่างข้อมูลหมายเลข 211 ซึ่งค่าของคุณลักษณะส่งผลให้ค่า SHAP Value ของทั้งสองคุณลักษณะมีค่าเป็นลบ เมื่อทำการรวมค่า SHAP Value เข้ากับค่า Base Value แล้วจะได้ผลรวมออกมาเป็น -0.54 ซึ่งน้อยกว่า 0 จึงทำให้แบบจำลองทำนายออกมาเป็น Negative Class



ภาพประกอบ 86 กราฟ Waterfall Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 211



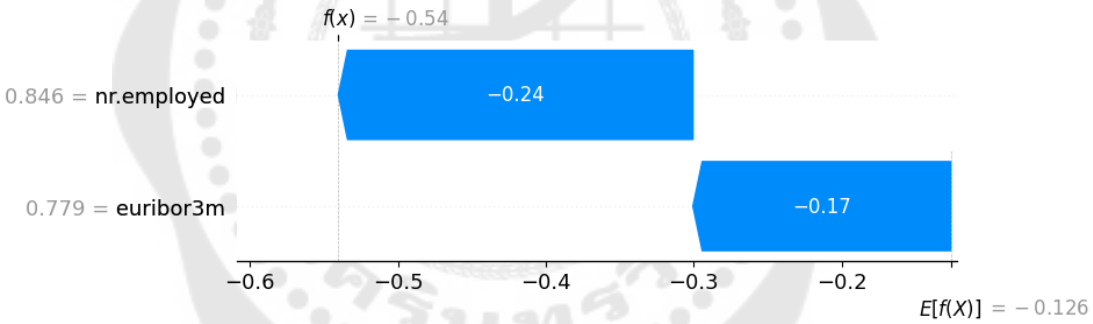
ภาพประกอบ 87 กราฟ Additive Force Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 211

ตัวอย่างข้อมูลหมายเลข 2889 จากชุดข้อมูลสำหรับทดสอบมีข้อมูลค่าจริงของคุณลักษณะ ผลลัพธ์จากการทำนาย และผลลัพธ์จริงของตัวอย่างข้อมูลเป็นไปตามภาพประกอบที่ 88 โดยมีค่าของคุณลักษณะ euribor3m ที่ 4.97 และค่าของคุณลักษณะ nr.employed เป็น 5228.1 ซึ่งทั้งสองคุณลักษณะมีค่าเท่ากันกับตัวอย่างข้อมูลหมายเลข 211 แต่ผลลัพธ์จริงของตัวอย่างข้อมูลนี้จะเป็น Class 1 หรือ ทำการสมัครผลิตภัณฑ์ โดยแบบจำลองทำนายตัวอย่างข้อมูลหมายเลข 2889 ผิดพลาด

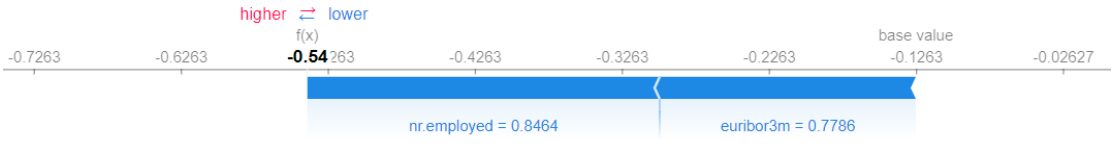
```
Test instance #2889
euribor3m    : 4.97
nr.employed  : 5228.1
predict class: 0
actual class : 1
```

ภาพประกอบ 88 รายละเอียดข้อมูลของตัวอย่างข้อมูลหมายเลข 2889 ประกอบด้วยค่าของคุณลักษณะ euribor3m และ nr.employed, ผลลัพธ์จากการทำนายของแบบจำลอง และผลลัพธ์จริงของตัวอย่างข้อมูล

จากภาพประกอบที่ 89 และ 90 จะสังเกตได้ว่ากราฟที่ได้จาก SHAP Value ของตัวอย่างข้อมูลหมายเลข 2889 มีลักษณะที่เหมือนกันกับกราฟที่ได้จาก SHAP Value ของตัวอย่างข้อมูลหมายเลข 211 โดยสาเหตุเนื่องมาจากว่าค่าจริงของทั้งสองคุณลักษณะมีค่าเท่ากันในทั้งสองตัวอย่างข้อมูล ส่งผลให้แบบจำลองทำนายตัวอย่างข้อมูลหมายเลข 2889 ออกมาผิดพลาดเป็น Negative Class แทนที่ Positive Class



ภาพประกอบ 89 กราฟ Waterfall Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 2889



ภาพประกอบ 90 กราฟ Additive Force Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางลบต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 2889

จากข้อมูลเบื้องต้นจึงสามารถสรุปได้ว่าเนื่องจากทั้งสองตัวอย่างข้อมูลมีค่าของ euribor3m ที่สูงกว่า 3 และค่าของ nr.employed ที่สูงกว่า 5,150 จึงส่งผลให้แบบจำลองทำนายทั้งสองตัวอย่างข้อมูลออกมาเป็น Negative Class

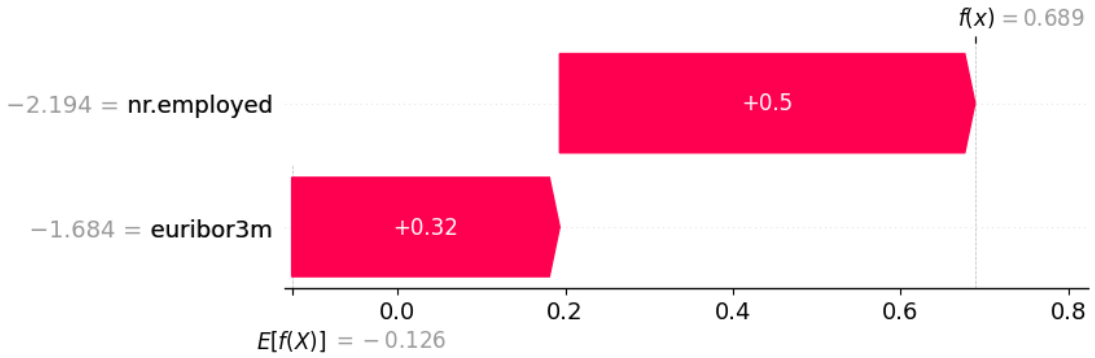
การทำนายผิดพลาดของกลุ่มลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (FP)

ตัวอย่างข้อมูลหมายเลข 8 จากชุดข้อมูลสำหรับทดสอบมีข้อมูลค่าจริงของคุณลักษณะ ผลลัพธ์จากการทำนาย และผลลัพธ์จริงของตัวอย่างข้อมูลเป็นไปตามภาพประกอบที่ 91 โดยมีค่าของคุณลักษณะ euribor3m ที่ 0.701 และค่าของคุณลักษณะ nr.employed เป็น 5008.7 โดยผลลัพธ์จริงของตัวอย่างข้อมูลคือ Class 1 หรือ ทำการสมัครผลิตภัณฑ์ โดยแบบจำลองสามารถทำนายตัวอย่างข้อมูลหมายเลข 8 ได้อย่างถูกต้อง

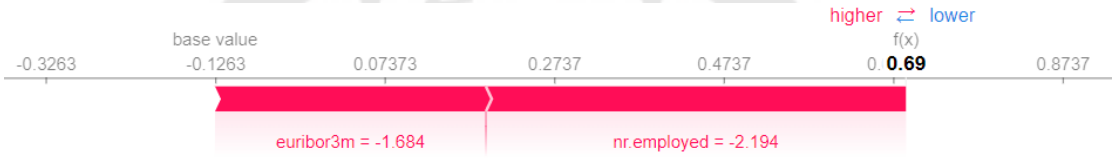
```
Test instance #8
euribor3m    : 0.701
nr.employed  : 5008.7
predict class: 1
actual class : 1
```

ภาพประกอบ 91 รายละเอียดข้อมูลของตัวอย่างข้อมูลหมายเลข 8 ประกอบด้วยค่าของคุณลักษณะ euribor3m และ nr.employed, ผลลัพธ์จากการทำนายของแบบจำลอง และผลลัพธ์จริงของตัวอย่างข้อมูล

ภาพประกอบที่ 92 และ 93 เป็นการวาดกราฟ Waterfall Plot และ Additive Force Plot จาก SHAP Value ของตัวอย่างข้อมูลหมายเลข 8 ซึ่งค่าของคุณลักษณะส่งผลให้ค่า SHAP Value ของทั้งสองคุณลักษณะมีค่าเป็นบวก เมื่อทำการรวมค่า SHAP Value เข้ากับค่า Base Value แล้วจะได้ผลรวมออกมาเป็น 0.689 ซึ่งมากกว่า 0 จึงทำให้แบบจำลองทำนายออกมาเป็น Positive Class



ภาพประกอบ 92 กราฟ Waterfall Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางบวกต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 8



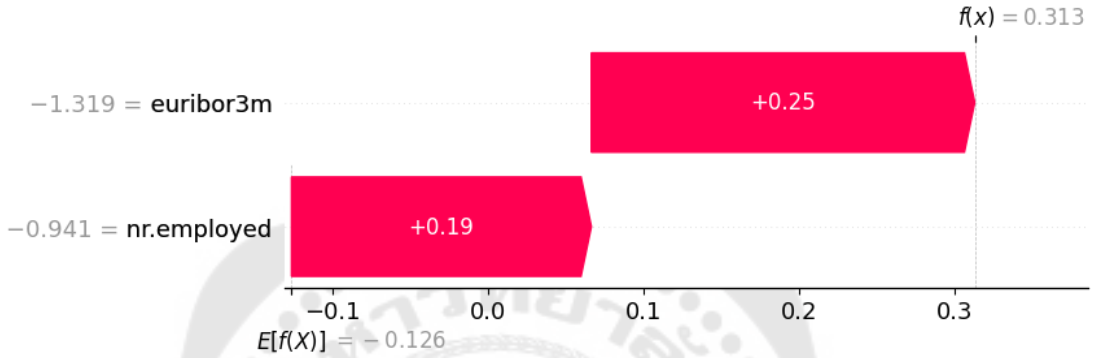
ภาพประกอบ 93 กราฟ Additive Force Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางบวกต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 8

ตัวอย่างข้อมูลหมายเลข 21 จากชุดข้อมูลสำหรับทดสอบมีข้อมูลค่าจริงของคุณลักษณะ ผลลัพธ์จากการทำนาย และผลลัพธ์จริงของตัวอย่างข้อมูลเป็นไปตามภาพประกอบที่ 94 โดยมีค่าของคุณลักษณะ euribor3m ที่ 0.739 และค่าของคุณลักษณะ nr.employed เป็น 5017.5 โดยผลลัพธ์จริงของตัวอย่างข้อมูลคือ Class 0 หรือ ไม่ทำการสมัครผลิตภัณฑ์ โดยแบบจำลองทำนายตัวอย่างข้อมูลหมายเลข 21 ผิดพลาดเป็น Class 1

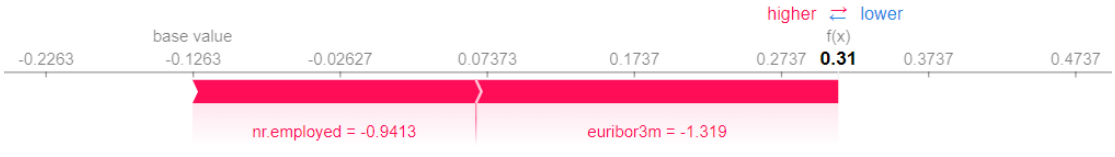
```
Test instance #21
euribor3m    : 0.739
nr.employed  : 5017.5
predict class: 1
actual class : 0
```

ภาพประกอบ 94 รายละเอียดข้อมูลของตัวอย่างข้อมูลหมายเลข 21 ประกอบด้วยค่าของคุณลักษณะ euribor3m และ nr.employed, ผลลัพธ์จากการทำนายของแบบจำลอง และผลลัพธ์จริงของตัวอย่างข้อมูล

จากภาพประกอบที่ 95 และ 96 จะสังเกตได้ว่ากราฟที่ได้จาก SHAP Value ของตัวอย่างข้อมูลหมายเลข 21 มีแนวโน้มที่เหมือนกันกับกราฟที่ได้จาก SHAP Value ของตัวอย่างข้อมูลหมายเลข 8 โดยสาเหตุเนื่องมาจากว่าค่าจริงของทั้งสองคุณลักษณะมีค่าที่ใกล้เคียงกันในทั้งสองตัวอย่างข้อมูล ส่งผลให้แบบจำลองทำนายตัวอย่างข้อมูลหมายเลข 21 ออกมาผิดพลาดเป็น Positive Class แทนที่ Negative Class โดยค่า SHAP Value อยู่ที่ 0.313



ภาพประกอบ 95 กราฟ Waterfall Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางบวกต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 21



ภาพประกอบ 96 กราฟ Additive Force Plot อธิบายความสำคัญของคุณลักษณะ nr.employed และ euribor3m ซึ่งส่งผลในทางบวกต่อการทำนายของแบบจำลองของตัวอย่างข้อมูลหมายเลข 21

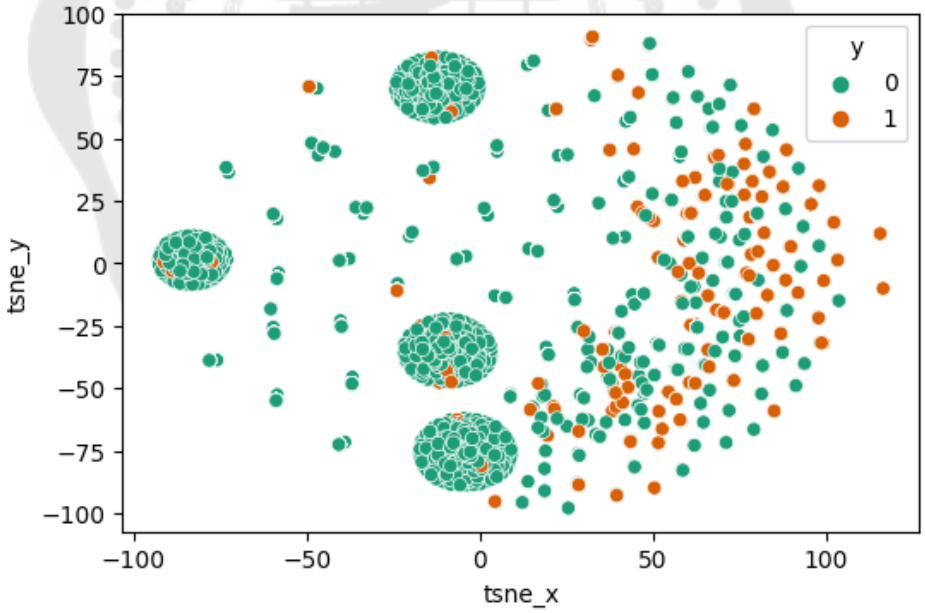
จากข้อมูลเบื้องต้นสามารถสรุปได้ว่าเนื่องจากทั้งสองตัวอย่างข้อมูลมีค่าของ euribor3m ที่ต่ำกว่า 3 และค่าของ nr.employed ที่ต่ำกว่า 5,150 จึงส่งผลให้แบบจำลองทำนายทั้งสองตัวอย่างข้อมูลออกมาเป็น Positive Class

5.2.3 การวิเคราะห์ความผิดพลาดด้วยกราฟ T-SNE

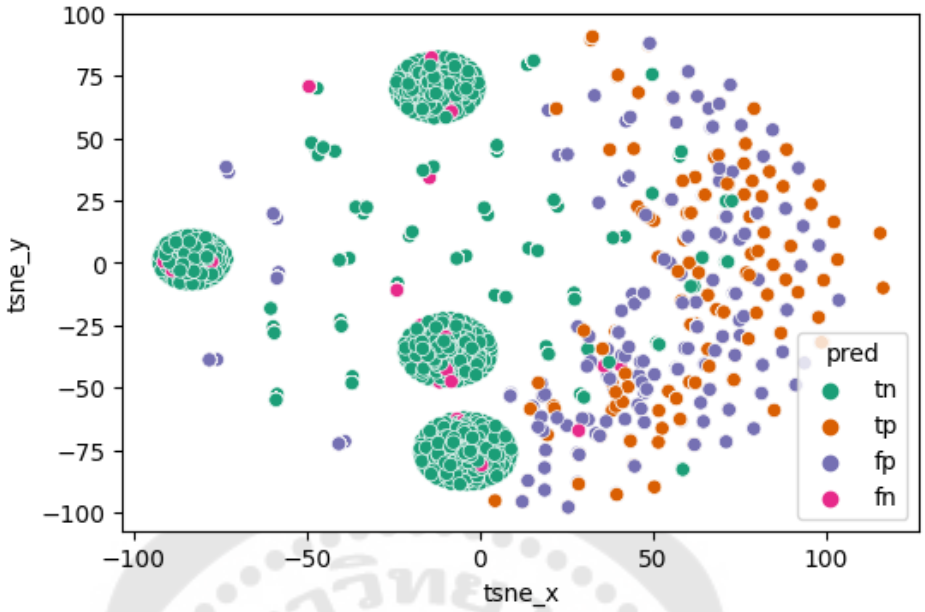
การใช้งาน T-SNE เป็นการลดมิติของข้อมูลแบบไม่เชิงเส้นและจัดกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันให้มาอยู่ในบริเวณเดียวกัน ส่วนข้อมูลที่มีลักษณะต่างกันจะถูกผลักออกเพื่อให้ห่างจากกัน โดยในการวาดกราฟ T-SNE ได้มีการดำเนินการกับชุดข้อมูลสำหรับการ

ทดสอบเพื่อดูแนวโน้มในการจัดกลุ่มของข้อมูล โดยมีการปรับแต่ง Hyperparameter ของ Perplexity เพื่อค้นหาการแสดงผลของการจัดกลุ่มข้อมูลที่เหมาะสมและมีประสิทธิภาพ

ภาพประกอบที่ 97 และ 98 เป็นการวาดกราฟ T-SNE โดยใช้งานค่า Perplexity เท่ากับ 2 เพื่อดำเนินการกับค่าจริงของคุณลักษณะของตัวอย่างข้อมูล โดยมีการใช้สีเพื่อแบ่งกลุ่มของข้อมูล ภาพประกอบที่ 97 เป็นการแบ่งกลุ่มด้วยสีโดยใช้ผลลัพธ์ของตัวอย่างข้อมูล ซึ่งจะเห็นได้ว่ากลุ่มข้อมูลของลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (สีเขียว) ส่วนใหญ่จะเกาะกลุ่มกันค่อนข้างหนาแน่นจำนวน 4 กลุ่ม ส่วนข้อมูลที่กระจัดกระจายออกไปทางฝั่งขวาจะมีการปะปนกันไประหว่างข้อมูลทั้งสองกลุ่ม ซึ่งส่งผลให้การทำนายตัวอย่างข้อมูลในบริเวณนั้นทำได้ยากและมีโอกาสผิดพลาดสูงขึ้น ส่วนภาพประกอบที่ 98 เป็นการแบ่งกลุ่มด้วยสีโดยใช้ผลลัพธ์จากการทำนายของแบบจำลอง โดยกลุ่มของตัวอย่างข้อมูลที่เกาะกลุ่มกันแน่นหนาซึ่งเป็นกลุ่มข้อมูลของลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ จะเห็นได้ว่าแบบจำลองสามารถทำนายได้ค่อนข้างถูกต้อง ส่วนตัวอย่างข้อมูลที่มีการปะปนกันของข้อมูลทั้งสองกลุ่มซึ่งอยู่ทางขวาของกราฟ จะเห็นได้ว่าผลการทำนายส่วนใหญ่ของแบบจำลองจะมีการทำนายเป็นกลุ่มลูกค้าที่สมัครผลิตภัณฑ์

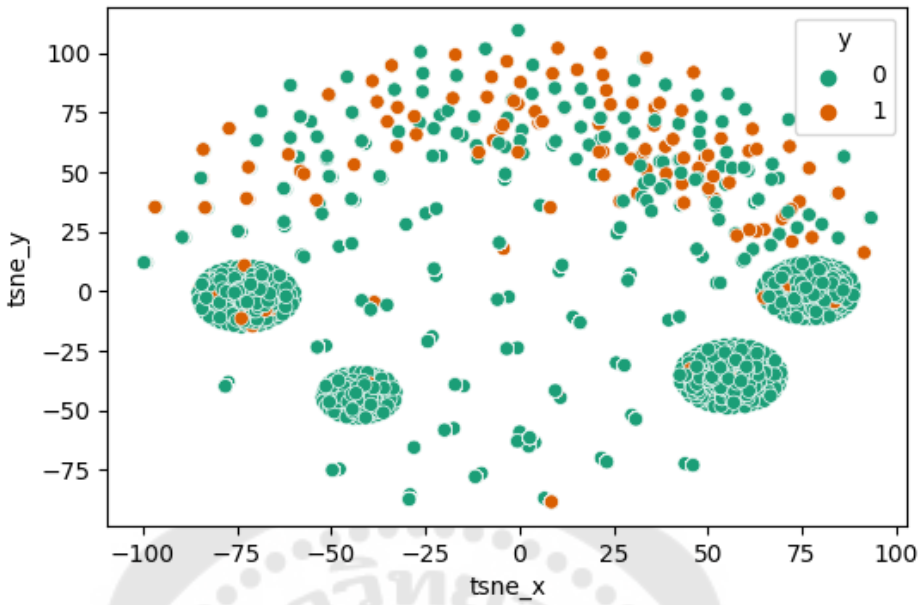


ภาพประกอบ 97 กราฟ T-SNE โดยใช้งานค่า Perplexity เท่ากับ 2 แสดงการกระจายตัวของข้อมูลจากค่าของคุณลักษณะของตัวอย่างข้อมูลโดยใช้ผลลัพธ์ของตัวอย่างข้อมูลในการแบ่งกลุ่ม

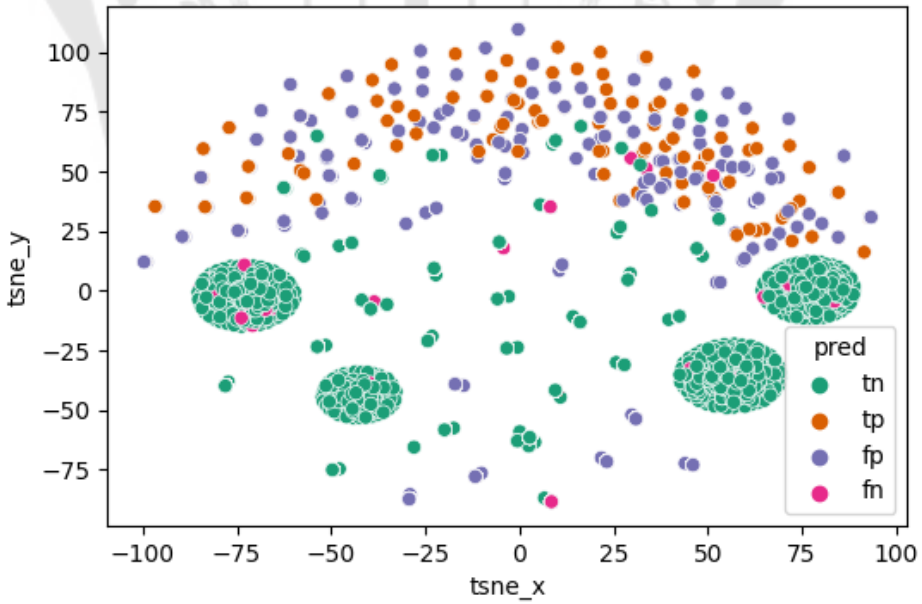


ภาพประกอบ 98 กราฟ T-SNE โดยใช้งานค่า Perplexity เท่ากับ 2 แสดงการกระจายตัวของข้อมูลจากค่าของคุณลักษณะของตัวอย่างข้อมูลโดยใช้ผลลัพธ์จากการทำนายของแบบจำลองในการแบ่งกลุ่ม

ภาพประกอบที่ 99 และ 100 เป็นการวาดกราฟ T-SNE โดยใช้งานค่า Perplexity เท่ากับ 2 เพื่อดำเนินการกับ SHAP Value ของคุณลักษณะของแต่ละตัวอย่างข้อมูล โดยมีการใช้สีเพื่อแบ่งกลุ่มของข้อมูล ภาพประกอบที่ 99 เป็นการแบ่งกลุ่มด้วยสีโดยใช้ผลลัพธ์ของตัวอย่างข้อมูล และภาพประกอบที่ 100 เป็นการแบ่งกลุ่มด้วยสีโดยใช้ผลลัพธ์จากการทำนายของแบบจำลอง จากการสังเกตจะพบว่ากราฟทั้งสองมีลักษณะที่คล้ายคลึงกับกราฟที่ได้จากการใช้งานค่าจริงของคุณลักษณะของตัวอย่างข้อมูล โดยกลุ่มข้อมูลของลูกค้าที่ไม่ทำการสมัครผลิตภัณฑ์ (สีเขียว) จะเกาะกลุ่มกันค่อนข้างหนาแน่นจำนวน 4 กลุ่ม และมีการกระจายของข้อมูลทั้งสองกลุ่มปะปนกันไปทางด้านบนของกราฟ ซึ่งทำให้ยากต่อการทำนายที่มีประสิทธิภาพ โดยแบบจำลองจะทำนายตัวอย่างข้อมูลในบริเวณดังกล่าวเป็นกลุ่มลูกค้าที่สมัครผลิตภัณฑ์ ซึ่งคล้ายคลึงกับกราฟที่ได้จากการใช้งานค่าจริงของคุณลักษณะของตัวอย่างข้อมูล



ภาพประกอบ 99 กราฟ T-SNE โดยใช้งานค่า Perplexity เท่ากับ 2 แสดงการกระจายตัวของข้อมูลจากค่า SHAP Value ของคุณลักษณะของตัวอย่างข้อมูลโดยใช้ผลลัพธ์ของตัวอย่างข้อมูลในการแบ่งกลุ่ม



ภาพประกอบ 100 กราฟ T-SNE โดยใช้งานค่า Perplexity เท่ากับ 2 แสดงการกระจายตัวของข้อมูลจากค่าของคุณลักษณะของตัวอย่างข้อมูลโดยใช้ผลลัพธ์จากการทำนายของแบบจำลองในการแบ่งกลุ่ม

5.3 อภิปรายผลการวิจัย

ในส่วนประสิทธิภาพของแบบจำลองที่มีการนำมาใช้งานในการวิจัย พบว่าแบบจำลองแบบ Logistic Regression ซึ่งมีความซับซ้อนที่ไม่สูงมากนัก สามารถทำงานได้ดีกับชุดข้อมูลการนำเสนอผลิตภัณฑ์เงินฝากประจำผ่านทางโทรศัพท์ที่ใช้ในการวิจัย ไม่แพ้แบบจำลองที่มีความซับซ้อนสูงกว่าไม่ว่าจะเป็นแบบจำลองแบบ LightGBM หรือ XGBoost โดยเป็นแบบจำลองที่ค่อนข้างมีความเสถียรในการทำงาน ซึ่งจากการประยุกต์ใช้ร่วมกับหลักการและวิธีการต่างๆ พบว่าแบบจำลองแบบ Logistic Regression ส่วนใหญ่ให้ผลประสิทธิภาพของค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ และค่า Accuracy ของแบบจำลองอยู่ที่ประมาณ 0.70 (70%) ถึง 0.72 (72%) นอกจากนี้ยังเป็นแบบจำลองที่สามารถทำนายกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ได้สูงที่สุดที่ 0.96 (96%) เมื่อประยุกต์ใช้งานร่วมกับ CatBoost Encoding และการจัดการความไม่สมดุลกันของข้อมูลแบบ SMOTE โดยนำไปใช้งานกับชุดของคุณลักษณะข้อมูลส่วนบุคคลของลูกค้าธนาคาร

จากการใช้งานกลุ่มของชุดข้อมูลย่อยทั้ง 3 ชุด ซึ่งประกอบด้วยชุดคุณลักษณะข้อมูลส่วนตัวของลูกค้าธนาคาร ชุดคุณลักษณะข้อมูลการติดต่อระหว่างธนาคารและลูกค้า และชุดคุณลักษณะข้อมูลทางเศรฐศาสตร์ในช่วงเวลานั้นๆ พบว่าชุดคุณลักษณะข้อมูลทางเศรฐศาสตร์ในช่วงเวลานั้นๆ ทำให้แบบจำลองมีความเสถียรมากที่สุดในทุกแบบจำลองที่มีการใช้งานชุดคุณลักษณะนี้ โดยมีค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ และค่า Accuracy ของแบบจำลองอยู่ที่ 0.71 (71%) และ 0.72 (72%) ตามลำดับ ส่วนการใช้งานชุดข้อมูลส่วนตัวของลูกค้าธนาคารจะส่งผลให้แบบจำลองมีค่าของ Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ที่สูงขึ้น ซึ่งในทางกลับกันการใช้งานชุดข้อมูลการติดต่อระหว่างธนาคารและลูกค้าจะทำให้แบบจำลองมีประสิทธิภาพของค่า Accuracy ที่สูงขึ้นแทน

การคัดเลือกคุณลักษณะด้วยวิธีการต่างๆ ทั้ง 6 แบบเพื่อหาคุณลักษณะที่มีความสำคัญสูงที่สุดสองอันดับ พบว่าคุณลักษณะจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) ปรากฏอยู่ในสองอันดับแรกของ 5 จาก 6 วิธีการ คุณลักษณะที่ปรากฏบ่อยรองลงมา คือ อัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) ซึ่งปรากฏเป็นจำนวน 4 ครั้ง นอกจากนี้มีชุดของคุณลักษณะที่ปรากฏร่วมกันของคุณลักษณะทั้งสองที่กล่าวไปอยู่ซ้ำกันถึง 4 ชุด โดยมาจากการคัดเลือกคุณลักษณะด้วยวิธี RFE ร่วมกับแบบจำลอง LightGBM และ XGBoost รวมถึงการใช้งาน SHAP ร่วมกับ LightGBM และ XGBoost จากข้อมูลดังกล่าวสามารถอนุมานได้ว่าสองคุณลักษณะนี้ค่อนข้างมีความสำคัญและมีอิทธิพลต่อผลลัพธ์ของตัวอย่างข้อมูล

นอกจากนี้การใช้งานชุดคุณลักษณะมีความสำคัญสูงที่สุดสองอันดับทำให้การทำงานของแบบจำลอง ทั้งในขั้นตอนการเรียนรู้และขั้นตอนการทำนายของแบบจำลองมีประสิทธิภาพสูงขึ้นในแง่ของการใช้เวลาที่ลดลง และการใช้งานทรัพยากรในการจัดเก็บและคำนวณข้อมูลที่ลดลงเช่นกัน โดยแบบจำลองที่ใช้งานชุดคุณลักษณะที่ประกอบด้วย จำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) ค่อนข้างมีความเสถียรและสามารถให้ประสิทธิภาพเทียบเท่ากับแบบจำลองที่มีการใช้งานคุณลักษณะทั้งหมดของชุดข้อมูลได้ โดยมีประสิทธิภาพของค่า Recall ของกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ และค่า Accuracy ของแบบจำลองอยู่ที่ 0.71 (71%) และ 0.72 (72%) ตามลำดับ

การใช้งานการเรียนรู้ด้วยเครื่องแบบอธิบายได้ด้วยวิธีการแบบ SHAP สามารถนำมาช่วยอธิบายได้ทั้งในระดับแบบจำลองและในระดับรายตัวอย่างข้อมูล โดยการอธิบายในระดับแบบจำลองจะเป็นการแสดงให้เห็นว่าคุณลักษณะใดที่มีค่า SHAP Value ที่สูงทั้งในทางบวกและลบและมีค่าเท่าใด ซึ่งจะส่งผลต่อการทำนายของแบบจำลองในการทำนายเป็น Positive Class หรือ Negative Class ซึ่งจะพบว่าคุณลักษณะจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) จะมีค่าของ SHAP Value ที่ค่อนข้างสูง ซึ่งเป็นการบ่งบอกว่าแบบจำลองมีการใช้งานสองคุณลักษณะนี้เป็นหลักในการทำนายผลลัพธ์ของตัวอย่างข้อมูล

การอธิบายแบบจำลองในระดับรายตัวอย่างข้อมูลจะมีความแม่นยำกว่าในการอธิบายหลักการในการทำนายผลลัพธ์แบบเฉพาะเจาะจงสำหรับแต่ละตัวอย่างข้อมูล โดยสามารถแสดงผลได้ว่าด้วยคุณลักษณะที่มีค่าเป็นเท่าใด ส่งผลต่อการทำนายของแบบจำลองไปในทางบวกหรือลบมากน้อยเพียงใด จากนั้นจะมีการคำนวณผลรวมจากทุกคุณลักษณะออกมาเพื่อสรุปว่าแบบจำลองจะทำนายเป็น Positive Class หรือ Negative Class ซึ่งการอธิบายแบบจำลองในระดับรายบุคคลนี้มีประโยชน์ในด้านของการนำไปประยุกต์ใช้งานเพื่อช่วยเพิ่มความน่าเชื่อถือของแบบจำลองในกระบวนการช่วยในการตัดสินใจเพื่อดำเนินการบางอย่าง หรือสามารถนำไปประยุกต์ใช้เพื่อให้สามารถดำเนินการขั้นตอนต่างๆ ที่เหมาะสมกับแต่ละรายตัวอย่างข้อมูลได้

การวิเคราะห์ความผิดพลาดของแบบจำลองเป็นการนำความผิดพลาดที่เกิดขึ้นของแบบจำลองมาเพื่อค้นหาสาเหตุที่ก่อให้เกิดความผิดพลาด เพื่อให้สามารถรู้และนำไปปรับปรุงพัฒนาแบบจำลองเพื่อให้มีประสิทธิภาพที่สูงขึ้น โดยลักษณะของการเรียนรู้ด้วยเครื่องแบบอธิบายได้จะเป็นการอธิบายที่แบบจำลอง ส่วนการวิเคราะห์ความผิดพลาดจะเน้นไปยังการอธิบายที่ตัว

ข้อมูลโดยตรง โดยจากการการวิเคราะห์ความผิดพลาดของแบบจำลองที่ใช้งานชุดคุณลักษณะจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) พบว่าตัวอย่างข้อมูลส่วนใหญ่ที่ทำนายผิดพลาด เกิดจากการที่ค่าของคุณลักษณะสองตัวข้างต้นไปตกอยู่ในบริเวณที่ตัวอย่างข้อมูลส่วนใหญ่มีผลลัพธ์ซึ่งแตกต่างออกไป ส่งผลให้แบบจำลองเกิดการทำนายผิดพลาด

การวิจัยในครั้งนี้แสดงให้เห็นเห็นว่าจากชุดข้อมูลที่นำมาใช้ในการวิจัยสามารถเลือกใช้งานชุดของคุณลักษณะย่อยที่แตกต่างกันไปได้เพื่อบรรลุมาตรวัดประสิทธิภาพที่เหมาะสมและสนใจที่แตกต่างกันไป โดยหากต้องการค้นหาหรือตรวจจับเพียงกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ให้ได้มากที่สุด สามารถเลือกใช้งานชุดคุณลักษณะข้อมูลส่วนตัวของลูกค้าธนาคาร หากต้องการให้แบบจำลองมีประสิทธิภาพความแม่นยำโดยรวมที่ดี สามารถเลือกใช้งานชุดคุณลักษณะข้อมูลการติดต่อระหว่างธนาคารและลูกค้า และหากต้องการประสิทธิภาพในการทำนายที่สามารถค้นหาและตรวจจับกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ให้ได้มากที่สุดโดยที่ความแม่นยำของแบบจำลองยังอยู่ในเกณฑ์ที่ดี สามารถเลือกใช้งานชุดคุณลักษณะข้อมูลทางเศรษฐศาสตร์ในช่วงเวลานั้นๆ ได้

นอกจากนี้ยังค้นพบว่าหากต้องการประสิทธิภาพในการทำนายที่สามารถค้นหาและตรวจจับกลุ่มลูกค้าที่ทำการสมัครผลิตภัณฑ์ให้ได้มากที่สุดโดยที่ความแม่นยำของแบบจำลองยังอยู่ในเกณฑ์ที่ดี สามารถเลือกใช้ชุดคุณลักษณะขนาดเล็กที่ประกอบด้วยคุณลักษณะจำนวนพนักงานเฉลี่ยรายไตรมาส (nr.employed) และอัตราดอกเบี้ยกู้ยืมระหว่างธนาคารในยุโรปรายวัน (euribor3m) ก็เพียงพอที่จะทำให้แบบจำลองมีประสิทธิภาพที่สูงได้ ซึ่งสามารถลดการใช้งานทรัพยากรคอมพิวเตอร์และเวลาได้เป็นอย่างมาก

ในส่วนของการใช้งาน SHAP สามารถทำให้เข้าใจถึงวิธีการทำงานหรือตัดสินใจของแบบจำลองได้ โดยสามารถแสดงถึงแนวโน้มของคุณลักษณะที่มีความสำคัญและปริมาณของความสำคัญที่คุณลักษณะนั้นๆ ส่งผลต่อการทำนายของแบบจำลอง ซึ่งสามารถนำไปช่วยในกระบวนการตั้งคำถามและจัดเก็บข้อมูลที่มีประสิทธิภาพมากยิ่งขึ้น โดยสามารถเลือกตัดออกการจัดเก็บคุณลักษณะที่ไม่มีความสำคัญหรือความจำเป็น และเพิ่มการจัดเก็บคุณลักษณะอื่นๆ ที่คาดว่าน่าจะมีความสำคัญแทนได้ และในส่วนของการใช้งานการอธิบายแบบรายตัวอย่างข้อมูลจะสามารถช่วยให้การตัดสินใจดำเนินการใดๆ กับลูกค้าแต่ละรายเป็นไปได้อย่างมีประสิทธิภาพและตรงกับความต้องการของลูกค้า โดยจะทำให้สามารถเพิ่มประสิทธิภาพของความพึงพอใจในการใช้งานของลูกค้าได้

5.4 ข้อเสนอแนะ

1. เนื่องจากชุดข้อมูลที่นำมาใช้งานในการวิจัยเป็นชุดข้อมูลที่มีการจัดเก็บตั้งแต่ปี พ.ศ. 2553 จึงอาจจะทำให้ข้อมูลมีความล้าสมัยอยู่บ้าง โดยในการนำไปประยุกต์ใช้งานอาจจะมีการเลือกใช้งานกับชุดข้อมูลที่จัดเก็บภายใน 5 ปีล่าสุด

2. การทำวิศวกรรมคุณลักษณะสามารถมีการปรับปรุงหรือประยุกต์ใช้หลักการอื่นๆ นอกเหนือจากที่ใช้งานในการวิจัยครั้งนี้ได้ เช่น วิธีการจัดการค่าว่างในชุดข้อมูล หรือการจัดการคุณลักษณะชนิดตัวเลขด้วยวิธีการแบบ Min-Max Scaling แทนการใช้งาน Standard Scaling ซึ่งอาจส่งผลให้แบบจำลองมีประสิทธิภาพที่สูงขึ้น

3. ชุดของคุณลักษณะย่อยที่คัดเลือกมาสองอันดับเพื่อนำมาใช้งานกับแบบจำลอง อาจมีการทดลองเพิ่มจำนวนของคุณลักษณะที่มีความสำคัญสูงที่สุดเพื่อนำมาใช้งาน ซึ่งอาจจะสามารถเพิ่มประสิทธิภาพของแบบจำลองได้

4. การใช้งานการเรียนรู้ด้วยเครื่องแบบอธิบายได้ ยังมีวิธีการอื่นๆ เช่น LIME ที่เน้นไปยังการวาดกราฟเพื่ออธิบายการทำนายระดับรายตัวอย่างข้อมูล ซึ่งสามารถนำมาประยุกต์ใช้งานได้เช่นกัน

5. อาจมีการเก็บคุณลักษณะของข้อมูลที่เพิ่มมากขึ้น เช่น ลักษณะของผลิตภัณฑ์ที่นำเสนอ อาจจะเป็นดอกเบี้ยเงินฝาก จำนวนปีที่ต้องทำการฝากเงิน หรือคุณลักษณะของข้อมูลอื่นๆ ที่คาดว่าน่าจะมีความสำคัญ เป็นต้น ซึ่งอาจจะสามารถช่วยเพิ่มประสิทธิภาพของแบบจำลองได้

บรรณานุกรม

AI, C. What is Local Interpretable Model-Agnostic Explanations (LIME)? Retrieved from https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

Aldraimli, M., Soria, D., Parkinson, J., Thomas, E., Bell, J., Dwek, M., & Chaussalet, T. (2020). Machine learning prediction of susceptibility to visceral fat associated diseases. *Health and Technology, 10*, 925-944.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, 70-79.

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, X. w., & Jeong, J. C. (2007, 13-15 Dec. 2007). *Enhanced recursive feature elimination.* Paper presented at the Sixth International Conference on Machine Learning and Applications (ICMLA 2007).

Choudhary, A. (2019). A Unique Method for Machine Learning Interpretability: Game Theory & Shapley Values! Retrieved from https://www.analyticsvidhya.com/blog/2019/11/shapley-value-machine-learning-interpretability-game-theory

Commons, W. (2021). Roc curve.svg. Retrieved from https://commons.wikimedia.org/wiki/File:Roc_curve.svg

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics, 35*(5), 352-359.

Equiskill. (2018). Understanding Logistic Regression. Retrieved from https://www.equiskill.com/understanding-logistic-regression

Gao, J., Tian, H., Yang, Y., Yu, X., Li, C., & Rao, N. (2014). A novel algorithm to enhance P300 in single trials: Application to lie detection using F-score and SVM. *PLoS*

One, 9(11), e109700.

Hasanin, T., & Khoshgoftaar, T. (2018, 6-9 July 2018). *The Effects of Random Undersampling with Simulated Class Imbalance for Big Data.* Paper presented at the 2018 IEEE International Conference on Information Reuse and Integration (IRI).

Hu, C., Tan, Q., Zhang, Q., Li, Y., Wang, F., Zou, X., & Peng, Z. (2022). Application of interpretable machine learning for early prediction of prognosis in acute kidney injury. *Computational and Structural Biotechnology Journal, 20*, 2861-2870.

Hulse, J. V., Khoshgoftaar, T. M., & Napolitano, A. (2007). *Experimental perspectives on learning from imbalanced data*. Paper presented at the Proceedings of the 24th international conference on Machine learning, Corvalis, Oregon, USA. https://doi.org/10.1145/1273496.1273614

IBM. (2020a). Machine Learning. Retrieved from https://www.ibm.com/cloud/learn/machine-learning

IBM. (2020b). Random Forest. Retrieved from https://www.ibm.com/cloud/learn/random-forest

Javatpoint. Supervised Machine Learning. Retrieved from https://www.javatpoint.com/supervised-machine-learning

Johannes, M., Brase, J. C., Fröhlich, H., Gade, S., Gehrmann, M., Fälth, M., . . . Beißbarth, T. (2010). Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics, 26*(17), 2136-2144.

Khadka, R. (2017). Machine Learning Types #2. Retrieved from https://towardsdatascience.com/machine-learning-types-2-c1291d4f04b1

Kumar, D. (2018). Introduction to Data Preprocessing in Machine Learning. Retrieved from https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

Kumawat, D. (2019). Introduction to Logistic Regression - Sigmoid Function, Code Explanation. Retrieved from https://www.analyticssteps.com/blogs/introduction-logistic-regression-sigmoid-function-code-explanation

Liu, Y., Liu, Z., Luo, X., & Zhao, H. (2022). Diagnosis of Parkinson's disease based on SHAP value feature selection. *Biocybernetics and Biomedical Engineering, 42*(3), 856-869.

López, F. (2021). Ensemble Learning: Bagging & Boosting. Retrieved from https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422

Lundberg, S. (2018). Welcome to the SHAP documentation. Retrieved from https://shap.readthedocs.io/en/latest/index.html

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Machado, M. R., Karray, S., & Sousa, I. T. d. (2019, 19-21 Aug. 2019). *LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry.* Paper presented at the 2019 14th International Conference on Computer Science & Education (ICCSE).

Manghani, A. (2017, September 18). A Primer on Machine Learning. *DCE Magazine,* (4), 18. Retrieved from https://ce.uci.edu/pdfs/magazine/2017fall.pdf

McGinnis, W. (2022). Category Encoders. Retrieved from https://contrib.scikit-learn.org/category_encoders/

Microsoft. What is machine learning? Retrieved from https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-machine-learning-platform

Microsoft. (2021a). Features. Retrieved from https://lightgbm.readthedocs.io/en/v3.3.2/Features.html

Microsoft. (2021b). Welcome to LightGBM's documentation! Retrieved from https://lightgbm.readthedocs.io/en/v3.3.3

Molnar, C. (2022). Interpretable Machine Learning. Retrieved from https://christophm.github.io/interpretable-ml-book/index.html

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22-31.

Novack, G. (2020). Building a One Hot Encoding Layer with TensorFlow. Retrieved from https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39

NVIDIA. (2022). XGBoost. Retrieved from https://www.nvidia.com/en-us/glossary/data-science/xgboost

Oracle. What is Machine Learning? Retrieved from https://www.oracle.com/th/artificial-intelligence/machine-learning/what-is-machine-learning

RapidMiner. Confusion Matrix. Retrieved from https://rapidminer.com/glossary/confusion-matrix

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939778

SagarDhandare. (2021). Nominal And Ordinal Encoding In Data Science! Retrieved from https://medium.com/nerd-for-tech/nominal-and-ordinal-encoding-in-data-science-c93872601f16

Saha, S. (2022). XGBoost vs LightGBM: How Are They Different. Retrieved from https://neptune.ai/blog/xgboost-vs-lightgbm

SAP. What is machine learning? Retrieved from https://www.sap.com/sea/insights/what-is-machine-learning.html

Saripuddin, M., Suliman, A., Sameon, S. S., & Jorgensen, B. N. (2021). *Random Undersampling on Imbalance Time Series Data for Anomaly Detection*. Paper presented at the 2021 The 4th International Conference on Machine Learning and Machine Intelligence, Hangzhou, China. https://doi.org/10.1145/3490725.3490748

Sarker, I. H., Kayes, A., & Watters, P. (2019). Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data, 6*(1), 1-28.

scikit-learn. Receiver Operating Characteristic (ROC). Retrieved from https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

Sharma, A. (2018). What makes LightGBM lightning fast? Retrieved from https://towardsdatascience.com/what-makes-lightgbm-lightning-fast-a27cf0d9785e

Song, Q., Jiang, H., & Liu, J. (2017). Feature selection based on FDA and F-score for multi-class classification. *Expert Systems with Applications, 81*, 22-27.

Tekouabou, S. C. K., Cherif, W., & Silkan, H. (2019). *A data modeling approach for classification problems: application to bank telemarketing prediction*. Paper presented at the Proceedings of the 2nd International Conference on Networking, Information Systems & Security, Rabat, Morocco. https://doi.org/10.1145/3320326.3320389

Tzorakoleftherakis, E. (2019). Reinforcement Learning: A Brief Guide. Retrieved from https://www.mathworks.com/company/newsletters/articles/reinforcement-learning-a-brief-guide.html

Wikipedia. (2013). Logistic regression. Retrieved from https://en.wikipedia.org/wiki/Logistic_regression

Wikipedia. (2017a). Feature engineering. Retrieved from https://en.wikipedia.org/wiki/Feature_engineering

Wikipedia. (2017b). Oversampling and undersampling in data analysis. Retrieved from https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis

XGBoost. (2021). XGBoost Documentation. Retrieved from https://xgboost.readthedocs.io/en/stable

Xia, W., Ma, C., Liu, J., Liu, S., Chen, F., Zhi, Y., & Duan, J. (2019). High-Resolution Remote Sensing Imagery Classification of Imbalanced Data Using Multistage Sampling Method and Deep Neural Networks. *Remote Sensing, 11*, 2523.

Zhou, Q., Zhou, H., Zhou, Q., Yang, F., & Luo, L. (2014). Structure damage detection based on random forest recursive feature elimination. *Mechanical Systems and Signal Processing, 46*(1), 82-90.

เบญจพร เอี่ยมประโคน. (2560). วิธีการเปรียบเทียบพื้นที่ใต้โค้ง *ROC* สำหรับข้อมูลชุดเดียวกัน: กรณีศึกษาแบบจำลองคะแนนเครดิต. (วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ). Retrieved from

http://cuir.car.chula.ac.th/handle/123456789/59846

ธนาคารแห่งประเทศไทย. (2563). แผนยุทธศาสตร์ธนาคารแห่งประเทศไทย พ.ศ. 2563-2565. Retrieved from https://www.bot.or.th/Thai/AboutBOT/RolesAndHistory/DocLib_StrategicPlan/BOT-StrategicPlan2020to2022.pdf

ประวัติผู้เขียน