การพัฒนาแบบจำลองการพิจารณาการอนุมัติสินเชื่อ โดยใช้เทคนิคการวิเคราะห์องค์ประกอบร่วมกับการเรียนรู้ด้วยเครื่อง

ชลลดา ม่วงธนัง[1*] สุรศักดิ์ มังสิงห์[2] นิเวศ จิระวิชิตชัย[3]

[1,2]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม, กรุงเทพมหานคร, ประเทศไทย*
[3]คณะวิศวกรรมศาสตร์และเทคโนโลยี สถาบันการจัดการปัญญาภิวัฒน์, นนทบุรี, ประเทศไทย

*ผู้ประพันธ์บรรณกิจ อีเมล : chonlada.mua@spulive.net

รับต้นฉบับ : 7 มีนาคม 2566; รับบทความฉบับแก้ไข: 30 เมษายน 2566; ตอบรับบทความ : 13 มิถุนายน 2566
เผยแพร่ออนไลน์ : 27 ธันวาคม 2566

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอแบบจำลองและทดสอบประสิทธิภาพของแบบจำลองการอนุมัติสินเชื่อในอุตสาหกรรมการเงิน โดยใช้เทคนิคการวิเคราะห์องค์ประกอบร่วมกับการเรียนรู้ด้วยเครื่อง ทดสอบกับอัลกอริทึม 4 วิธี ได้แก่ นาอีฟเบย์ (Naïve Bayes) ต้นไม้การตัดสินใจ (Decision Tree) แรนดอมฟอร์เรส (Random Forest) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยทดสอบกับชุดข้อมูลการปล่อยกู้จำนวน 441,335 ตัวอย่าง ทดสอบประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความครบถ้วน (Recall) และค่าการวัดประสิทธิภาพโดยรวม (F-Measure)

จากการทดลองพบว่าแบบจำลองแรนดอมฟอร์เรส ให้ประสิทธิภาพค่าความถูกต้องดีที่สุดคือร้อยละ 92.90 รองลงมาเป็น ซัพพอร์ตเวกเตอร์แมชชีน และต้นไม้การตัดสินใจ คือร้อยละ 87.00 และนาอีฟเบย์ คือร้อยละ 83.40 ตามลำดับ ซึ่งพบว่าการลดมิติของข้อมูลส่งผลให้ประสิทธิภาพของแบบจำลองดีขึ้นเนื่องจากตัดคุณลักษณะที่ไม่มีนัยสำคัญทิ้งไป และสามารถแก้ไขปัญหาของแบบจำลองการอนุมัติสินเชื่อแบบเดิมที่พิจารณาจากตัวแปรคุณลักษณะแบบเก่า โดยแบบจำลองได้พัฒนาประสิทธิภาพให้ดีขึ้นจากการพิจารณาตัวแปรคุณลักษณะแบบใหม่ ผลของค่าความครบถ้วน (Recall) และค่าการวัดประสิทธิภาพโดยรวม (F-Measure) จากการทดลองพบว่าแบบจำลองแรนดอมฟอร์เรส ให้ประสิทธิภาพดีที่สุด เช่นเดียวกันกับ ค่าความถูกต้อง (Accuracy) คือร้อยละ 99.64 และ 99.35 ตามลำดับ และค่าความแม่นยำ (Precision) มีค่าสูงสุด คือ แบบจำลองแรนดอมฟอร์เรส และต้นไม้การตัดสินใจ เท่ากันที่ร้อยละ 99.07

คำสำคัญ **:** อนุมัติสินเชื่อ การเรียนรู้ด้วยเครื่อง วัดประสิทธิภาพ

Developing a Credit Approval Determination Model Using Principal Component Analysis with Machine Learning Techniques

Chonlada Muangthanang[1*] Surasak Mungsing[2] Nivet Chirawichichai[3]

[1*,2]*School of Information Technology, Sripatum University, Bangkok, Thailand*
[3]*Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand*

*Corresponding Author. E-mail address: chonlada.mua@spulive.net

Received: 7 March 2023; Revised: 30 April 2023; Accepted: 13 June 2023
Published online: 27 December 2023

Abstract

This research presents a model and tests the performance of a credit approval determination model by using principal component analysis with machine learning techniques. Four algorithms: Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine were tested with 441,335 examples of lending data. There are four model performance test results: Accuracy, Precision, Recall, and F-Measure.

From the experiment, it was found that the random forest model provides the best accuracy performance of 92.90 percent, followed by support vector machine and decision tree is 87.00 percent, and Naïve Bayes is 83.40 percent, respectively. It was found that reducing the data dimensions resulted in improved model performance by eliminating insignificant features and solve the problem of the traditional credit approval model that considers the old attribute variables. The model's performance has improved by considering new attribute variables. The results of the completeness value (Recall) and the overall performance measurement value (F-Measure) from the experiment found that Random forest model provides the best performance as well as accuracy values were 99.64 and 99.35 percent, respectively, and the highest precision value was 99.07 percent for the random forest model and decision trees.

***Keywords*:** Credit approval, Machine learning, Performance

1) บทนำ

สถาบันการเงินที่บริการสินเชื่อใช้ข้อมูลที่ลูกค้ามายื่นขอรับสินเชื่อและข้อมูลที่ได้จากองค์กรกลางภายนอกมาตัดสินการให้สินเชื่อหรือการปฏิเสธการขอรับสินเชื่อ มีการเลือกใช้เจ้าหน้าที่ในการพิจารณาสินเชื่อของลูกค้า โดยสถาบันการเงินในแต่ละแห่งต่างมีแบบจำลองคะแนนเครดิต (credit scoring) ของตนเอง [1] ในการพิจารณาอนุมัติวงเงินสินเชื่อให้แก่ลูกค้า ซึ่งพัฒนาขึ้นมาจากฐานข้อมูลลูกค้าและประสบการณ์ในอดีตของตนเอง ทำให้สถาบันการเงินแต่ละแห่งมีความเข้มข้นของเกณฑ์แตกต่างกัน หากเกณฑ์การพิจารณาเข้มจนเกินไปอาจทำให้สูญเสียโอกาสทางธุรกิจได้ และในทางกลับกันเกณฑ์ที่ประนีประนอมมากไปก็อาจทำให้รับลูกค้ากลุ่มเสี่ยงเข้ามาได้เช่นกัน ด้วยเหตุนี้จึงส่งผลให้เกิดปัญหาความล่าช้าของกระบวนการพิจารณาสินเชื่อและการอนุมัติที่ไม่เหมาะสม การพิจารณาที่ไม่สมเหตุสมผลเป็นเหตุให้ลูกค้าเปลี่ยนไปขอรับสินเชื่อจากสถาบันการเงินที่บริการสินเชื่ออื่น ดังนั้นสถาบันการเงินที่บริการสินเชื่อจึงต้องการวิธีการพิจารณาสินเชื่อที่รวดเร็ว และมีประสิทธิภาพในการอนุมัติสินเชื่อ ซึ่งในการให้คะแนนเครดิตที่จัดระดับโดยสถาบันจัดอันดับความน่าเชื่อถือ (credit rating agencies) [2] ถูกวิพากษ์วิจารณ์ว่ามีความลำเอียงในการให้คะแนนทำให้เกิดความได้เปรียบเสียเปรียบเชิงการแข่งขันทางการค้าโดยเฉพาะนักลงทุนรายย่อยจึงจำเป็นต้องได้รับการแก้ไข จำเป็นที่ต้องมีวิธีการให้คะแนนเครดิตขั้นต้นที่จะนำไปสู่การประเมินเครดิตที่ทันเวลาและมีกระบวนการทำงานที่โปร่งใส มีความเป็นกลาง เพื่อลดความเสี่ยงของระบบการเงินในอนาคต

การได้มาซึ่งการพิจารณาสินเชื่อที่มีผลอย่างสมบูรณ์ เพื่อแสดงให้เห็นถึงผลที่ดีของการวิเคราะห์เชิงคาดการณ์ (predictive analytics) ซึ่งต้องสามารถวิเคราะห์ข้อมูลในปัจจุบันและอดีตเพื่อคาดการณ์เกี่ยวกับผลลัพธ์ในอนาคต นั่นคือ การทำความสะอาดข้อมูล (clean data) และเลือกข้อมูลมาอย่างระมัดระวังภายใต้การวิเคราะห์ โดยวิธีที่ทำให้โมเดลการคาดการณ์ (predictive model) นั้นมีประสิทธิภาพมากขึ้น คือ การใช้เทคนิควิศวกรรมคุณลักษณะ (feature engineering) เพื่อให้ได้ประโยชน์สูงสุดจากข้อมูลอย่างลึกซึ้ง อีกทั้งส่งผลให้ได้รับข้อมูลเชิงลึกเพิ่มเติมและระบุโมเดลการคาดการณ์ที่เหมาะสมที่สุด ทำให้เทคนิควิศวกรรมคุณลักษณะเป็นขั้นตอนที่สำคัญ โดยเป็นเทคนิคที่ต้องใช้ความรู้เกี่ยวกับกระบวนการ และผลลัพธ์ในการแยกคุณสมบัติ (properties) หรือคุณลักษณะ (features) ต่าง ๆ ที่ทำให้โมเดลการคาดการณ์ทำงานได้ดีขึ้น สู่การได้มาซึ่งโมเดลที่เหมาะสมและประสิทธิภาพที่มีความแม่นยำ

ดังนั้น จึงสามารถกล่าวได้ว่า เทคนิควิศวกรรมคุณลักษณะ (feature engineering) เป็นกระบวนการใช้แบบจำลองความรู้หลักการเฉพาะปัญหา (domain knowledge) ในการสร้างคุณลักษณะใหม่ขึ้นมา และทำการเลือกตัดคุณลักษณะที่ไม่เกี่ยวข้องออกไป เพื่อช่วยทำให้อัลกอริทึมที่ได้มาจากเทคนิคการเรียนรู้ด้วยเครื่อง (machine learning) เรียนรู้ได้ดียิ่งขึ้น โดยมีวัตถุประสงค์ของการวิจัย เพื่อทดสอบและเปรียบเทียบประสิทธิภาพของแบบจำลอง เป็นจำนวน 4 ตัวแบบ ในการพิจารณาการอนุมัติสินเชื่อโดยใช้เทคนิคการวิเคราะห์องค์ประกอบร่วมกับแรนดอมฟอร์เรส

จากความสำคัญที่กล่าวมาข้างต้น ได้มีแนวคิดเพื่อแก้ปัญหาแบบจำลองการอนุมัติสินเชื่อแบบเดิมที่พิจารณาจากตัวแปรคุณลักษณะประชากรแบบเก่า และประวัติในการทำธุรกรรมเกี่ยวกับสินเชื่อทางการเงินในอดีต ซึ่งพบว่าจะสามารถพัฒนาให้ดีขึ้นได้ เหตุผลเพราะข้อมูลธุรกรรมการเงินเกี่ยวกับสินเชื่อในอดีต เช่น NPL ratio และข้อมูลการผ่อนชำระหนี้แบบเดิมของลูกหนี้ อาจไม่สะท้อนให้เห็นถึงคุณภาพของสินเชื่อที่แท้จริง เนื่องจากไม่ได้พิจารณาปัจจัยสภาพแวดล้อมความเสี่ยงทางการเงินที่ครบถ้วน ดังนั้น จึงเกิดแนวคิดในการพัฒนาตัวชี้วัดใหม่ที่พิจารณาจากความเสี่ยง บริบทโดยรวมของข้อมูลทางการเงินทั้งหมดโดยประยุกต์ใช้กฎพิจารณาความเสี่ยงลูกหนี้ของธนาคารแห่งประเทศไทย เพื่อนำมาใช้ร่วมกับตัวชี้วัดเดิม โดยพิจารณารายละเอียดและอัตราการผิดชำระหนี้ (default rate) เข้าร่วมด้วย จึงเป็นเหตุผลให้แสวงหาคำตอบโดยพัฒนา Credit Scoring Model แบบใหม่ ในการพิจารณาการอนุมัติสินเชื่อโดยใช้เทคนิควิศวกรรมคุณลักษณะร่วมกับเทคนิคการเรียนรู้ด้วยเครื่อง เพื่อใช้ในการพยากรณ์คุณภาพของสินเชื่อ โดยใช้ตัวชี้วัดใหม่ที่พัฒนาขึ้นมานี้ร่วมด้วย ซึ่งแบบจำลองดังกล่าวสามารถแยกแยะคุณภาพลูกหนี้ดี - เสีย ออกจากกันได้ค่อนข้างดีกว่าวิธีการดั้งเดิม โดยใช้งานร่วมกับอัลกอริทึมการเรียนรู้ของเครื่องในการพยากรณ์ เพื่อจัดการความเสี่ยงที่จะเกิดเหตุการณ์การผิดนัดชำระหนี้ขึ้น และเพื่อใช้ประกอบการตัดสินใจในการดำเนินนโยบายและกำกับดูแลการปล่อยสินเชื่อให้กับผู้กู้รายย่อย

2) ทบทวนวรรณกรรม

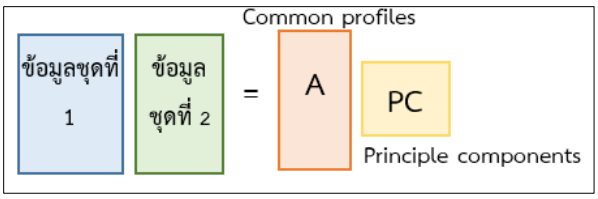
2.1) เทคนิคการวิเคราะห์หาองค์ประกอบหลัก (Principle Component Analysis: PCA)

PCA เป็นวิธีทางสถิติที่วิเคราะห์หาองค์ประกอบหลักของข้อมูลภายใต้เงื่อนไขค่าความแปรปรวน (variance) สูงสุด [3] ดังสมการ (1)

$$Y = PX \quad (1)$$

คำอธิบายสมการ เป็นกระบวนการแปลงข้อมูล X บนเมทริกซ์การแปลง P เพื่อเป็นข้อมูลเชิงเส้น และ Y แสดงความสัมพันธ์

เนื่องจากวิธี PCA เป็นวิธีที่พิจารณาความสัมพันธ์ร่วมกันของข้อมูล X หรือข้อมูลเพียงชุดเดียว หากมีชุดข้อมูลที่มากกว่าสองชุดขึ้นไป ข้อมูลเหล่านั้นจะถูกหลอมรวมก่อนในขั้นตอนแรกดังรูปที่ 1



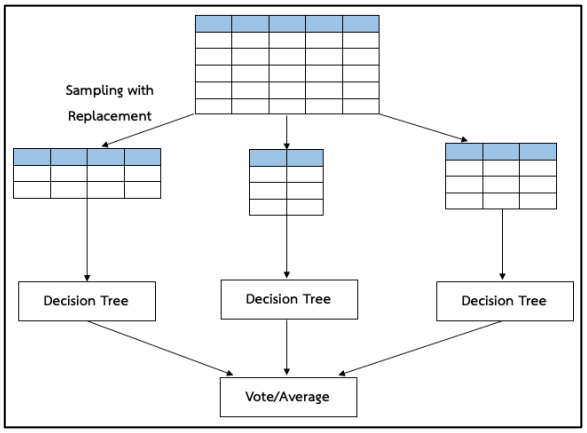
รูปที่ 1 : ขั้นตอนของชุดข้อมูลที่ถูกหลอมรวม [3]

จากนั้นข้อมูลจะหาความสัมพันธ์ร่วม Covariance และคำนวณหาองค์ประกอบหลักด้วย Eigen vector และ Eigen Value ที่สัมพันธ์กัน สำหรับการลดมิติข้อมูลด้วยวิธี PCA จะดำเนินการด้วยการจัดเรียงค่า Eigen จากมากไปหาน้อยหรือเรียกว่า diagonal ของค่า Eigen จากนั้นจะเลือกจำนวนขององค์ประกอบหลักจำนวน d ตัวเพื่อลดมิติลง การเลือกจำนวนขององค์ประกอบส่วนใหญ่จะเลือกจากค่าความเชื่อมั่นหรือค่าสะสมของ Eigen เมื่อได้จำนวนองค์ประกอบที่ จะนำค่า Eigen Vector เพื่อใช้เป็นค่า weight สำหรับการลดมิติข้อมูลก่อนเข้ากระบวนการประมวลผลต่อไป

2.2) แรนดอมฟอร์เรส (Random Forest)

Random Forest คือ อัลกอริทึมการเรียนรู้ของเครื่องที่ชนิดหนึ่งที่ใช้การสุ่มคุณลักษณะ เพื่อเพิ่มความหลายหลายของโมเดล โดยใช้โมเดลจาก Decision Tree หลาย ๆ โมเดลย่อย ๆ โดยแต่ละโมเดลจะได้รับชุดข้อมูลทดสอบที่ไม่เหมือนกันซึ่งเป็นชุดข้อมูลย่อย ของชุดข้อมูลทั้งหมด เมื่อทำการพยากรณ์ก็ให้แต่ละ Decision Tree ทำการพยากรณ์แยกเป็นคนละชุดไม่เกี่ยวข้องกัน และคำนวณผลการพยากรณ์โดยใช้ Vote Output ที่ถูกเลือกโดย Decision Tree มากที่สุด (กรณี Classification) หรือหาค่า mean จาก Output ของแต่ละ Decision Tree (กรณี Regression) [4]

Random Forest เป็นขั้นตอนวิธีพัฒนาต่อยอดมาจาก Decision Tree ต่างกันที่ Random Forest เป็นการเพิ่มจำนวนต้นไม้ (tree) เป็นหลาย ๆ ต้น ทำให้ประสิทธิภาพการทำงานและพยากรณ์สูงขึ้น Random Forest มีหลักการทำงาน คือ จะแบ่งข้อมูลออกเป็นต้นไม้ตัดสินใจ (decision tree) หลาย ๆ ต้น โดยแต่ละต้นจะได้รับคุณลักษณะ (feature) และข้อมูล (data) ที่ไม่เหมือนกันทั้งหมด เพื่อทำให้ได้ต้นไม้ที่มีความหลากหลายและมีความอิสระต่อกันมากขึ้น [5] ดังรูปที่ 2



รูปที่ 2 : หลักการทำงานของ Random Forest [6]

Random Forest เป็นการเรียนรู้อีกแบบหนึ่งในกลุ่ม Ensemble Learning ที่จะนำโมเดล Decision Tree หลาย ๆ อันมาใช้ร่วมกัน ซึ่งก็เปรียบได้กับต้นไม้หลาย ๆ ต้น เมื่อรวมกันก็จะกลายเป็นป่า (forest) นั่นเอง

ดังนั้น Random Forest จึงสามารถทำนายผลได้ทั้งแบบ Classification และ Regression เช่นเดียวกับ Decision Tree ซึ่งโดยทั่วไปแล้ว Random Forest มักจะมีความแม่นยำมากกว่า Decision Tree [6]

หลักการทำงานของ Random Forest มีแนวทางดังนี้ [6]

1) เนื่องจาก Random Forest ใช้วิธีแบบ Bagging ดังนั้นข้อมูลตัวอย่างที่นำมา Train Model จึงถูกแบ่งออกเป็นกลุ่ม ๆ ย่อย โดยใช้หลักการแบบ Sampling with Replacement ตามจำนวนการเรียนรู้ที่เรากำหนด

2) ข้อมูลแต่ละชุดจะถูกนำไป Train Model ตามหลักการของ Decision Tree

3) ผลลัพธ์จากแต่ละโมเดลจะถูกดำเนินการดังนี้

- หากเป็นการทำนายแบบ Classification ก็จะทำนายผลโดยวิธีการโหวต

- หากเป็นการทำนายผลแบบ Regression ก็จะทำนายผลโดยวิธีการหาค่าเฉลี่ย

เนื่องจาก Random Forest จะดำเนินการบนพื้นฐานของ Decision Tree ซึ่งข้อกำหนดต่าง ๆ จะเป็นไปตามหลักการของโมเดล

2.3) ทฤษฎีการนอร์มัลไลเซชัน (Normalization)

การทำ Normalization หมายถึง การทำให้ข้อมูลเป็นค่ามาตรฐาน เป็นเทคนิค Feature Scaling ที่จะนำข้อมูลแต่ละ Feature อยู่ในย่านจำนวนจริงที่ Machine Learning อ่านได้ง่ายขึ้น [7]

การปรับ scale ของข้อมูล ข้อมูลที่จะนำมา Train Model นั้น หากบางคอลัมน์มีค่าแตกต่างจากคอลัมน์อื่น ๆ มากเกินไป อาจทำให้คอลัมน์นั้นมีน้ำหนักมากกว่า หรือไม่สมดุลกับข้อมูลในคอลัมน์อื่น ๆ ซึ่งอาจส่งผลต่อการทำนาย และเกิดความคลาดเคลื่อนได้ [6]

วิธี StandardScaler จะคำนวณโดยใช้สมการ (2) [6]

$$Z = \frac{X - U}{S} \quad (2)$$

คำอธิบายสมการ

Z คือ ค่าใหม่หลังการปรับสเกล

X คือ ค่าเดิม หรือข้อมูลตัวอย่างแต่ละค่า

U คือ ค่าเฉลี่ยของข้อมูลในคอลัมน์นั้น

S คือ ส่วนเบี่ยงเบนมาตรฐานของข้อมูลในคอลัมน์นั้น

2.4) ตัววัดประสิทธิภาพ

2.4.1) ค่าความถูกต้อง (Accuracy) หรือ Accuracy Score [8] เป็นการวัดความถูกต้องของโมเดล โดยคำนวณรวมทุกคำตอบ สูตรในการคำนวณ คือ

$$\text{Accuracy} = \frac{\text{True Positives + True Negatives}}{\text{Total Number of Examples}} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

2.4.2) ค่าความแม่นยำ (Precision) [8] เป็นการคำนวณค่า Precision ของโมเดล โดยคำนวณแยกทีละคำตอบ

ค่า Precision วัดประสิทธิภาพของโมเดล ที่สนใจผลการทำนายแบบ FP (False Positive) ผลที่ได้เป็นเปอร์เซ็นต์ (%) สูตรในการคำนวณ คือ

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.4.3) ค่าความครบถ้วน หรือค่าความระลึก (Recall) [8] เป็นการคำนวณค่า Recall ของโมเดลโดยคำนวณแยกทีละคำตอบ

ค่า Recall วัดประสิทธิภาพของโมเดล ที่สนใจผลการทำนายแบบ FN (False Negative) ผลที่ได้เป็นเปอร์เซ็นต์ (%) สูตรในการคำนวณคือ

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

2.4.4) ค่าการวัดประสิทธิภาพโดยรวม (F-Measure) [8] เป็นการคำนวณค่าเฉลี่ยฮาร์โมนิก (Harmonic Mean) ของ Precision และ Recall โดยคำนวณแยกทีละลาเบลคำตอบ

ค่า F1-score ประสิทธิภาพของโมเดล ที่ไม่สนใจผลการทำนายแบบ FN (False Negative) หรือ FP (False Positive) เป็นหลัก ผลที่ได้เป็นเปอร์เซ็นต์ (%) สูตรในการคำนวณคือ

$$F_1 = 2 \text{ X } \frac{\text{Precision X Recall}}{\text{Precision + Recall}} \quad (7)$$

คำอธิบายสมการ

True Positive (TP) คือ จำนวนข้อมูลที่พยากรณ์ถูกว่าเป็นคลาส ซึ่งกำลังพิจารณา

True Negative (TN) คือ จำนวนข้อมูลที่พยากรณ์ถูกว่าเป็นคลาส ซึ่งไม่ได้พิจารณา

False Positive (FP) คือ จำนวนข้อมูลที่พยากรณ์ผิดมาเป็นคลาส ซึ่งกำลังพิจารณา

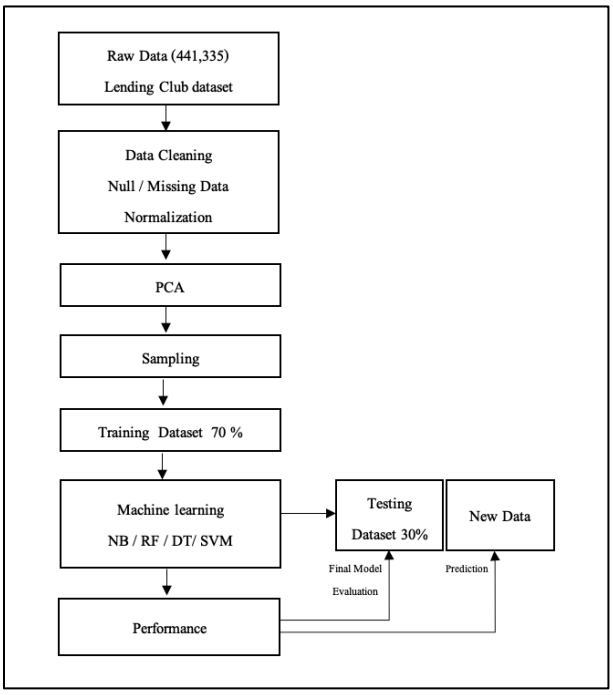
False Negative (FN) คือ จำนวนข้อมูลที่พยากรณ์ผิดมาเป็นคลาส ซึ่งไม่ได้พิจารณา

3) วิธีดำเนินการวิจัย

3.1) ประชากรและกลุ่มตัวอย่าง

กลุ่มข้อมูลเครดิตของพลเมืองประเทศสหรัฐอเมริกา (United Stated Credit) จำนวนกลุ่มตัวอย่าง 441,335 ตัวอย่าง [9] ซึ่งมีข้อมูลทั้งหมด 151 คุณลักษณะ (attributes) ประกอบด้วยตัวแปรต้น 150 คุณลักษณะ และตัวแปรตาม 1 คุณลักษณะ คือ สถานะปัจจุบันของหนี้ (loan_status)

3.2) กรอบงานวิจัย



รูปที่ 3 : กรอบงานวิจัย

3.3) เครื่องมือที่ใช้ในการวิจัย/รวบรวมข้อมูล

กระบวนการสุ่มตัวอย่าง (Sampling) แบ่งข้อมูลออกเป็น 2 ส่วน เพื่อใช้ทดสอบประสิทธิภาพโมเดล คือ

1) ข้อมูลฝึกสอน (Training Data) ที่ 70% ของข้อมูลทั้งหมด คือ กลุ่มตัวอย่าง 308,935 ตัวอย่าง

2) ข้อมูลทดสอบ (Testing Data) ที่ 30% ของข้อมูลทั้งหมด คือ กลุ่มตัวอย่าง 132,400 ตัวอย่าง

3.4) การวิเคราะห์ข้อมูล

3.4.1) กระบวนการจัดเตรียมข้อมูลกลุ่มตัวอย่าง สำหรับการประมวลผลด้วยวิธีการทำความสะอาดข้อมูล (Cleansing Data) คือ

1) การคัดกรองข้อมูลกลุ่มตัวอย่างที่มีคุณลักษณะ (Attributes) ของคุณภาพข้อมูลต่ำออก

2) แทนที่ข้อมูลที่เป็นค่าว่าง (Missing)

3) ลดมิติ (Dimension) ของข้อมูลขนาดใหญ่ด้วยวิธี การวิเคราะห์หาองค์ประกอบหลัก (Principal Component Analysis (PCA)) [3]

4) ลดความซ้ำซ้อนของข้อมูลด้วยกระบวนการ Normalization

3.4.2) กระบวนการของเทคนิควิศวกรรมคุณลักษณะ (Feature Engineering) [10]

1) ใช้เทคนิค Handling Outliers หรือ ค่าที่ผิดปกติ โดยจัดการแทนที่ข้อมูลที่มีค่าผิดปกติในข้อมูลที่มีค่าสูง หรือต่ำกว่าข้อมูลส่วนใหญ่ในชุดข้อมูลอย่างมาก ซึ่งการจัดการกับ Outlier จะพิจารณาจาก Standard Deviation และ Percentile [11]

2) ใช้เทคนิค One-Hot Encoding เป็นการเข้ารหัสข้อมูลแบบหนึ่ง โดยใช้วิธีแปลงข้อมูลที่ขาดหายไป

3.4.3) กระบวนการสกัดคุณลักษณะ (Feature Extraction) [10]

1) ใช้กระบวนการสกัดคุณลักษณะ ด้วยการเลือกเฉพาะคุณลักษณะที่สำคัญ (Selection) เพื่อให้มีมิติของข้อมูลลดลง จากคุณลักษณะทั้งหมด 151 คุณลักษณะ เหลือ 16 คุณลักษณะ โดยแบ่งออกเป็น

2) ตัวแปรต้นที่นำไปใช้งานได้ จำนวน 15 คุณลักษณะ (Attributes) คือ (1) issue_d (2) term (3) grade (4) sub_grade (5) emp_length (6) home_ownership (7) verification_status (8) purpose (9) initial_list_status (10) last_pymnt_d (11) last_credit_pull_d (12) pc_1 (13) pc_2 (14) pc_3 (15) pc_4 มีรายละเอียดคุณลักษณะ (Attributes) ดังแสดงในตารางที่ 1 ดังนี้

ตารางที่ 1 : รายละเอียดคุณลักษณะ (Attributes) ที่นำมาวิเคราะห์ข้อมูล

ข้อมูล	คำอธิบายข้อมูล
(1) issue_d	เดือนที่กู้เงิน
(2) term	จำนวนการชำระเงินกู้ ค่าเป็นเดือน และอาจเป็น 36 หรือ 60 ก็ได้
(3) grade	LC กำหนดเกรดสินเชื่อ (LC/Letter of Credit คือ สินเชื่อเพื่อเปิด)
(4) sub_grade	LC กำหนดลดเกรดเงินกู้ (LC/Letter of Credit คือ สินเชื่อเพื่อเปิด)
(5) empLength	อายุงานเป็นปี ค่าที่เป็นไปได้อยู่ระหว่าง 0 ถึง 10 โดยที่ 0 หมายถึงน้อยกว่าหนึ่งปี และ 10 หมายถึง 10 ปีขึ้นไป
(6) home_ownership	สถานะเจ้าของบ้านที่ผู้กู้ระบุระหว่างการลงทะเบียน ค่านิยมของเราคือ: เช่า, เป็นเจ้าของ, จำนอง, อื่นๆ
(7) verification_status	ระบุว่ารายได้ได้รับการยืนยันโดย LC หรือไม่ได้รับการยืนยัน หรือแหล่งที่มาของรายได้ได้รับการยืนยันหรือไม่ (LC/Letter of Credit คือ สินเชื่อเพื่อเปิด)

ตารางที่ 2 : รายละเอียดคุณลักษณะ (Attributes) ที่นำมาวิเคราะห์ข้อมูล (ต่อ)

ข้อมูล	คำอธิบายข้อมูล
(8) purpose	หมวดหมู่ที่ผู้กู้จัดเตรียมไว้สำหรับการขอสินเชื่อ
(9) initial_list_status	สถานะรายการเริ่มต้นของเงินกู้ ค่าที่เป็นไปได้คือ – W, F
(10) last_pymnt_d	ได้รับการชำระเงินเดือนล่าสุด
(11) last_credit_pull_d	เดือนล่าสุด ที่ LC ดึงเครดิตสำหรับเงินกู้นี้ (LC/Letter of Credit คือ สินเชื่อเพื่อเปิด)
(12) open_acc_6m	จำนวนการซื้อขายที่เปิดในช่วง 6 เดือนที่ผ่านมา
(13) open_il_12m	จำนวนบัญชีผ่อนชำระที่เปิดใน 12 เดือนที่ผ่านมา
(14) open_il_24m	จำนวนบัญชีผ่อนชำระที่เปิดใน 24 เดือนย้อนหลัง
(15) open_act_il	จำนวนการซื้อขายแบบผ่อนชำระที่ใช้งานอยู่ในปัจจุบัน

ตัวแปรตาม 1 คุณลักษณะ (attributes) คือ สถานะปัจจุบันของหนี้ (loan_status) ประกอบด้วย

1. สถานะชำระหนี้ปกติ (current)
2. สถานะชำระหนี้เต็มจำนวน (fully paid)
3. สถานะปิดการเรียกเก็บเงิน (charged off)

4) ผลการวิจัยและอภิปรายผล

กระบวนการพัฒนาแบบจำลอง ได้ผลการวิจัยจากการประเมินผลลัพธ์การทำนาย (Confusion Matrix) ขนาด 3×3 แสดงจำนวนที่พยากรณ์สถานะปัจจุบันของหนี้ (loan_status) ของแบบจำลองจาก 4 เทคนิค ดังนี้

ตารางที่ 2 : การประเมินผลลัพธ์การทำนาย (Confusion Matrix) แสดงจำนวนที่พยากรณ์ของตัวแบบจำลองนาอีฟเบย์ (Naïve Bayes)

	true Fully Paid	true Charged Off	true Current	class precision
pred. Fully Paid	51872	11702	8445	72.03%
pred. Charged Off	0	0	0	0.00%
pred. Current	805	453	55620	97.79%
class recall	98.47%	0.00%	86.82%	

จากตารางที่ 2 จะเห็นว่า โมเดลได้พยากรณ์ ค่าความครบถ้วน (Recall) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้เต็มจำนวน (Fully Paid) คิดเป็น 98.47% รองลงมาคือ สถานะชำระหนี้ปกติ (Current) คิดเป็น 86.82%

และ โมเดลได้พยากรณ์ ค่าความแม่นยำ (Precision) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้ปกติ (Current) คิดเป็น 97.79% รองลงมาคือ สถานะชำระหนี้เต็มจำนวน (Fully Paid) คิดเป็น 72.03%

ตารางที่ 3 : การประเมินผลลัพธ์การทำนาย (Confusion Matrix) แสดงจำนวนที่พยากรณ์ของตัวแบบจำลองต้นไม้การตัดสินใจ (Decision Tree)

	true Fully Paid	true Charged Off	true Current	class precision
pred. Fully Paid	28688	3427	4847	77.61%
pred. Charged Off	223	3270	473	82.45%
pred. Current	215	69	30217	99.07%
class recall	98.50%	48.33%	85.03%	

ตารางที่ 3 จะเห็นว่า โมเดลได้พยากรณ์ ค่าความครบถ้วน (recall) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้เต็มจำนวน (fully paid) คิดเป็น 98.50% รองลงมาคือ สถานะชำระหนี้ปกติ (current) คิดเป็น 85.03%

และ โมเดลได้พยากรณ์ ค่าความแม่นยำ (precision) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้ปกติ (current) คิดเป็น 99.07% รองลงมาคือ สถานะปิดการเรียกเก็บเงิน (charged off) คิดเป็น 82.45%

ตารางที่ 4 : การประเมินผลลัพธ์การทำนาย (Confusion Matrix) แสดงจำนวนที่พยากรณ์ของตัวแบบจำลองแรนดอมฟอร์เรส (Random Forest)

	true Fully Paid	true Charged Off	true Current	class precision
pred. Fully Paid	7240	502	620	86.58%
pred. Charged Off	5	1068	5	99.07%
pred. Current	21	110	8287	98.44%
class recall	99.64%	63.57%	92.99%	

ตารางที่ 4 จะเห็นว่า โมเดลได้พยากรณ์ ค่าความครบถ้วน (Recall) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้เต็มจำนวน (Fully Paid) คิดเป็น 99.64% รองลงมาคือ สถานะชำระหนี้ปกติ (Current) คิดเป็น 92.99%

และ โมเดลได้พยากรณ์ ค่าความแม่นยำ (Precision) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะปิดการเรียกเก็บเงิน (Charged

Off) คิดเป็น 99.07% รองลงมาคือ สถานะชำระหนี้ปกติ (Current) คิดเป็น 98.44%

ตารางที่ 5 จะเห็นว่า โมเดลได้พยากรณ์ ค่าความครบถ้วน (Recall) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้เต็มจำนวน (Fully Paid) คิดเป็น 95.28% รองลงมาคือ สถานะชำระหนี้ปกติ (Current) คิดเป็น 88.83%

และโมเดลได้พยากรณ์ ค่าความแม่นยำ (Precision) สถานะปัจจุบันของหนี้ มากที่สุดคือ สถานะชำระหนี้ปกติ (Current) คิดเป็น 93.04% รองลงมาคือ สถานะชำระหนี้เต็มจำนวน (Fully Paid) คิดเป็น 83.23%

ตารางที่ 6 จะเห็นว่า ค่าความถูกต้อง (Accuracy) สูงที่สุดคือ Random Forest ที่ 92.9% และเกิดข้อผิดพลาดในการจำแนกประเภท (Classification Error) สูงที่สุดคือ Naïve Bayes ที่ 16.6%

ตารางที่ 5 : การประเมินผลลัพธ์การทำนาย (Confusion Matrix) แสดงจำนวนที่พยากรณ์ของตัวแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

	true Fully Paid	**true Charged Off**	**true Current**	**class precision**
pred. Fully Paid	1677	107	231	83.23%
pred. Charged Off	74	165	6	67.35%
pred. Current	9	132	1884	93.04%
class recall	95.28%	40.84%	88.83%	

ตารางที่ 6 : ประสิทธิภาพของตัวแบบจำลอง

Model	**Accuracy Value**	**Classification Error**	**Standard Deviation**
Naïve Bayes	83.4%	16.6%	± 0.3%
Decision tree	87.0%	13.0%	± 0.1%
Random Forest	92.9%	7.1%%	± 0.1%
Support Vector Machine	87.0%	13.0%	± 0.6%

ตารางที่ 7 : การเปรียบเทียบค่าตัววัดประสิทธิภาพที่ได้จากการพยากรณ์ของตัวแบบจำลอง จำนวน 4 ค่า

Model	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Naïve Bayes	83.4%	97.79%	98.47%	98.13%
Decision tree	87.0%	**99.07%**	98.50%	98.78%
Random Forest	**92.9%**	**99.07%**	**99.64%**	**99.35%**
Support Vector Machine	87.0%	93.04%	95.28%	94.15%

ตารางที่ 7 พบว่า ตัวแบบจำลอง Random Forest มีค่าความถูกต้อง (Accuracy) ค่าความครบถ้วน หรือ ค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าการวัดประสิทธิภาพโดยรวม (F-Measure) สูงที่สุด คือ 92.9%, 99.07%, 99.64%, 99.35% ตามลำดับ

ตัวแบบจำลอง Decision tree มีค่าความแม่นยำ (Precision) สูงที่สุด คือ 99.07% เช่นกันกับ ตัวแบบจำลอง Random Forest ที่มีค่าสูงที่สุดได้เท่ากัน

ตารางที่ 8 : การเปรียบเทียบเวลาที่ใช้ในการประมวลผลตัวแบบจำลอง

Model	**Training Time (1,000 Rows)**	**Total Time**
Naïve Bayes	0.002 วินาที	26 วินาที
Decision tree	0.004 วินาที	20 วินาที
Random Forest	0.046 วินาที	1 นาที 39 วินาที
Support Vector Machine	0.792 วินาที	4 นาที 8 วินาที

ผลจากตารางที่ 7 ได้พบว่า ตัวแบบจำลอง Random Forest มีค่าความถูกต้อง (Accuracy) สูงที่สุด แต่สำหรับเวลาที่ใช้ในการประมวลผลตัวแบบจำลองนั้นไม่ได้ใช้น้อยหรือมากที่สุด ซึ่งในตารางที่ 8 นั้น พบว่า แบบจำลองที่ใช้เวลาน้อยที่สุด คือ ตัวแบบจำลอง Decision tree ใช้เวลา 20 วินาที และแบบจำลองที่ใช้เวลามากที่สุด คือ ตัวแบบจำลอง Support Vector Machine ใช้เวลา 4 นาที 8 วินาที

ตารางที่ 9 : การเปรียบเทียบค่าความน่าจะเป็นเกี่ยวกับสถานะปัจจุบันของหนี้จากการจำแนกข้อมูลของแบบจำลอง

สถานะของหนี้ / แบบจำลอง	**ชำระหนี้ปกติ**	**ชำระหนี้เต็มจำนวน**	**ปิดการเรียกเก็บเงิน**
Naïve Bayes	98.47%	0.00%	86.82%
Decision Tree	98.50%	48.33%	85.03%
Random Forest	**99.64%**	**63.57%**	**92.99%**
Support Vector Machine	95.28%	40.84%	88.83%

จากตารางที่ 9 พบว่า แบบจำลอง Random Forest มีความน่าจะเป็นเกี่ยวกับสถานะปัจจุบันของหนี้ โดย ลูกหนี้สถานะชำระหนี้ปกติ คือ 99.64% คิดเป็น 439,746 ตัวอย่าง ลูกหนี้สถานะชำระหนี้เต็มจำนวน คือ 63.57% คิดเป็น 280,557 ตัวอย่าง และลูกหนี้สถานะปิดการเรียกเก็บเงิน คือ 92.99% คิดเป็น 410,397 ตัวอย่าง จากจำนวนกลุ่มตัวอย่างทั้งหมด 441,335 ตัวอย่าง

เมื่อได้เปรียบเทียบกับผลการทดลองจากงานวิจัยผู้อื่นที่เกี่ยวข้องได้ผล ดังนี้

จากงานวิจัยระบบการพิจารณาอนุมัติเงินกู้สวัสดิการพนักงานอัตโนมัติด้วยเทคนิควิธีการเรียนรู้ด้วยเครื่อง [12] มีการศึกษาเทคนิค 4 เทคนิค คือ Naïve Bayes, K-Nearest Neighbors, Support Vector Machine และ Random Forest เพื่อพัฒนาและเปรียบเทียบแบบจำลอง ผลการวิจัยพบว่า เทคนิค Random Forest ให้ความแม่นยำ (Precision) สูงสุดที่ 99.56% ซึ่งผลของงานวิจัยที่ได้นั้นสอดคล้องกับงานวิจัยชิ้นนี้

Chitambira [13] ได้ศึกษา 6 เทคนิค คือ Decision Tree, Artificial Neural Network, Logistic Regression, Support Vector Machine, Random Forest และ K-Nearest Neighbors เพื่อพัฒนาและเปรียบเทียบแบบจำลอง ผลการวิจัยพบว่า เทคนิค Logistic Regression ให้ความแม่นยำสูงสุดที่ 81.50% รองลงมาคือ เทคนิค Random Forest ให้ความแม่นยำ (Precision) ที่ 81.30% ซึ่งผลของงานวิจัยที่ได้นั้นสอดคล้องกับงานวิจัยชิ้นนี้

ในงานวิจัยการพัฒนาแบบจำลองการพิจารณาให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูลของ Muangthanang *et al*. [14] ศึกษาอยู่ 5 วิธี คือ ต้นไม้การตัดสินใจ การจำแนกกลุ่มนาอีฟเบย์ โครงข่ายประสาท การถดถอยแบบโลจิสติก และซัพพอร์ตเวกเตอร์แมชชีน เพื่อทดสอบประสิทธิภาพของแบบจำลอง ผลการวิจัยพบว่า ค่าความแม่นยำ ของแบบจำลองที่ดีที่สุดคือ โครงข่ายประสาทเทียม ที่ 90.14% และ 90.08% จากชุดข้อมูล 2 กลุ่มข้อมูล ซึ่งผลของงานวิจัยที่ได้นั้นไม่สอดคล้องกับงานวิจัยชิ้นนี้

Bai *et al*. [15] ได้ศึกษาเทคนิค 4 เทคนิค คือ Naïve Bayes, Decision tree, Logistic regression และ Support Vector Machine เพื่อเปรียบเทียบความแม่นยำในการจำแนกประเภท ได้ผล Naïve Bayes และ Decision tree นั้น มีความแม่นยำ (Precision) สูงกว่า Logistic regression และ Support Vector Machine เป็นจำนวนร้อยละ 40 และเพื่อเปรียบเทียบค่า Precision – Recall ในการจำแนกประเภท ได้ผล Naïve Bayes และ Decision tree มีค่าที่ดีกว่า 40% เช่นกัน ซึ่งผลของงานวิจัยที่ได้นั้นสอดคล้องกับงานวิจัยชิ้นนี้

5) สรุปผล

งานวิจัยนี้ได้นำเสนอแบบจำลองและทดสอบประสิทธิภาพของแบบจำลองการอนุมัติสินเชื่อโดยใช้เทคนิคการวิเคราะห์องค์ประกอบร่วมกับการเรียนรู้ด้วยเครื่อง ทดสอบกับอัลกอริทึม 4 วิธี ได้แก่ นาอีฟเบย์ (Naïve Bayes) ต้นไม้การตัดสินใจ (Decision Tree) แรนดอมฟอร์เรส (Random Forest) และ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยทดสอบกับชุดข้อมูลการปล่อยกู้จำนวน 441,335 ตัวอย่างผลการทดสอบประสิทธิภาพของแบบจำลองจากค่าความถูกต้อง (Accuracy) จากการทดลองพบว่าแบบจำลองแรนดอมฟอร์เรส ให้ประสิทธิภาพดีที่สุดคือร้อยละ 92.90 รองลงมาเป็นแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้การตัดสินใจ ร้อยละ 87.00 และสุดท้ายนาอีฟเบย์ร้อยละ 83.40 ตามลำดับ และจากการทดลองพบว่าการลดมิติของข้อมูลส่งผลให้ประสิทธิภาพของแบบจำลองดีขึ้นเนื่องจากตัดคุณลักษณะที่ไม่มีนัยสำคัญทิ้งไป

แบบจำลองได้พยากรณ์ความน่าจะเป็นเกี่ยวกับสถานะปัจจุบันของหนี้จากกลุ่มตัวอย่างที่ดีที่สุด คือ ลูกหนี้สถานะชำระหนี้ปกติ 439,746 ตัวอย่าง ลูกหนี้สถานะชำระหนี้เต็มจำนวน 280,557 ตัวอย่าง และลูกหนี้สถานะปิดการเรียกเก็บเงิน 410,397 ตัวอย่าง ซึ่งสามารถกล่าวได้ว่า การใช้ตัวชี้วัดใหม่ที่ได้พัฒนาขึ้น สามารถแยกแยะคุณภาพของลูกหนี้ได้ค่อนข้างดี เพื่อจัดการความเสี่ยงที่จะเกิดเหตุการณ์การผิดนัดชำระหนี้ขึ้น สถาบันการเงินสามารถนำกลุ่มตัวอย่างมาใช้ร่วมกับแบบจำลองนี้เพื่อประกอบการตัดสินใจในการดำเนินนโยบายและกำกับดูแลการปล่อยสินเชื่อให้กับผู้กู้รายย่อย

REFERENCES

[1] Committee for the Protection of Credit Information. “Basic Information about Credit Scoring.” (in Thai), CREDITINFO COMMITTEE.or.th. https://www.creditinfocommittee.or.th/Public/BasicIntormation (accessed Mar. 1, 2023).

[2] Supervision Group, Bank of Thailand. “Credit scoring, things that need to be preserved.” (in Thai), BOT.or.th. https://www.bot.or.th/th/research-and-publications/articles-and-publications/articles/Article_22Jan201.html (accessed Mar. 1, 2023).

[3] P. Moeynoi. “Dimensionality Reduction.” (in Thai), MEDIUM.com. https://medium.com/@doohpim/02-การลดมิติข้อมูล-dimensionality-reduction-a5df2adb4f92 (accessed Sep. 7, 2022).

[4] S. Teerawittayakul, “Ransomware detection study using machine learning techniques from a sample attack dataset,” (in Thai), Dept. Data Commun. Netw., King Mongkut's Univ. Technol. North Bangkok, Bangkok, Thailand, 2019.

[5] PradyaSin. “What is Random Forest.” (in Thai), MEDIUM.com. https://medium.com/@pradyasin/random-forest-คืออะไร-74d2a0af3d7 (accessed Sep. 7, 2022).

[6] B. Pasilatesang, *Build Learning for AI with Python Machine Learning*. Bangkok, Thailand: SE-EDUCATION (in Thai), 2021.

[7] C. Ambi. “ Difference Between Normalization and Standardization.” TOWARDSAI.net. https://pub.towardsai.net/difference-between-normalization-and-standardization-745030eaf96f (accessed Sep. 7, 2022).

[8] E. Pacharawongsakda, *Practical Data Science with Rapid Miner Studio 1*. Nonthaburi, Thailand: IDC Premier (in Thai), 2022.

[9] *Lending Club Loan Data: Complete loan data for all loans issued through the 2007-2015, including status*, LendingClub 2007-2015, 2021. [Online]. Available: https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv

[10] N. Promrita and S. Waijanya, *Fundamental of DEEP LEARNING in Practice*. Nonthaburi, Thailand: IDC Premier (in Thai), 2021.

[11] T. Kaewsri. “Feature Engineering.” (in Thai), MEDIUM.com. https://grimreaperjunior.medium.com/สกัดข้อมูลให้ตรงใจด้วยการทำ-feature-engineering-bcdba80124bc (accessed Sep. 7, 2022).

[12] S. Tidta and T. Kangkachit, “Automated system for employee welfare loan approval through machine learning techniques,” (in Thai), *DPU Graduate Studies J.*, vol. 10, no. 2, pp. 93–107, 2022. [Online]. Available: https://grad.dpu.ac.th/upload/content/files/year10-2/10-7.pdf

[13] B. Chitambira, “Credit scoring using machine learning approaches,” M.S. thesis, Division Math. Phys., Malardalen Univ., Vasteras, Sweden, Jun. 2022. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1664698/FULLTEXT01.pdf

[14] C. Muangthanang, S. Mungsing, and N. Chirawichitchai, “Development of a credit scoring model using data mining techniques,” (in Thai), *Huachiew Chalermprakiet Sci, Technol. J.*, vol. 7, no. 1, pp. 82–93, 2021.

[15] R. Bai, H. Mo, and H. Li, “A comparison of credit rating classification models based on spark-evidence from lending-club,” *Procedia Comput. Sci.*, vol. 162, pp. 811–818, 2019.