การวิเคราะห์ในการปล่อยสินเชื่อธนาคารโดยใช้การเรียนรู้ของเครื่อง

Bank Lending Analysis by Using Machine Learning

ศรุต สมุทรโสภากุล*[1] และ ณัฏฐ์ ดิลกธนากุล[2]

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เลขที่ 1 ซอยฉลองกรุง 1 แขวงลาดกระบัง เขตลาดกระบัง กรุงเทพฯ 10520

บทคัดย่อ

การศึกษาอิสระชิ้นนี้เป็นการพัฒนาแบบจำลองโดยใช้การเรียนรู้ของเครื่องเพื่อเป็นเครื่องมือช่วยลดความเสี่ยงและช่วยคาดการณ์แนวโน้มบุคคลที่ผิดนัดชำระโดยใช้อัลกอรึทึมซึ่งมีวิธีการเรียนรู้แบบมีผู้สอนในการพัฒนาแบบจำลองและเพิ่มคลาสสัดส่วนน้อยด้วยวิธีการสุ่มเกินด้วยวิธีสโมท ผลการศึกษาพบว่าแบบจำลองการถดถอยโลจิสติก แรนดอมฟอเรสต์ เกรเดียนต์บูสต์ติง และซัพพอร์ตเวกเตอร์แมทชีน หลังจากมีการปรับเธรชโฮลด์ได้ค่าพรีซิชั่นเท่ากับ 0.04, 0.06, 0.04 และ 0.05 ตามลำดับ รีคอลเท่ากับ 0.70, 0.72, 0.70 และ 0.70 ตามลำดับและเอยูซีเท่ากับ 0.63, 0.73, 0.64 และ 0.69 ตามลำดับ การทำวิศวกรรมฟีเจอร์สามารถเพิ่มประสิทธิภาพแบบจำลองทั้ง 4 แบบจำลอง โดยรูปแบบบินนิ่งใช้ฟีเจอร์ Age ได้ค่าเฉลี่ยรีคอลและเอยูซีเพิ่มขึ้นร้อยละ 1.77 และ 4.43 ตามลำดับ รองลงมาคือรูปแบบผสมหมวดหมู่กับฟีเจอร์ ProductType และ Add02-HouseType ได้ค่าเฉลี่ยเอยูซีเพิ่มขึ้นร้อยละ 2.47 รูปแบบการเข้ารหัสเป้าหมายและการเข้ารหัสนับกับฟีเจอร์ Education ได้ค่าเฉลี่ยเอยูซีเพิ่มขึ้นร้อยละ 1.52 เมื่อนำวิศวกรรมฟีเจอร์ในแต่ละรูปแบบมาผสมกันสามารถเพิ่มประสิทธิภาพแบบจำลองได้เล็กน้อย

คำสำคัญ: การคาดการณ์ผิดนัดชำระ, การเรียนรู้แบบมีผู้สอน, วิศวกรรมฟีเจอร์

Abstract

This Independent study is about using machine learning for reducing risk and predicting tendency of people who default. Using algorithm with a supervised learning method and working with imbalanced dataset by SMOTE. The results show that after threshold adjustment of Logistic regression, Random forest, Gradient boosting and Support vector machine, Precisions were 0.04, 0.06, 0.04 and 0.05 respectively, Recalls were 0.70, 0.72, 0.70 and 0.70 respectively and AUCs were 0.63, 0.73, 0.64, and 0.69, respectively. Feature engineering was able to increase performance of all four models. Binning with Age increased average Recall and AUC by 1.77% and 4.43% respectively, Combining Categories with ProductType and Add02-HouseType increased average AUC by 2.47%, Target Encoding and Count Encoding with Education increased average AUC by 1.52%. Combining feature engineering can slightly improve model performance.

Keywords: Default prediction, Supervised learning, Feature engineering

---

* Author's Email: 64607063@it.kmitl.ac.th

1 นักศึกษาปริญญาโท สาขาวิชาปัญญาประดิษฐ์เพื่อการวิเคราะห์เชิงธุรกิจ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกลา้ เจ้าคุณทหาร ลาดกระบัง

2 อาจารย์ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้า เจ้าคุณทหารลาดกระบัง

1. บทนำ

1.1 ที่มาและความสำคัญ

การปล่อยสินเชื่อเป็นหนึ่งในหน้าที่หลักของธนาคาร ในปัจจุบันการลดความเสี่ยงในการปล่อยกู้ วิธีการหนึ่งคือมีการใช้บริการแบบจำลองของเครดิตบูโร ซึ่งวิธีการดังกล่าวมีผลลัพธ์จากการทำนายที่ดี แต่ก่อให้เกิดค่าใช้จ่ายที่สูง และทางผู้ใช้งานไม่สามารถทราบได้ว่าสิ่งที่ใช้งานอยู่มีการทำงานอย่างไร ดังนั้นการศึกษาอิสระชิ้นนี้มีวัตถุประสงค์ที่จัดทำเครื่องมือที่ช่วยในการลดความเสี่ยงโดยใช้แบบจำลองการเรียนรู้ของเครื่องมาช่วยคาดการณ์บุคคลเพื่อเป็นข้อมูลเสริมในการตัดสินใจให้กับการวิเคราะห์ในการปล่อยสินเชื่อของธนาคาร

1.2 วัตถุประสงค์

1.2.1 เพื่อนำเสนอแบบจำลองการเรียนรู้ของเครื่องสำหรับการคาดการณ์ในการปล่อยสินเชื่อ

1.3 ขอบการศึกษา

1.3.1 ชุดข้อมูลที่ได้นำมาศึกษาคือชุดข้อมูลเกี่ยวกับการขอสินเชื่อรายย่อย

1.3.2 พัฒนาแบบจำลองคาดการณ์เพื่อจำแนกว่าควรปล่อยหรือไม่ปล่อยสินเชื่อ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 เป็นเครื่องมือช่วยลดความเสี่ยงในการปล่อยสินเชื่อให้กับธนาคาร

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ความเสี่ยงด้านเครดิต

ความเสี่ยงด้านเครดิต หมายถึง ความเป็นไปได้ซึ่งลูกค้าหรือคู่สัญญาของธนาคารไม่สามารถปฏิบัติตามสัญญาในการจ่ายชำระหนี้คืน รวมทั้งความเสี่ยงที่เกิดจากโอกาสที่ลูกค้าอาจไม่สามารถชำระหนี้คืนได้ จนเป็นเหตุให้ถูกปรับลดอันดับความน่าเชื่อถือลงซึ่งอาจส่งผลกระทบต่อฐานะเงินกองทุน รายได้ของธนาคาร รวมถึงการสำรองเงินของธนาคาร [1]

2.2 การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่องอยู่ในสาขาวิชาวิทยาการคอมพิวเตอร์ ซึ่งสิ่งนี้พยายามทำความเข้าใจประเภทงานต่าง ๆ โดยใช้หลักการคณิตศาตร์และอิงจากข้อมูล อีกทั้งอัลกอริทึมที่สามารถเรียนรู้จากชุดข้อมูลได้และพัฒนาผลลัพธ์จากการเปรียบเทียบชุดข้อมูลการสอนและชุดข้อมูลการทดสอบ เป้าหมายของการเรียนรู้ของเครื่องคือการออกแบบให้แบบจำลองสามารถเรียนรู้ได้อย่างอัตโนมัติและสามารถทำความเข้าใจกับงานที่ได้รับมอบหมายได้ด้วยตัวเอง [2] การเรียนรู้ของเครื่องมีตัวแปรซึ่งบ่งบอกคุณลักษณะของเหตุการณ์ที่กำลังศึกษาอยู่ซึ่งเรียกว่าฟีเจอร์ (Feature)

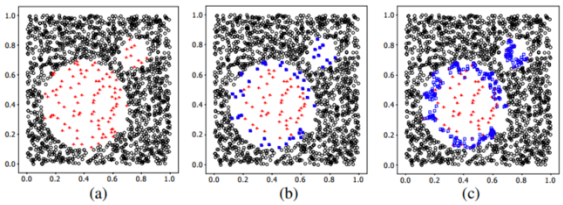
2.3 วิธีการเรียนรู้

วิธีการเรียนรู้ของเครื่องสำหรับการจำแนกในงานต่าง ๆ วิธีการซึ่งนิยมเป็นส่วนใหญ่คือวิธีการเรียนรู้ โดยในการศึกษาอิสระชิ้นนี้ใช้วิธีการเรียนรู้แบบมีผู้สอน (Supervised learning) ซึ่งเมื่อมีอินพุตและผลเฉลยของเอาต์พุต (Label) เป็นลักษณะการอนุมานฟังก์ชันจากผลเฉลยเอาต์พุตระหว่างชุดข้อมูลการสอนแบบจำลอง ซึ่งสามารถใช้ในการจำแนกกับข้อมูลที่ไม่มีผลเฉลยเอาต์พุต (Unlabel) ในอนาคตได้ [3]

2.4 การทำงานกับชุดข้อมูลไม่สมดุล

ในโลกแห่งความเป็นจริงจำนวนข้อมูลของคลาสที่สนใจมีจำนวนน้อยเมื่อเทียบกับชุดข้อมูลทั้งหมด เมื่อพิจารณาธุรกิจเกี่ยวกับชุดข้อมูลของการปล่อยสินเชื่อซึ่งประกอบด้วยเหตุกาณ์ที่จ่ายปกติและผิดนัดชำระ โดยเหตุการณ์ที่ผิดนัดชำระเป็นกรณีที่เกิดส่วนน้อย ส่งผลให้เกิดเป็นชุดข้อมูลไม่สมดุล ผลที่ตามมา เมื่อทำการสอนและวัดผลแบบจำลอง เช่น การวัดผลด้วยค่าความแม่นยำ (Accuracy) ซึ่งการวัดผลดังกล่าวไม่สามารถสะท้อนผลที่ออกได้อย่างแท้จริง เช่น หากมีชุดข้อมูลคลาสที่ไม่ได้สนใจ (Negative) และในคลาสที่สนใจ (Positive) เป็นสัดส่วนที่ 99% และ 1% ตามลำดับ หากสอนและวัดผลแบบจำลองซึ่งแบบจำลองจะจำลักษณะของข้อมูลในคลาสที่ไม่ได้สนใจได้มากกว่า ดังนั้นความแม่นยำจะอยู่ที่ 99% ซึ่งไม่สามารถยืนยันได้ว่าแบบจำลองนี้สามารถหาคุณลักษณะ

ของคลาสที่สนใจได้ [4] ดังนั้นการแก้ปัญหาของชุดข้อมูลไม่สมดุล ในการศึกษาอิสระชิ้นนี้จึงใช้วิธีการสุ่มเกิน (Oversampling) ด้วยวิธี SMOTE (Synthetic Minority Over-sampling Technique) เพื่อเพิ่มจำนวนสัดส่วนของคลาสที่สนใจ ซึ่งเป็นวิธีการทำงานโดยการค้นหาเพื่อนบ้านใกล้สุด k ตัว (k-nearest neighbors) สำหรับข้อมูลในคลาสที่มีสัดส่วนน้อย การเพิ่มจำนวนสัดส่วนใช้เป็นการสังเคราะห์ข้อมูลขึ้นมาใหม่โดยทำนายคลาสของข้อมูลพิจารณาจากคลาสของข้อมูลรอบ ๆ ข้างที่ใกล้ที่สุด k ข้อมูล ซึ่งอิงจากการคำนวณระยะห่างระหว่างข้อมูลและลักษณะของกระบวนการนี้เป็นดังรูปที่ 2.1 [5]



รูปที่ 2.1 (a) การกระจายตัวชุดข้อมูลแบบเดิม (b) จำนวนคลาสส่วนน้อย (c) จำนวนคลาสหลังจาก SMOTE [5]

2.5 การวัดประสิทธิภาพแบบจำลอง

การวัดประสิทธิภาพแบบจำลองได้ใช้เกณฑ์ในการวัดซึ่งแบ่งออกเป็นดังนี้ [3]

2.5.1 Confusion matrix เป็นวิธีการวัดผลของอัลกอริทึมการจำแนก (Classification algorithm) โดยแสดงผลลัพธ์ของค่าความแม่นยำอยู่ในรูปตาราง ดังแสดงในตารางที่ 2.1 จุดประสงค์เพื่อแสดงผลลัพธ์ที่ได้จากแบบจำลองมีการจำแนกเป็นคลาสใด ผลลัพธ์ที่ออกมามีค่าเป็นได้ทั้งหมด 4 ค่า ในตารางได้แก่ True positive (TP) คือจำนวนของคลาส Positive ที่ถูกแบบจำลองจำแนกได้อย่างถูกต้อง, True negative (TN) คือจำนวนของคลาส Negative ที่ถูกแบบจำลองจำแนกได้อย่างถูกต้อง, False positive (FP) คือจำนวนของคลาสที่ถูกแบบจำลองทำนายได้ไม่ถูกต้องซึ่งทำนายเป็น Positive เมื่อคลาสที่ถูกต้องเป็น Negative และ False negative (FN) จำนวนของคลาสที่ถูกแบบจำลองทำนายได้ไม่ถูกต้องซึ่งทำนายเป็น Negative เมื่อคลาสที่ถูกต้องเป็น Positive

ตารางที่ 2.1 ลักษณะของ Confusion matrix

	Predict positive	Predict negative
Positive	TP	FN
Negative	FP	TN

2.5.2 ค่าความแม่นยำ (Accuracy) เป็นการใช้ผลลัพธ์ที่ได้จาก Confusion matrix สามารถคำนวนค่าความแม่นยำได้ดังสมการที่ (1)

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{P+N}} \qquad (1)$$

ข้อบกพร่องเมื่อใช้มาตรวัดนี้ในการวัดผลแบบจำลองนำมาสู่เหตุการณ์ความขัดแย้งด้านความแม่นยำ (Accuracy paradox) ซึ่งเหตุการณ์นี้หมายถึงแบบจำลองมีค่าความแม่นยำที่สูงแต่ความสามารถในการทำนายได้อย่างถูกต้องที่ต่ำ

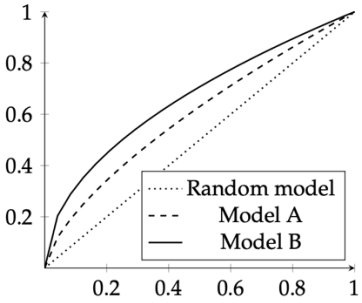
2.5.3 Precision and recall เป็นมาตรวัดทางสถิติที่นิยมใช้คู่กันในการวัดประสิทธิภาพแบบจำลอง โดยค่า Precision หรือค่าที่สนใจจากการทำนายเฉพาะคลาส Positive ซึ่งประกอบด้วยคลาส Positive ที่ทำนายได้อย่างถูกต้องและคลาสที่ถูกทำนายเป็น Positive แต่ค่าจริงคือ Negative สามารถคำนวนค่าได้ดังสมการที่ (2)

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2)$$

ในส่วนของ Recall หรืออัตราของ True positive (TP Rate)โดยเป็นการดูคลาส Positive ที่ทำนายได้อย่างถูกต้องและคลาสที่ถูกทำนายเป็น Negative แต่ค่าจริงคือ Positive สามารถคำนวนค่าได้ดังสมการที่ (3)

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (3)$$

2.5.4 Receiver Operating Characteristics (ROC) และ Area Under the ROC Curve (AUC) สำหรับ ROC ใช้เพื่อดูประสิทธิภาพแบบจำลองโดยเป็นลักษณะของกราฟเส้นระหว่างความสัมพันธ์อัตราของ True positive (TP/P) กับอัตราของ False positive (FP/N) โดยช่วงของผลลัพธ์ที่เป็นไปได้คือค่าระหว่าง 0 ถึง 1 ดังรูปที่ 2.2 [6]



รูปที่ 2.2 กราฟเส้นโค้ง ROC ประกอบด้วยแบบจำลองแบบสุ่ม (เส้นประ) และแบบจำลองที่มีความสามารถในการจำแนก [6]

จากรูปที่ 2.2 มีเส้น ROC ทั้งหมด 3 เส้นโค้ง ประกอบด้วย แบบจำลองแบบสุ่ม 1 แบบจำลอง และส่วนที่เหลือคือแบบจำลองที่มีความสามารถในการจำแนก สำหรับเส้น ROC แบบจำลองแบบสุ่มมีลักษณะเส้นโค้งแบบเส้นทแยงมุมโดยเป็นการลากเส้นจากจุดตัดที่ (0,0) ไปที่จุดตัด (1,1) หากแบบจำลองมีความสามารถในการจำแนกคลาสได้อย่างถูกต้องและมีความสามารถมากกว่าแบบจำลองแบบสุ่ม ค่าอัตราของ True positive จะมีค่ามากกว่าค่าอัตราของ False positive และลักษณะของเส้นโค้งจะอยู่เหนือกว่าเส้นทแยงมุม

สำหรับ AUC เป็นมาตรวัดที่ใช้วัดพื้นที่ใต้กราฟที่ได้จากเส้นโค้ง ROC ซึ่งค่าที่ได้จาก AUC เป็นค่าประมาณการ เพราะระหว่างในการคำนวณจะเกิดการสูญหายของข้อมูล สำหรับแบบจำลองแบบสุ่มค่า AUC มีค่าเท่ากับ 0.50 และแบบจำลอง B ค่า AUC มีค่าประมาณ 0.67 แบบจำลองที่มีความสามารถมากกว่าแบบจำลองแบบสุ่ม ค่าของ AUC จะมีค่ามากกว่า 0.50

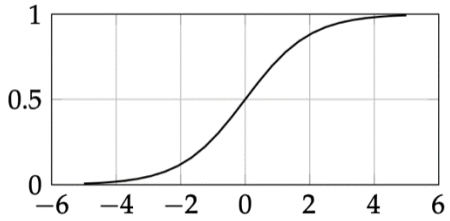
2.6 Cross-validation เป็นการวัดประสิทธิภาพของแบบจำลองเมื่อทำการสอนแบบจำลองแล้วเสร็จ โดยลักษณะของวิธีดังกล่าวคือชุดข้อมูลจะถูกแบ่งออกเป็น 2 ชุด คือข้อมูลฝึกสอน (Training set) และชุดข้อมูลทดสอบ (Test set) Cross-validation มีหลากหลายวิธีการโดยจุดประสงค์เพื่อลดความแปรปรวนของแบบจำลองซึ่งหากแบบจำลองมีความแปรปรวนมากเกิน จะก่อให้เกิดปัญหาแบบจำลองมีความซับซ้อนมากเกินไปจนเกิดมีพฤติกรรมท่องจำข้อมูลฝึกสอน (Overfitting) หนึ่งในวิธีการส่วนใหญ่ซึ่งนิยมใช้คือวิธีการแบบโฮลด์เอาท์ (Holdout method) ซึ่งวิธีการนี้เป็นการแบ่งชุดข้อมูลออกเป็นข้อมูลฝึกสอนและชุดข้อมูลทดสอบ โดยสัดส่วนที่มากถูกแบ่งอยู่ในชุดข้อมูลฝึกสอน สำหรับสัดส่วนที่นิยมใช้คืออยู่ในช่วงระหว่าง 30:70 ถึง 10:90 [7] ซึ่งข้อบกพร่องของวิธีการนี้คือเกิดการแบ่งชุดข้อมูลเพียงครั้งเดียว ดังนั้นแบบจำลองที่ถูกสอนจึงพบกับข้อมูลที่ไม่หลากหลายดังนั้นความแปรปรวนของแบบจำลองจึงเกิดแบบสุ่มซึ่งในบางครั้งอาจนำมาสู่ Overfitting ในการศึกษาอิสระชิ้นนี้ใช้วิธี Cross-validation แบบ k-fold cross-validation ซึ่งเป็นวิธีการที่ชุดข้อมูลถูกเปลี่ยนลำดับที่มีมาก่อนหน้าแบบสุ่ม (Shuffle) และถูกแบ่งออกเป็นจำนวน k ชุด จำนวนชุดข้อมูลฝึกสอนถูกใช้อยู่ที่ k-1 ชุด และส่วนคงเหลือถูกใช้เป็นข้อมูลทดสอบ โดยกระบวนการนี้กระทำเป็นจำนวน k รอบซึ่งชุดข้อมูลที่ถูกแบ่งไว้ก่อนหน้าจะถูกวนเวียนใช้เป็นชุดข้อมูลทดสอบ 1 ครั้ง การวัดประสิทธิภาพแบบจำลองเป็นไปในลักษณะของค่าเฉลี่ย [8]

2.7 Ensemble เป็นวิธีการที่รวบรวมผลลัพธ์ซึ่งได้จากการจำแนกจากหลาย ๆ แบบจำลองมาดูค่าเฉลี่ยของผลลัพธ์ จุดประสงค์คือการรวบรวมหลายแบบจำลองมีความเป็นไปได้ที่ได้ผลลัพธ์ที่ดีกว่าแบบจำลองเดียว [9]

2.8 Model description ในการศึกษาอิสระชิ้นนี้ใช้อัลกอริทึมในการสร้างแบบจำลองต่าง ๆ ดังนี้

Logistic regression เป็นอัลกอริทึมที่ถูกพัฒนามาจากการถดถอยเชิงเส้น (Linear regression) เพื่อประมาณค่าความน่าจะเป็นหรือทำนายคลาสที่สนใจ ลักษณะของอัลกอริทึมดังกล่าวมีความแตกต่างกับการถดถอยเชิงเส้นตรงตรงที่มีความน่าจะเป็นเข้าในเกี่ยวข้องโดยอยู่ในรูปของสมการซึ่งเรียกว่าโลจิตฟังก์ชัน (Logit function) ดังแสดงในสมการที่ (5) และรูปที่ 2.3 ซึ่งเป็นลักษณะของลอการิทึมธรรมชาติ (Natural logarithm) ของอัตราส่วนของความน่าจะเป็นของคลาสที่สนใจต่อความน่าจะเป็นของคลาสที่ไม่สนใจ (Odds) ซึ่งลักษณะของเอาต์พุตของโลจิตฟังก์ชันมีค่าระหว่าง 0 ถึง 1 [10]

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \qquad (5)$$

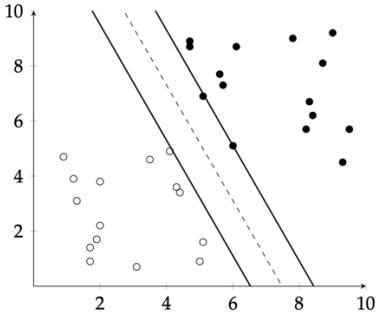


รูปที่ 2.3 กราฟโลจิตฟังก์ชันของการถดถอยโลจิสติก [10]

Random forest เป็นอัลกอรึทึมที่อ้างอิงมาจากต้นไม้ตัดสินใจแต่ลักษณะของ Random forest เป็นวิธีการแบบ Ensemble คือมีการสร้างต้นไม้ตัดสินใจหลาย ๆ แบบจำลองโดยเป็นการใช้ชุดข้อมูลการสอนชุดเดียวกัน ซึ่งอัลกอรึทึมดังกล่าว ยังมีการสุ่มฟีเจอร์ของในแต่ละแบบจำลองย่อยเพื่อให้เกิดความหลากหลายในขณะที่สอน จุดประสงค์ของการสุ่มฟีเจอร์เพื่อลดค่าสหสัมพันธ์ (Correlation) ของแบบจำลองย่อยเพื่อให้แบบจำลองย่อยมีความอิสระและความหลากหลายมากขึ้น [11]

Gradient boosting เป็นอัลกอรึทึมที่อ้างอิงมาจากต้นไม้ตัดสินใจและเป็นวิธีการแบบ Ensemble โดยลักษณะของอัลกอรึทึมนี้คือจะสร้างแบบจำลองต้นไม้ตัดสินใจหลาย ๆ แบบจำลอง ซึ่งลักษณะการทำงานคือแบบจำลองย่อยที่ถูกสร้างขึ้นใหม่จะถูกสอนบนค่าความผิดพลาด (Residual) ของแบบจำลองย่อยก่อนหน้า จุดประสงค์เพื่อลดค่าความผิดพลาดโดยมีแนวคิดที่ว่าค่าดังกล่าวควรจะลดลง [12]

Support vector machine (SVM) เป็นอัลกอรึทึมที่มีการสร้างเส้นแบ่งกลุ่มข้อมูลของแต่ละคลาสซึ่งเรียกว่าไฮเปอร์เพลน (Hyperplane) โดยเส้นไฮเปอร์เพลนควรมีระยะห่างกับกลุ่มข้อมูลซึ่งถูกแบ่งมากที่สุด ในส่วนของมิติไฮเปอร์เพลน ลักษณะของมิติมีค่าเท่ากับ N-1 กับมิติของชุดข้อมูล เช่น ชุดข้อมูลมีมิติเท่ากับ 2 มิติ เส้นมิติของไฮเปอร์เพลนมีค่าเท่ากับ 1 มิติ ดังรูปที่ 2.4 [13]



รูปที่ 2.4 SVM สร้างเส้นไฮเปอร์เพลน 1 มิติ แบ่งคลาสข้อมูลซึ่งมิติเท่ากับ 2 มิติ [13]

จากรูปที่ 2.4 เส้นประและเส้นทึบคือเส้นไฮเปอร์เพลนที่แบบจำลอง SVM สร้างเพื่อแบ่งคลาสของชุดข้อมูล โดยระยะห่างระหว่างเส้นทึบทั้ง 2 เส้นเรียกว่ามาร์จิน (Margin) โดยเส้นไฮเปอร์เพลนที่ดีควรมีมาร์จินที่สูงเพราะสามารถลดปัญหาการเกิด Overfitting ได้

2.9 Feature engineering

ประสิทธิภาพแบบจำลองส่วนใหญ่ขึ้นอยู่กับฟีเจอร์ในชุดข้อมูล การทำ Feature engineering คือการสังเคราะห์ฟีเจอร์ขึ้นมาใหม่โดยอิงจากฟีเจอร์เดิมที่มีอยู่ โดยฟีเจอร์ใหม่อาจอยู่ในรูปของอัตราส่วน ผลต่างหรือการแปลงทางคณิตศาสตร์ [14] วิธีการที่ใช้แบ่งเป็นดังนี้ [15]

2.9.1 Target Encoding คือการ Group by โดเมนของฟีเจอร์ซึ่งอาจเท่ากับหรือมากกว่า 1 ฟีเจอร์ และรวมผลลัพธ์ของคลาสเอาต์พุต (Target feature)

2.9.2 Count Encoding คือการ Group by โดเมนฟีเจอร์ซึ่งอาจเท่ากับหรือมากกว่า 1 ฟีเจอร์ และคำนวณผลลัพธ์อยู่ในรูปของความถี่

2.9.3 Combining Categories คือการนำฟีเจอร์ต่าง ๆ มารวมกันได้เป็นฟีเจอร์ใหม่ เนื่องจากฟีเจอร์ตัวมันเองไม่มีความสามารถในการทำนายคลาสเอาต์พุตแต่หากนำฟีเจอร์ต่าง ๆ รวมกัน สามารถเกิดเป็นรูปแบบให้โมเดลเรียนรู้ได้

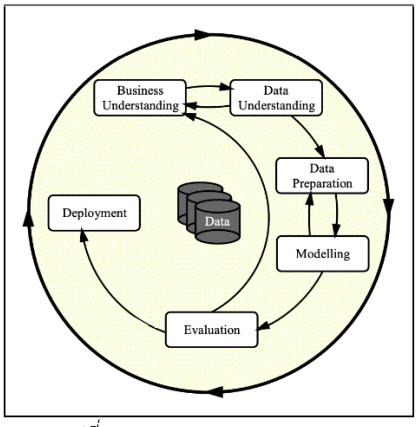
2.9.4 Binning คือการนำฟีเจอร์ซึ่งมีลักษณะเป็น Numerical หรือ Ordinal มาแบ่งเป็นกลุ่ม

2.10 งานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้พัฒนาระบบฐานข้อมูลสำหรับวิสาหกิจขนาดกลางและขนาดย่อมและพัฒนาแบบจำลองการเรียนรู้ของเครื่องโดยได้ผลลัพธ์ค่า AUC ของ Random forest อยู่ที่ 70-80% รองลงมาคือ Logistic regression แต่การใช้งานเลือกใช้แบบจำลองที่ 2 เพราะสามารถนำไปพัฒนาในเชิงของแอปพลิเคชันได้มากกว่า [16] การทบทวนงานวิจัยเรื่องการใช้การเรียนรู้สำหรับการจัดการความเสี่ยงด้านเครดิตทั้งหมด 136 งานวิจัยซึ่งพิจารณา 3 ปัจจัยซึ่งคาดว่าช่วยในการลดความเสี่ยงได้แก่ ความเสี่ยงด้านเครดิต การตรวจสอบอย่างพึงระมัดระวังและการตรวจสอบเชิงลึก โดยพบว่าใช้แบบจำลองโครงข่ายประสาทเทียมและ SVM ถูกพัฒนาเป็นส่วนใหญ่ [17] และบทความวิจัยเรื่องการปล่อยสินเชื่อบัตรเครดิตโดยใช้แบบจำลองการเรียนรู้ของเครื่อง ซึ่งใช้แบบจำลอง Logistic regression และ Random forest ได้ค่าความแม่นยำที่ 61% และ 60% ตามลำดับ [18]

3. วิธีการวิจัย

การศึกษาค้นคว้าอิสระชิ้นนี้ได้นำหลัก CRISP-DM มาช่วยเป็นกรอบในการทำงาน ดังแสดงในรูปที่ 3.1



รูปที่ 3.1 วงจรของ CRISP-DM [19]

CRISP-DM (Cross-industry standard process for data mining) เป็นกระบวนการเพื่อแสดงให้เห็นภาพรวมถึงวงจรสำหรับการทำเหมืองข้อมูล โดยแบ่งออกเป็นทั้งหมด 6 ระยะ โดยลำดับในแต่ละระยะไม่ได้ถูกกำหนดไว้คงที่ [19] ซึ่งสามารถแบ่งระยะออกเป็นดังนี้

3.1 ความเข้าใจทางธุรกิจ (Business Understanding) การบริการด้านสินเชื่อรายย่อยในธนาคารแบ่งออกเป็น 3 ประเภทหลัก ได้แก่

3.1.1 สินเชื่อเพื่อที่อยู่อาศัย คือสินเชื่อที่ให้แก่ลูกหนี้ที่เพื่อการจัดหาที่อยู่อาศัย โดยมีที่ดินหรือสิ่งปลูกสร้างนั้นจำนองเป็นหลักประกัน

3.2.2 สินเชื่อมีหลักประกัน คือสินเชื่อส่วนบุคคลที่ผู้ขอสินเชื่อนำทรัพย์สินทั้งที่เป็นสังหาริมทรัพย์หรืออสังหาริมทรัพย์มาเป็นหลักประกันในการกู้ยืม เช่น ที่อยู่อาศัย รถยนต์ บัญชีเงินฝาก บำเหน็จตกทอด พันธบัตร

3.3.3 สินเชื่อไม่มีหลักประกัน คือสินเชื่อเพื่ออุปโภคบริโภคแก่ลูกค้าประเภทบุคคลธรรมดา ตามประเภทผลิตภัณฑ์ที่ธนาคารกำหนดเป็นสินเชื่อที่ไม่มีหลักประกันที่นำมาคำนวณมูลค่าค้ำประกัน ลักษณะของผลิตภัณฑ์เช่น บัตรเครดิต บัตรกดเงินสด

ปัจจุบันธนาคารมีการใช้บริการแบบจำลองของเครดิตบูโร ซึ่งผลลัพธ์แบบจำลองออกมาในรูปแบบของเกรด เช่น AA, BB, CC ในการพิจารณาและปล่อยสินเชื่อธนาคารจะกำหนดช่วงเกรดที่ธนาคารสามารถรับความเสี่ยงได้ในแต่ละผลิตภัณฑ์ โดยหากเข้าเงื่อนไขธนาคารจะปล่อยสินเชื่อให้กับลูกค้ารายนั้น ๆ

3.2 ความเข้าใจเกี่ยวกับข้อมูล (Data Understanding) ข้อมูลที่ใช้ประกอบไปด้วยข้อมูลเกี่ยวกับคุณลักษณะของลูกค้าที่มาขอสินเชื่อกับทางสาขาธนาคารโดยมาจากระบบฐานข้อมูลธนาคาร ซึ่งผ่านการรวบรวมข้อมูลและทำการเปลี่ยนแปลงข้อมูลให้อยู่ในรูปแบบเดียวกันเพื่อนำไปใช้งานต่อได้ ในการกำหนดข้อมูลป้ายกำกับ (Label) จะกำหนดโดยดูลักษณะพฤติกรรมการชำระว่ามีการชำระตรงเวลาหรือไม่ โดยดูว่าธุรกรรมของลูกค้ารายนั้น ๆ ภายใน 6 เดือนล่าสุด มีการชำระไม่เกิน 7 วัน กำหนดเท่ากับ Good แต่หากเกิน 7 วัน กำหนดเท่ากับ Bad รายละเอียดข้อมูลดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 ข้อมูลเกี่ยวกับคุณลักษณะลูกค้า

ฟีเจอร์	ความหมาย	ชนิดข้อมูล
Age	อายุ	Integer
Sex	เพศ	Categorical
Nationality	สัญชาติ	Categorical
Religion	ศาสนา	Categorical
Country	ประเทศ	Categorical
Education	การศึกษา	Categorical
MaritalStatus	สถานภาพสมรส	Categorical
SpouseIncome	รายได้คู่สมรส	Integer
NumofChildren	จำนวนบุตร	Integer
StudyInCountry	บุตรศึกษาในประเทศ	Integer
StudyAbroad	บุตรศึกษานอกประเทศ	Integer
Graduate	บุตรสำเร็จการศึกษา	Integer
Occupation	อาชีพ	Categorical
Position	ตำแหน่งในอาชีพ	Categorical
IncomeSource	แหล่งที่มารายได้	Categorical
NumofEmployee	จำนวนพนักงาน	Integer
TotalExp-year	อายุงานรวม (ปี)	Integer
TotalExp-month	อายุงานรวม (เดือน)	Integer
RequiredCollateral	มีหลักประกัน/ไม่มี	Categorical
ProductType	ประเภทผลิตภัณฑ์	Categorical
CreditLimit	วงเงินสินเชื่อ	Integer
Term	ระยะเวลาผ่อนชำระ (เดือน)	Integer
Rate	ดอกเบี้ย	Float
Add02-HouseType	ประเภทที่อยู่	Categorical
Add02-HouseOwner	สถานะความเป็นเจ้า	Categorical

3.3 การเตรียมข้อมูล (Data Preparation)

3.3.1 การตรวจสอบค่าว่างข้อมูลและแทนที่ค่าว่าง กระบวนการการจัดการค่าว่างเป็นไปในลักษณะ การแทนค่าด้วยค่าฐานนิยม (Mode)

3.3.2 การแปลงฟีเจอร์เชิงหมวดหมู่ให้อยู่ในรูปของตัวเลข ฟีเจอร์เชิงหมวดหมู่เดิมไม่ได้มีลักษณะเป็นตัวเลขจึงต้องทำการแปลง โดยฟีเจอร์ที่เป็น Ordinal จะทำการแปลงเป็นตัวเลขให้มีลักษณะเป็นตามลำดับและฟีเจอร์ที่เป็น Nominal จะทำการแปลงเป็นไปในลักษณะของ One-Hot Encoding

3.3.3 การปรับช่วงข้อมูล ฟีเจอร์ที่เป็นช่วงตัวเลขเดิม ปรับช่วงข้อมูลสำหรับให้อยู่ในช่วงเดียวกัน ใช้เป็นในลักษณะของ Min-max scaler ดังสมการที่ (6)

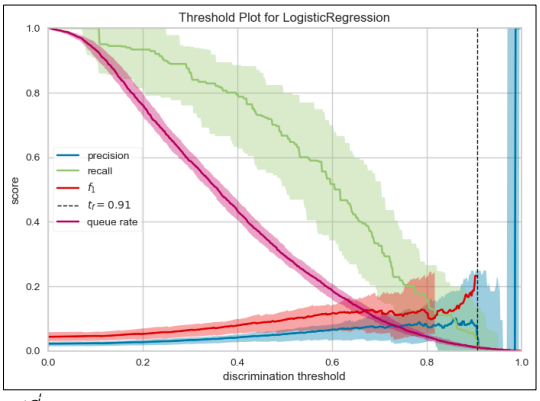
$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (6)$$

3.3.4 การทำ Feature engineering ทดลองด้วยการนำฟีเจอร์ต่าง ๆ มาสังเคราะห์เพื่อให้ได้ฟีเจอร์ใหม่ โดยการทำ Target Encoding และ Count Encoding ใช้ฟีเจอร์เป็น Categorical การทำ Combining Categories ใช้เป็นฟีเจอร์ Numerical และ Categorical และการทำ Binning ใช้เป็นฟีเจอร์ Numerical

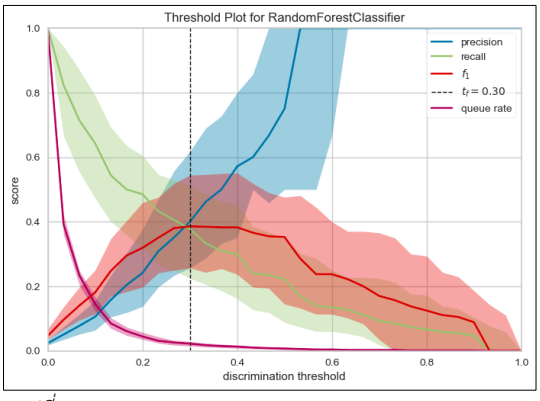
3.4 การสร้างแบบจำลอง (Modelling) แบ่งสัดส่วนชุดข้อมูลโดยชุดข้อมูลฝึกสอน ชุดข้อมูลตรวจสอบและชุดข้อมูลทดสอบในสัดส่วน 60:20:20 ตามลำดับ ใช้วิธีการ SMOTE เพื่อเพิ่มจำนวนข้อมูลในคลาสที่มีสัดส่วนน้อยก่อนทำการสอนแบบจำลอง ในการสอนจะทำการสอนแบบไม่ทำ Feature engineering และสอนแบบทำ Feature engineering หลังจากนั้นทำการวัดประสิทธิภาพแบบจำลอง

4. ผลการศึกษา

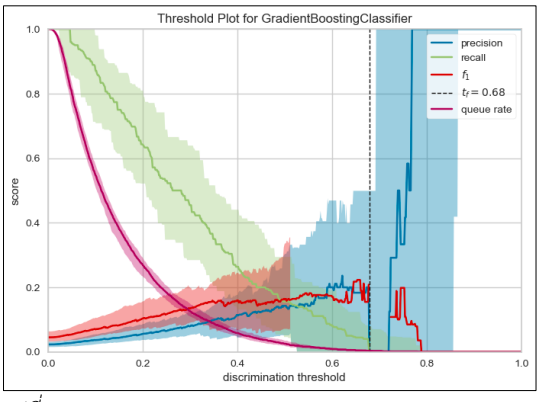
4.1 เมื่อให้แบบจำลอง Logistic regression, Random forest, Gradient boosting และ Support vector machine จำแนกคลาสซึ่งออกมาเป็นในลักษณะของความน่าจะเป็น โดยค่าเริ่มต้นของค่าดังกล่าวถูกกำหนดไว้เท่ากับ 0.5 หากความน่าจะเป็นของคลาสไหนที่มากกว่า 0.5 จะถูกจำแนกเป็นคลาสนั้น ๆ ดังนั้นจึงมีการกำหนดความน่าจะเป็นในการจำแนกคลาสโดยอิงจากกราฟ Discrimination threshold ของแต่ละแบบจำลอง ดังรูปที่ 4.1 – 4.4



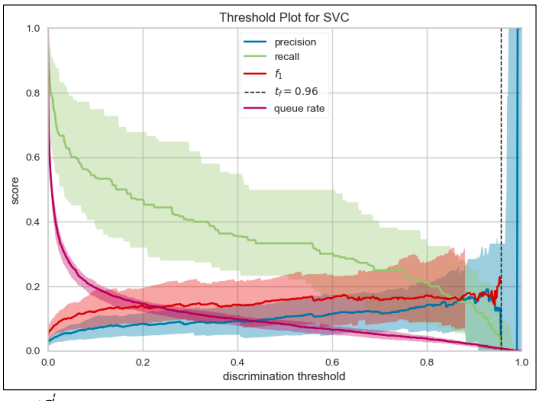
รูปที่ 4.1 Discrimination threshold ของ Logistic regression



รูปที่ 4.2 Discrimination threshold ของ Random forest



รูปที่ 4.3 Discrimination threshold ของ Gradient boosting



รูปที่ 4.4 Discrimination threshold ของ Support vector machine

การวัดประสิทธิภาพแบบจำลองกับชุดข้อมูลทดสอบโดยยังไม่ผ่านการทำ Feature engineering ได้ผลลัพธ์ดังตารางที่ 4.1 โดยกำหนดความน่าจะเป็นในแต่ละแบบจำลอง โดยให้แบบจำลองมีค่า Recall ประมาณ 0.70 เพราะให้ความสำคัญกับ False Negative

ตารางที่ 4.1 การเปรียบเทียบค่า Precision, Recall และ AUC หลังจากมีการปรับค่า Threshold โดยไม่ผ่าน Feature engineering

Model	Threshold	Precision	Recall	AUC
Logistic regression	0.39	0.04	0.70	0.63
Random forest	0.05	0.06	0.72	0.73
Gradient boosting	0.14	0.04	0.70	0.64
Support vector machine	0.03	0.05	0.70	0.69

จากการทดลองพบว่าการทำ Binning ที่ฟีเจอร์ Age โดยแบ่งช่วงข้อมูลออกเป็น 5 ช่วงข้อมูล นั้นช่วยเพิ่มประสิทธิภาพแบบจำลองที่ดีที่สุดเมื่อเทียบกับการทำ Binning ฟีเจอร์ในกลุ่ม Numerical เดียวกัน ได้ผลลัพธ์ดังตารางที่ 4.2

ตารางที่ 4.2 การเปรียบเทียบค่า Precision, Recall และ AUC หลังจากมีการปรับค่า Threshold โดยผ่านการ Binning ที่ฟีเจอร์: Age

Model	Threshold	Precision	Recall	AUC
Logistic regression	0.41	0.04	0.70	0.65
Random forest	0.059	0.07	0.77	0.76
Gradient boosting	0.148	0.04	0.70	0.67
Support vector machine	0.042	0.06	0.70	0.73

จากการทดลองพบว่าการทำ Target Encoding และ Count Encoding ที่ฟีเจอร์ Education นั้นช่วยเพิ่มประสิทธิภาพแบบจำลองที่ดีที่สุดเมื่อเทียบกับการทำ ในฟีเจอร์ในกลุ่ม Categorical เดียวกัน ได้ผลลัพธ์ดังตารางที่ 4.3

ตารางที่ 4.3 การเปรียบเทียบค่า Precision, Recall และ AUC หลังจากมีการปรับค่า Threshold โดยผ่านการ Target Encoding และ Count Encoding ที่ฟีเจอร์: Education

Model	Threshold	Precision	Recall	AUC
Logistic regression	0.442	0.04	0.70	0.66
Random forest	0.04	0.06	0.72	0.71
Gradient boosting	0.161	0.05	0.70	0.67
Support vector machine	0.028	0.05	0.70	0.68

จากการทดลองพบว่าการทำ Combining Categories ที่ฟีเจอร์ ProductType และ Add02-HouseType นั้นช่วยเพิ่มประสิทธิภาพแบบจำลองที่ดีที่สุดเมื่อเทียบกับการทำในฟีเจอร์ในกลุ่ม Numerical และ Categorical เดียวกัน ได้ผลลัพธ์ดังตารางที่ 4.4

ตารางที่ 4.4 การเปรียบเทียบค่า Precision, Recall และ AUC หลังจากมีการปรับค่า Threshold โดยผ่านการ Combining Categories ที่ฟีเจอร์: ProductType และ Add02-HouseType

Model	Threshold	Precision	Recall	AUC
Logistic regression	0.43	0.04	0.70	0.66
Random forest	0.05	0.07	0.72	0.75
Gradient boosting	0.141	0.04	0.70	0.65
Support vector machine	0.035	0.05	0.70	0.69

เมื่อนำวิธีการ Feature engineering ในแต่รูปแบบมารวมกัน โดยทำการทดลองกับฟีเจอร์ต่าง ๆ โดยภาพรวมสามารถเพิ่มประสิทธิภาพแบบจำลองขึ้นได้เล็กน้อย เมื่อเทียบกับการผ่าน Feature engineering แบบ 1 รูปแบบ ได้ผลลัพธ์ดังตารางที่ 4.5

ตารางที่ 4.5 การเปรียบเทียบค่า Precision, Recall และ AUC หลังจากมีการปรับค่า Threshold โดยผ่านการ Target Encoding และ Count Encoding ที่ฟีเจอร์: Add02-HouseOwner และ Binning ที่ฟีเจอร์: Age

Model	Threshold	Precision	Recall	AUC
Logistic regression	0.404	0.04	0.70	0.64
Random forest	0.04	0.06	0.74	0.73
Gradient boosting	0.168	0.05	0.70	0.69
Support vector machine	0.03	0.06	0.70	0.71

5. สรุปผลและข้อเสนอแนะ

5.1 สรุปผลการทดลอง

5.1.1 พัฒนาแบบจำลองทั้งหมด 4 แบบจำลองได้แก่ แบบจำลอง Logistic regression, Random forest, Gradient boosting และ Support vector machine ซึ่งวัดประสิทธิภาพแบบจำลองหลังจากมีการปรับค่า Threshold ซึ่งยังไม่ผ่านการ Feature engineering ได้ค่า Precision เท่ากับ 0.04, 0.06, 0.04 และ 0.05 ตามลำดับ ได้ค่า Recall เท่ากับ 0.70, 0.72, 0.70 และ 0.70 ตามลำดับ และค่า AUC ได้เท่ากับ 0.63, 0.73, 0.64 และ 0.69 ตามลำดับ

5.1.2 การทำ Feature engineering สามารถช่วยประสิทธิภาพแบบจำลองโดยการ Binning ที่ฟีเจอร์ Age สามารถเพิ่มประสิทธิภาพแบบจำลองได้ดีที่สุด โดยค่าเฉลี่ย Recall และ AUC ทั้ง 4 แบบจำลอง เพิ่มขึ้น 1.77% และ 4.43% ตามลำดับ รองลงมาคือ Combining Categories ที่ฟีเจอร์ ProductType และ Add02-HouseType ได้ค่าเฉลี่ย AUC เพิ่มขึ้น 2.47% Target

Encoding, Count Encoding ที่ฟีเจอร์ Education ได้ค่าเฉลี่ย AUC เพิ่มขึ้น 1.52% และเมื่อนำวิธีการ Feature engineering ในแต่รูปแบบมาผสมกันสามารถเพิ่มประสิทธิภาพแบบจำลองได้เล็กน้อย

5.1.3 การปรับค่า Threshold ในแต่ละแบบจำลองเพื่อเพิ่มค่า Recall เป็นหลักแต่ในขณะเดียวกัน Precision มีค่าค่อนข้างต่ำ โดยหากนำผลลัพธ์ที่ได้ไปใช้งานโดยไม่ผ่านการวิเคราะห์ร่วมกับปัจจัยอื่น ๆ จะส่งผลให้พนักงานสาขาที่ดูแลลูกค้ามีภาระงานที่สูงขึ้น

5.2 ข้อเสนอแนะ

การศึกษาอิสระชิ้นนี้ได้ใช้ข้อมูลคุณลักษณะของลูกค้าโดยฟีเจอร์ส่วนใหญ่ไม่มีลักษณะเป็นอนุกรมเวลาและทดลองรูปแบบ Feature engineering ทั้งหมด 4 รูปแบบ งานชิ้นต่อไปเสนอว่าควรใช้ลักษณะฟีเจอร์ที่มีลักษณะเป็นอนุกรมเวลามาเป็นฟีเจอร์ในการสอนโมเดลเพื่อให้ได้ผลดียิ่งขึ้น เช่น พฤติกรรมการชำระในแต่ละเดือน และทดลองการทำ Feature engineering ในรูปแบบอื่น ๆ เช่น Logarithm, Difference, Ratio

6. บรรณานุกรม

[1] ธนาคารแห่งประเทศไทย. 2561. **แนวทางการตรวจสอบแบบเน้นธุรกรรมที่สำคัญของสถาบันการเงิน.** [Online] Available:https://www.bot.or.th/Thai/FinancialInstitutions/PruReg_HB/RiskMgt_Manual/Documents/SA-Framework.pdf

[2] Blum, A. 2007. **Machine Learning Theory.** [Online] Available: https://www.cs.cmu.edu/~avrim/Talks/mlt.pdf

[3] Cornelissen, M. 2018. **Applying machine learning to the prediction of defaults in loans.**[Online].Available:https://essay.utwente.nl/75060/1/Cornelissen_MA_BMS.pdf

[4] Chawla, N., Japkowicz, N. and Kotcz, A. 2004. “Special issue on learning from imbalanced data sets.” **ACM Sigkdd Explorations Newsletter.** 16(1): 1–6.

[5] Hui Han., Wen-Yuan Wang. and Bing-Huan Mao. 2005. “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.” 878–887. In **ICIC'05: Proceedings of the 2005 international conference on Advances in Intelligent Computing.** Berlin:Springer

[6] Fawcett, T. 2006. “An introduction to ROC analysis” **Pattern Recognition Letters.** 27(8): 861–874.

[7] Zhang, G.P. 2009. “Neural Networks For Data Mining” 17–44 In Maimon, O. and Rokach, L. **Soft Computing for Knowledge Discovery and Data Mining.** Boston:Springer.

[8] Rodriguez, J.D., Perez, A. and Lozano, J. A. 2010. “Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation” **IEEE Transactions on Pattern Analysis and Machine Intelligence.** 32(3): 569-575.

[9] Rokach, L. 2009. “Ensemble Methods in Supervised Learning” 959–979 In Maimon, O. and Rokach, L. **Data Mining and Knowledge Discovery Handbook.** Boston:Springer.

[10] กาญจน์เขจร ชูชีพ. 2561. **การถดถอยโลจีสติก (Logistic Regression).** [Online] Available:https://forest-admin.forest.ku.ac.th/304xxx/?q=system/files/book/5%282018%29%20Logistic%20Regression.pdf

[11] Breiman, L. 2001. “Random Forests.” **Machine Learning.** 45(1): 5–32.

[12] Friedman, J.H. 2001. “Greedy Function Approximation: a Gradient Boosting Machine.” **The Annals of Statistics.** 29(5): 1189-1232

[13] Shmilovici, A. 2005. “Support Vector Machines.” 231–247 In Maimon, O. and Rokach, L. **Data Mining and Knowledge Discovery Handbook.** Boston:Springer.

[14] Heaton, J. “An Empirical Analysis of Feature Engineering for Predictive Modeling” 1-6. In **SoutheastCon 2016.** Norfolk: IEEE

[15] ACM RecSys. 2020. RecSys 2020 **Tutorial: Feature Engineering for Recommender Systems.** [Online] Available: https://www.youtube.com/watch?v=uROvhp7cj6Q

[16] ธนาคารแห่งประเทศไทย. 2564. **Credit Risk Database: Credit Scoring Models for Thai SMEs.** กรุงเทพฯ: สถาบันวิจัยเศรษฐกิจป๋วย อึ๊งภากรณ์

[17] Bhatore, S., Mohan, L. and Reddy, Y.R. 2020. “Machine learning techniques for credit risk evaluation: a systematic literature review.” **Journal of Banking and Financial.** 4(1): 111–138.

[18] SriLaxmi, K., Divya, N, Lakshmi, P., Vidya, A. and Hameeda, S. 2020. “Credit Card Customer Predicting using Machine Learning.” **International Journal for Research in Applied Science and Engineering Technology.** 8(5): 2697–2701.

[19] Wirth, R. and Hipp, J. 2000. “CRISP-DM: Towards a standard process model for data mining.” 29-40. In **International conference; 4th, Practical application of knowledge discovery and data mining.** Manchester: Practical Application Co.