การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้ของเครื่อง

เครือวัลย์ เนตรพนา[1*], ศิริสรรพ เหล่าหะเกียรติ[2]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อการศึกษาการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการทดลองกับชุดข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิตซึ่งประกอบด้วย ข้อมูลจำนวนทั้งหมด 307,511 แถว และคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะเว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มใหญ่ๆ คือ กลุ่มลูกหนี้ปกติ คือกลุ่มลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และกลุ่มลูกหนี้ที่ไม่ปกติ คือกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร เครื่องมือหลักที่นักวิจัยใช้ ได้แก่ Machine Learning Algorithms เช่น Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting เป็นต้น โดยอาศัยการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งเป็นเครื่องมือสำหรับการพัฒนาแบบจำลองในการเรียนรู้แบบผู้สอน (Supervised Learning) โดยมีการทำงานแบบการแบ่งแยกประเภท (Classification) ซึ่งการเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนรู้ของเครื่องในการเรียนรู้ข้อมูล โดยอาศัยชุดข้อมูลที่ใช้ในการฝึกฝนเพื่อทำการพัฒนาแบบจำลองและชุดข้อมูลที่ใช้ในการทดสอบสำหรับใช้ในการการทดสอบแบบจำลอง โดยเราสามารถนำผลลัพธ์ที่ได้ ไปตรวจสอบกับชุดข้อมูลที่ใช้ในการทดสอบที่เรามีอยู่แล้ว ว่าแบบจำลองที่ถูกพัฒนาขึ้นนั้น มีประสิทธิภาพและความถูกต้อง (Accuracy) มากน้อยเพียงใด แต่จากชุดข้อมูลที่นำมาใช้ในการวิเคราะห์ข้อมูล พบว่าข้อมูลมีความไม่สมดุลกันของชุดข้อมูล (Imbalance data) สูงมาก ซึ่งทำให้ค่าความถูกต้อง (Accuracy) ที่ได้อาจมีค่าที่สูงมาก แต่มีประสิทธิภาพที่ไม่เพียงพอ เพราะค่า precision, recall และ F1-Score ที่ได้มีค่าที่ต่ำมาก โดยเราต้องอาศัยเทคนิคต่างๆ มาช่วยในการแก้ปัญหาความไม่สมดุลของชุดข้อมูล เช่น Oversampling, Under sampling และ Synthetic Minority Oversampling Technique (SMOTE) เพื่อทำให้แบบจำลองที่ได้มีประสิทธิภาพที่ดี

ผลการศึกษาพบว่า การพัฒนาแบบจำลองโดยการใช้เทคนิควิธี XGBoost ให้ค่าความไว (Recall) ที่มากที่สุด ซึ่งมีค่าเท่ากับ 0.97 มีค่าความถูกต้อง (Accuracy) เท่ากับ 0.37 และมีค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองเท่ากับ 0.51 แต่เทคนิควิธี K-Nearest Neighbors (KNN) ให้ค่าความไว (Recall) ที่น้อยที่สุด ซึ่งมีค่าเท่ากับ 0.57 มีค่าความถูกต้อง (Accuracy) เท่ากับ 0.55 และมีค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองเท่ากับ 0.46 ซึ่งมีค่าที่น้อยที่สุด

คำสำคัญ : การเรียนรู้ของเครื่อง, การเรียนรู้แบบผู้สอน, การแบ่งแยกประเภท, ความไม่สมดุลกันของชุดข้อมูล

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 088-1444224 E-mail address: Kruewan.net@g.swu.ac.th

Analysis of credit card debt default risk analysis by using machine learning algorithm

Kruewan Netphana[1*], Sirisup Laohakiat[2]

Abstract

This research aims to study the prediction of debtors who are likely to default on their payments to the bank, using a dataset of credit card transactions. The dataset consists of 307,511 rows and 122 columns, sourced from a public data site. The data is divided into two main groups, normal debtors who comply with their payments, and abnormal debtors who default on their payments. The primary tool used by the researchers is Machine Learning Algorithms, such as Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest, Support Vector Classifier (SVC), and Gradient Boosting. Machine Learning is a tool used to develop models, with the help of supervised learning, which involves classification. The researchers used a training set to develop the model and a testing set to evaluate the model's performance. However, they found that the data was significantly imbalanced, which affected the model's accuracy, causing the model's precision, recall and F1-Score values to be low. To overcome this problem, they employed techniques such as Oversampling, under sampling, and Synthetic Minority Oversampling Technique (SMOTE), to improve the model's performance.

The study found that developing a model using XGBoostClassifier technique provides the highest value of Recall, which is equal to 0.97. However, the Accuracy value is only 0.37 and the F1-Score is 0.51, which is used to measure the effectiveness of the model. On the other hand, K-Nearest Neighbors (KNN) technique provides the lowest value of Recall, which is 0.57. But the Accuracy value is 0.55 and the F1-Score is 0.46, which are the lowest values.

Keywords : Machine Learning, Supervised Learning, Classification, Clustering, Imbalance data

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 088-1444224 E-mail address: Kruewan.net@g.swu.ac.th

บทนำ

ปัจจุบันประชาชนส่วนใหญ่มีการใช้งานบัตรเครดิตที่เพิ่มมากขึ้น และมีการนำบัตรเครดิตมาใช้เป็นเครื่องมือหลักๆ ในการใช้จ่ายในชีวิตประจำวัน จึงทำให้ผู้คนส่วนใหญ่มีความสะดวกรวดเร็วในการใช้จ่ายมากยิ่งขึ้น และยังมีความปลอดภัยในการใช้จ่ายมากกว่าการพกเงินสดติดตัวเป็นจำนวนมาก ซึ่งจะทำให้เสี่ยงต่อการสูญหายหรือโจรกรรมมากยิ่งขึ้น โดยสมัยนี้ร้านค้า ห้างสรรพสินค้า และศูนย์การค้าต่างๆ ที่ให้บริการในหลายๆ แห่งทั่วประเทศ ได้มีการรับชำระเงินผ่านบัตรเครดิตที่เพิ่มมากขึ้น ทำให้ผู้คนส่วนใหญ่สามารถใช้บัตรเครดิตในการชำระค่าสินค้าและค่าบริการ แทนการชำระเงินด้วยเงินสด

บัตรเครดิต คือ ผลิตภัณฑ์ทางการเงินรูปแบบหนึ่ง ที่เป็นการกู้ยืมเงินจากทางธนาคาร หรือจากสถาบันการเงินต่างๆ มาใช้จ่ายล่วงหน้า โดยใช้ในการชำระค่าสินค้าและค่าบริการแทนการชำระเงินด้วยเงินสด โดยวงเงินในการใช้จ่ายนั้นจะต้องไม่เกินยอดวงเงินที่สถาบันการเงินแต่ละสถาบันการเงินอนุมัติ ซึ่งจะต้องทำการชำระคืนในภายหลัง ซึ่งมีให้เลือกทั้งในรูปแบบของการชำระคืนแบบเต็มจำนวน ชำระคืนแบบจ่ายขั้นต่ำ หรือการผ่อนชำระผ่านบัตรเครดิต ซึ่งบัตรเครดิตในปัจจุบันมีให้เลือกหลากหลายรูปแบบและหลากหลายประเภท

ด้วยการสมัครบัตรเครดิตสมัยนี้นั้นเป็นเรื่องที่ง่ายมากยิ่งขึ้น จึงทำให้ผู้คนส่วนใหญ่หันมาใช้จ่ายผ่านบัตรเครดิตที่เพิ่มมากยิ่งขึ้น ซึ่งเอกสารในการสมัครบัตรเครดิต ใช้เพียงแค่สลิปเงินเดือนหรือหนังสือรับรองเงินเดือนภายใน 3 เดือน ก็สามารถสมัครบัตรเครดิตได้แล้ว และยังรองรับกับบุคคลที่ทำงานในหลากหลายอาชีพ ซึ่งไม่ว่าจะทำงานอาชีพไหนก็สามารถสมัครบัตรเครดิตได้ และด้วยเงื่อนไขการอนุมัติวงเงินที่สามารถทำได้สะดวกและรวดเร็วมากยิ่งขึ้น โดยไม่จำเป็นต้องมีหลักทรัพย์ค้ำประกัน และบัตรเครดิตบางประเภทยังมีเงื่อนไขในการยกเว้นค่าธรรมเนียมแรกเข้าและค่าธรรมเนียมรายปี การแลกคะแนนสะสมเพื่อแลกรับสิทธิ์ ของรางวัล หรือบัตรกำนัลต่างๆ นั้นจึงเป็นสาเหตุหลัก ที่ทำให้ประชาชนส่วนใหญ่หันมาใช้จ่ายผ่านบัตรเครดิตแทนการใช้จ่ายด้วยเงินสดเป็นจำนวนมากยิ่งขึ้น

แต่จากการเติบโตของการใช้จ่ายเงินสดผ่านบัตรเครดิต ซึ่งเติบโตขึ้นอย่างรวดเร็ว จึงทำให้ผู้คนส่วนใหญ่หันมาใช้จ่ายผ่านบัตรเครดิตแทนการชำระเงินด้วยเงินสด ถึงแม้ว่าจะช่วยทำให้ผู้คนส่วนใหญ่มีความสะดวกและรวดเร็วในการใช้จ่ายมากยิ่งขึ้น แต่ยังก่อให้เกิดปัญหาต่าง ๆ อีกมากมาย เช่น ปัญหาหนี้ที่ไม่ก่อให้เกิดรายได้หรือที่เรียกว่าหนี้เสีย อันเนื่องมาจากการใช้จ่ายที่ฟุ่มเฟือยจนเกินความสามารถในการที่จะชำระเงินคืนให้กับทางธนาคาร

งานวิจัยนี้ เน้นการศึกษาการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการทดลองในครั้งนี้ เราทดลองกับ ข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิตซึ่งประกอบด้วย ข้อมูลจำนวนทั้งหมด 307,511 แถว และคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะเว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มใหญ่ๆ คือ กลุ่มลูกหนี้ปกติ คือกลุ่มลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และกลุ่มลูกหนี้ที่ไม่ปกติ คือกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร

งานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับการพัฒนาแบบจำลองในการตรวจจับการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระ ซึ่งงานวิจัยส่วนใหญ่ที่พบจะเกี่ยวข้องกับการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต

งานวิจัยของ (Awoyemi, Adetunmbi et al. 2017) กล่าวถึง การตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต โดยการใช้เทคนิค Machine Learning บทความนี้ได้ทำการตรวจสอบประสิทธิภาพของการพัฒนาแบบจำลองโดยการใช้อัลกอริทึม naïve bayes, k-nearest neighbor และ logistic regression กับชุดข้อมูลธุรกรรมการใช้งานบัตรเครดิตที่ได้มาจากผู้ถือบัตรเครดิตในยุโรป ซึ่งมีธุรกรรมการทำรายการทั้งหมด 284,807 รายการ โดยธุรกรรม 492 รายการเป็นธุรกรรมที่เป็นการฉ้อโกง โดยจะใช้เทคนิค under-sampling และ over-sampling เพื่อแก้ไขปัญหาความไม่สมดุลของชุดข้อมูล และมีการใช้เทคนิค Feature Selection เพื่อคัดเลือก Feature ที่สำคัญที่จะนำมาใช้ในการพัฒนาแบบจำลอง และแสดงการเปรียบเทียบประสิทธิภาพของการพัฒนาแบบจำลองกับงานวิจัยอื่นๆที่เกี่ยวข้อง ซึ่งผลลัพธ์ประสิทธิภาพของการพัฒนาแบบจำลองจะดูจากค่า accuracy, sensitivity, specificity, precision, Matthews correlation coefficient และ balanced classification rate ซึ่งแบบจำลองที่ให้ผลลัพธ์ค่าความแม่นยำที่เหมาะสมที่ดีที่สุด คือ แบบจำลองที่พัฒนาโดยอัลกอริทึม k-nearest neighbor ซึ่งให้ค่าความแม่นยำเท่ากับ 97.92%

งานวิจัยของ (Varmedja, Karanovic et al. 2019) กล่าวถึง การตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต โดยการใช้เทคนิค Machine Learning บทความนี้ จะแสดง Machine Learning หลายๆอัลกอริทึมที่นำมาใช้ในการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต มีธุรกรรมการทำรายการทั้งหมด 284,807 รายการ โดยธุรกรรม 492 รายการเป็นธุรกรรมที่เป็นการฉ้อโกง มีการใช้เทคนิค Feature Selection เพื่อลดการเกิด overfitting ซึ่งได้มีการลด Feature ทำให้เหลือ 27 Feature ที่ถูกนำมาใช้ในการพัฒนาแบบจำลอง ซึ่งจากชุดข้อมูลที่นำมาใช้ในการพัฒนาแบบจำลอง ชุดข้อมูลมีความไม่สมดุลกันของชุดข้อมูลสูงมาก จึงได้มีการใช้เทคนิค SMOTE (Synthetic Minority Over-sampling) เพื่อนำมาช่วยในการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล โดยอัลกอริทึมที่ใช้ในการพัฒนาแบบจำลอง ได้แก่ Logistic Regression, Random Forest, Naive Bayes และ Multilayer Perceptron ซึ่งเกณฑ์ที่ใช้ในการพิจารณาผลลัพธ์ของอัลกอริทึมการเรียนรู้ของเครื่อง คือค่า Accuracy, Recall และ Precision ซึ่งแบบจำลองที่ให้ผลลัพธ์ค่าความแม่นยำที่เหมาะสมที่ดีที่สุด คือ แบบจำลองที่พัฒนาโดยอัลกอริทึม Random Forest ซึ่งให้ค่าความแม่นยำ เท่ากับ 99.96%

งานวิจัยของ (Xuan, Liu et al. 2018) กล่าวถึง Random Forest เป็นหนึ่งในวิธีในการนำมาใช้ในการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต โดย Random Forest จะใช้ในการจำแนกประเภท ซึ่งเป็นอัลกอริทึมที่นิยมในการทำต้นไม้ตัดสินใจ เนื่องจากมีความยืดหยุ่นในการจัดการคุณลักษณะข้อมูลประเภทต่างๆ อย่างไรก็ตามแบบจำลองต้นไม้ตัดสินใจต้นเดียวอาจมีประสิทธิภาพที่ไม่เพียงพอและทำให้เกิดการ Overfit ซึ่งจะใช้เทคนิค Ensemble ในการแก้ปัญหาเหล่านี้ โดยการรวมกลุ่มกัน

ของต้นไม้ตัดสินใจหลายๆต้น จะช่วยเพิ่มความแม่นยำมากกว่าการทำต้นไม้ตัดสินใจต้นเดียว ซึ่งข้อดีของ Random Forest คือมีความแข็งแกร่งต่อสัญญาณรบกวน (noise) และมีความแข็งแกร่งต่อค่าที่ผิดปกติ (outlier)

งานวิจัยของ (Kiran, Guru et al. 2018) กล่าวถึง Bayesian network classifiers ได้รับความนิยมอย่างมากในด้านการเรียนรู้ของเครื่อง และเป็นตัวในการจำแนกประเภทแบบการเรียนรู้แบบมีผู้สอน (Supervised Learning) Naïve Bayes เป็นเทคนิคที่ใช้ทฤษฎีความน่าจะเป็นตามกฎของเบย์ โดยอาศัยหลักการของความน่าจะเป็นเข้ามาช่วยในการหาคำตอบของเหตุการณ์หนึ่งๆที่สนใจ โดยจะพยายามทำนายคลาสที่เรียกว่าคลาสผลลัพธ์ จะมีการกำหนดค่าความน่าจะเป็นขั้นต่ำและค่าความน่าจะเป็นสูงสุดไว้ล่วงหน้าของธุรกรรมที่เป็นการฉ้อโกงหรือธุรกรรมที่เป็นธุรกรรมที่ถูกกฎหมาย จากนั้นสำหรับธุรกรรมที่เข้ามาใหม่ เราจะคำนวณค่าความน่าจะเป็นของธุรกรรมนั้น หากค่าความน่าจะเป็นที่ถูกกฎหมายน้อยกว่าค่าต่ำสุดที่กำหนดไว้สำหรับธุรกรรมที่ถูกกฎหมาย และมากกว่าค่าสูงสุดที่กำหนดไว้สำหรับธุรกรรมที่เป็นการฉ้อโกง หากเป็นจริงธุรกรรมที่เข้ามาใหม่จะถูกจัดประเภทเป็นการฉ้อโกง

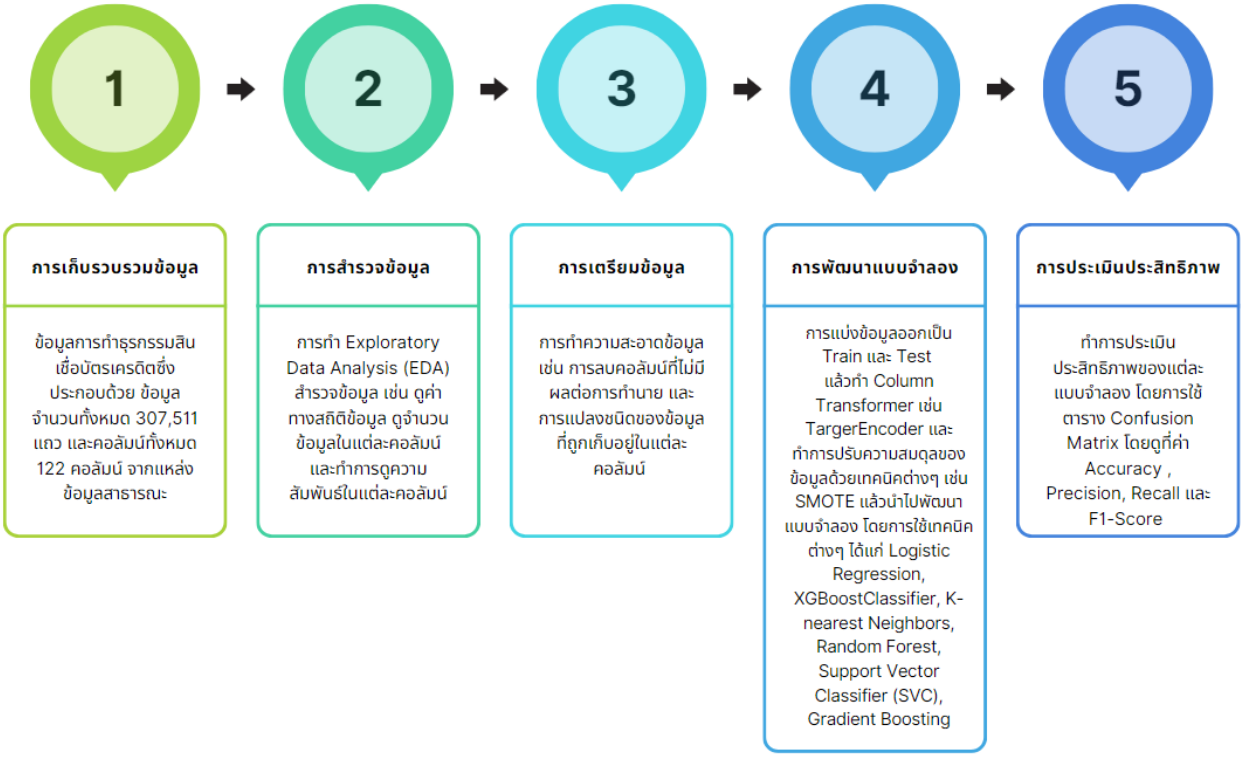
งานวิจัยของ (Kiran, Guru et al. 2018) กล่าวถึง K-Nearest Neighbor อัลกอริทึมนี้มีการเรียนรู้ที่ช้ามาก ถ้าเปรียบเทียบกับอัลกอริทึมอื่นๆ เนื่องจาก K-Nearest Neighbor ต้องมีการคำนวณระยะห่างระหว่างข้อมูลที่ต้องการใช้ในการพิจารณากับชุดข้อมูลตัวอย่าง จำนวน k ชุด หลักการทำงานของ K-Nearest Neighbor คือมีการแบ่งข้อมูลออกเป็นกลุ่มต่างๆ โดยจะทำการวิเคราะห์ข้อมูลใหม่จากข้อมูลเดิม ที่อยู่ในบริเวณใกล้เคียงกัน โดยจะกำหนดค่า k ซึ่งค่า k คือค่าที่ใช้ในการกำหนดว่าจะวิเคราะห์ข้อมูลที่อยู่ใกล้กับข้อมูลที่ต้องการจำแนกที่สุดกี่ข้อมูล ซึ่งการกำหนดค่า k ที่แตกต่างกันนั้น จะทำให้ได้ค่าความแม่นยำที่ไม่เท่ากัน ซึ่งนั้นก็เป็นส่วนหนึ่งในการหาค่า k ที่ดีที่สุด และทำการพิจารณาตัวแปรคลาสที่มีคะแนนโหวตสูงสุด และนำค่าตัวแปรคลาสนั้นมาเป็นคำตอบของปัญหานั้นๆ ข้อดีของ K-Nearest Neighbor คือเป็นเทคนิคที่เรียบง่าย สามารถใช้ในการแก้ปัญหาที่มีความซับซ้อนได้ และมีประสิทธิภาพสูง

งานวิจัยของ (Jain, Tiwari et al. 2019) กล่าวถึง Logistic Regression เป็นเทคนิคในการวิเคราะห์สถิติเชิงคุณภาพ เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสความน่าจะเป็นที่จะเกิดเหตุการณ์หนึ่งๆที่สนใจ หรือไม่เกิดเหตุการณ์หนึ่งๆที่สนใจ โดยอาศัยสมการ Logistic ที่สร้างขึ้นจากชุดตัวแปรทำนาย โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ Logistic Regression มีเป้าหมายในการประมาณค่าของสัมประสิทธิ์ของพารามิเตอร์โดยการใช้ sigmoid function เมื่อธุรกรรมดำเนินไปค่าของ attributes จะทำการตรวจสอบและบอกว่าธุรกรรมดังกล่าวควรดำเนินการต่อไปหรือไม่ ถ้าเป็นธุรกรรมที่เป็นธุรกรรมที่ถูกกฎหมายทั่วไปธุรกรรมจะยังคงดำเนินต่อไป แต่ถ้าเป็นธุรกรรมที่เป็นการฉ้อโกงก็จะหยุดการดำเนินการ

งานวิจัยของ (Wang, Deng et al. 2020) กล่าวถึง XGBoost เป็นอัลกอริทึมที่ถูกพัฒนาขึ้นมาจาก Gradient Tree Boosting ซึ่งสามารถจัดการงานที่มีข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ XGBoost เป็นอัลกอริทึมที่มีประสิทธิภาพ และใช้เวลาในการพัฒนาแบบจำลองที่ไม่นาน ซึ่งมีการประยุกต์ใช้งานในด้านการวิจัยที่หลากหลายตั้งแต่การวินิจฉัยโรคมะเร็งไปจนถึงการประเมินความเสี่ยงด้านการใช้งานบัตรเครดิต ซึ่งในปัจจุบัน XGBoost ได้กลายเป็นวิธีการทางเลือกแรกสำหรับใช้ในการพัฒนา

แบบจำลองของข้อมูลขนาดใหญ่ และยังเป็นวิธีการในการพัฒนาแบบจำลองที่ได้รับความนิยมมากที่สุด ถึงแม้ว่า XGBoost ประสบความสำเร็จอย่างมาก แต่ประสิทธิภาพของมันมักจะลดลง เมื่อชุดข้อมูลที่ถูกนำมาใช้ในการวิเคราะห์มีปัญหาความไม่สมดุลกันของชุดข้อมูล (Imbalance Data) แต่มีหลายงานวิจัยที่บอกว่า XGBoost สามารถใช้ในการจัดการกับปัญหาความไม่สมดุลกันของชุดข้อมูลได้ดี ซึ่งสามารถทำงานได้อย่างมีประสิทธิภาพเหนือกว่าวิธีการอื่นๆ ในการจัดการกับปัญหาความไม่สมดุลกันของชุดข้อมูล ซึ่งในงานวิจัยนี้ได้มีการแนะนำ imbalance-XGBoost ซึ่งเป็นแพ็คเกจ Python ที่ใช้ XGBoost ในการแก้ไขปัญหาความไม่สมดุลกันของชุดข้อมูล

วิธีดำเนินการ



ภาพประกอบ 1 Flow Chart วิธีดำเนินการพัฒนาแบบจำลอง

ในภาพประกอบที่ 1 จะแสดงขั้นตอนวิธีการดำเนินการพัฒนาแบบจำลอง ประกอบไปด้วย 5 ขั้นตอนหลักๆ คือ การเก็บรวบรวมข้อมูล การสำรวจข้อมูล การเตรียมข้อมูล การพัฒนาแบบจำลอง และการประเมินประสิทธิภาพของแบบจำลอง

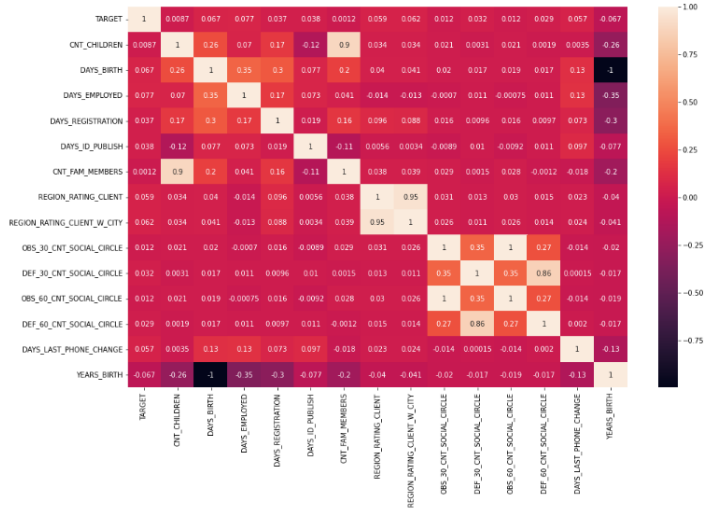
การเก็บรวบรวมข้อมูล

ข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิต ประกอบด้วย ข้อมูลจำนวนทั้งหมด 307,511 แถว และประกอบด้วยคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะ Kaggle.com จากเว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มใหญ่ๆ กลุ่มลูกหนี้ปกติ คือกลุ่มลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และกลุ่มลูกหนี้ที่ไม่ปกติ คือกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร

การสำรวจข้อมูล Exploratory Data Analysis (EDA)

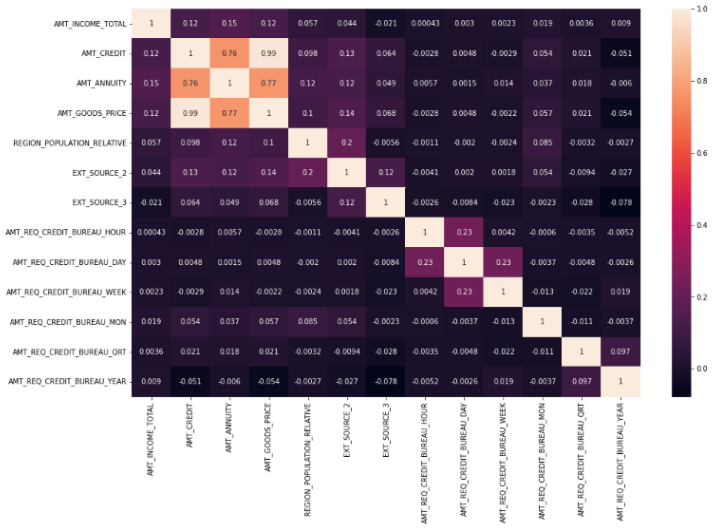
โดยการวิเคราะห์ระดับความสัมพันธ์ของตัวแปร ค้นหาว่าตัวแปรแต่ละตัวมีความสัมพันธ์กันมากน้อยเพียงใด วัตถุประสงค์เพื่อลดจำนวนตัวแปรที่ใช้ในการพัฒนาแบบจำลอง เพื่อความรวดเร็วในการพัฒนาแบบจำลอง และเพื่อเพิ่มประสิทธิภาพในการเรียนรู้ข้อมูลของแบบจำลอง และยังทำให้แบบจำลองที่ได้มีประสิทธิภาพในการทำนายที่มีความแม่นยำมากยิ่งขึ้น โดยการทำการสำรวจความสัมพันธ์ของแต่ละคอลัมน์ โดยการใช้เทคนิคที่เรียกว่า correlation ในการแสดงค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ โดยค่า Correlation จะมีค่าอยู่ระหว่าง -1 ถึง 1

เนื่องจากชุดข้อมูลที่นำมาใช้ในการพิจารณามีจำนวนคอลัมน์ที่มากมหาศาล หากนำคอลัมน์ทั้งหมดมาดูค่า Correlation อาจจะทำให้ดูไม่รู้เรื่อง จึงทำการแสดงค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ โดยการแบ่งแยกตามชนิดของข้อมูล



ภาพประกอบ 2 ค่าความสัมพันธ์ของข้อมูลชนิด int64

ซึ่งจากการหาค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ ที่จัดเก็บข้อมูลชนิด int64 จะเห็นได้ว่ามีบางคอลัมน์ที่มีค่าความสัมพันธ์สูงมากๆ โดย CNT_CHILDREN และ CNT_FAM_MEMBERS มีค่าความสัมพันธ์ เท่ากับ 0.9 และ REGION_RATING_CLIENT และ REGION_RATING_CLIENT_W_CITY มีค่าความสัมพันธ์ เท่ากับ 0.95



ภาพประกอบ 3 ค่าความสัมพันธ์ของข้อมูลชนิด float64

ซึ่งจากการหาค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ ที่จัดเก็บข้อมูลชนิด float64 จะเห็นได้ว่ามีบางคอลัมน์ที่มีค่าความสัมพันธ์สูงมากๆ โดย AMT_CREDIT และ AMT_GOODS_PRICE มีค่าความสัมพันธ์ เท่ากับ 0.99

การเตรียมข้อมูล (Preparing Data)

1. จากชุดข้อมูลที่เรานำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง มีข้อมูลที่ขาดหายไป (Missing Value) เป็นจำนวนมาก ซึ่งจะจัดการกับข้อมูลที่ขาดหายไปก่อนที่จะนำข้อมูลเหล่านี้ไปพัฒนาเป็นแบบจำลอง เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น โดยจะจัดการกับข้อมูลที่ขาดหายไป ทั้งในแนวระดับแถวและในแนวระดับคอลัมน์ โดยจะกำจัดคอลัมน์ที่มีข้อมูลที่ขาดหายไป ที่มากกว่า 45% ซึ่งเราจะไม่นำคอลัมน์เหล่านั้น มาใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลอง จากในตอนแรกที่มีคอลัมน์ที่นำมาใช้ในการวิเคราะห์ข้อมูลทั้งหมด 122 คอลัมน์ แล้วทำการลบคอลัมน์ที่มีจำนวนข้อมูลที่ขาดหายไป ที่มากกว่า 45% จะทำให้เหลือคอลัมน์ที่นำมาใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลองทั้งหมด 73 คอลัมน์
2. ลบคอลัมน์ที่เป็น “FLAG_DOCUMENT” ออกทั้งหมด เนื่องจากเป็นคอลัมน์ที่ใช้ในการเก็บข้อมูลที่เกี่ยวข้องกับการเก็บเอกสาร ซึ่งไม่น่าจะมีประโยชน์ที่จะนำคอลัมน์เหล่านี้ไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลอง โดยในขั้นตอนนี้ ได้มีการลบคอลัมน์ที่เกี่ยวข้องกับการเก็บเอกสาร จะทำให้เหลือคอลัมน์ที่จะนำไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลองทั้งหมด 52 คอลัมน์
3. ลบข้อมูลบางแถวในชุดข้อมูลออก เพราะข้อมูลในแถวนั้นส่วนใหญ่ จะเก็บข้อมูลที่เป็นค่าว่าง (NaN) เนื่องจากมีข้อมูลจำนวน 307,511 แถว จึงสามารถลบข้อมูลส่วนน้อย ที่เก็บข้อมูลที่เป็นค่าว่างออกได้ แล้วเมื่อทำการทำความสะอาดข้อมูล (Data cleansing) เรียบร้อยแล้ว จะพบว่าข้อมูลที่ถูกจัดเก็บอยู่ในคอลัมน์ทั้งหมด ไม่มีคอลัมน์ไหนที่มีการจัดเก็บข้อมูลที่เป็นค่าว่างอีกแล้ว

4. ลบข้อมูลบางแถวที่เก็บข้อมูลเพศเป็นค่า “XNA” ออก เนื่องจากเมื่อทำการนับจำนวนข้อมูลในคอลัมน์เพศแล้ว พบว่ามีข้อมูลเพศทั้งหมด 3 ค่า คือ M, F, XNA ซึ่งข้อมูลจริงๆที่ถูกต้องที่ควรจัดเก็บ ควรมีแค่ 2 ค่า คือ เพศชายและเพศหญิง (M, F)

5. ลบบางคอลัมน์ที่เก็บข้อมูลส่วนใหญ่ที่เป็นค่า “1” ออก เพราะคอลัมน์เหล่านี้ ไม่ค่อยมีความแตกต่าง เมื่อนำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง คอลัมน์เหล่านี้จะไม่มีผลต่อการทำนาย ที่จะสามารถนำไปใช้ในการแบ่งแยกคลาสได้

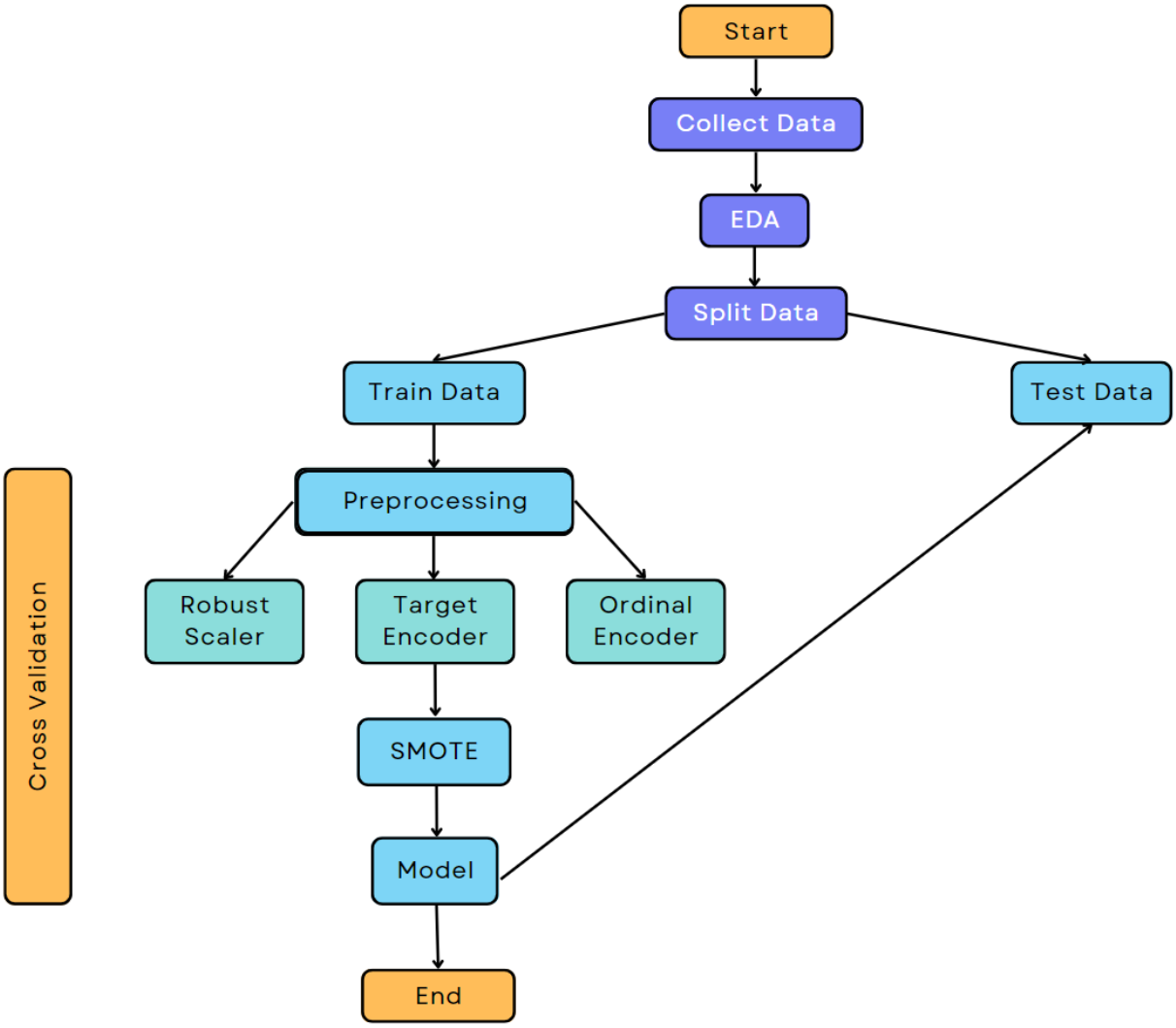
6. เพิ่มคอลัมน์ที่เก็บอายุของลูกหนี้ โดยการแปลงข้อมูลจากข้อมูลในคอลัมน์ DAYS_BIRTH ที่เก็บข้อมูลวันเกิดของลูกหนี้ แต่เปลี่ยนจากการเก็บในลักษณะของวันให้กลายเป็นปี เพื่อนำไปเป็นปัจจัยที่ช่วยในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร

7. ทำการแปลงชนิด (Type) ของข้อมูล เนื่องจากเมื่อเราทำการดูชนิดของข้อมูลแล้ว ยังมีบางคอลัมน์ที่เก็บชนิดของข้อมูลไม่ตรงกับลักษณะของข้อมูลที่ถูกจัดเก็บอยู่จริงๆ จึงได้มีการแปลงชนิดของข้อมูล ให้มีชนิดของข้อมูลที่ถูกต้อง ก่อนนำไปพัฒนาแบบจำลอง เพื่อทำให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น

8. ลบบางคอลัมน์ที่มีค่าความสัมพันธ์สูงๆออก และใช้เพียงคอลัมน์เดียวในการเรียนรู้ข้อมูลเพื่อพัฒนาแบบจำลอง เพื่อลดความซับซ้อนและลดเวลาในการเรียนรู้ข้อมูลเพื่อพัฒนาแบบจำลอง

9. ตัดบางคอลัมน์ออก โดยการใช้เทคนิคที่เรียกว่า Feature Selection เนื่องจากมีจำนวนคอลัมน์ที่เยอะมาก หากนำคอลัมน์ทั้งหมดเหล่านี้ ไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลอง จะทำให้แบบจำลองที่ได้มีความซับซ้อนมาก ซึ่งจะทำให้แบบจำลองมีประสิทธิภาพที่ลดน้อยลง และที่สำคัญการใช้ Feature Selection ยังสามารถช่วยลดการเกิด Overfitting ของแบบจำลองได้ เพราะมีจำนวน Dimension ที่มากเกินไป โดยการใช้เทคนิคที่เรียกว่า “SelectKBest” ในการคัดเลือกคุณลักษณะที่สำคัญ โดยได้มีการกำหนดค่า K เท่ากับ 25 ซึ่งค่า K คือค่าคุณลักษณะที่ต้องการ ว่าต้องการใช้ทั้งหมดกี่คุณลักษณะที่จะนำไปใช้ในการพัฒนาแบบจำลอง

การพัฒนาแบบจำลอง (Model)



ภาพประกอบ 4 กระบวนการของการพัฒนาแบบจำลอง

อธิบายถึงกระบวนการของการพัฒนาแบบจำลอง โดยมีการเก็บข้อมูลจากแหล่งข้อมูลสาธารณะ และทำการสำรวจข้อมูล เช่น ดูค่าทางสถิติของข้อมูล ดูจำนวนของข้อมูลที่ถูกเก็บอยู่ในแต่ละคอลัมน์ และดูความสัมพันธ์ระหว่างข้อมูลที่ถูกเก็บอยู่ในแต่ละคอลัมน์ แล้วทำการแบ่งข้อมูลออกเป็น Train 80% และ Test 20% และทำการเตรียมข้อมูล (Preprocessing) โดยมีการแปลงข้อมูลให้อยู่ใน 3 รูปแบบ คือ Robust Scaler Target Encoder และ Ordinal Encoder แล้วนำไปปรับความไม่สมดุลของข้อมูลด้วยเทคนิคต่างๆ เช่น SMOTE หลังจากนั้นนำข้อมูล Train ที่ได้ไปทำการ Fit เพื่อพัฒนาแบบจำลอง โดยทำการตรวจสอบกับข้อมูล Train ในทุกๆส่วน โดยการทำ Cross validation เพื่อปรับปรุงประสิทธิภาพของแบบจำลองให้มีความถูกต้องแม่นยำมากยิ่งขึ้น หลังจากนั้นนำข้อมูลที่แบ่งแยกไว้ในส่วนของ Test ไปทำการทดสอบกับแบบจำลองที่ถูกสร้างขึ้น โดยทำการ Predict กับแบบจำลอง แล้วนำผลลัพธ์ที่ได้มาทำการประเมินประสิทธิภาพของแบบจำลอง (Evaluation model)

การพัฒนาแบบจำลอง ประกอบด้วยขั้นตอนหลักๆ ดังต่อไปนี้

1. การแบ่ง training data และ test data (Data splitting) โดยจะมีการแบ่งชุดข้อมูลออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลที่ใช้ในการฝึกฝน (Train Data) เพื่อใช้ในการพัฒนาแบบจำลอง และชุดข้อมูลที่ใช้ในการทดสอบ (Test Data) เพื่อใช้ในการทดสอบแบบจำลอง โดยจะแบ่งข้อมูลออกเป็น 2 ส่วน ด้วยอัตราส่วนละ 70 : 30 ซึ่งจะได้ข้อมูลเท่ากับ 23,482 : 10,064
2. การทำ Column Transformer ประกอบไปด้วย Robust Scaler Target Encoder และ Ordinal Encoder
3. การปรับความไม่สมดุลของชุดข้อมูล (Imbalance Data) ด้วยเทคนิค Oversampling, Under sampling และ Synthetic Minority Oversampling Technique (SMOTE)
4. การพัฒนาแบบจำลอง มีการใช้ Cross Validation ซึ่งเป็นเทคนิคที่ใช้ในการหาค่า Hyperparameter ด้วยการลองใช้พารามิเตอร์ที่ได้มีการกำหนดไว้ พารามิเตอร์แต่ละตัวจะถูกนำมาใช้ในการพัฒนาแบบจำลอง และประเมินประสิทธิภาพหรือหาค่าความแม่นยำของแต่ละแบบจำลอง แบบจำลองไหนที่ให้ค่าความแม่นยำสูงสุดจะถือว่าดีที่สุด โดยการใช้เทคนิคที่เรียกว่า “GridSearchCV” เพื่อนำมาใช้ในการ Tunning Hyperparameter เพื่อหา Hyperparameter ที่ดีที่สุด เพื่อนำไปพัฒนาเป็นแบบจำลองที่ดีที่สุด

   ซึ่งในงานวิจัยนี้ อัลกอริทึมที่จะถูกนำมาใช้ในการพัฒนาแบบจำลอง ได้แก่

   4.1 Logistic Regression

   เทคนิคการวิเคราะห์การถดถอย มาใช้ในการพัฒนาแบบจำลอง ซึ่งเป็นเทคนิคการวิเคราะห์ตัวแปรเพื่อประมาณค่าหรือทำนายเหตุการณ์ที่สนใจว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร หรือไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร

   4.2 XGBoostClassifier

   เทคนิคต้นไม้ตัดสินใจหลายๆต้นมาช่วยกันในการตัดสินใจ มาช่วยกันในการทำนายเพื่อทำให้ได้ผลลัพธ์ที่ดียิ่งขึ้น ซึ่งการทำงานของแบบจำลอง XGBoost คือการสร้างต้นไม้ตัดสินใจหลายๆต้น โดยที่ต้นไม้ตัดสินใจแต่ละต้นจะถูกสร้างขึ้นมาจากการปรับปรุงประสิทธิภาพของแบบจำลองที่ถูกสร้างขึ้นก่อนหน้า แล้วจะพยายามแก้ไขความผิดพลาด (error) ของแบบจำลองที่ถูกสร้างขึ้นก่อนหน้า ให้แบบจำลองที่ถูกสร้างขึ้นในครั้งถัดๆไป มีความถูกต้องแม่นยำในการทำนายมากยิ่งขึ้นเรื่อย ๆ เมื่อมีการเรียนรู้ของต้นไม้ตัดสินใจต่อเนื่องกันจนมีความลึกมากพอ แบบจำลองจะหยุดการเรียนรู้ก็ต่อเมื่อไม่เหลือค่าความผิดพลาดจากต้นไม้ตัดสินใจก่อนหน้าให้เรียนรู้แล้ว

   4.3 K-nearest Neighbors

   เทคนิคเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุด โดยวิธีการทำงานของเทคนิคเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุด คือการค้นหาเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุด แล้วแบ่งกลุ่มข้อมูลและทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว ซึ่งค่า K คือค่าที่แบบจำลองนำมาใช้ในการพิจารณา ว่าต้องการดูเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุดจำนวนกี่จุดข้อมูล แล้วผลลัพธ์สุดท้ายของการทำนาย จะนำผลลัพธ์ทุกค่าที่ได้จากการทำนายมาหาผลลัพธ์โดยการทำ Majority vote หรือการใช้เสียงข้างมาก ซึ่งนั้นจะเป็นคำตอบสุดท้ายหรือผลลัพธ์ที่ได้จากการทำนาย

   4.4 Random Forest

เทคนิคต้นไม้ตัดสินใจหลายๆต้นมาช่วยกันในการตัดสินใจ มาช่วยกันในการทำนายเพื่อทำให้ได้ผลลัพธ์ที่ดียิ่งขึ้น ซึ่งการทำงานของแบบจำลอง Random Forest คือการสร้างต้นไม้ตัดสินใจหลายๆต้น โดยที่ต้นไม้ตัดสินใจแต่ละต้น จะถูกสร้างขึ้นมาจาก 2 วิธี คือ การทำ Bagging และ random feature projection

4.5 Support Vector Classifier (SVC)

เทคนิค Support Vector Machine (SVC) ซึ่งเป็นเทคนิคที่มีความยืดหยุ่นและทำงานได้ดี โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีความซับซ้อน และมีข้อมูลหลายๆ Feature โดยการทำงานของ Support Vector Machine (SVC) คือการพยายามหาเส้นแบ่งระหว่างคลาสต่างๆ ในข้อมูล ให้เส้นแบ่งที่ได้มีความกว้างมากที่สุด และยอมให้มีข้อมูลบางจุดอยู่ระหว่างเส้นแบ่ง เพื่อไม่ให้เส้นมันเพี้ยนมากจนเกินไป โดยมันพยายามที่จะแยกข้อมูลของทั้ง 2 คลาสออกจากกัน

4.6 Gradient Boosting

เทคนิคต้นไม้ตัดสินใจหลายๆต้นมาช่วยกันในการตัดสินใจ มาช่วยกันในการทำนายเพื่อทำให้ได้ผลลัพธ์ที่ดียิ่งขึ้น ซึ่งการทำงานของแบบจำลอง Gradient Boosting คือเทคนิคการเรียนรู้ของเครื่องมือสำหรับการแก้ปัญหา โดยจะใช้เทคนิคการเพิ่มการรวมจำนวนต้นไม้ตัดสินใจที่มีความแม่นยำต่ำ เพื่อสร้างเป็นต้นไม้ตัดสินใจต้นใหม่ โดยต้นไม้ตัดสินใจต้นใหม่จะถูกสร้างขึ้นจากข้อผิดพลาด (error) จากการคำนวณของต้นไม้ตัดสินใจก่อนหน้า

การประเมินประสิทธิภาพของแบบจำลอง (Evaluate)

การวัดประสิทธิภาพของการพัฒนาแบบจำลอง เพื่อประเมินประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้ Confusion matrix โดยดูที่ค่า Accuracy , Precision, Recall และ F1-Score ซึ่ง Confusion matrix จะมีลักษณะเป็นตาราง หากข้อมูลที่ต้องการจำแนกมี 2 ประเภท คือ ทายถูก (Positive) และ ทายผิด (Negative) โดยเราจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก เนื่องจากใช้เป็นตัววัดความสามารถของแบบจำลอง โดยที่ตาราง Confusion Matrix จะมีลักษณะตามตาราง

ตารางที่ 1 Confusion Matrix

	ค่าจริง (Actual)	
ค่าการทำนาย (Predict)	True positive (TP)	False positive (FP)
	False negative (FN)	True negative (TN)

จากรูป เมื่อมีการทำนาย 2 ประเภท ผลลัพธ์ของการทำนายทั้งหมดที่เป็นไปได้ จะมีทั้งหมด 4 ค่า ดังต่อไปนี้

1. True positive (TP) คือ การทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระ ถูก
2. True negative (TN) คือ การทำนายลูกหนี้ปกติที่ไม่ได้มีการผิดนัดชำระ ถูก

3. False positive (FP) คือ การทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระ ผิด
4. False negative (FN) คือ การทำนายลูกหนี้ปกติที่ไม่ได้มีการผิดนัดชำระ ผิด

ซึ่งสามารถนำค่าเหล่านี้ มาคำนวณหาประสิทธิภาพได้ ดังต่อไปนี้

1. ค่าความไว (Recall) คือ ค่าความถูกต้องของการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าจริง เทียบกับจำนวนครั้งของเหตุการณ์ ทั้งการทำนายและการเกิดขึ้นจริง ว่าเป็นจริง
   สูตรในการคำนวณ คือ TP / (TP+FN)
2. ค่าความถูกต้อง (Accuracy) คือ ค่าความถูกต้องและความแม่นยำของแบบจำลอง
   สูตรในการคำนวณ คือ (TP+TN) / (TP+FP+TN+FN)
3. ค่าความแม่นยำ (Precision) คือ การเปรียบเทียบการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าจริง แล้วเกิดขึ้นจริง (TP) เทียบกับการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าจริง แต่สิ่งที่เกิดขึ้นไม่จริง (FP)
   สูตรในการคำนวณ คือ TP / (TP+FP)
4. F1-Score คือ ค่าเฉลี่ยระหว่างค่า precision และ recall เพื่อใช้ในการวัดความสามารถของแบบจำลอง
   สูตรในการคำนวณ คือ 2 x (Precision x Recall) / (Precision + Recall)

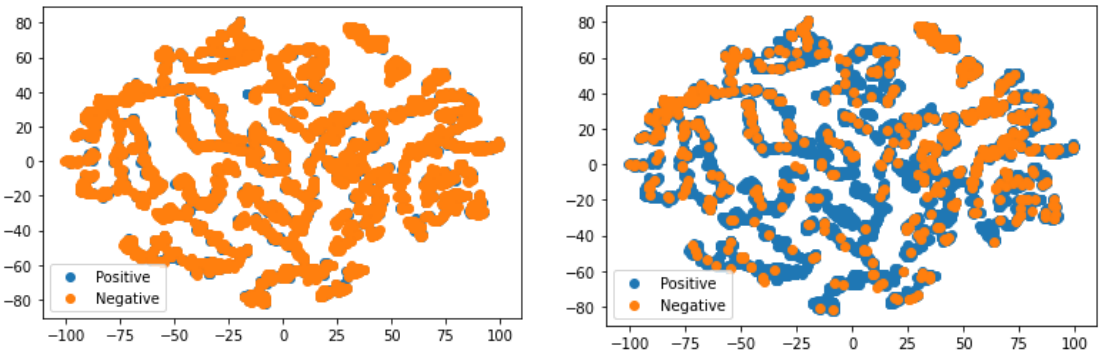
ผลการวิจัยและอภิปรายผลการวิจัย

การวิจัยนี้ เป็นงานวิจัยในการศึกษาการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยอาศัยการเรียนรู้ของเครื่องมือที่ช่วยในการตัดสินใจ ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามขบวนการและขั้นตอนต่างๆ จนกระทั่งประเมินประสิทธิภาพของแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร

การทำ Error Analysis วิเคราะห์ความผิดพลาดของการพัฒนาแบบจำลอง

จากผลลัพธ์ของการพัฒนาแบบจำลอง โดยการใช้วิธีการเทคนิคอัลกอริทึมต่างๆ จะเห็นได้ว่าคุณลักษณะของข้อมูล (Feature) ต่างๆ มีคุณลักษณะของข้อมูลที่ไม่เพียงพอในการแบ่งแยกความแตกต่างระหว่างแต่ละคลาสได้ ซึ่งค่าของคุณลักษณะข้อมูลเหล่านั้นไม่สัมพันธ์กับข้อมูลตัวแปรเป้าหมาย (Target class) และค่าของคุณลักษณะข้อมูลเหล่านั้นไม่แตกต่างกันมากนักระหว่าง Positive class และ Negative class ซึ่งค่าของคุณลักษณะข้อมูลไม่สามารถตรวจจับความแตกต่างระหว่าง Positive class และ Negative class ได้ ทำให้ค่าคุณลักษณะของข้อมูลเหล่านั้นไม่สามารถใช้ในการทำนายผลลัพธ์ได้อย่างมีประสิทธิภาพ โดยจะทำการวิเคราะห์ Error Analysis หรือข้อผิดพลาดของการพัฒนาแบบจำลอง โดยการใช้เทคนิค T-distributed Stochastic Neighbor Embedding (t-SNE) ซึ่งเป็นเทคนิคที่ใช้สำหรับแสดงผลข้อมูลให้อยู่ในรูปแบบสองมิติหรือสามมิติ โดยมีวัตถุประสงค์

เพื่อช่วยให้เราเข้าใจและวิเคราะห์ข้อมูลที่มีมิติสูงๆ ทำให้อยู่ในรูปแบบของมิติต่ำ ๆ ทำให้เราเข้าใจข้อมูลได้ดีมากยิ่งขึ้น โดยการใช้เทคนิคนี้จะช่วยให้เราเห็นภาพรวมของข้อมูลได้ง่ายมากยิ่งขึ้น และมีประสิทธิภาพในการวิเคราะห์ข้อมูลที่มีมิติสูงได้ดีมากยิ่งขึ้น



ภาพประกอบ 5 ภาพสองมิติจากการทำนายคลาสตัวแปรเป้าหมายระหว่าง Positive class และ Negative class

ในภาพประกอบ 5 เมื่อทำการแสดงผลของข้อมูลของเราให้อยู่ในรูปแบบของสองมิติ จะเห็นได้ว่าข้อมูลของเรา ไม่ได้มีการกระจายตัวของข้อมูลที่ดี ข้อมูลตัวแปรเป้าหมายของ Positive class ปะปนอยู่กับข้อมูลตัวแปรเป้าหมายที่เป็น Negative class ทำให้เราไม่สามารถแบ่งแยก Positive class และ Negative class ออกจากกันได้อย่างชัดเจน เมื่อเรานำข้อมูลเหล่านั้นมาใช้ในการพัฒนาแบบจำลอง ทำให้แบบจำลองที่ได้จึงมีประสิทธิภาพที่ไม่ค่อยดี และไม่เพียงพอต่อการนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้

ภาพประกอบ 6 ภาพสองมิติจากการทำนายคลาสตัวแปรเป้าหมายระหว่าง FP, TP, FN และ TN

ในภาพประกอบ 6 จากข้อมูลที่เรานำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง แล้วนำแบบจำลองที่ได้ ไปใช้ในการทำนาย ซึ่งจากผลลัพธ์ที่ได้จะเห็นได้ว่าค่าที่ได้จากการทำนาย Positive Class จะเห็นได้ว่าคลาสที่เป็น True Positive และ False Positive อยู่ในตำแหน่งเดียวกัน ปะปนกันอยู่ ไม่สามารถแบ่งแยกคลาสหรือแยกความแตกต่างออกจากกันได้อย่างชัดเจน และค่าที่

ได้จากการทำนาย Negative Class จะเห็นได้ว่าคลาสที่เป็น True Negative และ False Negative อยู่ในตำแหน่งเดียวกัน ปะปนกันอยู่ ไม่สามารถแบ่งแยกคลาสหรือแยกความแตกต่างออกจากกันได้อย่างชัดเจนเช่นเดียวกัน

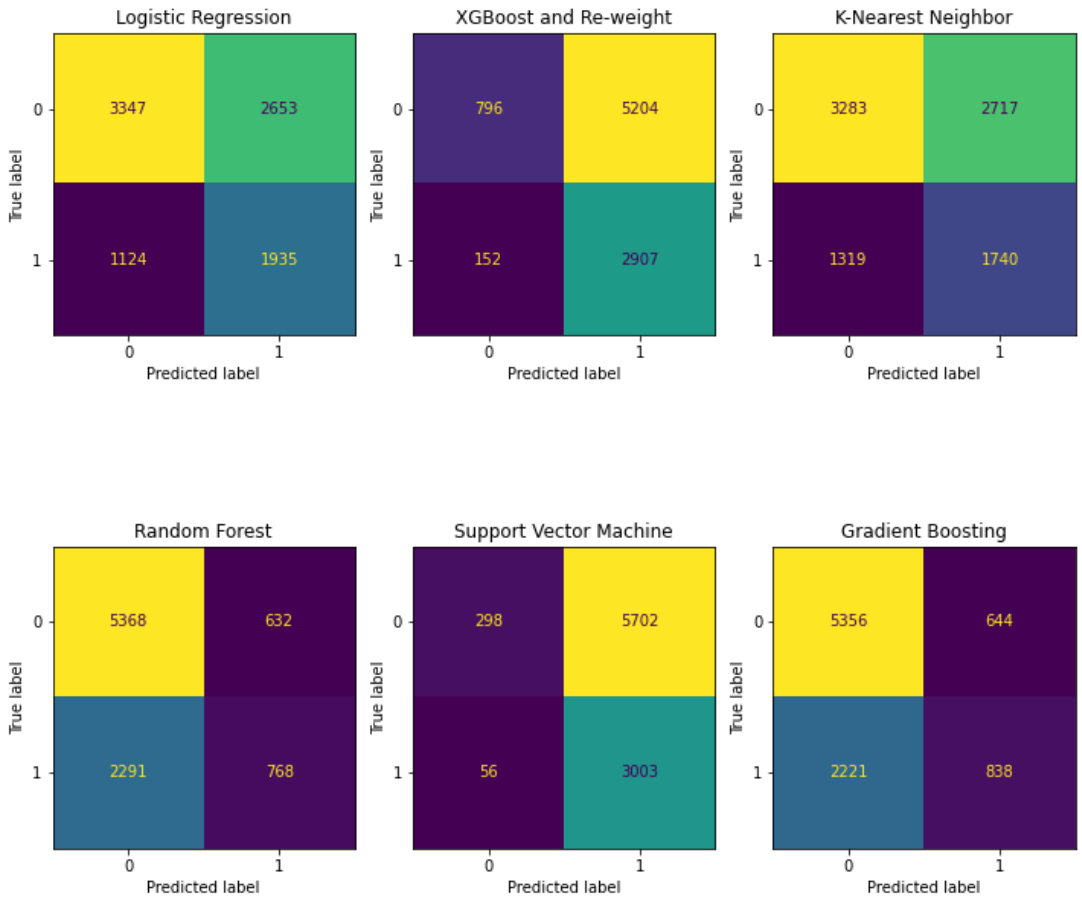
สรุปผลการวิจัย

การเปรียบเทียบผลของการพัฒนาแบบจำลอง

ในการเปรียบเทียบค่า Accuracy, Precision, Recall และ F1-Score ระหว่างแบบจำลอง Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting จะได้ค่าดังตารางที่ 2

ตารางที่ 2 แสดงประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้อัลกอริทึม Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting

Matrix	Logistic Regression	XGBoost	K-Nearest Neighbor	Random Forest	Support Vector Machine	Gradient Boosting
Accuracy	0.58	0.40	0.55	0.67	0.36	0.68
Precision	0.42	0.35	0.39	0.54	0.34	0.56
Recall	0.63	0.95	0.56	0.25	0.98	0.27
F1-Score	0.50	0.52	0.46	0.34	0.51	0.36



ภาพประกอบ 7 Confusion Matrix ของการพัฒนาแบบจำลอง โดยการใช้อัลกอริทึม Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting

จากผลการศึกษาการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการใช้เทคนิคต่าง ๆ เมื่อเปรียบเทียบประสิทธิภาพจากการพัฒนาแบบจำลองของหลายๆอัลกอริทึม จะพบว่าเทคนิควิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling เมื่อนำมาใช้ในการปรับความไม่สมดุลของการพัฒนาแบบจำลองจะทำให้การพัฒนาแบบจำลองที่ได้มีประสิทธิภาพที่ดีที่สุด และจะเห็นได้ว่าการพัฒนาแบบจำลองโดยการใช้เทคนิควิธี XGBoost ให้ค่าความไวที่มากที่สุด ซึ่งมีค่าเท่ากับ 0.97 มีค่าความถูกต้องเท่ากับ 0.37 และมีค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองเท่ากับ 0.51 แต่เทคนิควิธี K-Nearest Neighbors (KNN) ให้ค่าความไวที่น้อยที่สุด ซึ่งมีค่าเท่ากับ 0.57 มีค่าความถูกต้องเท่ากับ 0.55 และมีค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองเท่ากับ 0.46 ซึ่งมีค่าที่น้อยที่สุด

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] Awoyemi, J. O., et al. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 international conference on computing networking and informatics (ICCNI), IEEE.

[2] Jain, Y., et al. (2019). "A comparative analysis of various credit card fraud detection techniques." Int J Recent Technol Eng 7(5S2): 402-407.

[3] Kiran, S., et al. (2018). "Credit card fraud detection using Naïve Bayes model based and KNN classifier." International Journal of Advance Research, Ideas and Innovations in Technoloy 4(3): 44.

[4] Varmedja, D., et al. (2019). Credit card fraud detection-machine learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), IEEE.

[5] Wang, C., et al. (2020). "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost." Pattern Recognition Letters 136: 190-197.

[6] Xuan, S., et al. (2018). Random forest for credit card fraud detection. 2018 IEEE 15th international conference on networking, sensing and control (ICNSC), IEEE.