การเรียนรู้ของเครื่องเพื่อการทำนายการผิดนัดชำระของลูกหนี้บัตรเครดิต

สกุลกาญจน์ ทองคำ[1*], นุวีย์ วิวัฒนวัฒนา[2]

บทคัดย่อ

งานวิจัยนี้มุ่งศึกษาการทำนายลูกหนี้บัตรเครดิตที่มีโอกาสผิดนัดชำระ โดยใช้การเรียนรู้ของเครื่อง (Machine Learning) เป็นเครื่องมือสร้างแบบจำลองการแบ่งกลุ่มลูกค้าแบบมีผู้สอน (Supervised Learning) ประเภท Classification ด้วยการทดสอบกับข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต โดยมีข้อมูลรายการธุรกรรม จำนวน 1,048,575 แถว และข้อมูลลูกค้า จำนวน 438,557 แถว จากเว็บไซต์ Kaggle.com

ผู้วิจัยสร้างแบบจำลองเพื่อแบ่งกลุ่มลูกค้าที่มีความสามารถในการชำระหนี้บัตรเครดิต เป็น 2 ประเภท ได้แก่ กลุ่มลูกค้าปกติและกลุ่มลูกค้าผิดนัดชำระ ด้วยการใช้เทคนิค Machine Learning แบบ Supervised Learning ประเภท Classification ประกอบด้วยอัลกอลิทึม 3 วิธี ได้แก่ 1.) Logistic Regression 2.) Random Forest และ 3.) Catboost เพื่อหาแบบจำลองที่มีประสิทธิภาพมากที่สุดในการจัดกลุ่มลูกหนี้

ผลการศึกษาพบว่า วิธีทำนายแบบ XGBoost ให้ค่าความถูกต้อง 98% ที่จำนวนต้นไม้ 15 ต้น กับอัตราการเรียนรู้ที่ 0.1 วิธีทำนายแบบ Catboost ให้ค่าความถูกต้อง 97% ที่จำนวนต้นไม้ 7 ต้น กับอัตราการเรียนรู้ที่ 0.1 และวิธีทำนายแบบ Logistic Regression ให้ค่าความถูกต้อง 62% เมื่อเปรียบเทียบค่า Confusion Matrix พบว่าแบบจำลอง Random Forest และ Catboost ให้ผลลัพธ์สูงสุดใกล้เคียงกัน

คำสำคัญ : Machine Learning, Credit Card Approval, Prediction Algorithm, Non-Performing Loan

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 080-6134437 E-mail address: sakulkran.thk@g.swu.ac.th

MACHINE LEARNING MODELS FOR CREDIT CARD DEFAULT PREDICTION

Sakulkran Thongkham[1*], Nuwee Wiwatwattana[2]

Abstract

This thesis aimed to study the predictive analytic among credit card holders who could create non-performing loan; by using the Machine Learning to set up the Supervised Learning in the Classification character. The Learning Machine was tested to credit card loan transaction data with 1,048,575 rows of transaction list and 438,557 rows of credit card customer's data selected from Kaggle.com.

The process functioned by designing the model to divide credit card customers into 2 groups : normal customers and non-performing loan customers with the aid of Machine Learning in type of Classification Supervised Learning. This Machine Learning carried 3 algorithms : 1.) Logistic Regression 2.) XGBoost and 3.) Catboost, to explore the most effective model to analyze the credit card customers.

The study depicted that XGBoost algorithm provided 98% of accuracy at 15 Depth with 0.1 degree of learning rate, Catboost algorithm provided 97% of accuracy with 7 Depth and 0.1 degree of learning rate, and Logistic Regression algorithm provided 62% of accuracy. The output from Confusion Matrix table pointed that XGBoost algorithm and Catboost algorithm maintain the most effective outcome in close proximity. .

Keywords : Machine Learning, Credit Card Approval, Prediction Algorithm, Non-Performing Loan

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

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

* Corresponding author:Tel.: 080-6134437 E-mail address: sakulkran.thk@g.swu.ac.th

I บทนำ

ธุรกิจบัตรเครดิตเป็นการให้บริการของสถาบันการเงินต่างๆ ซึ่งได้รับความนิยมกันอย่างแพร่หลายในปัจจุบัน เนื่องจากมีความสะดวกสำหรับใช้จ่ายอีกทั้งมีความปลอดภัยในการพกพาเงินสดเป็นจำนวนมาก โดยสามารถใช้ในการชำระค่าสินค้า และค่าบริการหรือหนี้แทนการชำระด้วยเงินสด รวมทั้งผู้ใช้บัตรเครดิตสามารถเบิกถอนเงินสดได้โดยไม่จำเป็นต้องมีเงินสด การเปลี่ยนแปลงด้านพฤติกรรมการใช้จ่ายนี้สะท้อนให้เห็นว่าบัตรเครดิตเข้ามามีบทบาทตามยุคสมัยสังคมไร้เงินสด

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

งานวิจัยเน้นการศึกษาศึกษาการทำนายลูกหนี้ที่มีโอกาสผิดนัดชำระ โดยใช้การเรียนรู้ของเครื่อง (Machine Learning) เป็นเครื่องมือสำหรับสร้างแบบจำลองในการแบ่งกลุ่มลูกค้าแบบมีผู้สอน (Supervised Learning) ประเภท Classification ทดลองกับข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต ประกอบด้วย 2 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 1,048,575 แถวและข้อมูลลูกค้าประกอบด้วย 17 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 438,557 แถว

II งานวิจัยที่เกี่ยวข้อง

งานวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารงานวิจัยที่เกี่ยวข้องกับการพัฒนาแบบจำลอง ซึ่งยังไม่พบงานวิจัยที่ทำนายลูกหนี้ผิดนัดชำระ โดยใช้เทคนิคทำนายแบบ Catboost แต่เป็นงานที่ใช้เทคนิคการเรียนรู้ของเครื่องในการจำแนกประเภทต้นไม้เช่น XGBoost , Random Forests เป็นต้น

โดยบทความวิจัยที่ผู้วิจัยศึกษาได้ใช้เทคนิค สร้างแบบจำลองในการแบ่งกลุ่มลูกค้าแบบมีผู้สอน (Supervised Learning) ประเภท Classification ซึ่งมีตัวอย่างดังต่อไปนี้ Yue Yu [1] ได้กล่าวถึงการทำนายลูกหนี้ที่มีโอกาสเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ ซึ่งงานวิจัยนี้ใช้ข้อมูลของผู้ใช้บัตรเครดิตจากไต้หวัน ช่วงเดือนเมษายนถึงกันยายน 2005 โดยนำข้อมูลทั้งหมดมาวิเคราะห์ข้อมูลซึ่งพิจารณาจาก Customer Value หลังนำข้อมูลมาวิเคราะห์พบว่าข้อมูลที่ใช้ระบุกลุ่มเป้าหมายลูกค้า (Target) มีความไม่สมดุลของข้อมูลจึงได้นำเทคนิค Weighted Model มาใช้ปรับข้อมูลให้มีความเหมาะสมก่อนนำข้อมูลมาทดสอบแบบจำลอง หลังจากนั้นจึงได้นำเทคนิค Classification Algorithm มาใช้ทั้งหมดสี่วิธีได้แก่ Logistic Regression , Decision Tree, AdaBoost , Random Forest สำหรับการวัดผลประสิทธิภาพแบบจำลองได้ใช้วิธีวัดผลทั้งหมดสิงค่าได้แก่ Accuracy , Precision ซึ่งการวัดผลแบ่งเป็นสองกลุ่มชุดข้อมูลได้แก่ ข้อมูลก่อนทำการปรับความสมดุลของข้อมูล และ ข้อมูลที่ทำการปรับความสมดุลของข้อมูลด้วยเทคนิค

Weighted Model จากผลการทดสอบวัดประสิทธิภาพแบบจำลองด้วยวิธีก่อนและหลังปรับความสมดุลของข้อมูล Random Forest สามารถทำนายได้ดี Accuracy : 99.27% ทั้ง 2 แบบไม่ว่าจะเป็นข้อมูลก่อนทำการปรับความสมดุลของข้อมูล (Original Dataset) หรือข้อมูลที่ทำการปรับความสมดุลของข้อมูลด้วยเทคนิค Weighted Model

งานวิจัยของ Salma Khaled Shaheen & Essam ElFakharany [2] กล่าวถึงการทำนายลูกหนี้ที่มีโอกาสเป็นลูกหนี้ไม่ก่อให้เกิดรายได้ ซึ่งงานวิจัยนี้ใช้ข้อมูลลูกค้าธนาคารอียิปต์ ช่วงเดือนพฤษภาคม 2005 จนถึง ธันวาคม 2017 โดยมีข้อมูลทั้งหมด 2,954,168 แถว ซึ่งชุดข้อมูลดังกล่าวเป็นข้อมูลชั้นความลับจึงไม่สามารถเปิดเผยข้อมูลได้ โดยเป้าหมายของงานวิจัยนี้เพื่อจำแนกลูกหนี้ออกเป็น 2 ประเภทโดยผู้วิจัยได้นำแบบจำลองเพื่อใช้ในการทำนายผลมาใช้ทั้งหมดสี่วิธีได้แก่ K-NN , Logistic Regression , Random Forest , Gradient Boosting อัตราส่วนใช้ทดสอบแบบจำลอง 70:30 ซึ่งผลการวัดประสิทธิภาพแบบจำลองพบว่า Random Forest มีค่า Accuracy : 91.7% และ Precision : 95.83% และ Gradient Boosted มีค่า Accuracy : 91.7% และ Precision : 95.83%

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

III ชุดข้อมูล

งานวิจัยนี้ศึกษาการแบ่งกลุ่มลูกค้าที่มีคุณลักษณะและพฤติกรรมที่ใกล้เคียงกันหรือเหมือนกันออกเป็น 1. ลูกค้าปกติ 2. ลูกค้าผิดนัดชำระโดยใช้เทคนิค Machine Learning เป็นเครื่องมือสำหรับสร้างแบบจำลองในการแบ่งกลุ่มลูกค้า โดยข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต ประกอบด้วย 2 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 1,048,575 แถวและข้อมูลลูกค้า ประกอบด้วย 17 ตัวแปรและมีจำนวนข้อมูลทั้งหมด 438,557 แถว จากแหล่งข้อมูลสาธารณะ Kaggle.com

ตารางแสดงตัวแปรของข้อมูลลูกค้าที่ใช้สำหรับพัฒนาแบบจำลอง

Field Name	Data Type	คำอธิบายข้อมูล
ID	Int64	รหัสลูกค้า
CODE_GENDER	Object	เพศ (M = ชาย, F = หญิง)
FLAG_OWN_CAR	Object	Flag ระบุรายละเอียดลูกค้ามีรถหรือไม่ (Y = มี , N = ไม่มี)
FLAG_OWN_REALTY	Object	Flag ระบุรายละเอียดลูกค้ามีอสังหาริมทรัพย์หรือไม่ (Y = มี , N = ไม่มี)

CNT_CHILDREN	Int64	จำนวนบุตร
AMT_INCOME_TOTAL	Float64	รายได้ต่อปี
NAME_INCOME_TYPE	Object	ประเภทรายได้
NAME_EDUCATION_TYPE	Object	ระดับการศึกษา
NAME_FAMILY_STATUS	Object	สถานภาพการสมรส
NAME_HOUSING_TYPE	Object	ประเภทที่อยู่อาศัย
DAYS_BIRTH	Int64	วันเกิด (Date)
DAYS_EMPLOYED	Int64	วันที่เริ่มทำงาน (Date)
FLAG_MOBIL	Object	Flag ระบุรายละเอียดลูกค้ามีโทรศัพท์มือถือหรือไม่ (Y = มี , N = ไม่มี)
FLAG_WORK_PHONE	Object	Flag ระบุรายละเอียดลูกค้ามีเบอร์โทรที่ทำงานหรือไม่ (Y = มี , N = ไม่มี)
FLAG_PHONE	Object	Flag ระบุรายละเอียดลูกค้ามีเบอร์โทรศัพท์หรือไม่ (Y = มี , N = ไม่มี)
FLAG_EMAIL	Object	Flag ระบุรายละเอียดลูกค้ามี E-mail หรือไม่ (Y = มี , N = ไม่มี)
OCCUPATION_TYPE	Object	อาชีพ
CNT_FAM_MEMBERS	Float64	จำนวนสมาชิกครอบครัว

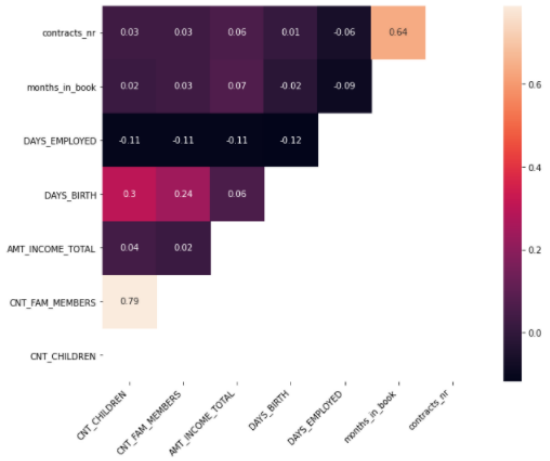
Field Name	**Data Type**	**คำอธิบายข้อมูล**
ID	Int64	รหัสลูกค้า
MONTHS_BALANCE	Int64	จำนวนเดือนสะสม
STATUS	Object	สถานะ

IV โมเดลและวิธีการ

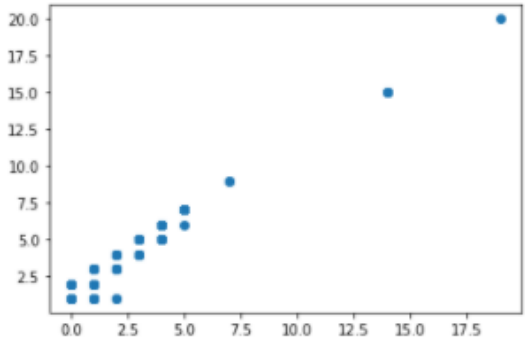
งานวิจัยนี้สร้างแบบจำลองเพื่อทำนายลูกหนี้ผิดนัดชำระ โดยอาศัยการเรียนรู้ของเครื่อง มีหลักการสร้างและเปรียบเทียบประสิทธิภาพแบบจำลองดังต่อไปนี้

A. สำรวจข้อมูล Exploratory Data Analysis (EAD)

วิเคราะห์ระดับความสัมพันธ์ของตัวแปร พบว่าตัวแปรแต่ละคอลัมน์มีความสัมพันธ์กันมากน้อยเพียงใดวัตถุประสงค์เพื่อลดจำนวนตัวแปรในการทดสอบแบบจำลอง และทำให้แบบจำลองมีประสิทธิภาพการทำนายที่ขึ้น

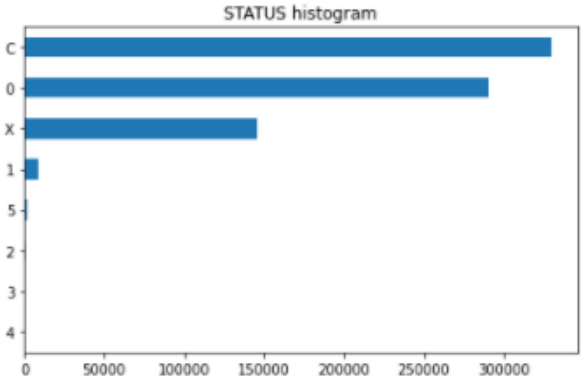


ภาพประกอบ 1 แสดงการวิเคราะห์ความสัมพันธ์ของแต่ละคอลัมน์



ภาพประกอบ 2 แสดง Correlation ระหว่าง CNT_FAM_MEMBERS และ CNT_CHILDREN

จากภาพประกอบที่ 2 ความสัมพันธ์ระหว่าง CNT_FAM_MEMBERS และ CNT_CHILDREN นั้นเป็นไปในทิศทางเดียวกัน เนื่องจากจำนวนสมาชิกครอบครัวที่มากจะทำให้จำนวนบุตรมากตามไปด้วย ดังนั้นจึงไม่นำ CNT_CHILDREN (จำนวนบุตร) มาใช้ในการพัฒนาแบบจำลอง และจะเลือกใช้ CNT_FAM_MEMBERS (จำนวนสมาชิกครอบครัว) แทน



ภาพประกอบ 3 แสดงการตรวจสอบข้อมูลการกระจายข้อมูลของตัวแปรสถานะ

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

B การเตรียมข้อมูล (Preparing Data)

- จัดการกับข้อมูลที่ขาดหายไป (Missing Value) โดยพบข้อมูลที่ขาดหายไปจำนวนมาก
- จัดการกับตารางข้อมูล ID.credit_record จัดกลุ่มลูกค้าหาก โดยหากลูกค้ามี Status[2,3,4,5] จะกำหนดลูกค้าเท่ากับ 1: ลูกค้าสงสัยจะสูญ และรายอื่นๆจะกำหนดเป็น 0 :ลูกค้าปกติ
- จากนั้นทำการ Group by และใช้ค่าสูงสุดของสถานะ (Max) เพื่อระบุลูกค้าจากการจัดกลุ่มใหม่โดยจะได้ลูกค้า 2 กลุ่ม คือ (0 ลูกค้าปกติ :1 ลูกค้าสงสัยจะสูญ)
- นำตารางมา Join กับโดยใช้ Key ระหว่าง ID.application_record : ID.credit_record จะได้ข้อมูลลูกค้าทั้งหมด 36,457 คนที่สามารถระบุสถานะลูกค้าเพื่อไว้ใช้กำหนดเป้าหมายในการทำนาย
- จัดการกับชนิดข้อมูลที่เป็นตัวอักษรให้อยู่ในรูปแบบตัวเลขด้วยวิธี Pipeline เพื่อให้เหมาะสมกับการนำไปสร้างแบบจำลอง

- แบ่งข้อมูลลสำหรับฝึกสอน (Train Data) และสำหรับทดสอบ (Test Data) ด้วยอัตราส่วน 80:20 จะได้ 29,165 : 7,292
- จัดการกับข้อมูลที่ไม่สมดุลกัน เนื่องจากปัญหาของข้อมูลที่นำมาวิเคราะห์นี้เป็นปัญหาที่ข้อมูลไม่สมดุลกัน ผู้วิจัยได้จัดการกับปัญหาดังกว่าวด้วยวิธีสังเคราะห์ข้อมูลเพิ่ม หรือ Synthetic Minoritory Oversampling Technique : SMOTE เป็นเทคนิคการสุ่มตัวอย่างแบบสุ่มเพิ่มแทนที่จะสุ่มเพิ่มจากข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิม หลักการคือการสุ่มจากข้อมูลเพื่อนบ้านที่ใกล้ที่สุด

C. การสร้างโมเดล(Modeling)

เนืองจากงานวิจัยนี้เป็นปัญหาแบบ Classification ผู้วิจัยได้เลือกแบบจำลองการทำนายทั้งหมด 3 เทคนิค เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองดังต่อไปนี้

เพื่อเปรียบเทียบแบบจำลอง ดังต่อไปนี้

- เทคนิค Logistic Regression ใช้แก้ปัญหา Binary Classification ทำนายคลาสที่มี 2 คลาสโดยมี Sigmoid Function เพื่อมา Normalize ซึ่งผลลัพธ์ที่ ได้จาก Sigmoid Function คือความน่าจะเป็นที่สามารถทำนายได้
- Extreme Gradient Boosting เป็นอัลกอริธึมที่มีโครงสร้างพื้นฐานมากจากอัลกอริธึมตระกูลต้นไม้ เช่น Decision Tree , Random Forest , Bagging , Boosting , Gradient Boosting โดยวัตถุประสงค์ของอัลกอริทึมเพื่อต้องการแก้ปัญหาความแปรปรวนของข้อมูล เนื่องจากมีข้อมูลใหม่เข้ามา หรือข้อมูลมีค่าว่าง ลักษณะการทำงานของอัลกอริทึมคือการสร้างต้นไม้ที่อ่อนแอตามลำดับ หลังจากนั้นทำการสร้างต้นไม้ใหม่ ซึ่งต้นไม้ที่ถูกสร้างใหม่จะเรียนรู้จากต้นไม้ก่อนหน้าที่มีความแม่นยำในการทำนายต่ำ แล้วนำน้ำหนักของข้อมูลมาใช้ปรับข้อมูลใหม่ เพื่อให้ได้การถ่วงน้ำหนักใหม่
- CatBoost ป็นวิธี Gradient Boosting ชนิดหนึ่ง ซึ่งถูกพัฒนาเพื่อจัดการกับข้อมูลจำพวกตัวแปรประเภทและเพิ่มประสิทธิภาพการทำงานในแบบจำลองสามารถทำนายได้รวดเร็ว โดยหลักการของอัลกอริทึมสมมุติว่ามีชุดข้อมูล (Data Set) D={(Xi,Yi)}, i=1, ..., n. และ Yi$\in$R คือชุดของเป้าหมายของการทำนาย (Label Set) ขั้นแรก CatBoost จะสุ่มเรียงลำดับข้อมูลทั้งหมด สำหรับค่าบางค่าในหมวดหมู่ของแต่ละตัวอย่าง เมื่อแปลงเป็นค่าตัวเลข ค่าเฉลี่ยจะถูกใช้ตามค่าเป้าหมายของการทำนาย ตัวอย่างก่อนหน้านี้ และน้ำหนักลำดับความสำคัญจะถูกเพิ่ม

D. การประเมินผล (Evaluate)

งานวิจัยนี้เป็นการทำนายผลลัพธ์ที่เป็น classificaion ในทางธนาคารจึงจำเป็นต้องมีการประเมินผลลัพธ์ให้แม่นยำที่สุด โดยใช้ตาราง Confusion Matrixประเมินผลของแบบจำลอง โดยมีรายละเอียดดังต่อไปนี้

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

ภาพประกอบ 4 แสดงภาพตาราง Confusion Matrix

จากภาพประกอบที่ 4 แสดงผลตัวแปรผลของการทำนายอัลกอริที่มเพื่อใช้ในการวัดผลประสิทธิโดยมีรายละเอียดดังนี้

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

True Negative (TN) True Negative (TN) คือ สิ่งที่โแบบจำลองทำนายว่า “ไม่จริง” และ มีค่า “ ไม่จริง”

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

False Negative (FN) False Negative (FN) คือ สิ่งที่แบบจำลองทำนายว่า “ไม่จริง” แต่ มีค่าเป็น “จริง”

โดยแต่ละค่าจะเก็บผลของการทำนายเพื่อนำไปคำนวนหาค่าประสิทธิของการทำนายซึ่งปัจจุบันนิยม 3 ค่าได้แก่

ค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้องของ Model โดยพิจารณารวมทุกคลาส หากข้อมูลคลาสมีความสมดุลค่าความถูกต้องจะมีประสิทธิภาพ แต่หากข้อมูลคลาสไม่มีความสมดุลค่าความถูกต้องจะมีประสิทธิภาพต่ำ แสดงได้ดังสมการที่ (3)

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

ค่าเรียกคืน (Recall) เป็นการวัดความถูกต้องของ Model โดยพิจารณาแยกทีละคลาส โดยใช้เปอร์เซ็น ของความถูกต้อง (True Positive) แสดงได้ดังสมการที่ (4)

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

ค่าความแม่นยำ (Precision) เป็นการวัดความแม่นยำของข้อมูล โดยพิจารณาแยกทีละคลาส แสดงได้ดังสมการที่ (5)

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

V การทดลองและผลลัพธ์

การทดลองงของงานวิจัยรี้ได้สร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง 3 เทคนิค ได้แก่ Logistic Regression , XGBoost , Catboost ผลลัพธ์ของการสร้างแบบจำลองได้เปรียบเทียบผลของการประเมินแบบจำลองแต่ละแบบจำลอง

ตารางที่ 1 ตารางผลการทดสอบประสิทธิภาพแบบจำลอง

Algorithm	Parameter Tuning	Best Parameter	%accuracy	%recall	%precision
Logistic Regression	Penalty : L1 , L2 C :100,10,1.0,0.1,0.01	Penalty : L1 C : 1.0	62	56	92
Gradient Boosting	max_depth : [5,10,15] learning_rate : [0.001, 0.01, 0.1]	max_depth : 15 learning_rate : 0.1	98	97	95
CatBoost	depth : [4, 5, 6,7] learning_rate : [0.001, 0.01, 0.1]	depth : 7 learning_rate : 0.1	97	96	95

จากผลการทดลองพบว่าแบบจำลอง Catboost และ XGBoost ให้ค่าความถูกต้องสูงสุด ด้วยการปรับพารามิเตอร์ตามตารางผลสรุปการทดสอบ และที่ทางธนาคารจะดูค่า Recall เป็นหลักเพราะผลลัพธ์เปรียบเทียบค่าความเป็นจริง ซึ่งจากตารางที่ 1 ค่า Recall ของ Catboost และ XGBoost ให้ค่าที่สูง จึงสรุปได้ว่าแบจำลอง Catboost และ XGBoost เป็นแบบจำลองที่ดีเมื่อใช้ทำนายกับชุดข้อมูล

สรุปผลการวิจัย

งานวิจัยนี้สร้างแบบจำลองเพื่อทำนายลูกหนี้ที่มีโอกาสผิดนัดชำระ กรณีลูกหนี้มาขอกู้สินเชื่อบัตรเครดิต โดยอาศัยการเรียนรู้ของเครื่องจากชุดข้อมูล Kaggle.com งานวิจัยนี้มีข้อมูลไว้ทั้งหมด 2 ชุดคือ ข้อมูลรายการธุรกรรมสินเชื่อบัตรเครดิต โดยมีข้อมูลรายการธุรกรรม จำนวน 1,048,575 แถว และข้อมูลลูกค้า จำนวน 438,557 แถว ซึ่งตรวจสอบค่าที่ขาดหายไปพร้อมแปลงข้อมูลให้พร้อมสำหรับนำไปสร้างแบบจำลองการทำนายในลำดับต่อไปจากนั้นทำการแบ่งข้อมูลสำหรับการฝึกสอนและข้อมูลสำหรับทดสอบในอัตรา 80:20 จัดการกับปัญหาข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE แล้วทำการประเมินด้วยตาราง Confusion Matrix จากตารางที่ 1 ผลการทดลองของงานวิจัยได้เปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง 3 เทคนิคได้แก้ Logistic Regression ให้ค่าความถูกต้อง 62% Catboost ให้ค่าความถูกต้อง 97% ที่จำนวนต้นไม้ 7 ต้น กับอัตราการเรียนรู้ที่ 0.1 วิธี XGBoost ให้ค่าความถูกต้อง 98% ที่จำนวนต้นไม้ 15 ต้น กับอัตราการเรียนรู้ที่ 0.1 นอกจากนี้จากตารางยังบอกถึง Percision การกระจุกตัวตัวของการทำนายผลซึ่งเมื่อเทียบกับค่าความถูกต้องผลของการกระจุกตัวสอดคล้องกับค่าความถูกต้อง และในส่วนของค่า Recall เนื่องจากแบบจำลองสามารถแยกประเภทของทั้ง 2 คลาสได้เป็นอย่างดีจึงทำให้ผลของ Recall สูงตามสอดคล้องกับผลการทำนายด้วย

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

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

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

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

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

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

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

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

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

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

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

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

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

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

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