วารสารวิทยาศาสตร์บูรพา ปีที่ 28 (ฉบับที่ 3) กันยายน – ธันวาคม พ.ศ. 2566
BURAPHA SCIENCE JOURNAL Volume 28 (No.3) September – December 2023 บทความวิจัย

การประยุกต์การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค และเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยสำหรับการทำนายการผิดนัดชำระสินเชื่อเพื่อการอุปโภคบริโภค

Application of Binary Logistic Regression Analysis and the Synthetic Minority Oversampling Technique for Predicting Consumer Loan Default

กัญญาณัฐ มากสูงเนิน และ ธิดารัตน์ อารีรักษ์*

Kanyanat Maksungnern and Tidarut Areerak*

สาขาวิชาคณิตศาสตร์ สำนักวิชาวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ประเทศไทย

School of Mathematics, Institute of Science, Suranaree University of Technology, Thailand

Received : 6 January 2023

Revised : 29 March 2023

Accepted : 25 April 2023

บทคัดย่อ

งานวิจัยนี้ศึกษาการวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค และการแก้ปัญหาข้อมูลที่ไม่สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย เพื่อประยุกต์ใช้ในการทำนายการผิดนัดชำระสินเชื่อซึ่งเป็นข้อมูลลูกค้าที่สมัครสินเชื่อเพื่อการอุปโภคบริโภคของประเทศอินเดียในปี 2021 งานวิจัยนี้แบ่งชุดข้อมูลออกเป็นข้อมูลฝึกหัด และข้อมูลทดสอบทั้งหมด 3 อัตราส่วนที่แตกต่างกัน โดยในแต่ละอัตราส่วนจะสร้างตัวแบบจากข้อมูลที่ยังไม่ปรับให้สมดุล และข้อมูลที่ปรับให้สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย สำหรับการศึกษาเปรียบเทียบประสิทธิภาพของตัวแบบ งานวิจัยนี้เลือกใช้การเรียกกลับในการเปรียบเทียบประสิทธิภาพของตัวแบบ ผลการศึกษาพบว่าสำหรับการแบ่งชุดข้อมูลในทุกอัตราส่วน ตัวแบบที่สร้างจากข้อมูลที่มีการแก้ปัญหาข้อมูลที่ไม่สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยมีการเรียกกลับมากกว่าตัวแบบที่สร้างจากข้อมูลที่ไม่สมดุล ส่งผลให้สามารถทำนายผู้ที่ผิดนัดชำระสินเชื่อได้ดีขึ้น

คำสำคัญ : การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค ; ข้อมูลไม่สมดุล ;
เทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย

Abstract

The binary logistic regression analysis and the synthetic minority oversampling technique (SMOTE) are studied in this paper to predict India's consumer loan default in 2021. SMOTE is used to solve the imbalanced problem. The ratio of training data to testing data is divided into three different groups. In each ratio, the models are constructed from imbalanced data and balanced data. The recall of the models is used for comparative study. The recall of the model based on balanced data is higher than the recall of the model based on imbalanced data in each ratio of the dataset. The prediction of loan default was improved.

Keywords : Binary Logistic Regression Analysis ; imbalanced data ; Synthetic Minority Oversampling Technique

***Corresponding author. E-mail** : tidarut@sut.ac.th

บทนำ

สินเชื่อเพื่อการอุปโภคบริโภค (Consumer Loan) หมายถึง สินเชื่อที่ให้แก่บุคคลธรรมดาที่มีวัตถุประสงค์เพื่อนำมาใช้ในการอุปโภคบริโภคในชีวิตประจำวัน ไม่มีวัตถุประสงค์เพื่อนำไปใช้ในการประกอบธุรกิจ (ตามความหมายที่ให้โดยธนาคารแห่งประเทศไทย) ซึ่งประกอบไปด้วย สินเชื่อเพื่อที่อยู่อาศัย สินเชื่อเพื่อการเช่าซื้อรถยนต์และรถจักยานยนต์ สินเชื่อบัตรเครดิต และสินเชื่อส่วนบุคคล เป็นต้น สินเชื่อเพื่อการอุปโภคบริโภคมีความสำคัญ เนื่องจากเป็นสินเชื่อที่แพร่หลายสำหรับประชาชนทั่วไป ดังนั้น การคาดการณ์ที่แม่นยำเกี่ยวกับพฤติกรรมการชำระสินเชื่อของลูกค้าเป็นรายบุคคลจึงมีความสำคัญ เพื่อช่วยให้สถาบันการเงินหรือหน่วยงานที่เป็นผู้ให้สินเชื่อสามารถประเมินความน่าเชื่อถือของผู้ขอสินเชื่อ โดยวิเคราะห์ความเสี่ยงในการผิดนัดชำระเงินกู้ของผู้ขอสินเชื่อ เพื่อใช้ประกอบการตัดสินใจตอบรับหรือปฏิเสธการให้สินเชื่อได้อย่างเหมาะสม

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) เป็นเครื่องมือทางสถิติที่สำคัญสำหรับสร้างตัวแบบเพื่อทำนายโอกาสของการเกิดเหตุการณ์ที่สนใจจากปัจจัยที่เกี่ยวข้อง โดยนิยมใช้กับข้อมูลซึ่งตัวแปรที่ต้องการทำนายเป็นตัวแปรเชิงคุณภาพ (Vanichbuncha, 2007; Sinsomboonthong, 2016) นอกจากนี้การวิเคราะห์การถดถอยลอจิสติกนิยมนำไปประยุกต์ใช้ในปัญหาการจำแนก (Classification Problem) ของงานหลากหลายด้านอีกด้วย จากการสำรวจปริทัศน์วรรณกรรมในอดีต พบว่าการวิเคราะห์การถดถอยลอจิสติกได้นำไปใช้ในการวิเคราะห์ลักษณะเฉพาะของผู้ทำธุรกรรมทางธนาคารผ่านอินเทอร์เน็ตของประเทศตุรกี (Serener, 2016) นอกจากนี้การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคได้นำไปใช้ในการวิเคราะห์และตรวจสอบพื้นที่ที่อาจมีการตัดไม้ทำลายป่าในประเทศอินเดีย (Bera *et al*., 2020) และศึกษาปัจจัยที่มีผลต่อการผิดนัดชำระสินเชื่อสำหรับการเงินรายย่อยของประเทศกานา (Boateng & Oduro, 2018) สำหรับการวิเคราะห์การถดถอยลอจิสติกแบบพหุกลุ่ม (Polytomous Logistic Regression) ได้นำไปใช้ศึกษาผลกระทบของปัจจัยทางเศรษฐกิจ สังคม และประชากร สำหรับเด็กที่อายุต่ำกว่า 5 ปี ในประเทศบังคลาเทศที่มีน้ำหนักน้อยกว่าเกณฑ์ น้ำหนักมากกว่าเกณฑ์ และน้ำหนักที่อยู่ในเกณฑ์ปกติ (Khan *et al*., 2020) ดังนั้น ในงานวิจัยนี้จึงสนใจนำการวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค (Binary Logistic Regression Analysis) มาประยุกต์ใช้กับการทำนายเกี่ยวกับพฤติกรรมของการผิดนัดชำระสินเชื่อเพื่อการอุปโภคบริโภครายบุคคล ซึ่งเป็นข้อมูลปี 2021 ของประเทศอินเดีย ข้อมูลตัวอย่างนำมาจากเว็บไซต์ kaggle.com โดยเป็นข้อมูลที่แสดงรายละเอียดรายบุคคลของลูกค้าที่ใช้บริการสินเชื่อเพื่อการอุปโภคบริโภค ประกอบด้วยข้อมูลส่วนตัวของลูกค้า และพฤติกรรมการชำระสินเชื่อของลูกค้าแต่ละรายโดยแบ่งออกเป็น กลุ่มลูกค้าที่ไม่ผิดนัดชำระสินเชื่อและกลุ่มลูกค้าที่ผิดนัดชำระสินเชื่อ

ข้อมูลตัวอย่างที่นำมาศึกษาในงานวิจัยนี้ พบว่าข้อมูลกลุ่มลูกค้าที่ไม่ผิดนัดชำระสินเชื่อมีจำนวนมากกว่าข้อมูลกลุ่มลูกค้าที่ผิดนัดชำระสินเชื่อค่อนข้างมากโดยมีอัตราส่วนประมาณ 7:1 ซึ่งทำให้เกิดปัญหาข้อมูลไม่สมดุล (Imbalanced Data) และเมื่อนำการวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคมาใช้จะส่งผลให้ผลการทำนายมักจะทำนายออกมาเป็นลูกค้าไม่ผิดนัดชำระสินเชื่อ ซึ่งหากสถาบันการเงินหรือหน่วยงานผู้ให้สินเชื่อนำผลของการทำนายดังกล่าวไปใช้จะทำให้ไม่สามารถตรวจจับผู้ที่ผิดนัดชำระสินเชื่อได้ และอาจส่งผลต่อการพิจารณาการให้สินเชื่อที่ไม่เหมาะสม จนอาจทำให้สูญเสียสภาพคล่อง

ทางการเงินจากการตัดสินใจดังกล่าวได้ การแก้ปัญหาของการวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคจะใช้วิธีการปรับเปลี่ยนจุดแบ่ง (cut-off point) ที่ใช้ในการตัดสินใจ แต่การพิจารณาดังกล่าวจำเป็นต้องใช้ดุลยพินิจหรือประสบการณ์ส่วนตัวของผู้ที่ต้องการทำนาย งานวิจัยนี้จึงนำเสนอการแก้ปัญหาในอีกแนวทางหนึ่ง โดยใช้การแก้ปัญหาข้อมูลไม่สมดุลซึ่งสามารถทำได้หลากหลายวิธี เช่น วิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน ในงานวิจัยนี้นำเสนอการแก้ปัญหาข้อมูลไม่สมดุลโดยวิธีสุ่มเกินที่ใช้เทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย ซึ่งเทคนิคนี้จะทำการสังเคราะห์ข้อมูลที่ใกล้เคียงกับข้อมูลเดิมขึ้นมาใหม่ ซึ่งช่วยลดข้อด้อยของเทคนิคอื่น ๆ โดยไม่ทำให้สูญเสียข้อมูลที่สำคัญบางข้อมูลไป และไม่เป็นการสุ่มเพิ่มข้อมูลเดิมขึ้นมา (Chawla *et al*., 2002) สำหรับเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยมีการนำมาใช้แก้ปัญหาสำหรับข้อมูลที่ไม่สมดุลอย่างแพร่หลายในงานหลากหลายด้าน เช่น การจำแนกข้อมูลสเปกตรัมของรังสีแกมมาโดยใช้โครงข่ายประสาทเทียม (Bellinger *et al*., 2015) การทำนายผลผลิตข้าวสาลีในประเทศฝรั่งเศสโดยใช้การเรียนรู้ด้วยเครื่อง (Chemchem *et al*., 2019) การคาดคะเนระดับความเสี่ยงของยาจากอาการไม่พึงประสงค์โดยใช้การเรียนรู้ด้วยเครื่อง (Wei *et al*., 2020) และการคาดคะเนการรอดชีวิตของผู้ป่วยโรคหัวใจโดยใช้เทคนิคเหมืองข้อมูล (Ishaq *et al*., 2021)

ความรู้พื้นฐาน

1. การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค (Binary Logistic Regression Analysis)

การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคเป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรอธิบาย (Explanatory variable) x_j เมื่อ $j = 1, 2, \cdots, k$ และตัวแปรตอบสนอง (Response variable) y ซึ่งอธิบายโอกาสหรือความน่าจะเป็นของการเกิดเหตุการณ์ของตัวแปรตอบสนอง ทั้งนี้ ตัวแปรตอบสนองเป็นตัวแปรเชิงคุณภาพที่มีค่าเพียง 2 ค่าหรือกลุ่ม 2 กลุ่มเท่านั้น โดยที่เหตุการณ์ $y = 1$ หมายถึง การเกิดเหตุการณ์ที่สนใจซึ่งมีความน่าจะเป็นเท่ากับ p ส่วนเหตุการณ์ $y = 0$ หมายถึง ไม่เกิดเหตุการณ์ที่สนใจซึ่งมีความน่าจะเป็นเท่ากับ $1 - p$ เนื่องจากตัวแปรตอบสนองมีค่าแค่เพียงสองค่าเท่านั้น ทำให้ความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรตอบสนองไม่สอดคล้องกับความสัมพันธ์เชิงเส้น ดังนั้น จึงนิยามอัตราส่วนออดส์ (Odds Ratio) และลอจิต (Logit) ดังสมการที่ (1) และ (2) ตามลำดับ

$$\text{Odds Ratio} = \frac{p}{1 - p} \tag{1}$$

$$\text{Logit} = \ln\left(\text{Odds Ratio}\right) = \ln\left(\frac{p}{1 - p}\right) \tag{2}$$

และเมื่อพิจารณาความสัมพันธ์ระหว่างตัวแปรอธิบายและลอจิตพบว่ามีความสัมพันธ์เชิงเส้น สอดคล้องตามสมการที่ (3)

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3}$$

โดยที่ β_j หมายถึง สัมประสิทธิ์การถดถอย เมื่อ $j = 0,1,\cdots,k$ ดังนั้น จะสามารถหาความน่าจะเป็นของเหตุการณ์ที่สนใจได้ดังสมการที่ (4)

$$p = \frac{e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}} \tag{4}$$

สมการที่ (4) เป็นสมการถดถอยลอจิสติกแบบทวิภาคที่ใช้สำหรับการคาดการณ์ตัวแปรตอบสนองหรือเหตุการณ์ที่สนใจ เมื่อแทนค่าตัวแปรอธิบาย x_j แต่ละค่าลงในสมการที่ (4) จะสามารถหาโอกาสการเกิดของเหตุการณ์ที่สนใจได้ ซึ่งโดยปกติถ้า $p \geq 0.5$ ตัวแบบจะคาดการณ์ว่าเกิดเหตุการณ์ที่สนใจขึ้น ส่วนถ้า $p < 0.5$ ตัวแบบจะคาดการณ์ว่าไม่เกิดเหตุการณ์ที่สนใจ โดยค่า 0.5 นี้จะเรียกว่าจุดแบ่งที่ใช้ในการตัดสินใจ มักนิยมเลือกจุดแบ่งดังกล่าวมีค่าเท่ากับ 0.5 แต่ในบางครั้งอาจมีการปรับค่าจุดแบ่งดังกล่าว เพื่อให้ตัวแบบสามารถทำนายผลลัพธ์ได้ดีขึ้น ซึ่งการกำหนดจุดแบ่งดังกล่าวจะขึ้นอยู่กับประสบการณ์และวิจารณญาณของผู้ใช้ตัวแบบ

การสร้างสมการถดถอยลอจิสติกแบบทวิภาคจำเป็นต้องมีการประมาณค่าสัมประสิทธิ์การถดถอย β_j เมื่อ $j = 0,1,\cdots,k$ ซึ่งในการประมาณค่าพารามิเตอร์ดังกล่าวจะใช้วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method) โดยสมมติว่ามีกลุ่มตัวอย่างขนาด n ของตัวแปรตอบสนอง y_i เมื่อ $i = 1,2,\cdots,n$ ที่เป็นอิสระต่อกัน และมีการแจกแจงแบบแบร์นูลลี (Bernoulli Distribution) โดยที่ $y_i = 1$ ด้วยความน่าจะเป็นเท่ากับ p และ $y_i = 0$ ด้วยความน่าจะเป็นเท่ากับ $1-p$ สำหรับทุก $i = 1,2,\cdots,n$ และตัวแปรอธิบาย $x_j^{(i)}$ เมื่อ $i = 1,2,\cdots,n$ และ $j = 1,2,\cdots,k$ ดังนั้น ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) และฟังก์ชันล็อกภาวะน่าจะเป็น (Log Likelihood Function) สอดคล้องตามสมการที่ (5) และ (6) ตามลำดับ

$$L\left(\beta_0,\beta_1,\cdots,\beta_k\right) = \prod_{i=1}^{n}\left(\frac{e^{\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_k x_k^{(i)}\right)}}{1+e^{\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_k x_k^{(i)}\right)}}\right)^{y_i}\left(\frac{1}{1+e^{\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_k x_k^{(i)}\right)}}\right)^{1-y_i} \tag{5}$$

$$\ln L\left(\beta_0,\beta_1,\cdots,\beta_k\right)=\sum_{i=1}^{n}\left\{y_i\ln\frac{e^{\left(\beta_0+\beta_1x_1^{(i)}+\cdots+\beta_kx_k^{(i)}\right)}}{1+e^{\left(\beta_0+\beta_1x_1^{(i)}+\cdots+\beta_kx_k^{(i)}\right)}}+\left(1-y_i\right)\ln\frac{1}{1+e^{\left(\beta_0+\beta_1x_1^{(i)}+\cdots+\beta_kx_k^{(i)}\right)}}\right\} \quad (6)$$

สำหรับการประมาณค่า $\beta_0,\beta_1,\cdots,\beta_k$ ด้วยวิธีภาวะน่าจะเป็นสูงสุด จะหา $\beta_0,\beta_1,\cdots,\beta_k$ ที่ทำให้ฟังก์ชันล็อกภาวะน่าจะเป็นมีค่าสูงสุด โดยหาอนุพันธ์ย่อยของสมการที่ (6) เทียบกับ $\beta_0,\beta_1,\cdots,\beta_k$ และให้มีค่าเท่ากับศูนย์ ซึ่งจะทำให้ได้สมการทั้งหมด $k+1$ สมการ และเรียกว่าสมการภาวะน่าจะเป็น เนื่องจากสมการดังกล่าวไม่เป็นสมการเชิงเส้นจึงต้องใช้วิธีเชิงตัวเลข (Numerical Method) ในการหาผลเฉลยของระบบสมการ ทั้งนี้ค่าประมาณของ $\beta_0,\beta_1,\cdots,\beta_k$ แทนด้วย $b_0,b_1,\cdots,b_k$ ตามลำดับ และเรียก $b_0,b_1,\cdots,b_k$ ว่าตัวประมาณค่าภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator) ของ $\beta_0,\beta_1,\cdots,\beta_k$ ตามลำดับ (Vanichbuncha, 2007; Sinsomboonthong, 2016)

การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคเป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรอธิบายที่ส่งผลต่อตัวแปรตอบสนอง ดังนั้น การคัดเลือกตัวแปรอธิบาย (Selection of Explanatory Variables) ที่เหมาะสมจะส่งผลกระทบต่อประสิทธิภาพของตัวแบบ การคัดเลือกตัวแปรอธิบายทำได้หลายวิธี (Boonmeekham, 2018) ดังนี้

1) วิธีข้างหน้าทีละขั้น (Forward Stepwise Method) เป็นวิธีการคัดเลือกตัวแปรอธิบายเข้าสู่ตัวแบบครั้งละหนึ่งตัวแปร โดยเลือกเพิ่มตัวแปรอธิบายที่อธิบายการเปลี่ยนแปลงของตัวแปรตอบสนองได้มากที่สุดอย่างมีนัยสำคัญทางสถิติก่อน จากนั้นจึงเลือกเพิ่มตัวแปรอธิบายที่อธิบายการเปลี่ยนแปลงของตัวแปรตอบสนองในลำดับถัดไป ทำเช่นนี้จนกระทั่งไม่มีตัวแปรอธิบายใดที่อธิบายการเปลี่ยนแปลงของตัวแปรตอบสนองได้อย่างมีนัยสำคัญทางสถิติอีก
2) วิธีย้อนหลังทีละขั้น (Backward Stepwise Method) เป็นวิธีการที่ตรงกันข้ามกับวิธีข้างหน้าทีละขั้น โดยเริ่มต้นจากการคัดเลือกตัวแปรอธิบายทั้งหมดเข้าสู่ตัวแบบ จากนั้นพิจารณาลดตัวแปรอธิบายออกครั้งละหนึ่งตัวแปร โดยเลือกตัดตัวแปรที่อธิบายการเปลี่ยนแปลงของตัวแปรตอบสนองได้น้อยที่สุดอย่างมีนัยสำคัญทางสถิติออกก่อน จากนั้นจึงเลือกตัดตัวแปรอธิบายที่อธิบายการเปลี่ยนแปลงของตัวแปรตอบสนองในลำดับถัดไป ทำเช่นนี้จนกระทั่งไม่สามารถตัดตัวแปรอธิบายใดได้อีก
3) วิธีข้างหน้า-ย้อนหลังทีละขั้น (Forward-Backward Stepwise Method) เป็นวิธีที่ผสมผสานระหว่างวิธีข้างหน้าทีละขั้น และวิธีย้อนหลังทีละขั้น

งานวิจัยนี้ได้เปรียบเทียบวิธีการคัดเลือกตัวแปรอธิบาย 3 วิธี ประกอบด้วย วิธีข้างหน้าทีละขั้น วิธีย้อนหลังทีละขั้น และวิธีข้างหน้า-ย้อนหลังทีละขั้น โดยใช้ในการสร้างตัวแบบ จากนั้นจึงพิจารณาผลกระทบของวิธีการคัดเลือกตัวแปรอธิบายวิธีต่าง ๆ ต่อตัวแบบและประสิทธิภาพของตัวแบบที่สร้างขึ้น

2. *ข้อมูลไม่สมดุล (Imbalanced Data)*

สำหรับปัญหาเกี่ยวกับการจำแนกข้อมูลที่มีการแบ่งกลุ่มของข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มส่วนมาก (Majority Class) และกลุ่มส่วนน้อย (Minority Class) อาจพบปัญหาข้อมูลไม่สมดุลได้ซึ่งเกิดจากการที่กลุ่มส่วนมากมีจำนวนข้อมูลแตกต่างกันเป็นจำนวนมากเมื่อเทียบกับกลุ่มส่วนน้อย การวัดระดับความไม่สมดุลของข้อมูลจะพิจารณาจากอัตราส่วนความไม่สมดุลของข้อมูล (Imbalanced Ratio:IR) ได้ตามสมการที่ (7)

$$\mathrm{IR} = \frac{N_{\mathrm{major}}}{N_{\mathrm{minor}}} \tag{7}$$

เมื่อ N_{major} หมายถึง จำนวนข้อมูลในกลุ่มส่วนมาก

N_{minor} หมายถึง จำนวนข้อมูลในกลุ่มส่วนน้อย

การนำข้อมูลไม่สมดุลมาวิเคราะห์จะส่งผลต่อประสิทธิภาพของตัวแบบที่สร้างขึ้น ทั้งนี้ การแก้ปัญหาข้อมูลไม่สมดุลสามารถทำได้หลายวิธี ดังนี้

1) วิธีสุ่มเกิน (Oversampling Method) เป็นวิธีการสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อยโดยสุ่มเลือกจากข้อมูลเดิมที่มีอยู่จนข้อมูลในกลุ่มส่วนน้อยมีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มส่วนมาก
2) วิธีสุ่มลด (Undersampling Method) เป็นวิธีการสุ่มลดข้อมูลในกลุ่มส่วนมากออกไปจนข้อมูลของกลุ่มส่วนมากมีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในกลุ่มส่วนน้อย
3) วิธีผสมผสาน (Hybrid Method) เป็นวิธีที่ผสมผสานระหว่างวิธีสุ่มเกินและวิธีสุ่มลด โดยสุ่มลดข้อมูลในกลุ่มส่วนมาก พร้อมกับสุ่มเพิ่มข้อมูลในกลุ่มส่วนน้อยให้จำนวนข้อมูลทั้งสองกลุ่มมีจำนวนใกล้เคียงหรือเท่ากัน

งานวิจัยนี้สนใจศึกษาการแก้ปัญหาข้อมูลไม่สมดุล โดยใช้วิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique : SMOTE) ซึ่งเป็นวิธีสุ่มเกินที่มีการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมในกลุ่มส่วนน้อยให้มีความใกล้เคียงกับข้อมูลเดิม ซึ่งแตกต่างกับวิธีสุ่มเกินที่นำข้อมูลเดิมมาเพิ่มจำนวนให้มากขึ้น สำหรับการเพิ่มจำนวนข้อมูลในกลุ่มส่วนน้อยนั้นจะทำให้การกระจายของข้อมูลมีความสมดุลมากขึ้น วิธีการสังเคราะห์ข้อมูลสำหรับกลุ่มส่วนน้อยทำได้โดยข้อมูลแต่ละข้อมูลที่อยู่ในกลุ่มส่วนน้อย จะพิจารณาหาค่าระยะทางแบบยุคลิด (Euclidean distance) ระหว่างค่าดังกล่าวกับ

ข้อมูลทั้งหมดในกลุ่มส่วนน้อย แล้วทำการเลือกข้อมูลที่มีระยะทางน้อยที่สุดมาจำนวน K ข้อมูล ซึ่งเรียกว่า ข้อมูลใกล้เคียง K ข้อมูล (K-nearest neighbor) จากนั้นสุ่มเลือกข้อมูลจากข้อมูลใกล้เคียงเพื่อสังเคราะห์ข้อมูลใหม่ โดยข้อมูลใหม่ที่สังเคราะห์ขึ้นจะมีค่าอยู่บนส่วนของเส้นตรงที่เชื่อมระหว่างข้อมูลดังกล่าวกับข้อมูลใกล้เคียงตัวที่สุ่มเลือกมา ทั้งนี้จำนวนข้อมูลใกล้เคียงที่สุ่มเลือกมาจะขึ้นอยู่กับจำนวนข้อมูลที่ต้องการสังเคราะห์เพิ่มนั่นเอง (Chawla *et al*., 2002; He & Garcia, 2009)

3. *การวิเคราะห์ประสิทธิภาพของตัวแบบ*

เมทริกซ์ความสับสน (Confusion Matrix) เป็นเครื่องมือที่สำคัญและมักนิยมใช้สำหรับการประเมินประสิทธิภาพของตัวแบบแสดงได้ใน Figure 1 (Chawla *et al*., 2002)

		Predicted Class: Positive	Predicted Class: Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Figure 1 Confusion Matrix

เมื่อ TP (True Positive) หมายถึง จำนวนค่าสังเกตที่เป็นบวกซึ่งตัวแบบจำแนกถูกต้อง นั่นคือ จำนวนค่าสังเกตที่ตัวแบบทำนายถูกว่าเกิดเหตุการณ์ที่สนใจ

TN (True Negative) หมายถึง จำนวนค่าสังเกตที่เป็นลบซึ่งตัวแบบจำแนกถูกต้อง นั่นคือ จำนวนค่าสังเกตที่ตัวแบบทำนายถูกว่าไม่เกิดเหตุการณ์ที่สนใจ

FP (False Positive) หมายถึง จำนวนค่าสังเกตที่เป็นลบซึ่งตัวแบบจำแนกไม่ถูกต้อง นั่นคือ จำนวนค่าสังเกตที่ตัวแบบทำนายผิดว่าเกิดเหตุการณ์ที่สนใจ แต่ข้อมูลจริงไม่เกิดเหตุการณ์ที่สนใจ

FN (False Negative) หมายถึง จำนวนค่าสังเกตที่เป็นบวกซึ่งตัวแบบจำแนกไม่ถูกต้อง นั่นคือ จำนวนค่าสังเกตที่ตัวแบบทำนายผิดว่าไม่เกิดเหตุการณ์ที่สนใจ แต่ข้อมูลจริงเกิดเหตุการณ์ที่สนใจ

โดยสามารถนำค่าต่าง ๆ ในเมทริกซ์ความสับสน มาคำนวณมาตรวัดต่าง ๆ เพื่อใช้ในการวัดประสิทธิภาพของตัวแบบได้ดังนี้

1) ความแม่น (Accuracy) คือ อัตราส่วนของจำนวนค่าสังเกตที่ทำนายถูกต้องเทียบกับจำนวนค่าสังเกตทั้งหมด แสดงได้ดังสมการที่ (8)

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

2) การเรียกกลับ (Recall) หรือเรียกได้อีกอย่างว่าความไว (Sensitivity) คือ อัตราส่วนของจำนวนค่าสังเกตที่ทำนายเหตุการณ์ที่สนใจถูกต้องเทียบกับจำนวนค่าสังเกตของเหตุการณ์จริงที่สนใจทั้งหมด แสดงได้ดังสมการที่ (9)

$$\text{recall} = \frac{TP}{TP + FN} \tag{9}$$

3) ความเที่ยง (Precision) คือ อัตราส่วนของจำนวนค่าสังเกตที่ทำนายเหตุการณ์ที่สนใจถูกต้องเทียบกับจำนวนค่าสังเกตที่ทำนายว่าเกิดเหตุการณ์ที่สนใจทั้งหมด แสดงได้ดังสมการที่ (10)

$$\text{precision} = \frac{TP}{TP + FP} \tag{10}$$

4) F1-Score คือ ค่าเฉลี่ยแบบฮาร์มอนิกของความเที่ยงและการเรียกกลับ แสดงได้ดังสมการที่ (11)

$$\text{F1-Score} = 2\left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right) \tag{11}$$

งานวิจัยนี้เลือกใช้การเรียกกลับในการวัดประสิทธิภาพของตัวแบบ เนื่องจากให้ความสำคัญกับการคาดการณ์ผู้ที่ผิดนัดชำระสินเชื่อที่ถูกต้องจากกลุ่มผู้ที่ผิดนัดชำระสินเชื่อที่แท้จริง

วิธีดำเนินการวิจัย

1. <u>สารสนเทศของชุดข้อมูล</u>

ชุดข้อมูลที่ใช้ในงานวิจัยนี้นำมาจากเว็บไซต์ kaggle.com ชื่อชุดข้อมูลคือ “Loan Prediction Based on Customer Behavior Predict who Possible Defaulters are for the Consumer Loans Product” โดยเป็นข้อมูลของลูกค้าที่สมัครสินเชื่อเพื่อการอุปโภคบริโภคในประเทศอินเดีย ซึ่งบันทึกไว้ในเดือนสิงหาคม ค.ศ. 2021 มีจำนวนข้อมูลทั้งหมด 252,000

ข้อมูล ซึ่งประกอบด้วยข้อมูลของผู้ที่ผิดนัดชำระสินเชื่อจำนวน 30,996 ข้อมูล และข้อมูลของผู้ที่ไม่ผิดนัดชำระสินเชื่อจำนวน 221,004 ข้อมูล

ในงานวิจัยนี้พิจารณาตัวแปรอธิบายทั้งหมด 8 ตัวแปร ได้แก่ รายได้ อายุ จำนวนปีในการทำงาน สถานภาพ (โสด / แต่งงาน) สถานะการเป็นเจ้าของบ้าน (เจ้าของ / ผู้เช่า / ผู้อาศัย) สถานะการเป็นเจ้าของรถ (เจ้าของ / ไม่ได้เป็นเจ้าของ) จำนวนปีในการประกอบอาชีพปัจจุบัน จำนวนปีที่อาศัยในที่อยู่ปัจจุบัน และตัวแปรตอบสนอง 1 ตัวแปร คือ สถานะการผิดนัดชำระสินเชื่อ (ไม่ผิดนัดชำระสินเชื่อ / ผิดนัดชำระสินเชื่อ)

2. *การตรวจสอบข้อมูล*

งานวิจัยนี้ใช้ภาษา Python ในการประมวลผล ซึ่งก่อนการนำข้อมูลไปใช้ได้มีการตรวจสอบข้อมูล โดยตรวจสอบประเภทข้อมูลของตัวแปรทั้งหมด และตรวจสอบข้อมูลสูญหาย (Missing Value) จากการตรวจสอบข้อมูลดังกล่าวไม่พบกรณีที่ข้อมูลสูญหาย

3. *การเตรียมข้อมูล*

การเตรียมข้อมูลของงานวิจัยนี้ดำเนินการโดยจำแนกออกเป็น 2 กรณี กรณีแรกถ้าตัวแปรอธิบายเป็นตัวแปรเชิงคุณภาพจะทำการปรับเป็นตัวแปรหุ่น (Dummy Variable) ส่วนกรณีที่สองถ้าตัวแปรอธิบายเป็นตัวแปรเชิงปริมาณที่มีช่วงของข้อมูลที่กว้างจะทำการปรับช่วงข้อมูลให้อยู่ในช่วง [0,1] โดยใช้วิธี MinMaxScaler จากนั้นจึงนำชุดข้อมูลดังกล่าวแบ่งออกเป็นข้อมูลฝึกหัด (Training Data) และข้อมูลทดสอบ (Testing Data) เนื่องจากการแบ่งอัตราส่วนข้อมูลไม่มีวิธีการหรือแนวทางที่ชัดเจน ซึ่งในการนำไปใช้งานจึงมีการเลือกอัตราส่วนที่แตกต่างกัน (Roshan, 2022) ดังนั้น ในงานวิจัยนี้จึงพิจารณาการแบ่งข้อมูลออกเป็น 3 อัตราส่วน (ข้อมูลฝึกหัด:ข้อมูลทดสอบ) ดังนี้ 70:30 75:25 และ 80:20

4. *การจัดการข้อมูลที่ไม่สมดุล*

เนื่องจากกลุ่มของผู้ที่ไม่ผิดนัดชำระสินเชื่อ (กลุ่มส่วนมาก) และกลุ่มของผู้ที่ผิดนัดชำระสินเชื่อ (กลุ่มส่วนน้อย) มีจำนวนข้อมูลที่แตกต่างกันมาก งานวิจัยนี้จึงใช้เทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย (SMOTE) ในการแก้ปัญหาข้อมูลไม่สมดุลที่เกิดขึ้น โดยใช้เทคนิคดังกล่าวสำหรับข้อมูลฝึกหัด ในงานนี้ใช้ฟังก์ชัน “SMOTE” ซึ่งเป็นฟังก์ชันสำเร็จรูปใน Python ทั้งนี้จำนวนข้อมูลก่อนและหลังการใช้ฟังก์ชัน SMOTE แสดงได้ตาม Table 1

5. *การคัดเลือกตัวแปรอธิบาย*

ขั้นตอนนี้จะนำข้อมูลฝึกหัดที่ปรับให้สมดุลเรียบร้อยแล้ว มาดำเนินการคัดเลือกตัวแปรอธิบายโดยเปรียบเทียบระหว่างวิธีข้างหน้าทีละขั้น วิธีย้อนหลังทีละขั้น และวิธีข้างหน้า-ย้อนหลังทีละขั้น ทั้งนี้ ตัวแปรอธิบายจะถูกคัดเข้าหรือคัดออกจากตัวแบบทีละตัว เพื่อทดสอบความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรตอบสนอง ในงานวิจัยนี้พิจารณาระดับนัยสำคัญทางสถิติที่ระดับ 0.05

Table 1 Result of using SMOTE to solve imbalanced data

	70:30		75:25		80:20	
	Number of majority class	Number of minority class	Number of majority class	Number of minority class	Number of majority class	Number of minority class
Before using SMOTE	154,675	21,725	165,796	23,204	176,857	24,743
After using SMOTE	154,675	154,675	165,796	165,796	176,857	176,857

6. *การสร้างตัวแบบโดยใช้การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค*

เมื่อได้ตัวแปรอธิบายทั้งหมดที่สอดคล้องกับการคัดเลือกตัวแปรอธิบายในหัวข้อก่อนหน้า จึงนำมาสร้างตัวแบบโดยใช้การวิเคราะห์การถดถอยลอจิสติกแบบทวิภาค สำหรับการสร้างตัวแบบจะแบ่งออกเป็น 2 กรณี คือ ข้อมูลที่ไม่มีการปรับให้สมดุล และข้อมูลที่มีการปรับให้สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย

7. *การวิเคราะห์ประสิทธิภาพของตัวแบบ*

ตัวแบบที่ได้ในขั้นตอนก่อนหน้าจะถูกนำไปใช้กับข้อมูลทดสอบ เพื่อวิเคราะห์ประสิทธิภาพของตัวแบบ งานวิจัยนี้เลือกใช้การเรียกกลับในการวัดประสิทธิภาพของตัวแบบ

ผลการวิจัย

การศึกษาการสร้างตัวแบบด้วยการวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคนั้น ความหมายของตัวแปรอธิบายต่าง ๆ ในตัวแบบแสดงไว้ใน Table 2

สำหรับกรณีข้อมูลที่ไม่มีการปรับให้สมดุลในอัตราส่วนข้อมูลเดียวกันพบว่า การคัดเลือกตัวแปรอธิบายทั้ง 3 วิธี (วิธีข้างหน้าทีละขั้น วิธีย้อนหลังทีละขั้น และวิธีข้างหน้า-ย้อนหลังทีละขั้น) จะได้สมการถดถอยลอจิสติกแบบทวิภาคที่เหมือนกัน ซึ่งสมการถดถอยลอจิสติกแบบทวิภาคสำหรับแต่ละอัตราส่วนของข้อมูลแสดงใน Table 3 ทั้งนี้ ความหมายของตัวแปรอธิบายแสดงใน Table 2 และเมื่อพิจารณาข้อมูลในทุกอัตราส่วนพบว่า ตัวแปรอธิบายเหมือนกันทั้ง 3 อัตราส่วนของข้อมูล

Table 2 Definition of explanatory variables

Explanatory variables	Meaning
x_1	experience
x_2	house ownership (rented)
x_3	car ownership (owned)
x_4	status (Single)
x_5	age
x_6	current job years
x_7	current house years
x_8	income
x_9	house ownership (owned)

Table 3 The comparison of the binary logistic regression model for imbalanced data

Training data: Testing data	Binary logistic regression model (forward stepwise / backward stepwise / forward-backward stepwise)
70:30	$P(y) = \frac{e^{-2.1432-0.4425x_1+0.2669x_2-0.1677x_3+0.2419x_4-0.1680x_5+0.1454x_6-0.0553x_9}}{1+e^{-2.1432-0.4425x_1+0.2669x_2-0.1677x_3+0.2419x_4-0.1680x_5+0.1454x_6-0.0553x_9}}$
75:25	$P(y) = \frac{e^{-2.1334-0.4412x_1+0.2563x_2-0.1599x_3+0.2450x_4-0.1727x_5+0.1333x_6-0.0893x_9}}{1+e^{-2.1334-0.4412x_1+0.2563x_2-0.1599x_3+0.2450x_4-0.1727x_5+0.1333x_6-0.0893x_9}}$
80:20	$P(y) = \frac{e^{-2.1209-0.4369x_1+0.2476x_2-0.1600x_3+0.2403x_4-0.1797x_5+0.1380x_6-0.1191x_9}}{1+e^{-2.1209-0.4369x_1+0.2476x_2-0.1600x_3+0.2403x_4-0.1797x_5+0.1380x_6-0.1191x_9}}$

กรณีข้อมูลที่มีการปรับให้สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยในอัตราส่วนข้อมูลเดียวกัน พบว่า การคัดเลือกตัวแปรอธิบายทั้ง 3 วิธี (วิธีข้างหน้าทีละขั้น วิธีย้อนหลังทีละขั้น และวิธีข้างหน้า-ย้อนหลังทีละขั้น) จะได้สมการถดถอยลอจิสติกแบบทวิภาคที่เหมือนกัน สำหรับกรณีนี้สมการถดถอยลอจิสติกแบบทวิภาคสำหรับแต่ละอัตราส่วนของข้อมูลแสดงใน Table 4 ทั้งนี้ ความหมายของตัวแปรอธิบายแสดงใน Table 2 และเมื่อพิจารณาข้อมูลในทุกอัตราส่วนข้อมูล พบว่า ตัวแปรอธิบายที่มีผลต่อตัวแบบของทั้ง 3 อัตราส่วน ประกอบไปด้วยตัวแปรอธิบาย x_1, x_2, x_3, x_4, x_5, x_6, x_7

และ x_9 แต่จะมีความแตกต่างเล็กน้อยสำหรับตัวแบบในอัตราส่วนข้อมูล 75:25 และ 80:20 ที่จะมีตัวแปรอธิบายที่เพิ่มจากอัตราส่วนข้อมูล 70:30 มาอีก 1 ตัวแปรคือ x_8

Table 4 The comparison of the binary logistic regression model for balanced data with SMOTE

Training data: Testing data	Binary logistic regression model (forward stepwise / backward stepwise / forward-backward stepwise)
70:30	$P(y)=\frac{e^{-0.1862-0.4521x_1+0.2987x_2-0.1728x_3+0.2313x_4-0.1664x_5+0.1534x_6-0.0285x_7-0.0123x_9}}{1+e^{-0.1862-0.4521x_1+0.2987x_2-0.1728x_3+0.2313x_4-0.1664x_5+0.1534x_6-0.0285x_7-0.0123x_9}}$
75:25	$P(y)=\frac{e^{-0.1196-0.4684x_1+0.2416x_2-0.1623x_3+0.2469x_4-0.1697x_5+0.1637x_6-0.0276x_7-0.0448x_8-0.1018x_9}}{1+e^{-0.1196-0.4684x_1+0.2416x_2-0.1623x_3+0.2469x_4-0.1697x_5+0.1637x_6-0.0276x_7-0.0448x_8-0.1018x_9}}$
80:20	$P(y)=\frac{e^{-0.0789-0.4544x_1+0.2021x_2-0.1594x_3+0.2482x_4-0.1856x_5+0.1655x_6-0.0293x_7-0.0479x_8-0.1807x_9}}{1+e^{-0.0789-0.4544x_1+0.2021x_2-0.1594x_3+0.2482x_4-0.1856x_5+0.1655x_6-0.0293x_7-0.0479x_8-0.1807x_9}}$

การวิเคราะห์ประสิทธิภาพของตัวแบบโดยเปรียบเทียบระหว่างข้อมูลที่ไม่มีการปรับให้สมดุล และข้อมูลที่ปรับให้สมดุลโดยใช้เทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อย แยกตามข้อมูลที่มีการแบ่งข้อมูลฝึกหัดและข้อมูลทดสอบด้วยอัตราส่วนที่แตกต่างกัน พบว่าค่า TP , TN , FP และ FN เป็นไปตาม Table 5 และการเรียกกลับของตัวแบบเป็นไปตาม Table 6

Table 5 The comparison of TP , TN , FP and FN

Training data: Testing data	Number of testing data	Imbalanced data				Balanced data with SMOTE			
		TP	TN	FP	FN	TP	TN	FP	FN
70:30	75,600	0	66,329	0	9,271	5,323	33,288	33,041	3,948
75:25	63,000	0	55,208	0	7,792	4,457	27,458	27,750	3,335
80:20	50,400	0	44,147	0	6,253	3,601	21,827	22,320	2,652

Table 6 The comparison of model performance

Training data: Testing data	Recall	
	Imbalanced data	Balanced data with SMOTE
70:30	0%	57.42%
75:25	0%	57.20%
80:20	0%	57.59%

วิจารณ์ผลการวิจัย

เมื่อพิจารณาวิธีการคัดเลือกตัวแปรอธิบายที่แตกต่างกัน 3 วิธี คือ วิธีข้างหน้าทีละขั้น วิธีย้อนหลังทีละขั้น และวิธีข้างหน้า-ย้อนหลังทีละขั้น พบว่า วิธีการคัดเลือกตัวแปรอธิบายที่แตกต่างกันไม่มีผลกระทบต่อตัวแบบที่สร้างขึ้น เนื่องจากสมการถดถอยลอจิสติกแบบทวิภาคของทั้ง 3 วิธี เป็นสมการเดียวกัน กรณีข้อมูลที่ไม่มีการปรับให้สมดุลสังเกตได้จาก Table 3 และกรณีข้อมูลที่มีการปรับให้สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยสังเกตได้จาก Table 4

จากผลการวิจัยพบว่า ข้อมูลที่นำมาวิจัยเกิดปัญหาความไม่สมดุล ทำให้ตัวแบบที่สร้างขึ้นไม่สามารถทำนายผู้ที่ผิดนัดชำระสินเชื่อได้เลย ซึ่งจะเห็นได้จากค่า TP และ FP ที่มีค่าเท่ากับ 0 สำหรับทุกอัตราส่วนของข้อมูลซึ่งแสดงใน Table 5 ส่งผลให้การเรียกกลับของตัวแบบที่ยังไม่ได้ปรับข้อมูลให้สมดุลมีค่าเท่ากับ 0% เสมอในทุกอัตราส่วนของข้อมูลซึ่งแสดงใน Table 6 แต่เมื่อทำการปรับข้อมูลให้มีความสมดุลโดยใช้เทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยพบว่า ค่า TP และ FP มีค่าเพิ่มขึ้นในทุกอัตราส่วนข้อมูลซึ่งแสดงใน Table 5 ซึ่งหมายถึง ตัวแบบสามารถทำนายผู้ที่ผิดนัดชำระสินเชื่อได้แล้ว ซึ่งส่งผลให้การเรียกกลับของตัวแบบมีค่ามากขึ้นโดยมีค่าประมาณ 57% จากผลดังกล่าวสังเกตได้ว่า เมื่อมีการปรับข้อมูลให้สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยแล้วตัวแบบที่ได้สามารถทำนายการผิดนัดชำระสินเชื่อได้ถูกต้องมากขึ้นกว่าเดิม

สรุปผลการวิจัย

งานวิจัยนี้ได้นำการวิเคราะห์การถดถอยลอจิสติกแบบทวิภาคและการปรับข้อมูลที่ไม่สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยมาใช้ในการคาดการณ์การผิดนัดชำระสินเชื่อ โดยใช้คำสั่งสำเร็จรูปจากโปรแกรมภาษา Python สำหรับการคัดเลือกตัวแปรอธิบายด้วยวิธีข้างหน้าทีละขั้น วิธีย้อนหลังทีละขั้น และวิธีข้างหน้า-ย้อนหลังทีละขั้น โดยเปรียบเทียบข้อมูลที่มีการแบ่งอัตราส่วนข้อมูลฝึกหัดต่อข้อมูลทดสอบออกเป็น 3 กลุ่ม ได้แก่ 70:30 75:25 และ 80:20 งานวิจัยนี้พบว่าวิธีการคัดเลือกตัวแปรไม่มีผลต่อตัวแบบที่ได้ ซึ่งเห็นได้จาก ไม่ว่าจะเลือกใช้วิธีการคัดเลือกแบบใดยังคงได้ตัวแบบเดียวกันเสมอ และเมื่อพิจารณาข้อมูลในอัตราส่วนเดียวกัน เปรียบเทียบระหว่างกรณีข้อมูลที่ไม่มีการปรับให้สมดุลและกรณี

ข้อมูลที่มีการปรับให้สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยพบว่าตัวแปรอธิบายมีความแตกต่างกันเล็กน้อย ส่งผลให้ตัวแบบมีความแตกต่างกัน แต่เมื่อพิจารณาประสิทธิภาพของตัวแบบโดยใช้การเรียกกลับพบว่า การปรับข้อมูลที่ไม่สมดุลด้วยเทคนิคการสังเคราะห์ข้อมูลเพิ่มจากกลุ่มส่วนน้อยทำให้ตัวแบบที่ได้มีการเรียกกลับมากขึ้นกว่าการไม่ปรับข้อมูลในทุกอัตราส่วนข้อมูล ซึ่งการเรียกกลับที่มากขึ้นนี้จะส่งผลให้สามารถทำนายผู้ที่ผิดนัดชำระหนี้ได้ดียิ่งขึ้น ทั้งนี้ สำหรับการพัฒนางานวิจัยต่อไปในอนาคต จากเดิมที่แบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลฝึกหัด และข้อมูลทดสอบ อาจเพิ่มเติมการแบ่งข้อมูลออกเป็น 3 ส่วน คือ ข้อมูลฝึกหัด ข้อมูลตรวจสอบ และข้อมูลทดสอบ โดยเปรียบเทียบผลการทำนายกับกรณีที่แบ่งข้อมูลออกเป็น 2 ส่วน

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยเทคโนโลยีสุรนารีที่สนับสนุนในด้านต่าง ๆ สำหรับการทำวิจัย

เอกสารอ้างอิง

Bellinger, C., Japkowicz, N., & Drummond, C. (2015). Synthetic Oversampling for Advanced Radioactive Threat Detection. In *IEEE 14th International Conference on Machine Learning and Applications*. (pp. 948-953). United States: IEEE.

Bera, B., Saha, S., & Bhattacharjee, S. (2020). Forest Cover Dynamics (1998 to 2019) and Prediction of Deforestation Probability using Binary Logistic Regression (BLR) Model of Silabati Watershed, India. *Trees, Forests and People*, *2*, 100034.

Boateng, E., & Oduro, F. (2018). Predicting Microfinance Credit Default: A Study of Nsoatreman Rural Bank, Ghana. *Journal of Advances in Mathematics and Computer Science*, *26*(1), 1-9.

Boonmeekham, A. (2018). *Predictive Models for Lapse of Life Insurance policy with Logistic Regression Model and Cox Proportional Hazard Model*. Master's Degree Thesis of Kasetsart University. (in Thai)

Chawla, V.N., Bowyer, W.K., Hall, O.L., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence research*, *16*, 321-357.

Chemchem, A., Alin, F., & Krajecki, M. (2019). Combining SMOTE Sampling and Machine Learning for Forecasting Wheat Yields in France. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering.* (pp. 9-14). United States: IEEE.

He, H., & Garcia, A.E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and DATA Engineering*, *21*(9), 1263-1284.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021) Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. *IEEE Access*, *9*, 39707-39716.

Khan, S., Halder, H., Rashid, M., Afroja, S., & Islam, M. (2020). Impact of Socioeconomic and Demographic Factors for Underweight and Overweight Children in Bangladesh: A Polytomous Logistic Regression Model. *Clinical Epidemiology and Global Health*, *8*, 1348-1355.

Roshan, V. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 15*(4), 531-538.

Serener, B. (2016). Statistical Analysis of Internet Banking Usage with Logistic Regression. *Procedia Computer Science*, *102*, 648-653.

Sinsomboonthong, S. (2016). *Multivariate Analysis.* (1). Bangkok: Chamchuree Products Company Limited. (in Thai)

Vanichbuncha, K. (2007). *Multivariate Analysis*. (2). Bangkok: Dharmasarn Printing Company Limited. (in Thai)

Wei, J., Lu, Z., Qiu, K., Li, P., & Sun, H. (2020). Predicting Drug Risk Level from Adverse Drug Reactions Using SMOTE and Machine Learning Approaches. *IEEE Access*, *8*, 185761-185775.