Feature Selection and Imbalanced Data Problem Solving in Classification of Banking Fraud Prevention

Rachatawan Virakul and Kitsana Waiyamai

September 3, 2022

การคัดเลือกตัวแปรและแก้ปัญหาข้อมูลไม่สมดุลสำหรับจำแนกประเภทลูกค้า กรณีศึกษา : การป้องกันการทุจริตในธนาคาร

Feature Selection and Imbalanced Data Problem Solving in Classification of Banking Fraud Prevention

รชตวรรณ วีรกุล
Rachatawan Virakul
ภาควิชาวิศวกรรมคอมพิวเตอร์ (เทคโนโลยีสารสนเทศ) คณะวิศวกรรมศาสตร์
Department of Computer Engineering (Information Technology) Faculty of Engineering
มหาวิทยาลัยเกษตรศาสตร์
Kasetsart University
Bangkok / Thailand
rachatawan.vi@ku.th

กฤษณะ ไวยมัย
Kitsana Waiyamai
ภาควิชาวิศวกรรมคอมพิวเตอร์ (เทคโนโลยีสารสนเทศ) คณะวิศวกรรมศาสตร์
Department of Computer Engineering (Information Technology) Faculty of Engineering
มหาวิทยาลัยเกษตรศาสตร์
Kasetsart University
Bangkok / Thailand
fengknw@ku.ac.th

บทคัดย่อ — **การคัดเลือกตัวแปรและแก้ปัญหาข้อมูลไม่สมดุล เป็นปัญหาสำคัญสำหรับเทคนิควิธีการจำแนกประเภท ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรและการแก้ปัญหาข้อมูลไม่สมดุลสำหรับจำแนกประเภทลูกค้า จากกรณีศึกษาการป้องกันการทุจริตในธนาคาร โดยการคัดเลือกตัวแปร จากวิธีการหาความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม โดยให้ค่าน้ำหนักของตัวแปรอิสระ ที่เรียกว่า Weight Of Evidence (WOE) เพื่อจัดอันดับความสำคัญของตัวแปรอิสระที่มีผลกับตัวแปรตาม โดยพิจารณาจากค่า Information Value (IV) ซึ่งเป็นเทคนิคที่สำคัญในการเลือกตัวแปร เพื่อเปรียบเทียบประสิทธิภาพเทคนิควิธีการแก้ปัญหาข้อมูลไม่สมดุล 4 วิธี คือ 1. Random Undersampling 2. SMOTE 3. Borderline-SMOTE และ 4. SMOTE-ENN และเปรียบเทียบประสิทธิภาพเทคนิควิธีการจำแนกประเภทลักษณะของลูกค้าที่มีแนวโน้มทุจริต 2 วิธี คือการวิเคราะห์ถดถอยโลจิสติก (Logistic Regression) และต้นไม้ตัดสินใจ (Decision Tree) จากผลการทดลองแสดงว่าการคัดเลือกตัวแปร โดยให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) กับเทคนิควิธีการสุ่มตัวอย่างแบบ RUS + SMOTE สำหรับการจำแนกประเภทลูกค้าทุจริตด้วยเทคนิค Logistic Regression จะให้ประสิทธิภาพในการจำแนกประเภทกลุ่มลูกค้าทุจริตได้ดีที่สุด**

คำสำคัญ — ***การคัดเลือกตัวแปร, ความไม่สมดุล, การจำแนกประเภท, การธนาคาร, การทุจริต***

ABSTRACT — **Feature selection and imbalanced data are important problems for classification techniques. Therefore, this research aims to compare the efficiency of feature selection and imbalanced data problem solving for customer classification in the case study of banking fraud prevention. We perform feature selection to find the relationship between each of the independent variables and the dependent variable. The Weight Of Evidence (WOE) and the Information Value (IV) are used to rank variables based on their importance to affect the dependent variable. For solving imbalanced data, Random Undersampling, SMOTE (Synthetic Minority Oversampling Technique), Borderline-SMOTE, and SMOTE-ENN (Synthetic Minority Oversampling Technique-EditedNearestNeighbours) are used to pre-process data and compared their accuracy with logistic regression and decision tree. Our experiment results show that WOE-based feature selection with sampling methods RUS + SMOTE using logistic regression provides the best accuracy.**

Keywords — ***Feature selection, Imbalance, Classification, Banking, Fraud***

1. บทนำ

ปัจจุบันการป้องกันการทำทุจริต จากข้อมูลการทำธุรกรรมในภาคธุรกิจการเงินการธนาคาร เป็นเรื่องสำคัญที่จะช่วยให้ธนาคารสามารถตรวจสอบการเกิดสิ่งผิดปกติที่เกิดขึ้นในกระบวนการทำทุจริตหรือการส่อเจตนาที่อาจมีความเสี่ยงในการทำทุจริตในอนาคตได้ ซึ่งปัจจุบันข้อมูลลูกค้าทุจริตของธนาคาร เกิดปัญหาความไม่สมดุลกันระหว่างข้อมูลของทั้ง 2 คลาส ส่งผลให้ประสิทธิภาพในการจำแนกประเภทข้อมูลมีความไม่แม่นยำเท่าที่ควร เพราะจะให้ผลลัพธ์ที่ดีกับจำนวนกลุ่มข้อมูลที่มากกว่า ดังนั้นจึงต้องมีการใช้เทคนิควิธีการคัดเลือกตัวแปรที่เหมาะสมควบคู่ไปกับการคัดเลือกวิธีที่จัดการกับข้อมูลที่มีความไม่สมดุลให้มีประสิทธิภาพมากยิ่งขึ้น รวมไปถึงการใช้แบบจำลองการจำแนกประเภทที่สามารถตรวจสอบพฤติกรรมหรือลักษณะของลูกค้าที่มีแนวโน้มทุจริตได้ ดังนั้นผู้วิจัยต้องการ

เปรียบเทียบเทคนิควิธีการคัดเลือกตัวแปรที่เหมาะสมสำหรับชุดข้อมูลที่ไม่สมดุล และเปรียบเทียบประสิทธิภาพเทคนิคการพยากรณ์ในการจำแนกประเภทลูกค้าที่มีแนวโน้มทุจริต จากข้อมูลลักษณะประจำตัว การถือครองผลิตภัณฑ์ และลักษณะพฤติกรรมจากการทำธุรกรรมผ่านช่องทางต่างๆ เพื่อใช้ในการป้องกันการทำทุจริตจากลูกค้าของธนาคารในอนาคตต่อไป

2. วัตถุประสงค์ของการวิจัย

2.1. คัดเลือกตัวแปรที่เหมาะสม และเปรียบเทียบประสิทธิภาพระหว่างตัวแปรที่ให้ค่าน้ำหนักของตัวแปรอิสระที่เรียกว่า Weight Of Evidence (WOE) [1] กับตัวแปรเดิมที่ไม่ให้ค่าน้ำหนักของตัวแปรอิสระ ส่งผลกับการจำแนกประเภทลูกค้าที่มีแนวโน้มทุจริตในธนาคาร

2.2 เปรียบเทียบเทคนิควิธีการแก้ปัญหาข้อมูลไม่สมดุล จากการสุ่มเพิ่มและลดข้อมูล เพื่อหาวิธีที่ดีที่สุดในการจำแนกประเภทลูกค้า

2.3 เปรียบเทียบประสิทธิภาพเทคนิคการพยากรณ์ในการจำแนกประเภทลูกค้าที่มีแนวโน้มทุจริตกับลูกค้าปกติ จากข้อมูลที่มีการจัดการความไม่สมดุล

3. วิธีดำเนินการวิจัย

การวิจัยนี้เป็นการศึกษาวิธีการคัดเลือกตัวแปรให้เหมาะสมกับชุดข้อมูลที่มีความไม่สมดุล และเปรียบเทียบเทคนิควิธีการแก้ไขปัญหาข้อมูลไม่สมดุลสำหรับจำแนกประเภทลูกค้าที่มีแนวโน้มทุจริตของธนาคาร โดยมีข้อมูลและขั้นตอนการดำเนินงานดังนี้

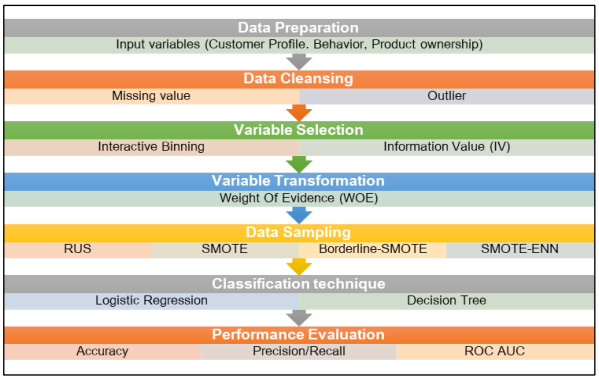
3.1 ข้อมูล

3.1.1 ข้อมูลลูกค้ารายย่อย ประกอบด้วยลักษณะประจำตัวของลูกค้า การถือครองผลิตภัณฑ์ ในช่วงระยะเวลา มี.ค., มิ.ย. และ ก.ย. 2562

3.1.2 ข้อมูลการทำธุรกรรมของลูกค้าผ่านช่องทางต่างๆ ของธนาคาร ในช่วงระยะเวลา เม.ย. 2561 - ก.ย. 2562

3.1.3 ข้อมูลลูกค้าทุจริตในช่วงระยะเวลา มี.ค., มิ.ย. และ ก.ย. 2562

3.2 ขั้นตอนการดำเนินงาน



รูปที่ 1. ภาพรวมขั้นตอนการดำเนินงานของงานวิจัย

จากรูปที่ 1 แสดงขั้นตอนการดำเนินงานของงานวิจัย มีรายละเอียดดังนี้

3.2.1 เตรียมข้อมูลลูกค้าในช่วงระยะเวลา มี.ค., มิ.ย. และ ก.ย. 2562 รวมกับข้อมูลการทำธุรกรรมของลูกค้าผ่านช่องทางต่างๆ ของธนาคาร ในเดือนปัจจุบันและช่วงระยะเวลาย้อนหลัง 3 เดือน 6 เดือน และ 12 เดือน

3.2.2 ทำความสะอาดข้อมูลโดยการกำจัดข้อมูลที่ผิดปกติ เช่น ข้อมูลซ้ำ, Outlier, Missing values

3.2.3 เลือกตัวแปรที่เหมาะสม โดยทำการแบ่งช่วงข้อมูลออกไม่เกิน 5 ส่วน (หรือน้อยกว่า ขึ้นอยู่กับการกระจายตัวของข้อมูล) จากตัวแปรอิสระทั้งตัวแปรต่อเนื่อง (Continuous variables) และตัวแปรแบ่งกลุ่ม (Categorical variables) เพื่อคำนวณหาค่าน้ำหนักของตัวแปรอิสระ (WOE) จากสัดส่วนของจำนวนเหตุการณ์ (event) ต่อจำนวนเหตุการณ์ทั้งหมด และสัดส่วนของจำนวนที่ไม่ใช่เหตุการณ์ (non-event) ต่อจำนวนที่ไม่ใช่เหตุการณ์ทั้งหมดในแต่ละกลุ่ม (bin) ดังสมการ

$$WOE = ln\ (\%\ of\ non\text{-}event\ /\ \%\ of\ event) \quad (1)$$

จากนั้นเลือกตัวแปรอิสระที่มีผลกับตัวแปรลูกค้าทุจริต โดยพิจารณาจากค่า Information Value (IV) [1] มากที่สุด ดังสมการ

$$IV = \sum (\%\ of\ non\text{-}event\ -\ \%\ of\ event)\ *\ WOE \quad (2)$$

จะได้ตัวแปรที่มีความสัมพันธ์กับตัวแปรลูกค้าทุจริต จำนวน 21 ตัวแปร เพื่อนำไปเปรียบเทียบประสิทธิภาพการแก้ไขปัญหาข้อมูลไม่สมดุล ดังตารางที่ 1

ตารางที่ 1. ตัวแปรอิสระที่มีค่า IV สูงสุด 10 อันดับ

Variables	IV	Predictive power
MTH_SINCE_LAST_COMMU	0.3949	Medium (>0.1-0.3)
RECENT_COMMU_TYPE	0.1908	Medium (>0.1-0.3)
AGE	0.1245	Medium (>0.1-0.3)
N_COMMU	0.1103	Medium (>0.1-0.3)
TENOR	0.1085	Medium (>0.1-0.3)
INCOME	0.1047	Medium (>0.1-0.3)
AUTOLOAN_LIMIT	0.0816	Weak (>0.02-0.1)
PREMIUM_LIFE_CS_AL	0.0581	Weak (>0.02-0.1)
OCCUPATION	0.0549	Weak (>0.02-0.1)
SECOND_PRODUCT	0.0536	Weak (>0.02-0.1)

ตารางที่ 2. ตัวอย่างผลลัพธ์ตัวแปร แสดงค่า WOE และ IV

AGE	EVENT COUNT	NON-EVENT COUNT	WOE	IV
AGE<=30	1,343	143,627	-0.7345	0.1245
31<AGE<=40	3,190	150,153	0.0861	0.1245
41<AGE<=51	3,103	140,201	0.1271	0.1245
52<AGE<=90	3,367	130,522	0.2803	0.1245
Total	11,003	564,503		

3.2.4 แก้ไขปัญหาข้อมูลไม่สมดุล โดยใช้วิธีการดังนี้

- วิธีสุ่มลดข้อมูลตัวอย่างในคลาสลบ (ลูกค้าไม่ทุจริต) โดยใช้เทคนิควิธี Random Undersampling [2-3] ในอัตราส่วน 5% เพื่อลดข้อมูลตัวอย่างกลุ่มมากให้มีสัดส่วนไม่ต่างจากข้อมูลตัวอย่างกลุ่มน้อยมากเกินไป ซึ่งวิธีการนี้ก็มีข้อเสีย

คือ การลดข้อมูลตัวอย่างของคลาสลบ อาจจะทำให้ข้อมูลที่สำคัญสูญหายจากชุดข้อมูลการสอนได้ [9] ดังตารางที่ 3

ตารางที่ 3. แสดงสัดส่วนของจำนวนลูกค้าทุจริตและไม่ทุจริต โดยใช้ Random Undersampling 5%

Target variable	Original		Undersampling 5%	
	Count	%	Count	%
Non-Fraud	4,100,049	99.7%	220,060	95%
Fraud	11,003	0.3%	11,003	5%
Total	4,111,052	100%	231,063	100%

แหล่งที่มา : ข้อมูลลูกค้ารายย่อยของธนาคาร (ก.ย. 2562)

- วิธีสุ่มเพิ่มข้อมูลตัวอย่างในคลาสบวก (ลูกค้าทุจริต) โดยใช้เทคนิควิธี SMOTE [4] และ Borderline-SMOTE [5-6] ในอัตราส่วน 50% หลังจากสุ่มลดข้อมูลตัวอย่างในคลาสลบ (ลูกค้าไม่ทุจริต) โดยใช้เทคนิควิธี Random Undersampling ในอัตราส่วน 5% เพื่อให้สัดส่วนของข้อมูลทั้ง 2 คลาส มีความสมดุลกันมากขึ้น ดังตารางที่ 4

ตารางที่ 4. แสดงสัดส่วนของจำนวนลูกค้าทุจริตและไม่ทุจริต โดยใช้ SMOTE และ Borderline-SMOTE 50%

Target variable	Undersampling 5%		SMOTE& Borderline-SMOTE 50%	
	Count	%	Count	%
Non-Fraud	220,060	95%	220,060	67%
Fraud	11,003	5%	110,030	33%
Total	231,063	100%	330,090	100%

แหล่งที่มา : ข้อมูลลูกค้ารายย่อยของธนาคาร (ก.ย. 2562)

- วิธีสุ่มลดข้อมูลตัวอย่างในคลาสลบ (ลูกค้าไม่ทุจริต) ควบคู่ไปกับสุ่มเพิ่มข้อมูลตัวอย่างในคลาสบวก (ลูกค้าทุจริต) โดยใช้เทคนิควิธี SMOTE-ENN [7-8] ในอัตราส่วน 50% หลังจากสุ่มลดข้อมูลตัวอย่างในคลาสลบ (ลูกค้าไม่ทุจริต) โดยใช้เทคนิควิธี Random Undersampling ในอัตราส่วน 5% เพื่อให้สัดส่วนของข้อมูลทั้ง 2 คลาส มีความสมดุลกันมากขึ้น ดังตารางที่ 5

ตารางที่ 5. แสดงสัดส่วนของจำนวนลูกค้าทุจริตและไม่ทุจริต โดยใช้ SMOTE-ENN 50%

Target variable	Undersampling 5%		SMOTE-ENN 50%	
	Count	%	Count	%
Non-Fraud	220,060	95%	161,286	68%
Fraud	11,003	5%	75,713	32%
Total	231,063	100%	236,999	100%

แหล่งที่มา : ข้อมูลลูกค้ารายย่อยของธนาคาร (ก.ย. 2562)

3.2.5 ทำการแบ่งชุดข้อมูลออกเป็นสองส่วนคือ ชุดข้อมูลสำหรับฝึกสอน (Training data) 70% และชุดข้อมูลสำหรับทดสอบ (Test data) 30%

3.2.6 เปรียบเทียบวิธีการเลือกตัวแปรกับการแก้ไขปัญหาข้อมูลไม่สมดุล สำหรับจำแนกประเภทลูกค้าทุจริตและไม่ทุจริต ด้วยเทคนิค Logistic Regression และ Decision Tree เพื่อหาแบบจำลองที่ให้ค่าความแม่นยำสูงที่สุด

3.2.7 ทดสอบประสิทธิภาพแบบจำลองโดยการใช้ชุดข้อมูลสัดส่วนเดิมของลูกค้าทุจริตและไม่ทุจริต เพื่อเปรียบเทียบค่าความแม่นยำ (Accuracy), ค่าความเที่ยง (Precision), ค่าระลึก (Recall) และ ค่าพื้นที่ใต้เส้นโค้ง (ROC AUC)

4. ผลการวิจัย

งานวิจัยนี้มีการเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปรโดยให้ค่าน้ำหนักของตัวแปรอิสระ กับเทคนิควิธีการแก้ไขปัญหาข้อมูลไม่สมดุล สำหรับการจำแนกประเภทลูกค้าทุจริตในธนาคาร โดยวัดผลจากค่าความแม่นยำ (Accuracy), ค่าความเที่ยง (Precision), ค่าระลึก (Recall) และ ค่าพื้นที่ใต้เส้นโค้ง (ROC AUC) [10]

4.1 เลือกตัวแปรอิสระที่มีผลกับตัวแปรลูกค้าทุจริต โดยไม่ให้ค่าน้ำหนักของตัวแปรอิสระ เปรียบเทียบเทคนิควิธีการแก้ไขปัญหาข้อมูลไม่สมดุล 4 วิธี คือ (1) RUS (5%) (2) RUS (5%) + SMOTE (50%) (3) RUS (5%) + Borderline-SMOTE (50%) และ (4) RUS (5%) + SMOTE-ENN (50%) กับเทคนิคการจำแนกประเภทลูกค้าทุจริต Logistic Regression ดังตารางที่ 6

ตารางที่ 6. แสดงประสิทธิภาพการคัดเลือกตัวแปรไม่ใช้ WOE กับการสุ่มตัวอย่างและเทคนิคการจำแนกประเภท Logistic regression

Sampling Method	Accuracy	Precision		Recall		F-measure		ROC AUC
		Non-Fraud	Fraud	Non-Fraud	Fraud	Non-Fraud	Fraud	
(1)	0.74	0.99	0.04	0.74	0.56	0.85	0.08	0.73
(2)	0.79	0.99	0.05	0.79	0.50	0.88	0.08	0.70
(3)	0.30	0.99	0.02	0.29	0.84	0.45	0.04	0.61
(4)	0.28	0.99	0.02	0.26	0.86	0.42	0.04	0.63

จากตารางที่ 6 แสดงให้เห็นว่าประสิทธิภาพการคัดเลือกตัวแปรแบบไม่ให้ค่าน้ำหนักของตัวแปรอิสระ ด้วยเทคนิคการจำแนกประเภทลูกค้าทุจริต Logistic Regression กับวิธีการสุ่มตัวอย่าง RUS (5%) + Borderline-SMOTE และ RUS (5%) + SMOTE-ENN (50%) เมื่อพิจารณาจากการทำนายกลุ่มลูกค้าทุจริต ให้ค่า Recall สูง คือ 0.84 และ 0.86 ตามลำดับ แต่ถ้าพิจารณาจากค่า Recall ในการทำนายกลุ่มลูกค้าปกติ จะมีค่าค่อนข้างน้อย ซึ่งมีผลทำให้ค่าความแม่นยำ (Accuracy) น้อยที่สุดคือ 0.30 และ 0.28 ตามลำดับ ดังนั้นเมื่อพิจารณาจากการทำนายกลุ่มลูกค้าทุจริต ค่า Recall, Accuracy และ ROC AUC ของแบบจำลอง แสดงให้เห็นว่าวิธีการสุ่มตัวอย่างแบบ RUS (5%) และ RUS (5%) + SMOTE (50%) จะสามารถวัดความถูกต้องของแบบจำลองในการจำแนกประเภทกลุ่มลูกค้าทุจริตได้ดีกว่าวิธีอื่น

4.2 เลือกตัวแปรอิสระที่มีผลกับตัวแปรลูกค้าทุจริต โดยให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) เปรียบเทียบเทคนิควิธีการแก้ไขปัญหาข้อมูลไม่สมดุล 4 วิธี คือ (1) RUS (5%) (2) RUS (5%) + SMOTE (50%) (3) RUS (5%) + Borderline-SMOTE (50%) และ (4) RUS (5%) + SMOTE-ENN (50%) กับเทคนิคการจำแนกประเภทลูกค้าทุจริต Logistic Regression ดังตารางที่ 7

ตารางที่ 7. แสดงประสิทธิภาพการคัดเลือกตัวแปร WOE กับการสุ่มตัวอย่างและเทคนิคการจำแนกประเภท Logistic Regression

Sampling Method	Accuracy	Precision		Recall		F-measure		ROC AUC
		Non-Fraud	Fraud	Non-Fraud	Fraud	Non-Fraud	Fraud	
(1)	0.74	0.99	0.04	0.74	0.60	0.85	0.08	0.76
(2)	0.74	0.99	0.04	0.74	0.60	0.88	0.08	0.76
(3)	0.77	0.99	0.05	0.77	0.57	0.87	0.09	0.75
(4)	0.75	0.99	0.05	0.76	0.58	0.86	0.08	0.74

จากตารางที่ 7 แสดงให้เห็นว่าประสิทธิภาพการคัดเลือกตัวแปรแบบให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) ด้วยเทคนิคการจำแนกประเภทลูกค้าทุจริต Logistic Regression กับวิธีการสุ่มตัวอย่างทั้ง 4 วิธี ให้ผลการทดสอบใกล้เคียงกัน เมื่อพิจารณาจากการทำนายกลุ่มลูกค้าทุจริต แสดงให้เห็นว่าวิธีการสุ่มแบบ RUS (5%) + SMOTE (50%) จะให้ค่า Recall สูงสุด คือ 0.60 ค่า Accuracy 0.74 และค่า ROC AUC สูงสุดคือ 0.76 รวมถึงสัดส่วนของข้อมูลตัวอย่างในคลาสบวก (ลูกค้าทุจริต) และสัดส่วนของข้อมูลตัวอย่างในคลาสลบ (ลูกค้าไม่ทุจริต) มีความสมดุลกันมากขึ้น ซึ่งเมื่อเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปร โดยพิจารณาจากค่า Recall, Accuracy และ ROC AUC ของการทำนายกลุ่มลูกค้าทุจริตทั้ง 4 วิธีการสุ่มตัวอย่าง จะเห็นว่าประสิทธิภาพของตัวแปรที่ให้ค่าน้ำหนักของตัวแปรอิสระจะสามารถทำนายกลุ่มลูกค้าทุจริตได้ดีกว่าวิธีการคัดเลือกตัวแปรแบบไม่ให้ค่าน้ำหนักของตัวแปรอิสระ

4.3 เลือกตัวแปรอิสระที่มีผลกับตัวแปรลูกค้าทุจริต โดยไม่ให้ค่าน้ำหนักของตัวแปรอิสระ เปรียบเทียบเทคนิควิธีการแก้ไขปัญหาข้อมูลไม่สมดุล 4 วิธี คือ (1) RUS (5%) (2) RUS (5%) + SMOTE (50%) (3) RUS (5%) + Borderline-SMOTE (50%) และ (4) RUS (5%) + SMOTE-ENN (50%) กับเทคนิคการจำแนกประเภทลูกค้าทุจริต Decision Tree ดังตารางที่ 8

ตารางที่ 8. แสดงประสิทธิภาพการคัดเลือกตัวแปรไม่ใช้ WOE กับการสุ่มตัวอย่างและเทคนิคการจำแนกประเภท Decision Tree

Sampling Method	Accuracy	Precision		Recall		F-measure		ROC AUC
		Non-Fraud	Fraud	Non-Fraud	Fraud	Non-Fraud	Fraud	
(1)	0.73	0.99	0.05	0.73	0.74	0.84	0.10	0.81
(2)	0.89	0.99	0.07	0.90	0.38	0.94	0.12	0.77
(3)	0.88	0.99	0.07	0.89	0.40	0.93	0.11	0.77
(4)	0.89	0.99	0.08	0.90	0.41	0.94	0.13	0.78

จากตารางที่ 8 แสดงให้เห็นว่าประสิทธิภาพการคัดเลือกตัวแปรแบบไม่ให้ค่าน้ำหนักของตัวแปรอิสระ ด้วยเทคนิคการจำแนกประเภทลูกค้าทุจริต Decision Tree กับวิธีการสุ่มตัวอย่างแบบ RUS (5%) เมื่อพิจารณาจากการทำนายกลุ่มลูกค้าทุจริต ให้ค่า Recall สูงสุดคือ 0.74 และค่า ROC AUC สูงสุดคือ 0.81 ถึงแม้ว่าค่า Accuracy จะมีค่าน้อยที่สุด หมายความว่าวิธีการสุ่มตัวอย่างแบบ RUS (5%) มีประสิทธิภาพในการทำนายกลุ่มลูกค้าทุจริตได้ดีกว่าวิธีการสุ่มตัวอย่างแบบอื่น

4.4 เลือกตัวแปรอิสระที่มีผลกับตัวแปรลูกค้าทุจริต โดยให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) เปรียบเทียบเทคนิควิธีการแก้ไขปัญหาข้อมูลไม่สมดุล 4 วิธี คือ (1) RUS (5%) (2) RUS (5%) + SMOTE (50%) (3) RUS (5%) + Borderline-SMOTE (50%) และ (4) RUS (5%) + SMOTE-ENN (50%) กับเทคนิคการจำแนกประเภทลูกค้าทุจริต Decision Tree ดังตารางที่ 9

ตารางที่ 9. แสดงประสิทธิภาพการคัดเลือกตัวแปร WOE กับการสุ่มตัวอย่างและเทคนิคการจำแนกประเภท Decision Tree

Sampling Method	Accuracy	Precision		Recall		F-measure		ROC AUC
		Non-Fraud	Fraud	Non-Fraud	Fraud	Non-Fraud	Fraud	
(1)	0.73	0.99	0.05	0.73	0.67	0.84	0.09	0.77
(2)	0.88	0.99	0.07	0.89	0.39	0.94	0.12	0.74
(3)	0.88	0.99	0.07	0.89	0.41	0.93	0.12	0.73
(4)	0.92	0.99	0.09	0.93	0.34	0.94	0.14	0.70

จากตารางที่ 9 แสดงให้เห็นว่าประสิทธิภาพการคัดเลือกตัวแปรแบบให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) ด้วยเทคนิคการจำแนกประเภทลูกค้าทุจริต Decision Tree กับวิธีการสุ่มตัวอย่างแบบ RUS (5%) ให้ประสิทธิภาพในการทำนายกลุ่มลูกค้าทุจริตได้ดีที่สุด เมื่อพิจารณาจากค่า Recall สูงสุดคือ 0.67 และค่า ROC AUC สูงสุดคือ 0.77 ซึ่งเมื่อเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปรกับเทคนิคการจำแนกประเภทลูกค้าทุจริต Decision Tree โดยพิจารณาจากค่า Recall, Accuracy และ ROC AUC ของการทำนายกลุ่มลูกค้าทุจริตทั้ง 4 วิธีการสุ่มตัวอย่าง จะเห็นว่าประสิทธิภาพของการคัดเลือกตัวแปรแบบไม่ให้ค่าน้ำหนักของตัวแปรอิสระกับการคัดเลือกตัวแปรแบบให้ค่าน้ำหนักของตัวแปรอิสระ ให้ผลการทดสอบที่ใกล้เคียงกัน ซึ่งการคัดเลือกตัวแปรแบบไม่ให้ค่าน้ำหนัก จะให้ผลได้ดีกว่าเล็กน้อย จากค่า Recall ที่สามารถทำนายกลุ่มลูกค้าทุจริตได้ถูกต้องแม่นยำกว่า

5. สรุปและอภิปรายผล

จากผลการเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปร โดยให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) กับเทคนิควิธีการสุ่มตัวอย่างแบบ RUS + SMOTE สำหรับการจำแนกประเภทลูกค้าทุจริตในธนาคาร ด้วยเทคนิค Logistic Regression เมื่อพิจารณาจากค่า Recall, Accuracy และ ROC AUC ของการทำนายกลุ่มลูกค้าทุจริต คือ 0.60, 0.74 และ 0.76 ตามลำดับ ในขณะที่ผลการเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปร โดยให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) กับเทคนิควิธีการสุ่มตัวอย่างแบบ RUS (5%) สำหรับการจำแนกประเภทลูกค้าทุจริตในธนาคาร ด้วยเทคนิค Decision Tree เมื่อพิจารณาจากค่า Recall, Accuracy และ ROC AUC

ของการทำนายกลุ่มลูกค้าทุจริต คือ 0.67, 0.73 และ 0.77 ตามลำดับ ดังนั้นวิธีการคัดเลือกตัวแปร โดยให้ค่าน้ำหนักของตัวแปรอิสระ (WOE) กับเทคนิควิธีการสุ่มตัวอย่างแบบ RUS + SMOTE สำหรับการจำแนกประเภทลูกค้าทุจริตในธนาคาร ด้วยเทคนิค Logistic Regression จะให้ประสิทธิภาพในการจำแนกประเภทกลุ่มลูกค้าทุจริตได้ดีที่สุด

กิตติกรรมประกาศ

ขอขอบพระคุณ รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย อาจารย์ที่ปรึกษาค้นคว้าอิสระ ที่ได้กรุณาให้คำแนะนำ ช่วยเหลือ ให้คำปรึกษา และชี้แนะการแก้ไขปัญหาต่างๆ จนทำให้งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี

เอกสารอ้างอิง

[1] Deepanshu Bhalla, 2015. WEIGHT OF EVIDENCE (WOE) AND INFORMATION VALUE (IV) EXPLAINED. [Online]. Available : https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html [2022, 27 July]

[2] Paula Branco, Luis Torgo, and Rita Ribeiro, “A Survey of Predictive Modelling under Imbalanced Distributions,” Universidade do Porto, 2015.

[3] Chawla, N. V., Japkowicz, N., and Kotcz, A., “Editorial: special issue on learning from imbalanced data sets,” ACM SIGKDD Explorations Newsletter, vol 6, issue 1, 2004, pp. 1–6.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmayer, “SMOTE: Synthetic Minority Over-Sampling Technique,” Journal of Artificial Intelligent Research, vol 16, 2002, pp. 321–357.

[5] H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced datasets Learning,” Lecture Notes in Computer Science, vol 3644, 2005, pp. 878–887.

[6] Nguyen Hien M., Cooper Eric W. and Kamei Katsuari, “Borderline Over-sampling for Imbalanced Data Classification,” International Journal of Knowledge Engineering and Soft Data Paradigms, vol 3, issue 1, 2011, pp. 4-21.

[7] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data,” ACM SIGKDD Explorations Newsletter, vol 6, 2004, pp. 20-29.

[8] Farzana Anowar, and Samira Sadaoui, “Detection of Auction Fraud in Commercial Sites,” Journal of Theoretical and Applied Electronic Commerce Research ISSN 0718–1876 Electronic Version, vol 15, issue 1, 2020, pp. 81-98.

[9] วันทนี ประจวบศุภกิจ, “วิธีการปรับข้อมูลตัวอย่างแบบผสมผสานเพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลที่มีจำนวนตัวอย่างในแต่ละคลาสไม่สมดุลกัน,” Science and Technology RMUTT Journal vol 8, no. 2, 2018, pp. 125-142.

[10] F. J. Provost, and T. Fawcett, “Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions,” In Knowledge Discovery and Data Mining, 1997, pp. 43–48.