การจัดการข้อมูลไม่สมดุลของการทำกลยุทธ์เสนอขายประกันต่อยอดสำหรับผู้ถือบัตรเครดิต

กิตติภพ แซ่เตีย[1] และ จิรภัทร์ หยกรัตนศักดิ์[1*]

[1]ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ
* jiraphat.yo@kmitl.ac.th

บทคัดย่อ

ปัจจุบันการเสนอขายกรมธรรม์ประกันชีวิตต่อยอด (Cross-sell) สำหรับผู้ถือบัตรเครดิตของธนาคารต่าง ๆ ได้รับความนิยมอย่างแพร่หลาย แต่ผลตอบรับในการเสนอขายกรมธรรม์ประกันชีวิตนั้นมีการตอบรับเอนเอียงไปทางปฏิเสธการขาย จึงทำให้เกิดความไม่สมดุลของข้อมูลในการเสนอขาย ดังนั้นงานวิจัยฉบับนี้จึงจัดทำขึ้นโดยมีวัตถุประสงค์จัดการข้อมูลที่ไม่สมดุลของผลการตอบรับ เพื่อการนำไปใช้งานสร้างแบบจำลองทำนายอัตราตอบรับการเสนอขายกรมธรรม์ ทั้งนี้ ผู้วิจัยได้สร้างชุดข้อมูลจำลองประกอบไปด้วยตัวแปรต่าง ๆ ของลูกค้าแต่ละคนพร้อมทั้งคำตอบในการตอบรับ โดยให้คำตอบในการตอบรับมีสัดส่วนไม่สมดุล (Imbalanced Data) คล้ายกับข้อมูลจริง แล้วได้ใช้เทคนิคการสังเคราะห์ข้อมูลด้วยวิธีการสุ่มลด (Under Sampling) วิธีการสุ่มเกิน (Over Sampling) และวิธีสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Over-sampling Technique: SMOTE) มาจัดการชุดข้อมูล แล้วใช้อัลกอริทึมช่วยเลือกปัจจัยที่เหมาะสม และใช้แบบจำลองถดถอยในการทำนายเพื่อเปรียบเทียบประสิทธิภาพเทคนิคการจัดการข้อมูล ผลการวิจัยพบว่า วิธีสังเคราะห์ข้อมูลเพิ่ม SMOTE มีประสิทธิภาพเหมาะสมกับการแก้ปัญหาข้อมูลไม่สมดุลของการเสนอขายกรมธรรม์

คำสำคัญ: กรมธรรม์ประกันชีวิต บัตรเครดิต ข้อมูลไม่สมดุล วิธีการสุ่ม

Managing the Imbalanced Data of Cross-sell Strategy for Credit Card Holders

Kittipob Saetia[1] and Jiraphat Yokrattanasak[1*]

[1] Department of Mathematics, Faculty of Science,
King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520
* Corresponding Author; email: jiraphat.yo@kmitl.ac.th

Abstract

Cross-selling insurance products, which is part of banking activities, has grown in importance in today's banking industry. Because of the negative customer responses, highly imbalanced data is taken into account. The study introduces the concept of management imbalanced data and a forecasting model for predicting customer satisfaction response rates, including with specific data. In this study, a simulated dataset was created that replicated the statistical properties of the original data. The data is then adjusted using Under Sampling, Over Sampling, and the Synthetic Minority Over-sampling Technique (SMOTE). In addition, an algorithm is used to select the precise factor. Furthermore, regression analysis is used to forecast the effectiveness of data analysis. This study demonstrates that the Synthetic Minority Over-sampling Technique (SMOTE) is an effective method for improving the imbalanced cross-selling insurance data.

Keywords: Life Insurance Policy, Credit Card, Imbalanced Data, Data Sampling

1. บทนำ

ในปัจจุบันนี้ กรมธรรม์ประกันชีวิตเป็นผลิตภัณฑ์การลงทุนรูปแบบหนึ่งที่ทุกคนควรให้ความสำคัญ ซึ่งถือเป็นการวางแผนระยะยาวอย่างหนึ่งที่ช่วยให้ลูกค้าลดความกังวลด้านค่าใช้จ่าย เมื่อต้องเข้ารับการรักษาพยาบาลเมื่อเกิดอุบัติเหตุหรือเหตุการณ์ไม่คาดคิดในอนาคต และเมื่อความต้องการของอุปสงค์ในกรมธรรม์มีมากขึ้น อาชีพและบริการเสนอขายกรมธรรม์ประกันชีวิตจึงเกิดขึ้นตามมา ซึ่งเจ้าหน้าที่ตัวแทนที่มีหน้าที่เสนอขายกรมธรรม์มักประสบปัญหาในการทำงาน เนื่องจากอัตราการตอบรับการเสนอขายกรมธรรม์ค่อนข้างต่ำ ดังนั้น เจ้าหน้าที่จึงต้องใช้เทคนิคการเสนอขายกรมธรรม์ โดยใช้เทคนิค 9 เทคนิค ดังนี้ 1.การขายด้วยความจริงใจ 2.การเตรียมตัวให้พร้อม 3.ความมั่นใจในความรู้ 4.เป็นผู้ฟังที่ดี 5.ตอบได้ทุกคำถามในเรื่องประกัน 6.มีศิลปะในการพูด 7.ให้ข้อเท็จจริง 8.ขายกรมธรรม์ด้วยความเข้าใจ และ 9.ทำให้การขายนั้นเป็นที่น่าสนใจ โดยการเสนอขายกรมธรรม์จะทำให้ประสบความสำเร็จได้ไม่ขึ้นอยู่แค่เทคนิคการเสนอขายเพียงอย่างเดียว ยังต้องเสนอขายให้ถูกกลุ่ม เวลา และ สถานที่ด้วย ซึ่งการเสนอขายกรมธรรม์ ณ ปัจจุบันมีประสิทธิภาพที่ค่อนข้างต่ำเนื่องจากการเสนอขายไม่ถูกกลุ่ม เวลา และสถานที่ งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพการเสนอขายกรมธรรม์ให้ได้รับผลตอบรับมากขึ้น โดยจะใช้เทคนิคการเตรียมข้อมูล การสำรวจ

ข้อมูล และการสร้างแบบจำลองในการเสนอขายกรมธรรม์ประกันชีวิต ซึ่งในส่วนข้อมูลการเสนอขายกรมธรรม์ เป็นข้อมูลการเสนอขายประกันชีวิตต่อยอด (Cross-sell) สำหรับผู้ถือบัตรเครดิตของธนาคาร ลักษณะข้อมูลมีอัตราการตอบรับการเสนอขายกรมธรรม์ค่อนข้างต่ำ เป็นข้อมูลที่มีความเอนเอียงไปทางไม่ตอบรับ ส่งผลให้เกิดความไม่สมดุลของข้อมูล

การแก้ปัญหาในระดับข้อมูล เป็นการแก้ไขในขั้นต้นก่อนที่จะมีการประมวลผล (Processing Stage) ซึ่งเป็นการแก้ไขที่ข้อมูลโดยตรง ทำการปรับปรุงข้อมูลที่มีความไม่สมดุลให้กลายเป็นข้อมูลที่มีความสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล (Data Sampling Technique) หรือเทคนิคการเลือกข้อมูล (Data Selection Technique) ต่าง ๆ เพื่อเพิ่มประสิทธิภาพสูงสุดของการนำข้อมูลไปใช้ ในงานวิจัยนี้ประยุกต์ใช้ 3 วิธี มาแก้ปัญหาความไม่สมดุลของข้อมูล คือ 1.วิธีการสุ่มลด (Under Sampling) เป็นเทคนิคหรือวิธีที่ใช้ในการลดจำนวนข้อมูลที่อยู่ในคลาสส่วนมาก (จำนวนข้อมูลการปฏิเสธ) ให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในคลาสส่วนน้อย (จำนวนข้อมูลการตอบรับ) 2.วิธีสุ่มเกิน (Over Sampling) เป็นเทคนิคหรือวิธีที่ใช้ในการเพิ่มข้อมูลที่อยู่ในคลาสส่วนน้อย (จำนวนข้อมูลการตอบรับ) ให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในคลาสส่วนมาก (จำนวนข้อมูลการปฏิเสธ) โดยการสร้างข้อมูลขึ้นมาใหม่ด้วยการสุ่มเลือกข้อมูลจากข้อมูลเดิม (Drummord and Holte, 2003) 3.วิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE) จะสร้างข้อมูลสังเคราะห์ให้อยู่ในขอบเขตของข้อมูลในคลาสส่วนน้อย (จำนวนข้อมูลการตอบรับ) ด้วยการสังเคราะห์ข้อมูลนั้นให้อยู่ระหว่างข้อมูลคลาสส่วนน้อยตัวที่ *n* ไปยังข้อมูลตัวที่ใกล้เคียง โดย Chawla et. al. (2002) เสนอวิธีการแก้ไขปัญหาชุดข้อมูลไม่สมดุลด้วยเทคนิคการสุ่มตัวอย่างแบบกลุ่มน้อยสังเคราะห์ (Synthetic Minority Over-sampling Technique : SMOTE) โดยใช้แบบจำลอง C4.5, Ripper และ Naive Bayes Classifier และวัดประสิทธิภาพโดยพื้นที่ใต้เส้นโค้ง ROC (Area Under the Receiver Operating Characteristic Curve) เพื่อพิสูจน์การแก้ไขปัญหาการจัดการของข้อมูลไม่สมดุล วิชญ์วิสิฐ เกสรสิทธิ์ และคณะ (2561) ได้ศึกษาการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลผู้ป่วยโรคเบาหวาน พิจารณาการแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล 4 วิธี คือ วิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE) โดยใช้เทคนิคการจำแนก คือ วิธีการถดถอยโลจิสติกแบบมัลติโนเมียลและวิธีต้นไม้การตัดสินใจในการจำแนกผู้ป่วยโรคเบาหวาน จากการเปรียบเทียบประสิทธิภาพของสถิติและอัลกอริทึมในการจำแนก พบว่าข้อมูลที่แก้ปัญหาความไม่สมดุลด้วยวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE) สามารถจำแนกผู้ป่วยโรคเบาหวานด้วยวิธีต้นไม้การตัดสินใจมีผลลัพธ์ดีที่สุด

งานวิจัยนี้จะประยุกต์ใช้เทคนิคการเพิ่มประสิทธิภาพในการจัดการข้อมูลไม่สมดุลด้วยวิธีการสุ่มลด (Under Sampling) เพื่อสุ่มข้อมูลคลาสส่วนมากให้มีจำนวนลดลง วิธีการสุ่มเกิน (Over Sampling) เพื่อสุ่มข้อมูลคลาสส่วนน้อยขึ้นมาใหม่ให้มีจำนวนข้อมูลเพิ่มขึ้น และวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE) เพื่อสังเคราะห์ข้อมูลคลาสส่วนน้อยให้มีความหลากหลายมากขึ้น รวมถึงการใช้อัลกอริทึมเพื่อเลือกปัจจัยที่เหมาะสม และการใช้แบบจำลองในการทำนายเพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจัดการข้อมูลทั้งสามเทคนิค และเลือกวิธีแก้ปัญหาที่เหมาะสม

2. วัตถุประสงค์และขอบเขตการวิจัย

2.2.1 วัตถุประสงค์การวิจัย

1) มุ่งเน้นที่การแก้ไขและปรับปรุงข้อมูลไม่สมดุล (Imbalanced Data) เพื่อเพิ่มประสิทธิภาพแบบจำลองการพยากรณ์การเสนอขายประกันต่อยอดสำหรับผู้ถือบัตรเครดิต
2) เปรียบเทียบผลการแก้ข้อมูลไม่สมดุล (Imbalanced Data) ด้วย 3 เทคนิค คือ วิธีการสุ่มลด (Under Sampling) วิธีสุ่มเกิน (Over Sampling) และ วิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE)
3) นำเสนอขั้นตอนในการแก้ปัญหาข้อมูลไม่สมดุลของผลการตอบรับในการขายกรมธรรม์ประกันชีวิตต่อยอด

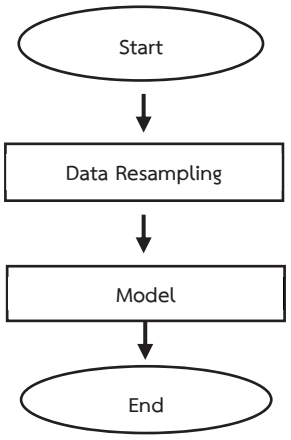
2.2.2 ขอบเขตการวิจัย

ในการศึกษาครั้งนี้เป็นการศึกษาข้อมูลการใช้งานบัตรเครดิตในช่วง 2 ปีย้อนหลัง ข้อมูลนี้ได้รับการสนับสนุนโดยบริษัท อยุธยา แคปปิตอล เซอร์วิสเซส จำกัด (AYCAP) แต่เนื่องจากข้อมูลของลูกค้าเป็นความลับของบริษัท ในงานวิจัยฉบับนี้ จะแสดงตัวอย่างการวิเคราะห์แก้ปัญหาข้อมูลไม่สมดุล โดยจำลองข้อมูลให้มีลักษณะเป็นข้อมูลแบบไม่สมดุลที่มีคำตอบในการตอบรับการขายกรมธรรม์ 2 แบบ คือ ตอบรับและปฏิเสธโดยจำลองให้คล้ายคลึงกับข้อมูลจริง โดยจำลองข้อมูลมาทั้งหมด 12,000 ข้อมูล และข้อมูลจำลองแบ่งออกเป็น 3 ชุดข้อมูล ดังนี้ ชุดที่ 1 ชุดข้อมูลการเรียนรู้ (Training Dataset) มีจำนวน 10,000 ข้อมูล (84% ของข้อมูลทั้งหมด) ชุดที่ 2 ชุดข้อมูลทดสอบ (Testing Dataset) จำนวน 2,000 ข้อมูล (16% ของข้อมูลทั้งหมด) และชุดที่ 3 ชุดข้อมูลตรวจสอบ (Validation Dataset) จำนวน 3,000 ข้อมูล (30% ของชุดข้อมูลการเรียนรู้)

3. วิธีดำเนินการวิจัย

3.1 แผนผังขั้นตอนการดำเนินงาน (Process Diagram)

ขั้นตอนการดำเนินงานวิจัย มี 2 ขั้นตอน ดังนี้ การสังเคราะห์ข้อมูล (Data Resampling) และการสร้างแบบจำลอง (Model) ซึ่งผังงานขั้นตอนการดำเนินงานวิจัยแสดงไว้ในภาพที่ 1



ภาพที่ 1 แผนผังขั้นตอนการดำเนินงาน (Process Diagram)

3.2 การสังเคราะห์ข้อมูล (Data Resampling)

การสังเคราะห์ข้อมูล เป็นการแก้ปัญหาในระดับข้อมูลขั้นต้นก่อนที่จะมีการประมวลผล (Processing Stage) โดยการแก้ไขนี้จะแก้ไขข้อมูลโดยตรง ทำการปรับปรุงข้อมูลที่มีความไม่สมดุลให้กลายเป็นข้อมูลที่มีความสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล (Data Sampling Technique) หรือเทคนิคการเลือกข้อมูล (Data Selection Technique) ในงานวิจัยนี้ได้พิจารณาใช้เทคนิค 3 แบบ คือ วิธีการสุ่มลด (Under Sampling) วิธีการสุ่มเกิน (Over Sampling) และวิธีการสังเคราะห์ข้อมูลเพิ่ม (SMOTE) และนำไปเปรียบเทียบประสิทธิภาพของแต่ละวิธี

3.2.1 วิธีการสุ่มลด (Under Sampling)

วิธีสุ่มลด (Under Sampling) เป็นเทคนิคหรือวิธีที่ใช้ในการลดจำนวนข้อมูลที่อยู่ในคลาสส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในคลาสส่วนน้อย

3.2.2 วิธีการสุ่มเกิน (Over Sampling)

วิธีสุ่มเกิน(Over Sampling) เป็นเทคนิคหรือวิธีที่ใช้ในการเพิ่มข้อมูลที่อยู่ในคลาสส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลในคลาสส่วนมาก โดยการสุ่มเลือกข้อมูลจากข้อมูลเดิม (Drummord and Holte, 2003)

3.2.3 การสังเคราะห์ข้อมูลเพิ่ม (SMOTE)

วิธีการสุ่มเกินด้วย SMOTE จะสร้างข้อมูลสังเคราะห์ให้อยู่ในขอบเขตของข้อมูลในคลาสส่วนน้อย โดยการสังเคราะห์ข้อมูลนั้นจะอยู่ระหว่างข้อมูลคลาสส่วนน้อยตัวที่ *n* ไปยังข้อมูลตัวที่ใกล้เคียง สังเคราะห์ข้อมูลให้อยู่ในช่วงนั้น และจะทำให้ได้ข้อมูลใหม่ที่ใกล้เคียงกับข้อมูลเก่าในคลาสส่วนน้อยดังสมการที่ 1 (Chawla et al, 2002)

$$\begin{aligned} c_1 &= a_1 + (unif(0,1) \times (b_1 - a_1)) \\ c_2 &= a_2 + (unif(0,1) \times (b_2 - a_2)) \\ &\mathrm{M} \\ c_n &= a_n + (unif(0,1) \times (b_n - a_n)) \end{aligned} \qquad (1)$$

โดยที่

$c_1, c_2, c_3, ..., c_n$ คือ ข้อมูลในค่าสังเกตที่จุด $\mathrm{N}_{\mathrm{point}}$ ของข้อมูลของคลาสส่วนน้อย (จำนวนข้อมูลการตอบรับ) ที่สร้างขึ้นมาใหม่

$a_1, a_2, a_3, ..., a_n$ คือ ข้อมูลในค่าสังเกตที่จุด $\mathrm{O}_{\mathrm{point}}$ ของข้อมูลของคลาสส่วนน้อย (จำนวนข้อมูลการตอบรับ) ที่นำไปใช้เป็นตัวตั้งต้น

$b_1, b_2, b_3, ..., b_n$ คือ ข้อมูลในค่าสังเกตที่จุด $\mathrm{M}_{\mathrm{point}}$ ของข้อมูลของคลาสส่วนน้อย (จำนวนข้อมูลการตอบรับ) ที่นำไปใช้เป็นจุดใกล้เคียงกับจุดตั้งต้น

$unif(0,1)$ คือ ตัวแปรสุ่มแบบยูนิฟอร์ม (Uniform) โดยการสุ่มค่าระหว่าง 0 ถึง 1

$b_n - a_n$ คือ ระยะห่างระหว่างจุดตั้งต้นกับจุดใกล้เคียงในแอตทริบิวต์ 1, 2 , 3, ... , n

3.3 การสร้างแบบจำลอง (Modeling)

3.3.1 แบบจำลองการถดถอยโลจิสติค (Logistic Regression Model)

แบบจำลองการถดถอยโลจิสติค (Peng et al., 2002) ถูกจัดเป็นแบบจำลองประเภทการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งเป็นแบบจำลองที่จะนำมาใช้ทำนายความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจจากชุดข้อมูลของตัวแปรอิสระ (Independent Variables) ที่เหมาะสม และเป็นแบบจำลองที่มีพื้นฐานการสร้างมาจากสมการทางคณิตศาสตร์ซึ่งแบ่งตามชนิดของตัวแปรอิสระได้ 2 ประเภท คือ แบบจำลองการถดถอยตัวแปรเดียว (Simple Logistic Regression) มีตัวแปรอิสระเพียงตัวเดียว และแบบจำลองการถดถอยหลายตัวแปร (Multiple Logistic Regression) มีตัวแปรอิสระหลายตัว โดยมีการนำมาประยุกต์ใช้อย่างแพร่หลาย อาทิเช่น การทำนายโอกาสที่ผู้บริโภคจะเกิดหนี้เสียให้กับบริษัทของธุรกิจประเภทธนาคาร การทำนายโอกาสที่ผู้บริโภคจะย้ายเครือข่ายของธุรกิจประเภทโทรศัพท์ การใช้ความสามารถของแบบจำลองในการประยุกต์ทางการตลาด พยากรณ์ส่วนแบ่งการตลาด (Market Segmentation) เพื่อเพิ่มโอกาสในการเสนอแคมเปญ (campaign) ให้กับส่วนของตลาดใน

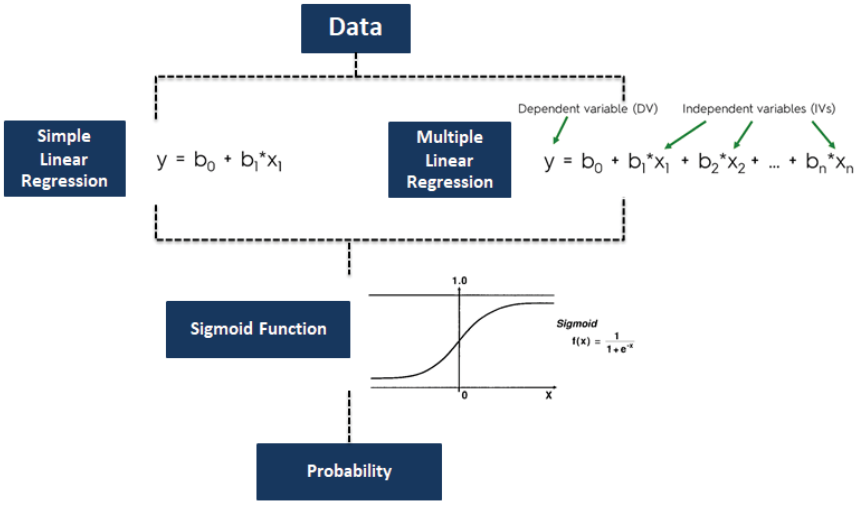
แต่ละส่วนได้

ขั้นตอนในการสร้างแบบจำลองหลักแสดงดังภาพที่ 2 และมีรายละเอียดดังนี้

1. การนำชุดข้อมูลมาสร้างสมการถดถอยตัวแปรเดียว (Simple Linear Regression) หรือสมการถดถอยหลายตัวแปร (Multiple Linear Regression) โดยขึ้นอยู่กับตัวแปรอิสระที่ถูกนำมาใช้ในประเภทของงานที่ทำ

2. นำสมการถดถอยที่ได้ไปเข้าสู่ฟังก์ชันซิกมอยด์ (Sigmoid) เพื่อปรับค่าที่ได้ให้อยู่ในช่วง 0-1 เนื่องจากสมการถดถอยที่ได้อาจมีค่ามากกว่า 1 หรือน้อยกว่า 0 ได้ ซึ่งกฎความน่าจะเป็นจะต้องอยู่ระหว่าง 0-1 เท่านั้น จึงได้มีการนำฟังก์ชันนี้มาใช้งาน

3. เมื่อผ่านฟังก์ชันซิกมอยด์จะได้ความน่าจะเป็นของเหตุการณ์ที่สนใจ



ภาพที่ 2 แสดงแผนภาพขั้นตอนการสร้างแบบจำลอง Logistic Regression

3.3.2 การประเมินผล (Evaluation)

ในการเรียนรู้ของเครื่องจักร (Machine Learning) การวัดประสิทธิภาพเป็นงานที่สำคัญ เมื่อต้องการตรวจสอบหรือแสดงภาพประสิทธิภาพของปัญหาการจำแนกประเภท (Classification) สามารถวัดได้จากพื้นที่ใต้เส้นโค้ง ROC (Area under the Receiver Operating Characteristics Curve)

Confusion Matrix เป็นเครื่องมือสำคัญในการประเมินผลลัพธ์ของการทำนายจากแบบจำลองที่สร้างขึ้น โดยมีหลักการจากการวัดว่า แบบจำลองทำนายกับสิ่งที่เกิดขึ้นจริงมีสัดส่วนเป็นอย่างไร แสดงดังตารางที่ 1 ซึ่งบ่งบอกถึงการทำนายผิดถูกของแบบจำลองเป็น 4 กรณี

ตารางที่ 1 Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positive (TP)	False Positive (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)

โดย TP,TN,FP,FN ในตารางจะแทนด้วยค่าความถี่

True Positive (TP) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง ในกรณีทำนายว่า “จริง” และสิ่งที่เกิดขึ้น คือ “จริง”

True Negative (TN) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า “ไม่จริง” และสิ่งที่เกิดขึ้น คือ “ไม่จริง”

False Positive (FP) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า “จริง” แต่สิ่งที่เกิดขึ้น คือ “ไม่จริง”

False Negative (FN) คือ สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่า “ไม่จริง” แต่สิ่งที่เกิดขึ้น คือ “จริง”

True Positive Rate (TPR) คือ วัดสิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้นจริง ส่วนด้วย เหตุการณ์ที่เกิดขึ้นเป็นจริงทั้งหมด (True positive (TP) + False Negative (FN))

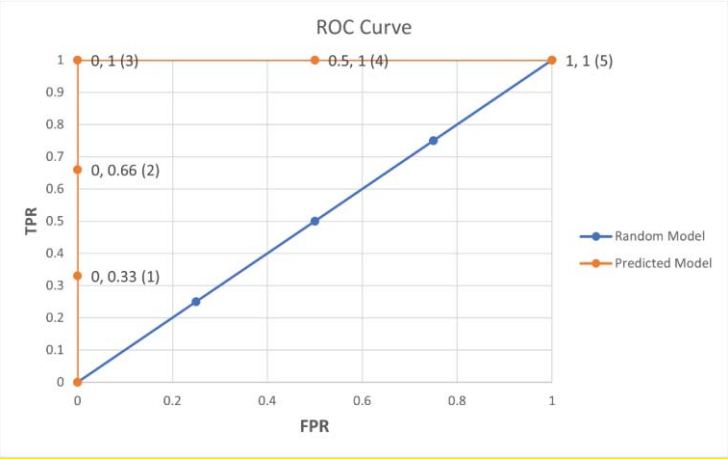
$$TPR=\frac{TP}{TP+FN} \quad (2)$$

False Positive Rate (FNR) คือ วัดสิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น ส่วนด้วย เหตุการณ์ที่เกิดขึ้นเป็นเท็จทั้งหมด (True Negative (TN) + False Positive (FP))

$$FPR=\frac{FP}{TN+FP} \quad (3)$$

เส้นโค้ง ROC (Receiver Operating Characteristic Curve) เป็นกราฟที่เกิดจากการพล็อตคู่อันดับ (TPR, FPR) ที่ได้มีการเปลี่ยนเกณฑ์คะแนนความน่าจะเป็นที่ไม่ซ้ำกัน สามารถดูเพิ่มเติมได้ในภาพที่ 3 ซึ่งแสดงตัวอย่างเส้นโค้ง ROC โดยมีการเปลี่ยนเกณฑ์คะแนนความน่าจะเป็นที่ไม่ซ้ำกัน 5 รายการ ดังนั้นจะมีคู่อันดับ (TPR, FPR) 5 คู่

จากภาพที่ 3 พื้นที่ที่ครอบคลุมโดยเส้นโค้งคือพื้นที่ระหว่างเส้นสีส้ม (ROC) และแกนนอน ยิ่งพื้นที่ครอบคลุมมากเท่าไร แบบจำลองยิ่งมีประสิทธิภาพดีขึ้น โดยค่าที่ดีที่สุดสำหรับพื้นที่ใต้เส้นโค้ง คือ 1



ภาพที่ 3 ภาพตัวอย่างเส้นโค้ง ROC (Receiver Operating Characteristic Curve)

ที่มา: https://towardsdatascience.com/performance-metrics-receiver-operating-characteristic-roc-area-under-curve-auc-79d6d5b0b977

4. ผลการวิจัย

เนื่องจากผลการตอบรับการเสนอขายกรมธรรม์มีการปฏิเสธ (False) ค่อนข้างสูงกว่าการตอบรับ (True) จึงทำให้เกิดข้อมูลไม่สมดุล ดังนั้นผู้วิจัยจึงได้นำวิธีการสังเคราะห์ข้อมูลมาเปรียบเทียบผลและเลือกวิธีการสังเคราะห์ข้อมูลที่เหมาะสม ซึ่งประกอบด้วย 3 เทคนิค ดังนี้ วิธีการสุ่มลด (Under Sampling) วิธีการสุ่มเกิน (Over Sampling) และวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE) และใช้ปัจจัยที่เหมาะสม เนื่องจากผลการศึกษาของข้อมูลจริงไม่สามารถนำมาเปิดเผยได้ทางผู้วิจัยจึงขอแสดงตัวอย่างของการวิเคราะห์ผลจากข้อมูลจำลองที่สร้างขึ้นให้มีค่าสถิติใกล้เคียงกับข้อมูลจริง โดยมีข้อมูลทั้งหมด 12,000 ข้อมูล เป็นจำนวนการปฏิเสธ 11,889 ข้อมูล และจำนวนการตอบรับ 111 ข้อมูล และข้อมูลจำลองแบ่งออกเป็น 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลการเรียนรู้ (Training Dataset) จำนวน 10,000 ข้อมูล โดยเป็นจำนวนการปฏิเสธ 9,909 ข้อมูล และจำนวนการตอบรับ 91 ข้อมูล ชุดข้อมูลตรวจสอบ (Validation Dataset) มีจำนวน 3,000 ข้อมูล โดยเป็นจำนวนการปฏิเสธ 2,976 ข้อมูล และจำนวนการตอบรับ 24 ข้อมูล และชุดข้อมูลทดสอบ (Testing Dataset) จำนวน 2,000 ข้อมูล โดยเป็นจำนวนการปฏิเสธ 1,980 ข้อมูล และจำนวนการตอบรับ 20 ข้อมูล

ตารางที่ 2 ผลการวิจัยด้วยวิธีการประเมินผลและการเปรียบเทียบประสิทธิภาพโดยใช้ Area Under Receiver Operating Characteristic Curve

Sampling Data	Area Under Receiver Operating Characteristic Curve			SD
	Training Dataset	Validation Dataset	Testing Dataset	
Under Sampling	0.7214	0.7441	0.7525	0.0131
Over Sampling	0.7358	0.7401	0.7458	0.0041
Over Sampling: SMOTE	0.7346	0.7393	0.7494	0.0062

ผลการศึกษาแสดงในตารางที่ 2 เมื่อพิจารณาค่าพื้นที่ใต้โค้ง ROC (Area Under Receiver Operating Characteristic Curve) ของแต่ละวิธีในการสังเคราะห์ข้อมูลและในแต่ละชุดข้อมูล ชุดข้อมูลการเรียนรู้ (Training Dataset) ชุดข้อมูลทดสอบ (Testing Dataset) และชุดข้อมูลตรวจสอบ (Validation Dataset) พบว่าวิธีการสุ่มลด (Under Sampling) ให้ค่าพื้นที่ใต้โค้ง ROC เป็น 0.7214, 0.7441 และ 0.7525 ตามลำดับ โดยให้ค่าส่วนเบี่ยงเบนมาตรฐาน (SD) เท่ากับ 0.0131 วิธีการสุ่มเกิน (Over Sampling) ให้ค่า พื้นที่ใต้โค้ง ROC เป็น 0.7358, 0.7401 และ 0.7485 ตามลำดับ โดยให้ค่าส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 0.0041 และวิธีสังเคราะห์ข้อมูลเพิ่ม (SMOTE) ให้ค่าพื้นที่ใต้โค้ง ROC เป็น 0.7346, 0.7393 และ 0.7494 ตามลำดับ โดยให้ค่าส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 0.0062

จากตารางที่ 2 สามารถสรุปได้ว่า การแก้ไขข้อมูลไม่สมดุลด้วยวิธีสุ่มเกินนั้นมีผลลัพธ์ดีที่สุด เนื่องจากมีค่าส่วนเบี่ยงเบนมาตรฐานที่น้อยที่สุด แสดงให้เห็นว่าข้อมูลที่ถูกแก้ไขแล้วทำให้ผลของแบบจำลองนั้นเมื่อทำนายในชุดข้อมูลที่แตกต่างกัน ค่าความคลาดเคลื่อนที่เกิดขึ้นก็จะไม่ห่างกันมากหรือไม่เกิด Overfitting มากเกินไป ข้อสรุปที่ได้เป็นผลลัพธ์จากชุดข้อมูลจำลองเท่านั้น ในส่วนของกรณีศึกษาของข้อมูลจริงที่ผู้วิจัยได้ดำเนินการวิเคราะห์ตามตัวอย่างข้างต้น ได้ข้อสรุปว่าวิธีสุ่ม SMOTE เป็นวิธีที่เหมาะสมที่สุดในการจัดการชุดข้อมูลไม่สมดุลของการขายประกันต่อยอด (Cross-sell) ของผู้ถือบัตรเครดิตของธนาคาร

5. สรุปผล อภิปรายผลการวิจัย และข้อเสนอแนะ

5.1 สรุปผลและอภิปรายผลการวิจัย

ผลจากการวิจัยแสดงถึงขั้นตอนในการแก้ปัญหาข้อมูลไม่สมดุลของผลการตอบรับในการขายกรมธรรม์ประกันชีวิตต่อยอด และนำข้อมูลที่ถูกปรับแล้วมาสร้างแบบจำลองการทำนายผลการตอบรับการขายกรมธรรม์สำหรับผู้ถือบัตรเครดิตของธนาคาร โดยนำเสนอการนำทฤษฎีการแก้ไขปัญหาข้อมูลไม่สมดุล (Imbalance Data) มาใช้ในงานวิจัย วิธีแก้ปัญหาข้อมูลไม่สมดุลที่ถูกนำมาใช้ ได้แก่ วิธีการสุ่มลด (Under Sampling) ซึ่งคือการลดจำนวนข้อมูลซึ่งอาจไม่เหมาะกับการสร้างแบบจำลองที่ต้องการความหลากหลายของข้อมูล วิธีการสุ่มเกิน (Over Sampling) ซึ่งคือการสุ่มข้อมูลที่มีอยู่แล้วเพิ่มขึ้นมาอาจส่งผลดีกับลักษณะข้อมูลที่ไม่ซับซ้อนมากเกินไป และวิธีการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Over-sampling Technique : SMOTE) ซึ่งคือการสังเคราะห์ข้อมูลเพิ่มโดยอ้างอิงจากข้อมูลที่มีอยู่ สามารถสร้างความหลากหลายของข้อมูลได้และส่งผลดีต่อการสร้างแบบจำลอง และผลการวิจัยปรากฏว่า วิธีแก้ปัญหาข้อมูลไม่สมดุลที่ดีที่สุดของข้อมูลที่จำลองขึ้นมาต่อแบบจำลองการพยากรณ์การเสนอขายประกันต่อยอด (Cross-sell) คือ วิธีการสุ่มเกิน (Over Sampling) เนื่องจากมีความผันผวนค่าพื้นที่ใต้โค้ง ROC ที่ต่ำกว่าอีก 2 วิธี และในส่วนของข้อมูลจริงที่ทางผู้วิจัยได้ศึกษาเป็นกรณีศึกษา (Case Study) ได้ผลสรุปว่า วิธีการสังเคราะห์ข้อมูลเพิ่ม (SMOTE) เป็นวิธีที่เหมาะสมที่สุดในการจัดการชุดข้อมูลไม่สมดุลของการขายประกันต่อยอด (Cross-sell) ของผู้ถือบัตรเครดิตของธนาคาร และจากการทำงานวิจัยในครั้งนี้นั้นสามารถช่วยแก้ไขปัญหาจัดการข้อมูลผลตอบรับการขายประกันต่อยอดของลูกค้าที่ไม่สมดุลให้อยู่ในรูปแบบที่ สามารถสรุปข้อมูลและพร้อมนำมาวิเคราะห์ต่อได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

5.2 ข้อเสนอแนะ

จากการวิจัยการจัดการข้อมูลไม่สมดุลของโอกาสในการขายกรมธรรม์ประกันโดยการสร้างแบบจำลองการทำนายผลสำหรับผู้ถือบัตรเครดิต ถ้าหากมีโอกาสครั้งถัดไปผู้วิจัยจะนำวิธีที่แก้ปัญหาข้อมูลไม่สมดุลอื่น ๆ มาทดสอบเพิ่มเติมเพื่อเพิ่มประสิทธิภาพการแก้ไขข้อมูลไม่สมดุลให้ดียิ่งขึ้น เช่น วิธีผสมผสาน (Hybrid Methods) วิธีสังเคราะห์ข้อมูลเพิ่ม ADASYN (Adaptive Synthetic Sampling Approach) เป็นต้น

6. กิตติกรรมประกาศ

จากการทำงานวิจัย ณ บริษัท อยุธยา แคปปิตอล เซอร์วิสเซส จำกัด ส่งผลให้ผู้วิจัยได้รับความรู้และประสบการณ์ที่มีประโยชน์มากมาย ในการทำงานวิจัยครั้งนี้สำเร็จลงได้ด้วยดีเนื่องจากความร่วมมือและสนับสนุนจากหลายฝ่าย ดังนี้ คุณอัญชีรา ชุมชัยเวทย์ รองประธานอาวุโสหัวหน้าฝ่าย Data intelligence and Customer Insights คุณทิพยวัณย์ ตรียาวธัญญู ผู้อำนวยการอาวุโสผู้บริหารฝ่าย Data Science & Big Data Infrastructure ทีม Big Data & Data Engineer, Data Literacy, Data Scientist และ Campaign Management ของบริษัท อยุธยา แคปปิตอล เซอร์วิสเซส จำกัด และขอขอบคุณ ดร.นรุตม์ สุนทรานนท์ ซึ่งเป็นพี่เลี้ยงที่คอยให้คำปรึกษาคำแนะนำในส่วนงานต่าง ๆ

เอกสารอ้างอิง

วิชญ์วิสิฐ เกสรสิทธิ์, วิชิต หล่อจีระชุณห์กุล และจิราวลัย จิตรถ. (2561). **การแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน**. วารสารวิจัย มข, 18(3), 11-21.

Bunkhumpornpat, C. & Subpaiboonkit, S. (2013). **Safe level graph for synthetic minority over-sampling techniques.** 13th International Symposium on Communications and Information Technologies (ISCIT), IEEE, 570-575.

Peng, C. J., Lee, K.L., & Ingersoll, G. M. (2002). **An introduction to logistic regression analysis and reporting**. Ph.D. dissertation, University of Indiana.

Drummord, C. & Holte, R.C. (2003). **C4.5, Class imbalance, and cost sensitivity: Why under-sampling beats over-sampling**. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets.

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