บทความวิจัย (Research Article)

การแก้ปัญหาข้อมูลไม่สมดุลแบบหลายคลาสสำหรับการทำนายการยกเลิกการใช้บริการอินเทอร์เน็ต

Multi-class Imbalanced Data Problem Solving for Predicting Internet Service Cancellation

รักชนก ปิยพาณิชยกุล และ กฤษณะ ไวยมัย*

ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเกษตรศาสตร์
* ผู้รับผิดชอบบทความ
fengknw@ku.ac.th

Received: 3 Aug 2023
Revised: 2 Oct 2023
Accepted: 18 Oct 2023

บทคัดย่อ

การยกเลิกการใช้บริการเป็นปัญหาที่พบได้ทั่วไปในหลายธุรกิจ โดยเฉพาะอย่างยิ่งกับธุรกิจโทรคมนาคม การจัดการเพื่อลดอัตราการยกเลิกใช้บริการอย่างมีประสิทธิภาพเป็นสิ่งสำคัญที่ธุรกิจต้องการ เพื่อสร้างความประทับใจให้กับลูกค้า ดังนั้นธุรกิจจำเป็นจะต้องเข้าใจถึงเหตุผลที่ลูกค้ายกเลิกการใช้บริการก่อน ในงานวิจัยนี้ได้ประยุกต์ใช้การเรียนรู้ของเครื่องโดยใช้เทคนิคการทำนายแบบหลายคลาสกับธุรกิจการให้บริการอินเทอร์เน็ต โดยเสนอการแก้ปัญหาเป็นสองส่วน 1) แก้ปัญหาข้อมูลไม่สมดุลในระดับข้อมูล โดยใช้เทคนิคการสุ่มเพิ่มข้อมูลด้วย SMOTE 2) แก้ปัญหาข้อมูลไม่สมดุลด้วยการปรับค่าน้ำหนักข้อมูลในกระบวนการเรียนรู้ร่วมกันด้วยเทคนิค XGBoost ผลลัพธ์ที่ได้จากการทำนายพบว่าสามารถทำนายคลาสที่มีข้อมูลขนาดใหญ่และข้อมูลขนาดเล็กได้อย่างเท่าเทียมและแม่นยำ

คำสำคัญ: การทำนายแนวโน้มการยกเลิกใช้บริการ ปัญหาข้อมูลไม่สมดุล การทำนายแบบหลายคลาส การเรียนรู้ของเครื่อง การเรียนรู้ร่วมกัน

Abstract

Churn is common problem across various industry, especially in the telecommunication sector. In order to reduce customer churn rates and complete effectively, many telecom companies are recognizing the need to improve customer experience and service. Therefore, understanding the reasons behind customer churn is essential. In this research, we apply machine learning multi-class classification technique for predicting customer churn to the internet service provider industry. We propose a two-fold approach to address the issue: 1) Handling imbalanced data at the dataset level by utilizing the SMOTE technique to synthetically generate additional data, and 2) Addressing imbalanced data at the learning process level by adjusting the data weights using the XGBoost technique. The resulting XGBoost based classifier is able to predict accurately the majority classes and as well as the minority classes. The combination of SMOTE and XGBoost techniques demonstrates effectiveness in solving imbalanced problem in the case of multi-class datasets.

Keywords: churn prediction, imbalanced data, multi-class classification, machine learning, ensemble learning

1. บทนำ

ในปัจจุบันที่ข้อมูลเป็นสิ่งสำคัญในการขับเคลื่อนธุรกิจให้เติบโตหรือก้าวหน้านำคู่แข่งรายอื่น ส่วนหนึ่งที่สำคัญมากที่แต่ละธุรกิจให้ความสำคัญคือการรักษาฐานลูกค้าเดิมไว้ให้ได้มากที่สุด เพื่อที่จะสามารถรักษาฐานของส่วนแบ่งการตลาดอย่างมั่นคง ข้อมูลการใช้บริการของลูกค้าจึงเป็นสิ่งที่แต่ละธุรกิจจะต้องให้ความสำคัญ และจำเป็นต้องนำมาใช้เพื่อเรียนรู้และวิเคราะห์พฤติกรรมของลูกค้าเพื่อทำนายการยกเลิกการใช้บริการ (Churn) ก่อนที่เหตุการณ์จะเกิดขึ้นจริง โดยนำการเรียนรู้ของเครื่อง (Machine Learning) ด้วยเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) เข้ามาประยุกต์ใช้ [1-3]

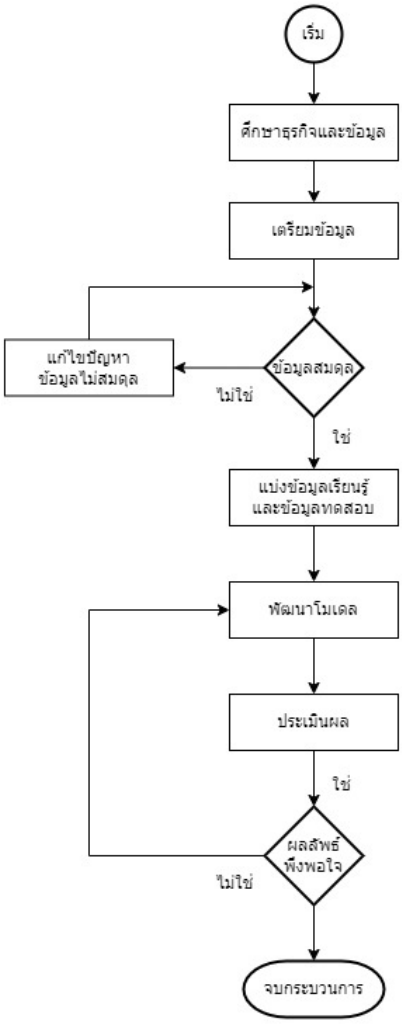
ประสิทธิภาพที่ดีของโมเดลนั้น ขึ้นอยู่กับหลายปัจจัย ซึ่งลักษณะและคุณภาพของข้อมูลก็เป็นปัจจัยหนึ่งที่มีความสำคัญเป็นอย่างมาก และโดยส่วนมากมักจะพบความไม่สมบูรณ์ของข้อมูลที่เกิดขึ้นบ่อยในกรณีของการยกเลิกบริการและจะต้องนำไปผ่านกระบวนการเตรียมข้อมูลต่อไป [1,3] เพื่อให้ได้ชุดข้อมูลที่มีสัดส่วนที่เท่ากันหรือใกล้เคียงกัน เพื่อป้องกันการทำนายที่โน้มเอียงไปยังคลาสใดคลาสหนึ่ง

และด้วยทุกคลาสมีความสำคัญเท่าเทียมกัน การทำนายแบบหลายคลาส (Multi-class classification) [4] จึงเป็นที่น่าสนใจเนื่องจากสามารถระบุได้ถึงผลลัพธ์ที่ชัดเจนกว่าแบบ 2 คลาส (Binary Classification) ทำให้ธุรกิจสามารถเข้าใจปัญหาของลูกค้าได้ตรงจุด ในงานวิจัยนี้จะมีความท้าทายในการแก้ปัญหาข้อมูลไม่สมดุลร่วมกับการทำนายแบบหลายคลาส [5] โดยใช้เทคนิคการสุ่มเพื่อทำให้ข้อมูลมีความสมดุล ได้แก่ การสุ่มเพิ่มข้อมูล (Over-sampling) [6,8,11] การสุ่มลดข้อมูล (Under-sampling) [7-8,11-12] หรือเทคนิคการสุ่มแบบผสม (Mixed-sampling) [8,11-12]

งานวิจัยชิ้นนี้เสนอการแก้ปัญหาข้อมูลไม่สมดุลกับการทำนายแบบหลายคลาสในระดับของข้อมูล [7] และในระดับการเรียนรู้ของเครื่อง โดยใช้เทคนิค SMOTE เพื่อสังเคราะห์ข้อมูลใหม่จากข้อมูลเดิม ซึ่งจะช่วยให้ค่าความสำคัญของข้อมูลไม่สูญหาย และจะช่วยให้ข้อมูลในการทำนายผลในคลาสขนาดเล็กมีความถูกต้องมากยิ่งขึ้น แล้วจึงนำข้อมูลที่สมดุลไปสร้างโมเดลโดยใช้วิธีการเรียนรู้ร่วมกันด้วยเทคนิคการจำแนกข้อมูลวิธี Bagging และการจำแนกข้อมูลวิธี Boosting เพื่อเปรียบเทียบและหาผลลัพธ์ของโมเดลที่ให้ค่าประสิทธิภาพและความแม่นยำมากที่สุด

จากผลการทดลองพบว่าเทคนิค SMOTE + XGBoost ซึ่งเป็นเทคนิคการจำแนกข้อมูลวิธี Boosting ให้ผลลัพธ์ที่ดีที่สุด โดยมีค่า Accuracy ที่ 0.89 และค่า Macro F1-score ที่ 0.80 ตามลำดับ ช่วยให้สามารถทำนายคลาสที่มีข้อมูลขนาดเล็กว่าลูกค้าจะยกเลิกบริการด้วยเหตุผลอะไรได้แม่นยำมากยิ่งขึ้น ซึ่งเป็นไปตามจุดประสงค์งานวิจัยนี้ที่ให้ความสำคัญกับการทำนายผลของทุกคลาสอย่างเท่าเทียม ทำให้ธุรกิจสามารถทำการตลาดในการรักษาฐานลูกค้าได้อย่างตรงจุดต่อไป

2. ขั้นตอนการดำเนินงานวิจัย



รูปที่ 1 ขั้นตอนการดำเนินงานวิจัย

จากรูปที่ 1 แสดงภาพรวมของขั้นตอนการดำเนินงานวิจัย ตั้งแต่เริ่มต้นจนจบกระบวนการวิจัย โดยมีรายละเอียดการดำเนินงานวิจัยในแต่ละขั้นตอน ดังนี้

2.1 แหล่งข้อมูล

ในงานวิจัยนี้ใช้การบูรณาการข้อมูลจากหลายแหล่งข้อมูล ตามรูปที่ 1 ได้แก่ 1) ข้อมูลการใช้บริการ อาทิ ข้อมูลส่วนบุคคลของลูกค้า ข้อมูลบริการที่ใช้งาน และข้อมูลการเรียกเก็บเงิน เป็นต้น 2) ข้อมูลลูกค้าสัมพันธ์ อาทิ ข้อมูลบริการ ข้อมูลการยกเลิกบริการ และข้อมูลใบแจ้งหนี้ เป็นต้น

2.2 การเตรียมข้อมูล

ในขั้นตอนการเตรียมข้อมูลนี้แบ่งการเตรียมข้อมูลออกเป็น 4 ขั้นตอนย่อย ดังต่อไปนี้

2.2.1 การทำความสะอาดข้อมูล (Data Cleansing)

- การจัดรูปแบบข้อมูลให้มีความถูกต้อง เหมือนกัน
- การจัดการค่าว่าง กรณีข้อมูลเป็นตัวเลขจะแทนที่ด้วยค่าเฉลี่ย หากข้อมูลเป็นตัวอักษรจะแทนที่ด้วยคำว่าอื่น ๆ
- การลบข้อมูลซ้ำซ้อนออกจากชุดข้อมูล
- การจัดการข้อมูลผิดปกติ (Outlier) โดยใช้วิธีการลบข้อมูลที่ต่ำกว่า upper percentile และ lower percentile ออกจากชุดข้อมูล

2.2.2 การเลือกข้อมูล (Data Selection)

- เลือกตัดตัวแปรที่มีความสำคัญน้อยออกจากชุดข้อมูล เช่น ข้อมูลที่สามารถระบุตัวตนของลูกค้า (Personal Data) เป็นต้น
- เลือกตัวแปรที่ให้ค่า Information Gain สูง (เข้าใกล้ 1) โดยคำนวณด้วยวิธีการหาค่า Entropy

2.2.3 การแปลงข้อมูล (Data Transformation)

- ตัวแปรที่อยู่ในรูปแบบวันที่ นำมาคำนวณเพื่อสร้างตัวแปรที่แสดงจำนวนวันที่ใช้บริการตั้งแต่สมัครใช้งานจนถึงปัจจุบัน
- สร้างตัวแปรคลาสเป้าหมาย (Label) แบบหลายคลาส เพื่อรวมกลุ่มของข้อมูลบางประเภทที่มีความหมายใกล้เคียงหรือเป็นสับเซ็ตเดียวกัน รวมเข้าด้วยกัน ส่งผลให้จากเดิมมีตัวแปรเป้าหมาย 17 คลาส เหลือจำนวน 4 คลาส
- การแปลงข้อมูลจัดกลุ่ม (categorical data) ให้อยู่ในลักษณะของ One-Hot Encoding โดยแปลงข้อมูล ให้เป็นค่า 0 และ 1 (binary values)

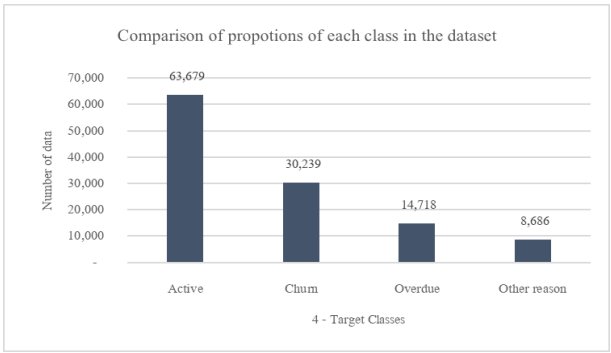
2.2.4 การรวมข้อมูล (Data Integration)

ทำการรวมข้อมูลที่ผ่านกระบวนการตามขั้นตอนที่ 2.2.1 ถึง 2.2.3 โดยใช้ตัวแปรหมายเลขบริการเป็นคีย์หลักในการเชื่อมข้อมูลหลายชุด เพื่อสร้างเป็นชุดข้อมูลตั้งต้นที่เตรียมจะนำไปสร้างโมเดลต่อไป

จะได้ชุดข้อมูลตั้งต้นจำนวน 167,603 รายการ ที่คงเหลือตัวแปรจำนวน 11 ตัว ได้แก่ ตัวแปรเป้าหมาย ประเภทลูกค้า จำนวนเงินที่จ่ายให้บริษัททั้งหมด จำนวนเงินที่จ่ายเดือนล่าสุด ประเภทข่ายสาย จำนวนความเร็วในการดาวน์โหลด จำนวนความเร็วในการอัปโหลด จำนวนวันที่ใช้บริการสะสม ประเภทกลุ่มลูกค้า ประเภทกลุ่มบริการอินเทอร์เน็ต และประเภทข้อตกลงการให้บริการ

2.3 การผสม (Mixed sampling)

ชุดข้อมูลตั้งต้นจำนวน 167,603 รายการ มีตัวแปรเป้าหมายที่นำมาทำนาย จำนวน 4 คลาส ได้แก่ ใช้บริการ (Active) ยกเลิกใช้บริการเนื่องจากการย้ายค่าย (Churn) ยกเลิกใช้บริการเนื่องจากการค้างชำระ (Overdue) และยกเลิกใช้บริการเนื่องจากเหตุผลอื่น ๆ (Other reason) โดยนำมาแบ่งในอัตราส่วน 70:30 เพื่อเป็นข้อมูลเรียนรู้ (Training dataset) จำนวน 117,322 รายการและข้อมูลทดสอบ (Testing dataset) จำนวน 50,281 รายการ



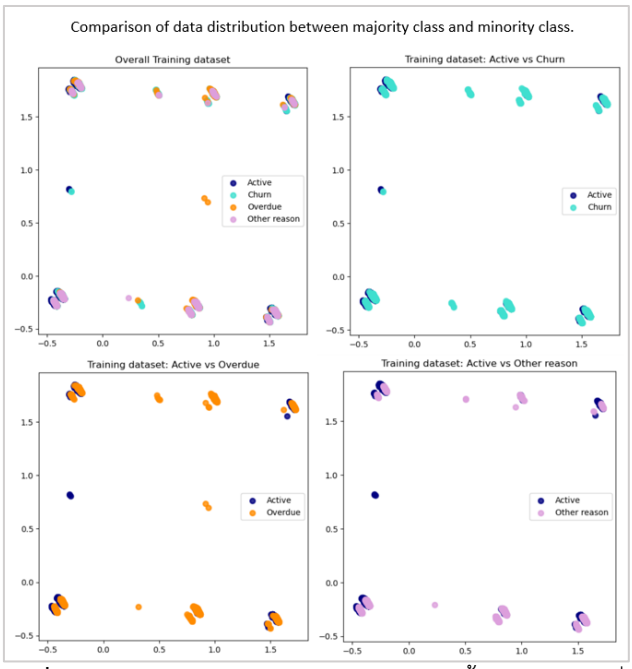
รูปที่ 2 กราฟแท่งเปรียบเทียบสัดส่วนข้อมูลในแต่ละคลาสของชุดข้อมูลตั้งต้น

จากการวิเคราะห์ข้อมูลตัวแปรเป้าหมายในชุดข้อมูลเรียนรู้ก่อนนำไปสร้างโมเดล พบว่าจำนวนข้อมูลแต่ละคลาสไม่สมดุลกัน และมีสัดส่วนที่แตกต่างกันอย่างชัดเจน โดยมีสัดส่วนของคลาส Active 63,679 รายการ คลาส Churn 30,239 รายการ คลาส Overdue 14,718 รายการ และคลาส Other reason 8,686 รายการ ตามลำดับ ดังรูปที่ 2

สามารถคำนวณหาอัตราความไม่สมดุล (Imbalance Ratio: IR) [11] เพื่อเปรียบเทียบสัดส่วนจำนวนข้อมูลระหว่างคลาสใหญ่และคลาสเล็ก โดยใช้สูตรการคำนวณดังนี้

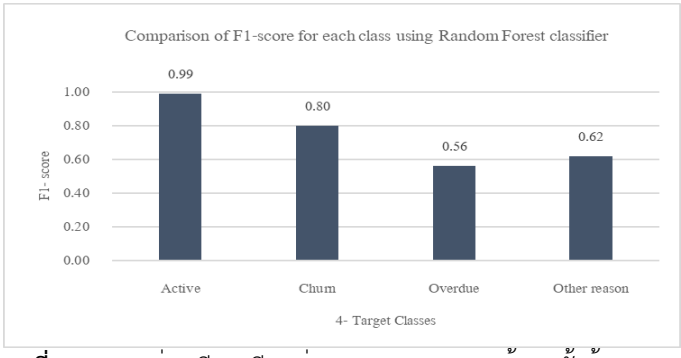
$$IR = \frac{Number\ of\ Majority\ class}{Number\ of\ Minority\ class} \quad (1)$$

เมื่อกำหนดให้คลาส Active เป็นคลาสใหญ่ และคลาสอื่น ๆ เป็นคลาสเล็กเพื่อเปรียบเทียบกับทุกคลาส พบว่าคลาส Active vs Churn มีค่า IR ที่ 2.10 คลาส Active vs Overdue มีค่า IR ที่ 4.33 และคลาส Active vs Other reason มีค่า IR ที่ 7.33 สามารถเปรียบเทียบการกระจายข้อมูลระหว่างคลาสใหญ่และคลาสเล็ก ดังรูปที่ 3



รูปที่ 3 กราฟเปรียบเทียบการกระจายตัวของข้อมูลคลาสใหญ่ Active กับทุก ๆ คลาสเล็ก

จากชุดข้อมูลตั้งต้นได้ทำนายการยกเลิกการใช้บริการอินเทอร์เน็ตโดยใช้เทคนิคการสุ่มป่าไม้ (Random Forest) เป็นโมเดลพื้นฐานในการทำนาย พบว่ากรณีที่ข้อมูลมีความไม่สมดุลจะส่งผลให้โมเดลมีความโน้มเอียงของการทำนายไปยังคลาสใหญ่และทำนายคลาสเล็กได้ไม่แม่นยำ เนื่องจากความไม่สมดุลของข้อมูลในแต่ละคลาส โดยมีค่า Accuracy ที่ 0.87 และ Macro F1-score ที่ 0.74 ซึ่งเป็นค่าเฉลี่ยของ F1-score ในแต่ละคลาส สามารถเปรียบเทียบค่า F1-score ของแต่ละคลาส ตามรูปที่ 4



รูปที่ 4 กราฟแท่งเปรียบเทียบค่า F1-score ของชุดข้อมูลตั้งต้น

เนื่องด้วยธุรกิจนี้ให้ความสำคัญกับทุกคลาสอย่างเท่าเทียม เพื่อที่จะสามารถทำนายเหตุผลการยกเลิกใช้บริการได้ จึงจำเป็นจะต้องแก้ปัญหาข้อมูลไม่สมดุลเพื่อให้ได้ผลการทำนายที่มีความแม่นยำในทุกคลาส

2.4 การแก้ปัญหาข้อมูลไม่สมดุลแบบหลายคลาส

เพื่อให้ได้ชุดข้อมูลที่มีความสมดุลเท่าเทียมกันในทุกคลาส และป้องกันการทำนายผลที่โน้มเอียง ในงานวิจัยนี้เสนอการแก้ปัญหาออกเป็น 2 ส่วน 1) การแก้ปัญหาในระดับข้อมูล และ 2) การใช้วิธีการเรียนรู้ร่วมกัน เพื่อช่วยในการทำนายผลแบบหลายคลาส ซึ่งจะกล่าวถึงในหัวข้อที่ 2.5

วิธีการแก้ปัญหาข้อมูลไม่สมดุลแบบหลายคลาสในระดับข้อมูล เป็นวิธีที่ทำให้ข้อมูลในคลาสขนาดใหญ่ (Majority Class) และคลาสขนาดเล็ก (Minority Class) มีขนาดเท่ากันหรือใกล้เคียงกัน ซึ่งเป็นวิธีที่นิยมใช้ในการแก้ปัญหาข้อมูลไม่สมดุล โดยแบ่งกลุ่มการแก้ปัญหาได้เป็น 3 วิธี [11] ดังนี้

2.4.1 การสุ่มลดข้อมูล (Under-sampling)

- RandomUnderSampler (RUS) [7-8,12] เป็นการสุ่มลดข้อมูลจากคลาสใหญ่ เพื่อให้จำนวนข้อมูลในคลาสใหญ่มีขนาดข้อมูลเท่ากับคลาสเล็ก
- NearMiss version1 [8] เป็นการสุ่มลดข้อมูลจากคลาสใหญ่ โดยมีตัวแปรควบคุม *k*-neighbors มาเกี่ยวข้อง ใช้การเลือกสุ่มตัวอย่างจากคลาสใหญ่ที่มีค่าระยะทางเฉลี่ยของ *k* ที่น้อยที่สุด โดยเปรียบเทียบกับเพื่อนบ้านคลาสเล็กที่ใกล้ที่สุด
- NearMiss version2 [8] เลือกสุ่มตัวอย่างจากคลาสใหญ่ที่มีค่าระยะทางเฉลี่ยของ *k* ที่น้อยที่สุด โดยเปรียบเทียบกับเพื่อนบ้านคลาสเล็กที่ไกลที่สุด
- EditedNearestNeighbors (ENN) [7-8,11] เป็นการใช้การเปรียบเทียบระหว่างข้อมูลตัวอย่างในคลาสใหญ่และผลลัพธ์ที่ได้จากการทำนายคลาสเป้าหมายด้วยเทคนิค *k* nearest neighbors (*k*-NN) โดยลบข้อมูลที่คลาสเป้าหมายมีผลลัพธ์ไม่ตรงกันออก ช่วยให้ข้อมูลที่เหลืออยู่มีความถูกต้องมากขึ้น
- Tomek's Links (T-Links) [7-8] เป็นเทคนิคการจับคู่ของข้อมูลต่างคลาสที่อยู่ใกล้กัน แล้วเลือกลบข้อมูลคลาสใหญ่ออก เพื่อให้เกิดช่องว่างระหว่างกลุ่ม ช่วยให้การแบ่งกลุ่มมีความชัดเจนมากขึ้น

2.4.2 การสุ่มเพิ่มข้อมูล (Over-sampling)

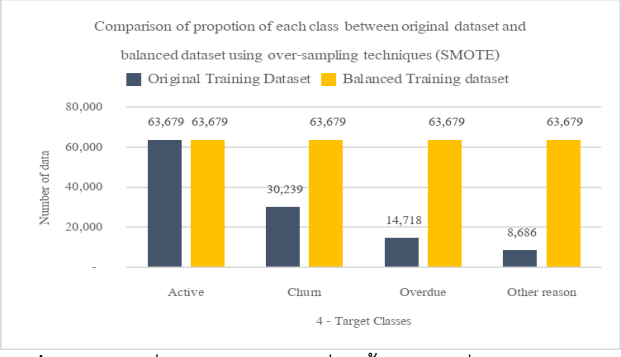
- RandomOverSampler (ROS) [8] เป็นการสุ่มเพิ่มข้อมูลจากคลาสเล็ก เพื่อให้จำนวนข้อมูลในคลาสเล็กมีขนาดข้อมูลเท่ากับคลาสใหญ่
- Synthetic Minority Oversampling Technique (SMOTE) [6,8-9,11-12] เป็นเทคนิคที่ใช้ในการสุ่มเพิ่มข้อมูลคลาสเล็กให้มีขนาดใกล้เคียงหรือเท่ากับคลาสใหญ่ โดยทำการสังเคราะห์ข้อมูลใหม่จากข้อมูลเดิม ทำให้ข้อมูลในคลาสเล็กมีการกระจายตัวที่ดีขึ้นและช่วยให้ค่าความสำคัญของข้อมูลไม่สูญหาย
- Adaptive Synthetic (ADASYN) [6] เป็นเทคนิคที่ปรับปรุงมาจาก SMOTE โดยจะไม่พิจารณาข้อมูลทุกตัวในคลาสเล็กแบบเดียวกับ SMOTE แต่จะใช้การแจกแจงแบบถ่วงน้ำหนักของข้อมูลในคลาสเล็ก ทำให้การปรับขอบเขตการตัดสินใจในการแบ่งกลุ่มดีขึ้น

2.4.3 การสุ่มข้อมูลแบบผสม (Mixed-Sampling)

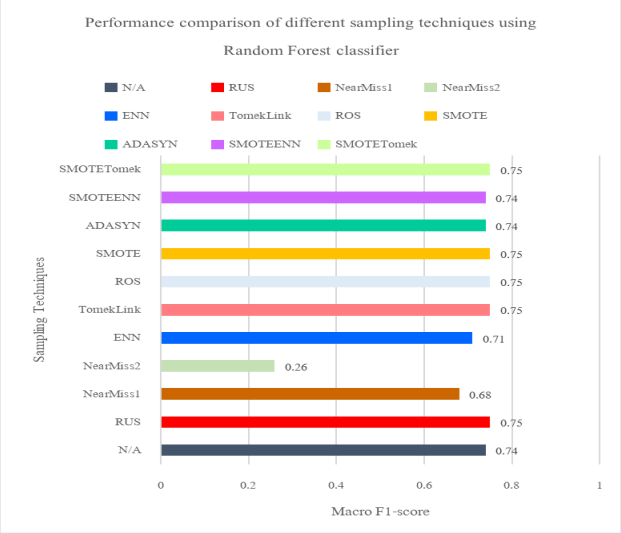
- SMOTEENN [8,11-12] เทคนิคผสมระหว่าง SMOTE และ ENN โดยจะทำการเพิ่มข้อมูลให้แต่ละคลาสมีขนาดเท่ากันก่อน แล้วจึงจะสุ่มลดข้อมูลให้มีความถูกต้องมากขึ้น
- SMOTETOMEK [8,11] เทคนิคผสมระหว่าง SMOTE และ Tomek's Link โดยจะทำการเพิ่มข้อมูลให้แต่ละคลาสมีขนาดเท่ากันก่อน แล้วจึงจะสุ่มลดข้อมูลให้มีความถูกต้องมากขึ้น

จากการทดลองแก้ปัญหาข้อมูลไม่สมดุลแบบหลายคลาสด้วยวิธีต่าง ๆ ร่วมกับการใช้เทคนิค Random Forest ซึ่งเป็นการจำแนกข้อมูลในกลุ่มวิธี Bagging ตามข้อ 2.5.1 เป็นโมเดลพื้นฐานในการทำนาย พบว่าการสุ่มเพิ่มข้อมูลโดยใช้เทคนิค SMOTE จะได้ชุดข้อมูลเรียนรู้ที่มีความสมดุลกันจำนวน 254,716 รายการ โดยแบ่งสัดส่วนอย่างเท่ากันคลาสละ 63,679 รายการ เปรียบเทียบสัดส่วนข้อมูลของแต่ละคลาสก่อนและหลังแก้ปัญหาข้อมูลไม่สมดุล ดังรูปที่ 5

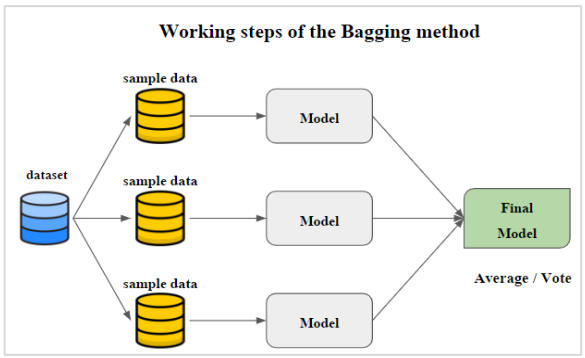
เมื่อเปรียบเทียบประสิทธิภาพโมเดลระหว่างเทคนิคการสุ่มแบบต่าง ๆ ผลการทดลองพบว่าวิธีการสุ่มเพิ่มข้อมูลด้วยเทคนิค SMOTE ทำให้ชุดข้อมูลมีความสมดุลและให้ผลการทำนายที่ดีที่สุด ซึ่งจะดีกว่าการเลือกใช้วิธีการสุ่มลดข้อมูลหรือการสุ่มแบบผสมที่อาจจะทำให้ความสำคัญของข้อมูลสูญหายเนื่องจากการลดข้อมูล ดังรูปที่ 6



รูปที่ 5 กราฟแท่งเปรียบเทียบสัดส่วนข้อมูลในแต่ละคลาสก่อนและหลังแก้ไขปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE



รูปที่ 6 กราฟแท่งเปรียบเทียบประสิทธิภาพของโมเดลแบบหลายคลาสตามแต่ละเทคนิคการแก้ไขปัญหาข้อมูลไม่สมดุล



รูปที่ 7 ขั้นตอนการทำงานของวิธี Bagging

2.5 วิธีการเรียนรู้ร่วมกัน (Ensemble Learning)

โดยทั่วไปการจำแนกข้อมูลแบบโมเดลเดียวให้ผลการทำนายโดยรวมที่ค่อนข้างดี แต่ยังไม่ดีเพียงพอสำหรับการทำนายแบบหลายคลาสที่มีความไม่สมดุลของตัวแปรเป้าหมาย โดยเฉพาะกับโมเดลที่ต้องการให้ความสำคัญกับทุกคลาสอย่างเท่าเทียม เมื่อพิจารณาผลการทำนายในคลาสขนาดเล็กพบว่าโมเดลทำนายได้ไม่ดีนัก มีความโน้มเอียงไปยังคลาสขนาดใหญ่

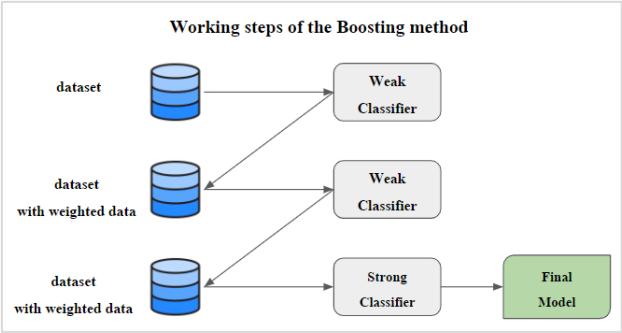
เพื่อที่จะสามารถจำแนกข้อมูลแบบหลายคลาสได้อย่างแม่นยำและลดความโน้มเอียงที่เกิดขึ้น การจำแนกข้อมูลด้วยวิธีการเรียนรู้แบบร่วมกันโดยใช้โมเดลหลาย ๆ ตัว (Multiple classifier) [11] จึงให้ประสิทธิภาพการทำนายที่ดีกว่าการจำแนกข้อมูลโดยใช้โมเดลเดียว

2.5.1 การจำแนกข้อมูลด้วยวิธี Bootstrap Aggregation (Bagging)

เป็นเทคนิคการเรียนรู้แบบร่วมกันโดยสุ่มสร้างโมเดลพร้อม ๆ กันด้วยวิธีการสุ่มข้อมูลแบบแทนที่ (sampling with replacement) ซึ่งจะทำให้การสุ่มข้อมูลทุกครั้งมีโอกาสเท่ากัน และแต่ละครั้งอาจจะสุ่มซ้ำกันได้ วิธีนี้จะช่วยสุ่มข้อมูลและสร้างโมเดลหลาย ๆ ตัวในการทำนาย แล้วใช้ค่าเฉลี่ยหรือการโหวตเพื่อตัดสินใจ ดังรูปที่ 7 โดยเทคนิคที่ใช้วิธีนี้ ได้แก่ Random Forest [9-11]

2.5.2 การจำแนกข้อมูลด้วยวิธี Boosting

เป็นเทคนิคการเรียนรู้ร่วมกันโดยใช้การเรียนรู้แบบเป็นลำดับ โดยโมเดลจะนำค่าที่ผิดพลาดที่ได้จากโมเดลก่อนหน้ามาเรียนรู้และปรับปรุงโมเดลในครั้งถัดไป จะช่วยลดความผิดพลาดและเพิ่มความถูกต้องในการสร้างโมเดล โดยจะทำเป็นลำดับไปเรื่อย ๆ จนกว่าจะได้โมเดลสุดท้ายที่มีผลลัพธ์ที่ดีที่สุด ดังรูปที่ 8 เทคนิคที่ใช้วิธีนี้ ได้แก่ ADABoost, Gradient Boost และ Extream Gradient Boost (XGBoost) [9-11]



รูปที่ 8 ขั้นตอนการทำงานของวิธี Boosting

จากวิธีการจำแนกข้อมูลแบบหลายคลาส ได้กำหนดให้เทคนิค Random Forest ในกลุ่มวิธี Bagging เป็นโมเดลพื้นฐานในการทำนาย โดยเปรียบเทียบกับโมเดลจากกลุ่มวิธี Boosting

พบว่าการใช้เทคนิค XGBoost เป็นอัลกอริทึมการเรียนรู้ของเครื่องที่มีความสามารถในการจัดการกับชุดข้อมูลที่ไม่สมดุลในระหว่างขั้นตอนการเรียนรู้ โดยจัดการกับปัญหาความไม่สมดุลของข้อมูลด้วยการกำหนดน้ำหนักสูงให้กับตัวอย่างของคลาสเล็กและน้ำหนักต่ำให้กับตัวอย่างของคลาสใหญ่ หลักการนี้ช่วยให้อัลกอริทึมให้ความสำคัญต่อตัวอย่างคลาสเล็กมากขึ้นและปรับ

กระบวนการการเรียนรู้ของโมเดลให้เหมาะสม และสามารถทำนายผลในแต่ละคลาสที่ธุรกิจให้ความสำคัญเท่าเทียมกันได้อย่างแม่นยำ

3. ผลการดำเนินงานวิจัย

3.1 การประเมินผล

3.1.1 ตารางวัดประสิทธิภาพ (Confusion Matrix)

ตารางวัดประสิทธิภาพของข้อมูลโดยเปรียบเทียบผลลัพธ์จากค่าจริง (Actual) และค่าที่ได้จากการทำนาย (Predicted) ในแต่ละคลาส ตามตารางที่ 1 รายละเอียดดังนี้

- True Positive: TP คือ ค่าจริงเป็น true และทำนายได้ true
- False Positive: FP คือ ค่าจริงเป็น false แต่ทำนายได้ true
- True Negative: TN คือ ค่าจริงเป็น false และทำนายได้ false
- False Negative: FN คือ ค่าจริงเป็น true แต่ทำนายได้ false

ตารางที่ 1 Confusion matrix

Confusion Matrix		Predict	
		False	True
Actual	False	TN	FP
	True	FN	TP

3.1.2 Accuracy

ค่าประสิทธิภาพโดยรวมของโมเดล มีสูตรการคำนวณ ดังนี้

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2)$$

3.1.3 Precision

ค่าความแม่นยำที่โมเดลทำนายในคลาสที่พิจารณาโดยเทียบกับค่าที่ทำนายได้ทั้งหมด มีสูตรการคำนวณ ดังนี้

$$Precision = \frac{TP}{(TP + FP)} \quad (3)$$

3.1.4 Recall

ค่าความแม่นยำที่โมเดลทำนายในคลาสที่พิจารณาโดยเทียบกับค่าจริงทั้งหมด มีสูตรการคำนวณ ดังนี้

$$Recall = \frac{TP}{(TP + FN)} \quad (4)$$

3.1.5 F1-score

ค่าประสิทธิภาพของโมเดลในคลาสที่พิจารณา เพื่อดูว่าโมเดลมีประสิทธิดีหรือไม่ โดยคำนวณจากค่าเฉลี่ยของ Precision กับ Recall มีสูตรการคำนวณ ดังนี้

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

3.1.6 Macro average F1-score (Macro F1-score)

ค่าประสิทธิภาพรวมของโมเดลในข้อมูลแบบหลายคลาส เนื่องจากทุกคลาสมีความสำคัญอย่างเท่าเทียม จึงใช้วิธีการหาค่าเฉลี่ยของ F1-score จากแต่ละคลาส มีสูตรการคำนวณ ดังนี้

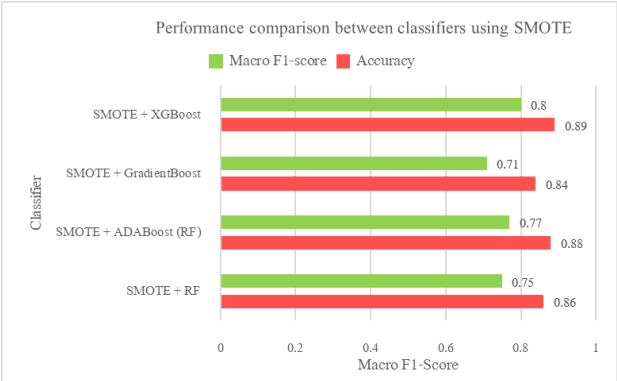
$$Macro\ F1 = \frac{Sum\ of\ F1\ score\ in\ all\ class}{number\ of\ class} \quad (6)$$

3.2 ผลการวิจัย

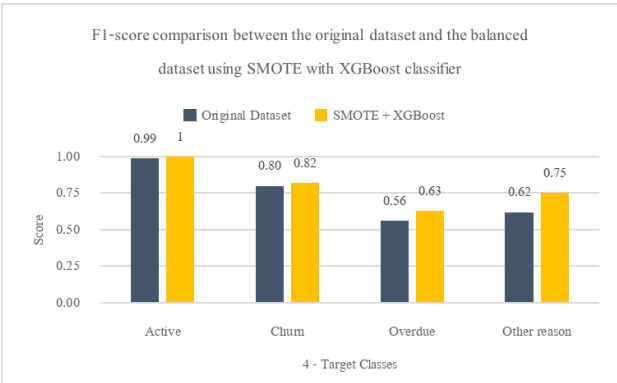
จากการเตรียมข้อมูลและวิเคราะห์ชุดข้อมูลตั้งต้นพบว่าชุดข้อมูลตั้งต้นมีความไม่สมดุลกันในแต่ละคลาส ส่งผลให้ผลลัพธ์ของการทำนายด้วยเทคนิค Random Forest มีความโน้มเอียงไปยังคลาสใหญ่และทำให้การทำนายข้อมูลในคลาสเล็กไม่แม่นยำเท่าใดนัก แต่เนื่องด้วยธุรกิจให้ความสำคัญกับข้อมูลทุกคลาสอย่างเท่ากัน จึงจำเป็นต้องมีการแก้ไขปัญหาข้อมูลไม่สมดุล ซึ่งได้ทดสอบการแก้ไขปัญหาข้อมูลไม่สมดุลด้วยเทคนิคที่หลากหลายแล้วจึงแบ่งข้อมูลเรียนรู้และข้อมูลทดสอบในอัตราส่วน 70:30 ตามลำดับเพื่อทำนายผลด้วยเทคนิค Random Forest อีกครั้ง ได้ผลลัพธ์ตามรูปที่ 5 พบว่าเทคนิคการสุ่มเพิ่มข้อมูล SMOTE ให้ผลลัพธ์ดีที่สุด โดยเปรียบเทียบจากค่า precision และ recall ของแต่ละคลาสเทียบกับเทคนิคการแก้ไขปัญหาข้อมูลไม่สมดุลอื่น ๆ ที่ได้ผลของ Macro F1-score ที่เท่ากัน โดย SMOTE ช่วยเพิ่มจำนวนข้อมูลในชุดข้อมูลเรียนรู้ให้มีความสมดุลและทำให้แต่ละคลาสมีสัดส่วนเท่ากัน ซึ่งการเพิ่มข้อมูลจะช่วยให้ไม่สูญเสียค่าความสำคัญของข้อมูล จะดีกว่าการเลือกใช้วิธีการสุ่มลดข้อมูลหรือการสุ่มแบบผสมที่อาจจะทำให้ความสำคัญของข้อมูลสูญหายเนื่องจากการลดข้อมูล

หลังจากแก้ไขปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE จึงได้ชุดข้อมูลเรียนรู้ที่มีความสมดุลและทำให้แต่ละคลาสมีสัดส่วนเท่ากัน ซึ่งจะได้ข้อมูลเรียนรู้จำนวน 254,716 รายการ

โดยแบ่งสัดส่วนอย่างเท่ากันทั้ง 4 คลาส คลาสละ 63,679 รายการ แต่ยังพบว่าผลการทำนายด้วยเทคนิค Random Forest ให้ผลการทำนายในคลาสเล็กยังไม่แม่นยำเท่าใดนัก จึงทดลองทำนายโดยใช้การสุ่มเพิ่มข้อมูลด้วยเทคนิค SMOTE ร่วมกับเทคนิคการจำแนกข้อมูลโดยใช้วิธีการเรียนรู้ร่วมกันแบบอื่น ๆ ได้แก่ ADABoost Gradient Boost และXGBoost เพื่อทำการเปรียบเทียบโมเดลที่ให้ประสิทธิภาพการทำนายสาเหตุการยกเลิกการใช้บริการอินเทอร์เน็ตที่ดีที่สุด



รูปที่ 9 กราฟแท่งเปรียบเทียบประสิทธิภาพโมเดลแบบหลายคลาสตามแต่ละเทคนิคการจำแนกข้อมูล



รูปที่ 10 กราฟแท่งเปรียบเทียบค่า F1-score ของแต่ละคลาสก่อนและหลังแก้ไขปัญหาข้อมูลไม่สมดุลด้วยเทคนิค SMOTE

จากผลการทดลองพบว่าเทคนิค SMOTE + XGBoost ให้ประสิทธิภาพผลการทำนายที่ดีที่สุด มีปัจจัยที่ส่งผลต่อตัวแปรเป้าหมาย 3 อันดับแรก คือ จำนวนเงินที่จ่ายให้บริษัททั้งหมด จำนวนเงินที่จ่ายเดือนล่าสุด และจำนวนวันที่ใช้บริการสะสม ส่งผลให้โมเดลมีค่า Accuracy ที่ 0.80 และมีค่า Macro F1-score ที่ 0.89 เป็นผลลัพธ์ที่ดีที่สุด ตามรูปที่ 9 โดย Macro F1-score เป็นการวัดผลประสิทธิภาพรวมของโมเดลของข้อมูลแบบหลายคลาส คำนวณจากค่าเฉลี่ยของ F1-score ในแต่ละคลาส ซึ่งเป็นวิธีที่เหมาะสมในการประเมินผลของโมเดลกรณีที่ทุกคลาสมีความสำคัญเท่ากัน

เมื่อเปรียบเทียบค่า F1-score ของแต่ละคลาสก่อนและหลังปรับข้อมูลให้สมดุล พบว่าโมเดลสามารถทำนายทุกคลาสได้อย่างแม่นยำมากยิ่งขึ้นโดยไม่ทำให้ผลการทำนายโน้มเอียงไปยังคลาสใดคลาสหนึ่ง ซึ่งเป็นการทำนายโดยให้ความสำคัญกับทุกคลาสอย่างเท่าเทียมกัน ตามรูปที่ 10

4. สรุปผลการดำเนินงานวิจัย

เนื่องด้วยงานวิจัยนี้ให้ความสำคัญกับการทำนายผลในทุกคลาสอย่างเท่าเทียม จึงเสนอการแก้ปัญหาข้อมูลไม่สมดุลร่วมกับการทำนายแบบหลายคลาส โดยใช้เทคนิค SMOTE + XGBoost ที่แบ่งการแก้ปัญหาเป็น 2 ส่วน คือ 1) การแก้ไขปัญหาข้อมูลไม่สมดุลในระดับข้อมูลด้วยการประยุกต์ใช้เทคนิคการสุ่มเพิ่มข้อมูล SMOTE เพื่อแก้ปัญหาข้อมูลไม่สมดุลแบบหลายคลาส โดยใช้วิธีการสังเคราะห์ข้อมูลใหม่จากข้อมูลเดิมเพื่อให้ข้อมูลไม่สูญเสียความสำคัญ โดยข้อมูลที่ถูกเพิ่มจำนวนจะทำให้สามารถแบ่งกลุ่มได้ดีมากยิ่งขึ้น และ 2) การทำนายด้วยวิธีการเรียนรู้ร่วมกันโดยใช้เทคนิค XGBoost ซึ่งเป็นการปรับปรุงโมเดลแบบลำดับขั้นทำให้ได้โมเดลที่มีประสิทธิภาพดีสุด พบว่าโมเดลสามารถทำนายผลในคลาสที่มีข้อมูลขนาดเล็กได้แม่นยำเช่นเดียวกับคลาสที่มีข้อมูลขนาดใหญ่ และให้ผลลัพธ์ที่ดีที่สุดเมื่อเปรียบเทียบในกลุ่มเทคนิคที่ใช้สำหรับการทำนายแบบหลายคลาส โดยมี 3 ปัจจัยหลักที่ส่งผลต่อการยกเลิกการใช้บริการ ได้แก่ จำนวนเงินที่จ่ายให้บริษัททั้งหมด จำนวนเงินที่จ่ายเดือนล่าสุด และจำนวนวันที่ใช้บริการสะสม

5. ข้อเสนอแนะ

จากผลลัพธ์ของโมเดล SMOTE + XGBoost ที่สามารถทำนายสาเหตุการยกเลิกใช้บริการอินเทอร์เน็ตได้อย่างแม่นยำ ได้ค้นพบ 3 ปัจจัย ที่ส่งผลต่อการทำนายคลาสเป้าหมาย ได้แก่ จำนวนเงินที่จ่ายให้บริษัททั้งหมด จำนวนเงินที่จ่ายเดือนล่าสุด และจำนวน

วันที่ใช้บริการสะสม ทำให้ธุรกิจสามารถนำผลที่ได้ไปวิเคราะห์เพื่อวางกลยุทธ์ทางการตลาดได้อย่างชาญฉลาด และตอบสนองต่อพฤติกรรมการใช้บริการที่เปลี่ยนแปลงได้อย่างทันท่วงที อาทิ การเสนอปรับแพ็คเกจให้สอดคล้องกับพฤติกรรมการใช้งานของลูกค้า ซึ่งจะช่วยสร้างความพึงพอใจให้กับลูกค้าหากสามารถลดค่าใช้จ่ายในส่วนที่ใช้บริการไม่ถึงตามแพ็คเกจ
ที่ซื้อลงได้ เพื่อป้องกันการยกเลิกการใช้บริการที่อาจจะเกิดขึ้นในอนาคตและก่อให้เกิดความรู้สึกเชิงบวกต่อธุรกิจว่ามีความใส่ใจลูกค้าอยู่เสมอ หรือเป็นการทำการตลาดเชิงรุกต่อลูกค้าในกลุ่มที่ทำนายแล้วคาดว่าจะยกเลิกการใช้บริการ โดยเสนอโปรโมชันเพื่อดึงดูดลูกค้าให้ยังคงใช้บริการอย่างต่อเนื่อง

อีกทั้ง เนื่องจากชุดข้อมูลตั้งต้นมีจำนวนคลาสเป้าหมาย 17 คลาส ซึ่งมีสัดส่วนของคลาสไม่สมดุลกัน จึงจำเป็นต้องลดคลาสเป้าหมายที่บ่งบอกถึงสาเหตุการยกเลิกการใช้บริการอินเทอร์เน็ตให้เหลือเพียง 4 คลาส ทำให้เกิดความเสียโอกาสในการทำความเข้าใจพฤติกรรมของลูกค้าที่ยกเลิกบริการด้วยเหตุผลอื่น หากต้องการที่จะทำนายทั้ง 17 คลาส จำเป็นที่จะต้องเก็บข้อมูลด้านพฤติกรรมเพิ่มเติม เช่น ปัญหาที่พบระหว่างใช้บริการ ระยะเวลาในการแก้ปัญหา ช่วงเวลาที่ใช้งาน จำนวนอุปกรณ์ที่ใช้งาน หรือสอบถามสาเหตุเชิงลึกของการตัดสินใจยกเลิกการใช้บริการอินเทอร์เน็ต เพื่อให้ได้ข้อมูลที่มีความสอดคล้องกับตัวแปรเป้าหมายที่ต้องการจะทำนายมากยิ่งขึ้น รวมถึงการรวบรวมข้อมูลที่จัดเก็บให้มีความสมดุลกันมากที่สุด

6. เอกสารอ้างอิง

[1] P. Bhanuprakash and G.S. Nagaraja, "A Review on Churn Prediction Modeling in Telecom Environment". 2017 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS). 21-23 Dec. Bengaluru India : pp. 8-12, 2017.

[2] A. Śniegula, A. Poniszewska-Marańda, M. Popović, "Study of machine learning methods for customer churn prediction in telecommunication company". The 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019). 2-4 Dec. Munich Germany : pp. 640-644, 2019.

[3] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques". 2018 International Conference on Advanced Computation and Telecommunication (ICACAT). 28-29 Dec. Bhopal India : pp. 1-5, 2018.

[4] F. Khan and S. S. Kozat, "Sequential churn prediction and analysis of cellular network users — A multi-class, multi-label perspective". 2017 25th Signal Processing and Communications Applications Conference (SIU). 15-18 May. Antalya Turkey : pp 1-4, 2017.

[5] M. M. Khamsan and R. Maskat, "Handling highly imbalanced output class label: a case study on Fantasy Premier League (FPL) virtual player price changes prediction using machine learning," Malaysian Journal of Computing (MJoC)., Vol. 4(2), pp. 304-316, 2019.

[6] พุทธิพร ธนธรรมเมธี และ เยาวเรศ ศิริสถิตย์กุล, "เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ," วารสารวิทยาศาสตร์และเทคโนโลยี, ปีที่ 27 (ฉบับที่ 6), หน้า 1164-1178, 2019.

[7] M. Bach, A. Werner and M. Palt, “The Proposal of Undersampling Method for Learning from Imbalanced Datasets,” Procedia Computer Science., Vol. 159, pp. 125-134, 2019.

[8] A. More. (11 January 2023). Survey of resampling techniques for improving classification performance in unbalanced datasets, [Online]. Available : ResearchGate, https://www.researchgate.net/

[9] H. R. Sanabila and W. Jatmiko, "Ensemble Learning on Large Scale Financial Imbalanced Data". 2018 International Workshop on Big Data and Information Security (IWBIS). 8-9 May. Jakarta Indonesia : pp. 93-98, 2018.

[10] P. Goyal, R. Rani and K. Singh, "Comparative Analysis of Machine Learning and Ensemble Learning Classifiers for Parkinson's Disease Detection". 2022 3rd International Conference on Computing, Analytics and Networks (ICAN), 18-19 Nov. Rajpura Punjab India : pp. 1-6, 2022.

[11] Nasritha, K., Kerdprasop, K., & Kerdprasop, N., "Comparison of Sampling Techniques for Imbalanced Data Classification," Journal of Applied Informatics and Technology., Vol. 1(1), pp. 20–37, 2018.

[12] Virakul, R., & Waiyamai, K. "Feature Selection and Imbalanced Data Problem Solving in Classification of Banking Fraud Prevention". 14th National Conference on Information Technology (NCIT-2022). 10-11 Nov. Nonthaburi Thailand : pp. 41-45, 2022.