

การวิเคราะห์เชิงทำนายพฤติกรรมการซื้อขายของลูกค้าโดยใช้อัลกอริทึม การเรียนรู้ของเครื่อง: กรณีศึกษาร้านซูปิดแฮนด์เมดจิวเอลรี Predictive Analytics of Consumer Buying Behavior using Machine Learning Algorithms: A Case Study of Sue's Beads Handmade Jewelry Shop

จิตติมารัตม์ บัคเลย์ และ นิติเศษฐ หมวดทองอ่อน*

Chittimarat Buckley and Nithizethe Mhuadthongon

สาขาวิชาวิทยาศาสตร์และเทคโนโลยี แขนงวิชาเทคโนโลยีดิจิทัล

มหาวิทยาลัยสุโขทัยธรรมมาธิราช จังหวัดนนทบุรี 11120

Program of Digital Technology, School of Science and Technology,
Sukhothai Thammathirat Open University, Nonthaburi Province, 11120

*Corresponding author E-mail: Nithizethe.Mhu@stou.ac.th

Received 24 August 2025, Accepted 18 February 2026, Published 15 June 2026

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) พัฒนาแบบจำลองการวิเคราะห์เชิงทำนายพฤติกรรมการซื้อขายของลูกค้า 2) ประเมินประสิทธิภาพของแบบจำลองที่พัฒนา และ 3) ประยุกต์ใช้แบบจำลองที่พัฒนาในการวิเคราะห์พฤติกรรมการซื้อขายของลูกค้า กรณีศึกษาร้านซูปิดแฮนด์เมดจิวเอลรี ผู้วิจัยได้พัฒนาแบบจำลองโดยใช้ 3 อัลกอริทึม ได้แก่ ต้นไม้ตัดสินใจ เกรเดียนต์บูสต์รี และการถดถอยโลจิสติก ร่วมกับการวิเคราะห์อาร์เอฟเอ็ม ซึ่งใช้ข้อมูลลูกค้าจำนวน 500 ราย ระหว่างเดือนมกราคม พ.ศ. 2566 ถึงเดือนธันวาคม พ.ศ. 2567 แบ่งตามลำดับเวลา (Time-Based Split) โดยใช้ข้อมูลในช่วงเวลาร้อยละ 80 สำหรับการฝึกแบบจำลอง และอีกร้อยละ 20 สำหรับการทดสอบ เพื่อป้องกันการรั่วไหลของข้อมูล และสะท้อนการใช้งานจริงในอนาคต ผลการวิจัยพบว่าแบบจำลองการถดถอยโลจิสติกให้ค่าความแม่นยำสูงสุดร้อยละ 100 ในทุกตัวชี้วัด ขณะที่แบบจำลองต้นไม้ตัดสินใจและแบบจำลองเกรเดียนต์บูสต์รี แม้มีวิธีการประมวลผลที่แตกต่างกันแต่ให้ค่าความแม่นยำเท่ากันที่ ร้อยละ 95.40 ผลการวิเคราะห์ ยังระบุว่าคะแนนอาร์เอฟเอ็มเป็นตัวแปรที่มีความสำคัญมากที่สุดในการทำนายพฤติกรรมการซื้อ โดยมีค่าน้ำหนักเฉลี่ยร้อยละ 19 ขณะเดียวกัน ผลการแบ่งกลุ่มลูกค้าชี้ให้เห็นว่า มีลูกค้า 458 ราย (ร้อยละ 91.60) มีแนวโน้มกลับมาซื้อซ้ำ และมีลูกค้า 42 ราย (ร้อยละ 8.40) ที่มีแนวโน้มไม่กลับมาซื้อซ้ำ ส่วนค่า ROC-AUC เฉลี่ยอยู่ในระดับสูงกว่า 0.95 ทั้งนี้ เมื่อพิจารณาลักษณะข้อมูลที่ไม่สมดุล (ลูกค้าซื้อซ้ำร้อยละ 91.60) ได้มีการประเมินเพิ่มเติมด้วยค่า Precision-Recall AUC (PR-AUC) และ Recall ซึ่งพบว่าตัวแปร RFM มีอิทธิพลสูงที่สุดในการพยากรณ์พฤติกรรมการซื้อของลูกค้า ผลลัพธ์ดังกล่าวสะท้อนให้เห็นว่าการบูรณาการ การวิเคราะห์อาร์เอฟเอ็มเข้ากับอัลกอริทึมการเรียนรู้ของเครื่องสามารถใช้เป็นแนวทางเชิงกลยุทธ์ที่มีประสิทธิภาพในการวิเคราะห์และจัดการพฤติกรรมลูกค้า โดยเฉพาะในธุรกิจขนาดเล็กด้านหัตถกรรม ซึ่งสามารถนำไปประยุกต์ใช้เพื่อสนับสนุนการวางแผนกลยุทธ์ทางการตลาดและการรักษาลูกค้าได้อย่างเป็นระบบ

คำสำคัญ: การเรียนรู้ของเครื่อง การวิเคราะห์แบบ RFM การวิเคราะห์เชิงทำนาย พฤติกรรมการซื้อ เครื่องประดับแฮนด์เมด

ABSTRACT

This study aims to 1) develop predictive models for customer purchasing behavior, 2) evaluate the performance of the developed models, and 3) apply the most effective model to a real business case of Sue's Beads Handmade Jewelry. Three algorithms were employed, namely Decision Tree (DT), Gradient Boosted Tree (GBT), and Logistic Regression (LR), in combination with RFM (Recency, Frequency, Monetary) analysis. The dataset consisted of purchase records of 500 customers collected between January 2023 and December 2024. Data were split chronologically using a time-based hold-out method, with 80% used for training and 20% for testing, to prevent data leakage and reflect real-world predictive conditions. The results show that Logistic Regression achieved the highest performance with 100% accuracy across all evaluation metrics, whereas Decision Tree and Gradient Boosted Tree achieved equal accuracy of 95.40%. The RFM score was identified as the most influential predictor of customer purchasing behavior, with an average importance weight of 19%. Customer segmentation revealed that 458 customers (91.60%) were likely to make repeat purchases, whereas 42 customers (8.40%) were unlikely to return to the store. The average ROC-AUC value exceeded 0.95. Considering the imbalanced nature of the dataset (91.6% repeat customers), additional evaluation metrics, namely Precision-Recall AUC (PR-AUC) and Recall, were applied to provide a more balanced and fair assessment of model performance. These findings indicate that integrating RFM analysis with machine learning algorithms provides a reliable and practical tool for analyzing and managing customer behavior. For small businesses, particularly in the handicraft sector, such predictive models can be applied to support marketing strategy development and customer retention.

Keywords: Machine Learning, RFM Analysis, Predictive Analytics, Purchasing Behavior, Handmade Jewelry

บทนำ

ในยุคของเศรษฐกิจดิจิทัลปัจจุบันที่การแข่งขันทางธุรกิจมีความเข้มข้นและเปลี่ยนแปลงอย่างรวดเร็วนั้น การเข้าใจและคาดการณ์พฤติกรรมของลูกค้าได้อย่างถูกต้องและแม่นยำ ถือเป็นปัจจัยสำคัญที่มีผลต่อความสำเร็จและความอยู่รอดขององค์กรในทุกระดับชั้น โดยเฉพาะอย่างยิ่ง ธุรกิจขนาดกลางและขนาดเล็ก (SMEs) ที่ต้องเผชิญกับข้อจำกัดทางด้านทรัพยากรและการพึ่งพาฐานลูกค้าเดิมเป็นสำคัญ จากการวิจัยในอดีตชี้ให้เห็นว่า ประเด็นในเรื่องของต้นทุนในการหาลูกค้าใหม่มีมูลค่าสูงกว่าการรักษาฐานลูกค้าเดิมถึง 5 เท่า ดังนั้นกลยุทธ์ทางการรักษาความสัมพันธ์กับลูกค้าเพื่อลดปริมาณต้นทุนที่ไม่จำเป็นจึงกลายเป็นประเด็นสำคัญที่ต้องได้รับการพิจารณากันอย่างจริงจังกับแวดวงธุรกิจในปัจจุบัน

การพัฒนาแบบก้าวกระโดดของเทคโนโลยีสารสนเทศและอัลกอริทึมการเรียนรู้ของเครื่องในช่วงทศวรรษที่ผ่านมาถือเป็นการเปิดโอกาสให้กับภาคธุรกิจต่าง ๆ สามารถนำข้อมูลมาวิเคราะห์และคาดการณ์พฤติกรรมของผู้บริโภคได้อย่างมีประสิทธิภาพเพิ่มมากขึ้น ซึ่งการนำเทคโนโลยีดังกล่าวเข้ามาประยุกต์ใช้งาน นอกจากจะช่วยลดความเสี่ยงจากการสูญเสียลูกค้าและลดต้นทุนที่ไม่จำเป็นแล้ว ยังช่วยส่งเสริมให้เกิดการวางแผนกลยุทธ์ใหม่ที่เหมาะสมและการบริหารความสัมพันธ์กับลูกค้าได้อย่างมีศักยภาพ (Munde and Kaur, 2024) โดยเฉพาะอย่างยิ่ง ธุรกิจที่มีลักษณะเฉพาะตัวสูง เช่น อุตสาหกรรมเครื่องประดับแฮนด์เมด ที่มีทั้งคุณค่าทางศิลปะความเป็นเอกลักษณ์ และความสัมพันธ์ที่ใกล้ชิดกับลูกค้า ซึ่งตลาดเครื่องประดับแฮนด์เมดในปัจจุบันมีมูลค่าสูงถึง 90,000 ล้านบาท และมีอัตราการเติบโตเฉลี่ยต่อปีอยู่ที่ประมาณร้อยละ 12 ซึ่งสะท้อนให้เห็นว่า ธุรกิจมีศักยภาพของการเติบโตได้ดีในระยะยาว อย่างไรก็ตาม ธุรกิจกลุ่มนี้ยังคงเผชิญหน้ากับ

ความท้าทายที่หลากหลาย ทั้งจากทางด้านการแข่งขันในตลาดโลกออนไลน์และการเปลี่ยนแปลงอย่างรวดเร็วของพฤติกรรมผู้บริโภค (Brüggenmann et al., 2025)

หนึ่งในเทคนิคสำคัญของการวิเคราะห์ข้อมูลเกี่ยวกับพฤติกรรมของลูกค้าคือ RFM Analysis (Recency, Frequency, Monetary) ซึ่งแบ่งกลุ่มลูกค้าตามความล่าสุดของการซื้อ ความถี่ และมูลค่าการใช้จ่าย โดยเทคนิคดังกล่าวเมื่อนำมาประยุกต์ใช้งานร่วมกับอัลกอริทึมการเรียนรู้ของเครื่องแบบต่าง ๆ จะสามารถเพิ่มประสิทธิภาพในการคาดการณ์พฤติกรรมลูกค้าได้อย่างมีนัยสำคัญ เช่น ต้นไม้ตัดสินใจ เกรเดียนต์บูสต์ทรี และการถดถอยโลจิสติก จากการวิจัยที่ผ่านมาพบว่า การประยุกต์ใช้งานร่วมกันระหว่าง 2 เทคนิค ส่วนใหญ่แล้วจะถูกนำมาประยุกต์ใช้งานในบริบทขององค์กรขนาดใหญ่ที่มีข้อมูลจำนวนมาก (Chen et al., 2012; Hossain et al., 2023) เป็นสำคัญ ในขณะที่ธุรกิจขนาดกลางและขนาดเล็ก โดยเฉพาะอย่างยิ่ง อุตสาหกรรมเครื่องประดับแฮนด์เมด ซึ่งยังมีข้อจำกัดทางด้านจำนวนข้อมูลและทรัพยากร ยังขาดแนวทางของการนำเทคนิคดังกล่าวมาประยุกต์ใช้งานให้เหมาะสมกับลักษณะข้อมูลและบริบทของธุรกิจ ทั้งในแง่ของการเตรียมข้อมูล การเลือกอัลกอริทึมที่สอดคล้องกับขนาดและคุณภาพของข้อมูล ตลอดจนการประเมินประสิทธิภาพของแบบจำลองที่เหมาะสมสำหรับธุรกิจที่มีฐานลูกค้าจำกัด (Chuang et al., 2017) ซึ่งงานวิจัยก่อนหน้านี้จะประยุกต์ใช้การวิเคราะห์อาร์เอฟเอ็มร่วมกับอัลกอริทึมการเรียนรู้ของเครื่องเพื่อพยากรณ์พฤติกรรมผู้บริโภคในภาคค้าปลีกและอีคอมเมิร์ซ (Hossain et al. (2023); Baisani et al. (2021) แต่ยังคงขาดการศึกษาในบริบทของธุรกิจขนาดกลางและขนาดย่อม (SME) โดยเฉพาะธุรกิจหัตถกรรมแฮนด์เมด ซึ่งมีข้อจำกัดด้านปริมาณข้อมูล ความไม่สม่ำเสมอของพฤติกรรมผู้บริโภค และทรัพยากรที่จำกัด อีกทั้งงานที่ผ่านมาไม่ค่อยให้ความสำคัญกับปัญหาทางเทคนิคที่ส่งผลต่อความแม่นยำของแบบจำลอง เช่น การรั่วไหลของข้อมูล (Kapoor and Narayanan 2023) การประเมินแบบตามเวลา (Cerqueira et al., 2020) และการจัดการข้อมูลไม่สมดุล (He and Garcia 2009) ดังนั้น งานวิจัยนี้จึงมุ่งพัฒนาและประเมินแบบจำลองการเรียนรู้ของเครื่องร่วมกับการวิเคราะห์อาร์เอฟเอ็ม เพื่อพยากรณ์พฤติกรรมผู้บริโภคซ้ำในธุรกิจหัตถกรรมแฮนด์เมด อันเป็นกรณีศึกษาของ SME เพื่อแสดงศักยภาพของการวิเคราะห์เชิงพยากรณ์ในการสนับสนุนการตัดสินใจเชิงกลยุทธ์ของธุรกิจขนาดเล็ก จากประเด็นปัญหาดังกล่าวข้างต้น การวิจัยนี้มีจึงมีวัตถุประสงค์สำคัญเพื่อวิเคราะห์และพัฒนาแบบจำลองสำหรับการทำนายพฤติกรรมผู้บริโภคซ้ำของลูกค้าในธุรกิจเครื่องประดับแฮนด์เมด โดยจะเป็นการใช้ข้อมูลประวัติการซื้อของลูกค้าจำนวน 500 ราย จากร้านชูปิดแฮนด์เมดจิ๋วเอเลรี ซึ่งเป็นธุรกิจขนาดกลางในประเทศไทย ที่มีการเก็บข้อมูลตั้งแต่ในช่วงเดือนมกราคม พ.ศ. 2566 จนถึงเดือนธันวาคม พ.ศ. 2567 ผ่านการสุ่มตัวอย่างแบบเฉพาะเจาะจงจากฐานลูกค้าทั้งหมด 2,000 ราย กระบวนการวิจัยจะเป็นการใช้เทคนิคการวิเคราะห์ RFM ร่วมกับอัลกอริทึมการเรียนรู้ของเครื่อง 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ เกรเดียนต์บูสต์ทรี และการถดถอยโลจิสติก ซึ่งแต่ละอัลกอริทึมที่นำมาใช้ในการวิจัยจะมีความสำคัญกับในด้านของการประมวลผลข้อมูลและความเหมาะสมกับลักษณะของข้อมูลที่แตกต่างกัน โดยอัลกอริทึมเหล่านี้สามารถนำมาประยุกต์ใช้งานร่วมกับการแก้ไขปัญหาทางด้านการจำแนกประเภทได้อย่างมีนัยสำคัญ ซึ่งจะช่วยให้สามารถเปรียบเทียบประสิทธิภาพและเลือกแบบจำลองที่เหมาะสมที่สุดกับสถานการณ์จริงได้ ดังนั้นการวิจัยนี้จึงไม่ใช่เพียงแค่ออกแบบและทดสอบเชิงทฤษฎีทางด้านการวิเคราะห์ข้อมูลเชิงทำนายเท่านั้นแต่ยังเป็นแนวปฏิบัติสำคัญที่สามารถนำไปประยุกต์ใช้งานได้จริง ในบริบทของธุรกิจขนาดกลางและขนาดเล็กที่ต้องอาศัยเครื่องมือสนับสนุนการตัดสินใจและการวางแผนกลยุทธ์ใหม่ เพื่อนำไปสู่การบริหารความสัมพันธ์กับลูกค้าและการพัฒนาธุรกิจได้อย่างยั่งยืนในบริบทของเศรษฐกิจดิจิทัลปัจจุบัน

วิธีดำเนินการวิจัย

การวิจัยนี้ดำเนินการพัฒนาและเปรียบเทียบประสิทธิภาพของแบบจำลองสำหรับการทำนายพฤติกรรมผู้บริโภคซ้ำของลูกค้า โดยบูรณาการร่วมกันระหว่างเทคนิคการวิเคราะห์ข้อมูล RFM และอัลกอริทึมการเรียนรู้ของเครื่อง Baisani et al. (2021) ที่ใช้ข้อมูลจากกรณีศึกษาร้านชูปิดแฮนด์เมดจิ๋วเอเลรี วิธีดำเนินการวิจัยประกอบด้วย การกำหนดประชากรและกลุ่มตัวอย่าง การเก็บรวบรวมและเตรียมข้อมูล การวิเคราะห์ข้อมูล ตลอดจนการสร้างและประเมินผลแบบจำลอง ตามลำดับ

1. ประชากรและกลุ่มตัวอย่าง

ประชากรที่ใช้ในการวิจัยนี้ คือ ลูกค้าของร้านซูบีตแฮนด์เมดจิวเวลรี่ (Sue's Beads Handmade Jewelry) ที่มีประวัติการซื้อสินค้า ตั้งแต่เดือนมกราคม พ.ศ. 2566 ถึงเดือนธันวาคม พ.ศ. 2567 เป็นจำนวนทั้งสิ้น 2,000 ราย สำหรับการคัดเลือกกลุ่มตัวอย่างจะเป็นการใช้วิธีการสุ่มตัวอย่างแบบเฉพาะเจาะจง (Purposive Sampling) เพื่อเลือกเฉพาะลูกค้าที่มีข้อมูลการซื้อครบถ้วนและมีความต่อเนื่องตามเกณฑ์ที่การวิจัยกำหนด จากผลการคัดเลือกได้กลุ่มตัวอย่างทั้งสิ้นเป็นจำนวน 500 ราย สำหรับการคุ้มครองข้อมูลส่วนบุคคลนั้นจะเป็นการแทนชื่อและข้อมูลเฉพาะบุคคลของลูกค้า ด้วยรหัสประจำตัวลูกค้า (Customer ID) ที่ไม่สามารถระบุตัวตนที่แท้จริงได้ สำหรับตัวแปรที่ใช้ในการวิจัยจะแบ่งออกเป็น 2 กลุ่มหลัก ได้แก่ ตัวแปรด้านประชากรศาสตร์ (เพศ อายุ และอาชีพ) และตัวแปรด้านพฤติกรรมการซื้อ (จำนวนวันระหว่างการซื้อ จำนวนครั้งในการซื้อมูลค่าการซื้อรวม ประเภทสินค้าที่ชื่นชอบและคะแนนสะสม) ซึ่งโครงสร้างและการเตรียมข้อมูลสามารถแสดงได้ดังตารางที่ 1

ตารางที่ 1 ข้อมูลที่ใช้ในการวิจัย

ตัวแปร	ประเภท	คำอธิบาย
Sex	เชิงกลุ่ม	เพศ
Age	จำนวนจริง	อายุ
Occupation	เชิงกลุ่ม	อาชีพ
Days	จำนวนจริง	จำนวนวันนับตั้งแต่จากที่ซื้อครั้งแรกถึงครั้งสุดท้าย
Total Purchases	จำนวนจริง	จำนวนครั้งการซื้อทั้งหมด
Total Spent	จำนวนจริง	มูลค่าการซื้อทั้งหมด (บาท)
Favourite Category	เชิงกลุ่ม	ประเภทของสินค้า
Royal Point	จำนวนจริง	คะแนนสะสม

2. เครื่องมือในการวิจัย

การดำเนินการวิจัยนี้ เครื่องมือที่ใช้จะถูกเลือกและจัดลำดับตามกระบวนการวิจัยอย่างเป็นระบบ เริ่มต้นจากขั้นตอนของการจัดการและเตรียมข้อมูล จะเป็นการใช้โปรแกรม Microsoft Excel สำหรับรวบรวม จัดหมวดหมู่ และปรับปรุงข้อมูลดิบให้มีความถูกต้องและครบถ้วนสมบูรณ์ รวมถึงแปลงข้อมูลให้อยู่ในรูปแบบไฟล์ CSV (Comma-Separated Values) เพื่อรองรับกับการประมวลผลในขั้นตอนต่อไป (Ashokan, 2023) สำหรับขั้นตอนของการวิเคราะห์ข้อมูลและพัฒนาแบบจำลองการเรียนรู้ของเครื่องนั้น จะเป็นการใช้ซอฟต์แวร์ RapidMiner Studio Altair RapidMiner เวอร์ชัน 2024.1.0 เอกสิทธิ์ (2566) ซึ่งเป็นเครื่องมือหลักในการสร้างแบบจำลอง วิเคราะห์ประสิทธิภาพ และเปรียบเทียบผลลัพธ์ของแบบจำลองสำหรับฮาร์ดแวร์ที่ใช้ในการประมวลผลในการวิจัยนี้ จะเป็นเครื่องคอมพิวเตอร์ส่วนบุคคลที่สามารถจัดการข้อมูลขนาดใหญ่และรองรับการพัฒนาแบบจำลองการเรียนรู้ของเครื่องในลักษณะการวิจัยเชิงประยุกต์ได้ โดยมีหน่วยประมวลผลกลาง Intel Core i7-10700H ความเร็ว 2.6 GHz (8 Cores, 16 Threads) หน่วยความจำ 16 GB DDR4 และหน่วยเก็บข้อมูล SSD ขนาด 512 GB ทำงานบนระบบ Windows 11 Pro 64-bit ซึ่งมีประสิทธิภาพเพียงพอสำหรับการประมวลผลข้อมูลลูกค้าจำนวน 500 รายและการพัฒนาแบบจำลองการเรียนรู้ของเครื่องตามมาตรฐานที่กำหนด (Altair Engineering Inc., 2024) รวมถึงการใช้งานซอฟต์แวร์ RapidMiner Studio ในการวิเคราะห์ข้อมูลลูกค้าสำหรับธุรกิจขนาดเล็กจำเป็นต้องใช้ทรัพยากรด้านฮาร์ดแวร์ที่เพียงพอ โดยเฉพาะหน่วยความจำไม่น้อยกว่า 8 GB และหน่วยประมวลผลแบบหลายคอร์ เพื่อสนับสนุนการประมวลผลอย่างมีประสิทธิภาพ (Alghamdi et al., 2024)

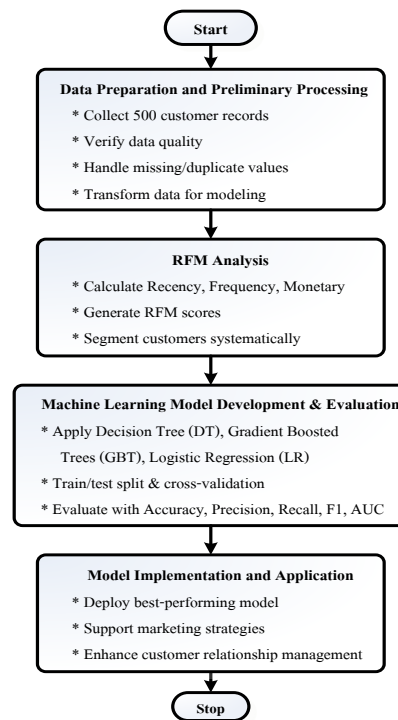
3. กระบวนการดำเนินการวิจัย แบ่งออกได้เป็น 4 ขั้นตอนสำคัญ คือ

(1) การเตรียมข้อมูลและการประมวลผลเบื้องต้น เพื่อให้ได้ข้อมูลที่ถูกต้อง ครบถ้วน และพร้อมสำหรับการวิเคราะห์โดยมุ่งเน้นการตรวจสอบคุณภาพของข้อมูล การแก้ไขปัญหาข้อมูลสูญหายและข้อมูลซ้ำซ้อน ตลอดจนการปรับให้อยู่ในรูปแบบที่เหมาะสมสำหรับการสร้างแบบจำลอง

(2) การวิเคราะห์ข้อมูลอาร์เอฟเอ็ม เพื่อสร้างตัวแปรนำเข้าที่สะท้อนพฤติกรรมการณ์ซื้อของลูกค้าในมิติที่สำคัญทั้งในด้านของความถี่ ความล่าสุด และมูลค่าการซื้อ เพื่อรองรับกับการจำแนกกลุ่มลูกค้าได้อย่างเป็นระบบ

(3) การพัฒนาและประเมินแบบจำลองการเรียนรู้ของเครื่อง เพื่อทำนายพฤติกรรมการณ์ซื้อซ้ำของลูกค้าด้วยอัลกอริทึมที่เหมาะสม พร้อมประเมินประสิทธิภาพของแต่ละแบบจำลองโดยใช้ตัวชี้วัดมาตรฐาน เพื่อให้ได้แบบจำลองที่แม่นยำและสามารถนำไปประยุกต์ใช้งานได้จริง

(4) การนำผลลัพธ์ไปประยุกต์ใช้งาน เพื่อสนับสนุนการตัดสินใจเชิงกลยุทธ์การพัฒนาแผนการตลาด และการบริหารความสัมพันธ์กับลูกค้าได้อย่างมีประสิทธิภาพในบริบทของธุรกิจขนาดกลางและขนาดเล็ก สำหรับภาพรวมของกระบวนการดำเนินการวิจัยทั้ง 4 ขั้นตอน สามารถแสดงได้ดังภาพที่ 1



ภาพที่ 1 กระบวนการดำเนินการวิจัย

3.1 การเตรียมข้อมูลและการประมวลผลเบื้องต้น (Data Preparation and Preliminary Processing)

ในขั้นตอนแรกของการวิจัย ได้ทำการรวบรวมข้อมูลประวัติการซื้อของลูกค้าจำนวน 500 ราย โดยข้อมูลประกอบด้วยตัวแปรด้านประชากรศาสตร์ (รหัสลูกค้า เพศ อายุ อาชีพ) พฤติกรรมการณ์ซื้อ (วันที่ซื้อครั้งแรก วันที่ซื้อครั้งล่าสุด จำนวนวันระหว่างการซื้อ จำนวนครั้งในการซื้อ และยอดใช้จ่ายรวม) รวมถึงข้อมูลด้านความชอบ (หมวดหมู่สินค้าที่สนใจ และคะแนนสะสม) สามารถแสดงได้ดังภาพที่ 2 จากนั้นดำเนินการสำรวจและวิเคราะห์ข้อมูลเบื้องต้น (Exploratory Data Analysis: EDA) ทั้งเชิงคุณภาพ (เพศ อาชีพ และประเภทสินค้าที่ชื่นชอบ) และเชิงปริมาณ (อายุ จำนวนวันระหว่างการซื้อ ยอดใช้จ่ายรวม และคะแนนสะสม) เพื่อทำความเข้าใจโครงสร้างข้อมูลและตรวจสอบความสมบูรณ์ สามารถแสดงได้ดังภาพที่ 3 ต่อมาทำการทำความสะอาดข้อมูล (Data Cleaning) เพื่อเพิ่มความถูกต้องและความน่าเชื่อถือของข้อมูล โดยจัดการกับ

ค่าที่หายไป ข้อมูลซ้ำซ้อน ข้อมูลที่ไม่สอดคล้องกัน และค่าผิดปกติของกระบวนการจัดการค่าที่หายไป (Missing Value) พบว่าในคอลัมน์ Date/Time มีการใช้เครื่องหมายคำถาม (?) แทนค่าที่ว่าง โดยเฉพาะในแถวที่ 1 ซึ่งอาจส่งผลให้โครงสร้างของชุดข้อมูลไม่ถูกต้อง และเพื่อแก้ไขปัญหาดังกล่าว ผู้วิจัยได้ดำเนินการกำหนดให้แถวที่ 1 เป็นส่วนของ Header ของชุดข้อมูล ส่งผลให้ข้อมูลในแถวถัดไปเลื่อนขึ้นมามีตำแหน่งที่ว่างอย่างเหมาะสม ซึ่งช่วยให้รูปแบบของชุดข้อมูลมีความเป็นระเบียบและพร้อมสำหรับการประมวลผลในขั้นตอนต่อไป ผลการดำเนินการดังกล่าวสามารถแสดงได้ในภาพที่ 4 และภาพที่ 5 เอกสิทธิ์ (2566) หลังจากนั้นจึงแปลงข้อมูล (Data Transformation) ด้วยวิธีการ Min-Max Scaling และ Standardization เพื่อปรับค่าตัวแปรให้อยู่ในรูปแบบมาตรฐานที่เหมาะสมสำหรับการสร้างแบบจำลองในขั้นตอนถัดไป สามารถแสดงได้ดังภาพที่ 6

Select the cells to import.

Sheet: Sue_beads_customer_data.csv Cell range: A:K Select All Define header row: 1

Customer ID	Firstname	Sex	Age	Occupation	First Purchase	Last Purchase	Total Purchase	Total Spent	Favorite Item	Royal points
CUS001	David	Male	37	Artist	May 28, 2023	Aug 1, 2023	16	28064	Earrings	3
CUS002	John	Male	67	Salesperson	Aug 7, 2023	Sep 2, 2023	18	50076	Earrings	3
CUS003	Lisa	Female	58	Salesperson	Mar 11, 2022	Mar 24, 2022	11	33386	Bracelets	3
CUS004	Jane	Female	67	Artist	Jan 2, 2022	Jan 11, 2022	2	4300	Rings	3
CUS005	Lisa	Female	30	Nurse	Feb 6, 2022	Feb 18, 2022	17	24667	Rings	3
CUS006	Lisa	Female	26	Manager	May 29, 2023	Oct 25, 2023	9	27873	Sets	3
CUS007	John	Male	39	Salesperson	Jul 5, 2023	Mar 16, 2024	19	65684	Bracelets	3
CUS008	Emily	Female	64	Teacher	Oct 1, 2022	Nov 3, 2023	2	5162	Rings	3
CUS009	Jane	Female	34	Lawyer	Dec 2, 2023	Aug 7, 2024	5	3430	Sets	3
CUS010	Chris	Male	40	Chef	Oct 26, 2022	May 13, 2023	9	23843	Rings	3
CUS011	Anna	Female	35	Accountant	Feb 1, 2022	Feb 1, 2023	14	56580	Earrings	3
CUS012	Chris	Male	56	Accountant	Feb 27, 2022	Apr 6, 2023	5	18910	Rings	3
CUS013	Mike	Male	26	Accountant	Jul 8, 2023	Dec 25, 2023	20	49200	Necklaces	3
CUS014	Lisa	Female	54	Manager	Sep 12, 2022	Mar 10, 2023	6	21300	Rings	3
CUS015	Lisa	Male	31	Manager	Dec 7, 2022	Jan 18, 2023	10	65408	Sets	3
CUS016	Tom	Male	44	Salesperson	Dec 19, 2022	Feb 17, 2023	8	30800	Rings	3
CUS017	Lisa	Female	64	Accountant	Aug 29, 2022	Feb 26, 2023	10	54900	Rings	3
CUS018	Tom	Female	64	Doctor	Oct 3, 2023	Nov 15, 2023	8	22680	Bracelets	3
CUS019	Scott	Male	50	Doctor	Aug 29, 2022	Feb 26, 2023	9	59300	Earrings	3
CUS020	Jane	Female	43	Nurse	Feb 20, 2022	Nov 15, 2022	8	27500	Rings	3

ภาพที่ 2 การเตรียมชุดข้อมูลในรูปแบบ Excel.CSV

Customer ID	Firstname	Sex	Age	Occupation	First Purchase	Last Purchase	Days	Total Purchase	Total Spent	Favorite Item	Royal points
CUS001	David	Male	37	Artist	May 28, 2023	Aug 1, 2023	65	16	28064	Earrings	3
CUS002	John	Male	67	Salesperson	Aug 7, 2023	Sep 2, 2023	26	18	50076	Earrings	3
CUS003	Lisa	Female	58	Salesperson	Mar 11, 2022	Mar 24, 2022	13	11	33386	Bracelet	2
CUS004	Jane	Female	67	Artist	Jan 2, 2022	Jan 11, 2022	9	2	4300	Rings	1
CUS005	Lisa	Female	30	Nurse	Feb 6, 2022	Feb 18, 2022	12	17	24667	Rings	2
CUS006	Lisa	Female	26	Manager	May 29, 2023	Oct 25, 2023	149	9	27873	Rings	2
CUS007	John	Male	39	Salesperson	Jul 5, 2023	Mar 16, 2024	255	19	65684	Bracelet	3
CUS008	Emily	Female	64	Teacher	Oct 1, 2022	Nov 3, 2023	33	2	5162	Sets	1
CUS009	Jane	Female	34	Lawyer	Dec 2, 2023	Aug 7, 2024	249	5	3430	Sets	1
CUS010	Chris	Male	40	Chef	Oct 26, 2022	May 13, 2023	199	9	23843	Rings	2
CUS011	Anna	Female	35	Accountant	Feb 1, 2022	Feb 1, 2023	365	14	56580	Rings	3
CUS012	Chris	Male	56	Accountant	Feb 27, 2022	Apr 6, 2023	38	5	18910	Rings	2

500 rows-12columns (5 nominal , 5 numerical, 2 Data

ภาพที่ 3 ชุดข้อมูลที่นำเข้าไว้ในระบบโปรแกรม RapidMiner

att1 Category	att2 Category	att3 Category	att4 Category	att5 Category	att6 Date / Time	att7 Date / Time	att8 Category	att9 Category	att10 Category	att11 Category
Customer ID	Firstname	Sex	Age	Occupation	?	?	Total Purchases	Total Spent	Favorite Category	Royal Points
CUS001	David	Male	37	Artist	May 28, 2023	Aug 1, 2023	16	28054	Earrings	3
CUS002	John	Male	67	Salesperson	Aug 7, 2023	Sep 2, 2023	18	50076	Earrings	3
CUS003	Lisa	Female	58	Salesperson	Jan 11, 2022	Dec 3, 2022	11	33836	Sets	3
CUS004	Jane	Female	67	Artist	Jan 2, 2022	Jan 11, 2022	2	4300	Rings	1
CUS005	Lisa	Female	30	Nurse	Feb 6, 2022	Feb 18, 2022	17	24667	Rings	3
CUS006	Lisa	Female	26	Manager	May 29, 2023	Oct 26, 2023	9	27873	Sets	3
CUS007	John	Male	39	Salesperson	Jul 5, 2023	Mar 16, 2024	19	65664	Bracelets	3
CUS008	Emily	Female	64	Teacher	Oct 1, 2022	Nov 3, 2022	2	5162	Rings	2
CUS009	Jane	Female	34	Lawyer	Dec 2, 2023	Aug 7, 2024	5	3430	Sets	1
CUS010	Chris	Male	40	Chef	Oct 26, 2022	May 13, 2023	9	23643	Rings	3

ภาพที่ 4 การตรวจสอบและแก้ไขข้อมูล

Customer ID	Firstname	Sex	Age	Occupation	First Purchase	Last Purchase	Days	Total Purchase	Total Spent	Favorite Item	Royal points
CUS001	David	Male	37	Artist	28/05/2023	01/08/2023	65	16	28054	Earrings	3
CUS002	John	Male	67	Salesperson	07/08/2023	02/09/2023	26	18	50076	Earrings	3

ภาพที่ 5 ผลลัพธ์การแก้ไขข้อมูลและค่าว่าง

Customer ID	Firstname	Sex	Age	Occupation	First Purchase	Last Purchase	Days	Total Purchase
CUS016	Chris	Female	18	Manager	19, 2022	Mar 22, 2023	93	5
CUS190	Lisa	Female	18	Artist	13, 2023	Mar 13, 2024	21	6
CUS204	David	Female	18	Engineer	-	-	21	9
CUS237	Anna	Female	18	Nurse	5, 2023	Nov 30, 2023	298	12
CUS261	Anna	Female	18	Nurse	13, 2022	May 1, 2023	201	10
CUS306	Emily	Male	18	Accountant	Mar 24, 2022	Jul 22, 2022	120	11
CUS368	David	Female	18	Artist	Jul 18, 2022	Jul 6, 2023	353	15
CUS452	Chris	Female	18	Engineer	Jan 5, 2022	Dec 5, 2022	339	2
CUS023	Lisa	Male	19	Accountant	Oct 3, 2022	Apr 23, 2023	202	6
CUS101	Emily	Male	19	Nurse	Feb 7, 2022	Jul 17, 2022	160	3
CUS152	Sarah	Female	19	Teacher	Jan 27, 2023	Nov 23, 2023	300	14
CUS193	Mike	Female	19	Artist	Oct 20, 2023	Sep 8, 2024	324	3

500 rows • 12 columns (5 nominal, 6 numerical, 2 date)

ภาพที่ 6 การแปลงข้อมูลเชิงคุณภาพให้เป็นตัวเลขจากลำดับน้อยไปมาก

3.2 การวิเคราะห์ข้อมูลอาร์เอฟเอ็ม (RFM Analysis)

ในการวิจัยครั้งนี้ ได้นำเทคนิค RFM Analysis ซึ่งเป็นวิธีมาตรฐานที่ใช้ในการสังเคราะห์คุณลักษณะเชิงพฤติกรรมของลูกค้าเพื่อสร้างแบบจำลองเชิงทำนาย (Ho et al., 2023; Putra et al., 2023) มาประยุกต์ใช้กับข้อมูลของลูกค้าแต่ละราย โดยใช้เครื่องมือ Pivot Table ในการสกัดและสรุปตัวแปรที่สำคัญสำหรับการวิเคราะห์ ซึ่งประกอบด้วย 3 องค์ประกอบหลัก ได้แก่

3.2.1 Recency (R) คือ จำนวนวันนับจากการซื้อครั้งล่าสุดจนถึงปัจจุบัน

Frequency (F) คือ จำนวนครั้งที่ลูกค้าทำการซื้อสินค้า

Monetary (M) คือ มูลค่าการซื้อรวมทั้งหมดภายในช่วงเวลาที่ศึกษา (Fader and Hardie, 2013)

กระบวนการดังกล่าวช่วยให้สามารถจำแนกและทำความเข้าใจพฤติกรรมการซื้อของลูกค้าได้อย่างเป็นระบบ โดยผลลัพธ์สามารถแสดงได้ดังภาพที่ 7(ก) และ(ข) ตามลำดับ

Customer ID	Recency	Frequency	AverageTotal Purchase
CUS001	65	16	16
CUS002	26	13	13
CUS003	13	19	19
CUS004	9	9	9
CUS005	285	17	17
CUS006	249	9	9
CUS007	249	5	5
CUS010	35	1,4	1,4

500 rows - 4 columns (1 nominal | 3 numerical)

Customer ID	Recency	Frequency	Monetary
CUS001	65	16	28064
CUS002	26	18	50014
CUS003	13	11	33368
CUS004	9	2	4300
CUS005	12	17	24067
CUS006	149	9	21873
CUS007	255	19	66864
CUS008	18	3	5182
CUS009	249	5	5544
CUS010	190	9	28463
CUS011	132	7	21394
CUS012	380	5	18910

(ก) การสร้างตัวแปร Recency Frequency และ Monetary

(ข) ผลลัพธ์การสร้างตัวแปร RFM

ภาพที่ 7 (ก) การสร้างตัวแปร Recency Frequency และ Monetary
และ (ข) ผลลัพธ์การสร้างตัวแปร RFM จนครบทุกคอลัมน์

ข้อมูลแต่ละองค์ประกอบของ RFM ได้ถูกแปลงค่าให้อยู่ในรูปแบบคะแนนมาตรฐานส่วน 1-5 โดยใช้วิธีการแบ่งแบบช่วง (Discretize) ในรูปแบบ Equal Size เพื่อให้ค่าข้อมูลมีความต่อเนื่องและอยู่ในช่วงมาตรฐานเดียวกันอันจะเอื้อต่อการนำไปประยุกต์ใช้ในการสร้างแบบจำลองเชิงทำนายและการแบ่งกลุ่มลูกค้าอย่างมีประสิทธิภาพ (Ho et al., 2023) ดังนั้นคะแนน RFM ที่ได้จากกระบวนการนี้จึงถูกกำหนดให้เป็นตัวแปรสำคัญในการทำนายพฤติกรรมการซื้อซ้ำของลูกค้าในขั้นตอนถัดไปซึ่งผลลัพธ์สามารถแสดงได้ดังภาพที่ 8(ก) และ(ข) ตามลำดับ

Customer ID	Recency	Frequency	Monetary
CUS001	66	8	2940
CUS002	20	16	1380
CUS003	8	1	1430
CUS004	8	1	2940
CUS005	149	8	2430
CUS006	26	1	2940
CUS007	29	14	4890
CUS008	249	1	2430
CUS009	29	14	2905
CUS011	30	1	5860
CUS012	308	14	1810

508 rows - 4 columns - 508 numerics / 3 numerics

(ก) การแบ่งออกเป็น 5 ช่วง ในการแบ่งแบบ Equal Sizes

(ข) ผลลัพธ์การแปลงข้อมูลในรูปแบบสเกลของ RFM

ภาพที่ 8 (ก) การแบ่งออกเป็น 5 ช่วง ในการแบ่งแบบ Equal Sizes
และ (ข) ผลลัพธ์การแปลงข้อมูลในรูปแบบสเกลของ RFM

3.2.2 การคำนวณคะแนนอาร์เอฟเอ็ม (RFM Score) เป็นขั้นตอนสำคัญในการจำแนกกลุ่มลูกค้า โดยอาศัยการประมวลผลจากตัวแปรหลัก 3 ด้าน ได้แก่ ความเร็วในการซื้อซ้ำ (Recency) ความถี่ในการซื้อ (Frequency) และมูลค่าการซื้อ (Monetary) การรวมคะแนนสามารถดำเนินการได้ทั้งในรูปแบบการบวกหรือการคูณ ทั้งนี้ขึ้นอยู่กับข้อกำหนดน้ำหนักความสำคัญตามกลยุทธ์ทางธุรกิจและวัตถุประสงค์ของการวิเคราะห์ สำหรับการวิจัยครั้งนี้ผู้วิจัยได้ประยุกต์ใช้วิธีการคำนวณคะแนนอาร์เอฟเอ็มตามแนวทางของ เอกสิทธิ์ (2566) โดยอาศัยสมการดังต่อไปนี้

$$\text{RFM Score} = (R \times 100) + (F \times 10) + M \quad (1)$$

โดยที่ R = คะแนนความเร็วล่าสุดในการซื้อ (1-5)

F = คะแนนความถี่ในการซื้อ (1-5)

M = คะแนนมูลค่าการซื้อ (1-5)

ตัวอย่าง: ลูกค้าที่มีคะแนน R=4, F=3, M=5 คะแนน RFM = $(4 \times 100) + (3 \times 10) + 5 = 400 + 30 + 5 = 435$
ดังนั้นคะแนนดังกล่าวแสดงว่าลูกค้ารายนี้มีการซื้อครั้งล่าสุดในระยะเวลาที่ค่อนข้างใกล้ (R=4) มีความถี่ในการซื้อปานกลาง (F=3) และมีมูลค่าการซื้อสูง (M=5)

3.2.3 การพัฒนาระบบการจัดกลุ่มลูกค้าเชิงกลยุทธ์ จากการศึกษาแนวคิด RFM แบบดั้งเดิมที่แบ่งลูกค้าออกเป็น 11 กลุ่ม การวิจัยครั้งนี้ได้ปรับกรอบการจัดกลุ่มให้เหมาะสมกับบริบทของธุรกิจขนาดเล็ก โดยลดจำนวนกลุ่มลงเหลือเพียง 4 กลุ่มหลัก ดังแสดงในภาพที่ 9(ก) การปรับเปลี่ยนดังกล่าวมีเหตุผลทั้งในเชิงทฤษฎีและเชิงปฏิบัติ ได้แก่ 1) ความยืดหยุ่นของระบบคะแนน RFM ช่วยให้สามารถรวมตำแหน่งต่าง ๆ บนแผนภูมิ RFM (25 ช่อง) เข้าด้วยกันเป็นกลุ่มใหญ่ได้อย่างเหมาะสม 2) การกำหนดกลุ่มตามวัตถุประสงค์ทางธุรกิจทำให้การบริหารจัดการลูกค้าเป็นไปอย่างมีประสิทธิภาพมากขึ้น และ 3) การลดจำนวนกลุ่มช่วยเพิ่มประสิทธิภาพในการบริหารจัดการ ทำให้ทีมการตลาดสามารถวางแผนและดำเนินกลยุทธ์ได้อย่างชัดเจนยิ่งขึ้น (Hosseini et al., 2010) ดังแสดงในภาพที่ 9(ข) จากคะแนน RFM ที่ได้ การวิจัยครั้งนี้ได้จำแนกกลุ่มลูกค้าออกเป็น 4 กลุ่มหลักตามระดับความสัมพันธ์และมูลค่าทางธุรกิจ ได้แก่ กลุ่มลูกค้าที่ต้องกระตุ้น (Activation Customers) มีค่า RFM ในระดับต่ำ กลุ่มลูกค้ามีศักยภาพ (Potential Customers) แสดงแนวโน้มการเติบโตที่ดี กลุ่มลูกค้าที่ต้องรักษา (Maintenance Customers) มีประวัติการซื้ออย่างต่อเนื่อง และกลุ่มลูกค้าชั้นนำ (Premium Customers) มีความสัมพันธ์ที่แข็งแกร่งและสร้างคุณค่าทางธุรกิจสูงสุดการจัดกลุ่มดังกล่าวสามารถแสดงได้ในภาพที่ 9(ค)



(ก) การวิเคราะห์กลุ่มลูกค้าด้วยวิธี RFM 4 กลุ่ม



(ข) การวิเคราะห์การแบ่งกลุ่ม RFM



(ค) ตารางการวิเคราะห์การจัดกลุ่มลูกค้า 4 กลุ่ม

ภาพที่ 9 (ก) การวิเคราะห์กลุ่มลูกค้าด้วยวิธี RFM 4 กลุ่ม (ข) การวิเคราะห์การแบ่งกลุ่ม RFM และ (ค) ตารางการวิเคราะห์การจัดกลุ่มลูกค้า 4 กลุ่ม

หลักการวิเคราะห์ที่ได้พิจารณาข้อมูลสำคัญ 4 ตัวแปร ได้แก่ คะแนนรวม RFM ความถี่ในการซื้อ (Frequency) มูลค่าการซื้อ (Monetary) และระยะเวลาการกลับมาซื้อซ้ำ (Recency) โดยโครงสร้างข้อมูลที่ใช้ในการวิเคราะห์ประกอบด้วยคอลัมน์หลัก คือ RFM_Score ซึ่งแสดงช่วงคะแนนที่กำหนดให้กับลูกค้าแต่ละกลุ่ม คอลัมน์ Frequency ที่แสดงจำนวนครั้งในการซื้อสินค้า คอลัมน์ Monetary ที่แสดงมูลค่าการซื้อสินค้า และคอลัมน์ Recency ที่แสดงระยะเวลาล่าสุดที่ลูกค้ากลับมาซื้อสินค้า การจัดโครงสร้างข้อมูลในลักษณะดังกล่าวช่วยให้สามารถวิเคราะห์และทำความเข้าใจพฤติกรรมของลูกค้าได้อย่างครบถ้วนในทุกมิติ ดังแสดงในภาพที่ 10(ก) ภายหลังการคำนวณค่า RFM_Score ผู้วิจัยได้ดำเนินการแปลงข้อมูลให้อยู่ในรูปแบบของการแบ่งกลุ่มลูกค้าออกเป็น 4 กลุ่ม ก่อนนำไปใช้ในกระบวนการพยากรณ์ในขั้นตอนถัดไป ดังแสดงในภาพที่ 10(ข) ทั้งนี้ได้ใช้ฟังก์ชัน Generate Attributes ในโปรแกรม RapidMiner โดยกำหนดสมการสำหรับการแบ่งกลุ่มลูกค้าดังต่อไปนี้

if([RFM_score] >= 400, "4", if([RFM_score] >= 300, "3", if([RFM_score] >= 200, "2", "1"))) (2)

ซึ่งสมการดังกล่าวถูกนำมาใช้ในการจัดประเภทลูกค้าตามค่า RFM_Score โดยผลลัพธ์ที่ได้จากการแปลงข้อมูลจะถูกแบ่งออกเป็นกลุ่มตามเกณฑ์ที่กำหนดไว้ ทำให้สามารถระบุได้อย่างชัดเจนว่าลูกค้าแต่ละรายอยู่ในกลุ่มใดจากคะแนน RFM_Score ที่สะท้อนพฤติกรรมการซื้อ ระบบการจัดกลุ่มนี้กำหนดให้ลูกค้าที่มีคะแนนตั้งแต่ 400 ขึ้นไปอยู่ในกลุ่มระดับ 4 ซึ่งถือเป็นกลุ่มลูกค้าคุณภาพสูงสุด ลูกค้าที่มีคะแนนระหว่าง 300–399 อยู่ในกลุ่มระดับ 3 ลูกค้าที่มีคะแนนระหว่าง 200–299 อยู่ในกลุ่มระดับ 2 และลูกค้าที่มีคะแนนต่ำกว่า 200 อยู่ในกลุ่มระดับ 1 ตามลำดับ

RFM.score Number	Recency Number	Monetary Number	Frequency Number
142	1	1	4
153	1	1	5
132	3	2	3
111	1	2	5
152	1	4	5
332	3	2	4
454	4	4	1
421	1	1	1
332	4	2	3
111	3	1	4
344	3	4	5

Group Category	RFM_score Number	Recency Number	Monetary Number	Frequency Number	Customer ID Category
1	111	1	1	4	CUS004
1	153	1	2	5	CUS005
2	332	3	2	3	CUS006
4	154	1	4	5	CUS007
4	332	3	2	5	CUS005
5	111	1	1	1	CUS008
1	421	4	2	1	CUS009
5	332	1	2	3	CUS010
1	332	1	2	3	CUS011
4	110	1	1	2	CUS012
3	344	3	4	4	CUS015

500 rows - 5 columns (1 nominal, 4 numerical)

(ก) ผลลัพธ์จากการแบ่งกลุ่ม 5 ช่วงของ 3 ตัวแปร

(ข) ผลลัพธ์การแปลงข้อมูล RFM_Score เป็นกลุ่ม

ภาพที่ 10 (ก) ผลลัพธ์จากการแบ่งกลุ่ม 5 ช่วงของ 3 ตัวแปร
และ (ข) ผลลัพธ์การแปลงข้อมูล RFM_Score เป็นกลุ่ม

3.2.4 การสร้างตัวแปรเป้าหมาย (Target Variable) ในขั้นตอนการพัฒนาแบบจำลองเพื่อทำนายพฤติกรรมการกลับมาซื้อของลูกค้า ผู้วิจัยได้ดำเนินการเริ่มต้นจากการแปลงค่า RFM Score ให้อยู่ในรูปแบบของกลุ่มลูกค้า (Group) โดยใช้ Operator “Transform” ในโปรแกรม RapidMiner เพื่อจัดกลุ่มลูกค้าตามคะแนน RFM ที่ได้รับ ซึ่งช่วยให้สามารถมองเห็นภาพรวมของการแบ่งกลุ่มลูกค้าได้อย่างชัดเจนยิ่งขึ้น ดังแสดงในภาพที่ 11 หลังจากได้ผลลัพธ์การจัดกลุ่มแล้ว จึงดำเนินการสร้างตัวแปรเป้าหมาย (Target Variable) สำหรับการทำนายพฤติกรรมการกลับมาซื้อซ้ำโดยกำหนดคอลัมน์ใหม่ชื่อ “WILL_BUY” และตั้งค่าเป็นประเภท nominal ที่มีเพียงสองค่า คือ 0 (ไม่กลับมาซื้อซ้ำ) และ 1 (กลับมาซื้อซ้ำ)

โดยใช้สมการ

IF (DATEDIFF (next_purchase_date, last_purchase_date) <= 30, 1, 0) (3)

จากสมการที่ (3) หมายความว่า การคำนวณระยะห่างระหว่างวันที่ลูกค้าซื้อครั้งล่าสุดและวันที่ซื้อครั้งถัดไป โดยถ้า ซื้อซ้ำภายใน 30 วัน ระบบจะกำหนดค่า WILL_BUY = 1 หมายถึง “มีการซื้อซ้ำ” แต่ถ้าเกิน 30 วันหรือไม่มีการซื้อซ้ำ

ระบบจะ กำหนดค่า WILL_BUY = 0 หมายถึง “ไม่มีการซื้อซ้ำ” โดยสมการนี้ถูกใช้ในการสร้างตัวแปรเป้าหมายสำหรับแบบจำลองการพยากรณ์พฤติกรรมการซื้อซ้ำของลูกค้าในอนาคต โดยไม่ถูกนำไปใช้ในขั้นตอนการทดสอบแบบจำลอง เพื่อป้องกันการรั่วไหลของข้อมูล (Label Leakage) และให้ผลลัพธ์สะท้อนพฤติกรรมจริงของลูกค้า แนวทางดังกล่าวสอดคล้องกับข้อเสนอของ Kapoor and Narayanan (2023) ที่ระบุว่า การรั่วไหลของข้อมูลเป็นปัจจัยสำคัญที่ทำให้แบบจำลองการเรียนรู้ของเครื่องมีความแม่นยำสูงเกินจริงและลดความสามารถในการทำซ้ำ (Reproducibility Crisis) สามารถแสดงได้ดังภาพที่ 12(ก) และ(ข) ตามลำดับ

Group Category	RFM_score Number	Recency Number	Monetary Number	Frequency Number	Customer ID Category
1	111	1	1	1	CUS004
1	152	1	2	5	CUS005
3	332	3	2	3	CUS006
4	454	4	4	5	CUS007
1	111	1	1	1	CUS008
4	421	4	1	2	CUS009
3	332	3	2	3	CUS010
4	543	5	3	4	CUS011
1	121	1	1	2	CUS012
4	433	4	3	3	CUS013
1	111	1	1	1	CUS014
3	344	3	4	4	CUS015

500 rows - 6 columns (2 nominal, 4 numerical)

ภาพที่ 11 การเตรียมข้อมูลกลุ่มลูกค้าเพื่อสร้างตัวแปรเป้าหมาย WILL_BUY สำหรับการทำนายพฤติกรรมการกลับมาซื้อซ้ำของลูกค้า

Group

Customer ID

Last Purchase Date

Next Purchase Date

Name: WILL_BUY

Formula:

IF(DATEDIFF(next_purchase_date, last_purchase_date) ≤ 30, 1, 0)

Customer ID	Last Purchase Date	Next Purchase Date	WILL_BUY	Group
CO01	2024-10-25	2024-11-15	1	1
CO02	2024-09-15	2024-09-05	0	1
CO03	2024-09-25	2024-09-05	1	1
CO04	2024-09-20	—	0	1
CO05	2024-07-20	2024-08-10	1	1
CO06	2024-09-15	2024-09-30	0	1
CO05	2024-07-20	2024-08-10	1	1
CUS001	2024-10-01	2024-10-25	1	1
CUS002	2024-09-10	2024-11-15	0	1
CUS003	2024-08-15	2024-09-05	1	1
CUS004	2024-09-25	—	0	1
CUS005	2024-07-20	2024-08-10	1	1
CUS006	2024-09-01	2024-09-25	1	2
CUS007	2024-10-05	2024-10-30	1	2
CUS008	2024-08-10	—	0	1
CUS009	2024-09-15	2024-09-30	1	2
CUS010	2024-07-30	2024-08-20	1	1

(ก) การสร้าง Target Variable

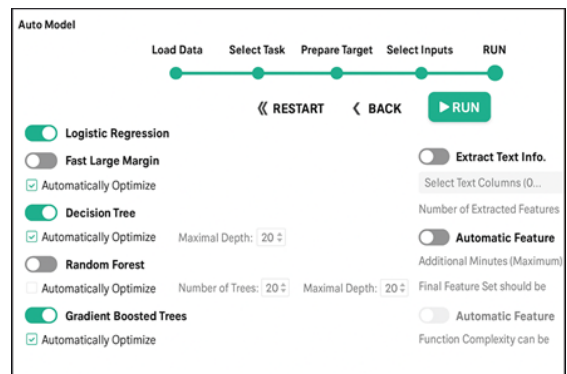
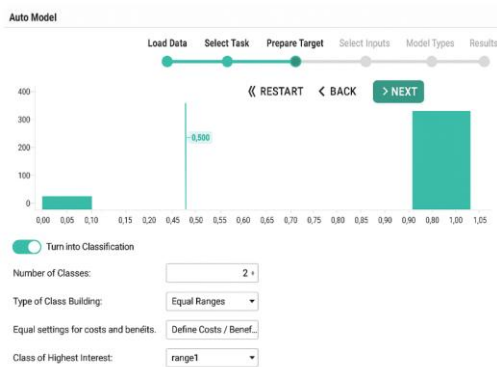
(ข) ผลลัพธ์การสร้าง Target Variable

ภาพที่ 12 (ก) การสร้าง Target Variable และ (ข) ผลลัพธ์การสร้าง Target Variable

3.3 การพัฒนาแบบจำลองการเรียนรู้ของเครื่อง

การวิจัยครั้งนี้ได้ดำเนินการตามลำดับขั้นตอนที่เป็นระบบ เริ่มจากการเตรียมข้อมูล โดยทำการทำความสะอาดข้อมูลและแปลงให้อยู่ในรูปแบบที่เหมาะสม ตลอดจนการประยุกต์ใช้เทคนิค RFM เพื่อวิเคราะห์ตัวแปรด้านความถี่ ความล่าสุด และมูลค่าการซื้อของลูกค้า รวมถึงการจัดการค่าผิดปกติและค่าที่หายไป ทั้งนี้ได้เลือกใช้การสุ่มตัวอย่างแบบเฉพาะเจาะจง (Stratified Sampling) เพื่อให้ได้กลุ่มข้อมูลที่มีความสมบูรณ์และเป็นตัวแทนที่เหมาะสม ในขั้นตอนการพัฒนาแบบจำลอง ผู้วิจัยได้แบ่งข้อมูลออกเป็นสองส่วน คือ ชุดข้อมูลสำหรับฝึกสอน (Training Set) ร้อยละ 80 ของข้อมูลทั้งหมด

(400 รายการ) และชุดข้อมูลสำหรับทดสอบ (Test Set) ร้อยละ 20 ของข้อมูลทั้งหมด (100 รายการ) จากทฤษฎีของ Kohavi (1995) ซึ่งการเลือกใช้สัดส่วนนี้มีเหตุผลสองประการ คือ ประการแรก สัดส่วน 80:20 มีความสมดุลที่เหมาะสมระหว่างข้อมูลสำหรับการเรียนรู้และการทดสอบสำหรับชุดข้อมูลขนาด 500 ตัวอย่าง และประการที่สอง สัดส่วนดังกล่าวสอดคล้องกับการประเมินแบบจำลองด้วยวิธีการทดสอบแบบไขว้ทบ 5 ส่วน (5-Fold Cross Validation) ซึ่งแต่ละรอบใช้ข้อมูลร้อยละ 80 สำหรับการฝึกสอน และร้อยละ 20 สำหรับการทดสอบ ทั้งนี้ผู้วิจัยได้กำหนดค่า Local Random Seed เท่ากับ 1992 เพื่อให้การสุ่มข้อมูลสามารถทำซ้ำได้อย่างมีประสิทธิภาพ สำหรับการกำหนดตัวแปรเป้าหมาย ระบบได้รับการตั้งค่าให้เป็นการจำแนกประเภท (Classification) โดยกำหนด Number of Classes เป็น 2 กลุ่ม และเลือกการแบ่งช่วงแบบ Equal Ranges เพื่อให้เกิดความสมดุลในการจัดกลุ่มข้อมูล ทั้งนี้ได้กำหนดตัวแปรเป้าหมายเป็น “จะกลับมาซื้อซ้ำ” ดังแสดงในภาพที่ 13(ก) ในขั้นตอนสุดท้ายก่อนการสร้างแบบจำลอง ผู้วิจัยได้ใช้เทคนิค Feature Auto Model ในโปรแกรม RapidMiner เพื่อเลือกแบบจำลองที่มีประสิทธิภาพสูงสุด โดยเลือกพัฒนาและทดสอบแบบจำลอง 3 อัลกอริทึม ได้แก่ ต้นไม้ตัดสินใจ (DT) สำหรับการสร้างกฎการตัดสินใจ เกรเดียนต์บูสท์ (GBT) สำหรับการเพิ่มประสิทธิภาพเชิงลำดับ และการถดถอยโลจิสติก (LR) สำหรับการจำแนกประเภท ดังแสดงในภาพที่ 13(ข)



(ก) การจำแนกประเภทลูกค้าจะกลับมาซื้อซ้ำหรือไม่กลับมาซื้อซ้ำ (ข) ขั้นตอนการเลือกแบบจำลองที่ต้องการนำมาวัดประสิทธิภาพ

ภาพที่ 13 (ก) การจำแนกประเภทลูกค้าจะกลับมาซื้อซ้ำหรือไม่กลับมาซื้อซ้ำ
และ (ข) ขั้นตอนการเลือกแบบจำลองที่ต้องการนำมาวัดประสิทธิภาพ

3.3.1 การแบ่งข้อมูลเชิงเวลาและการสร้างตัวแปรเป้าหมาย (Time-Based Hold-Out and Label Definition) การวิจัยนี้แบ่งข้อมูลตามลำดับเวลาเพื่อป้องกันปัญหา Label Leakage โดยใช้ข้อมูลพฤติกรรมการซื้อขายของลูกค้าช่วงเดือนมกราคม พ.ศ. 2566 ถึงเดือนตุลาคม พ.ศ. 2567 สำหรับการฝึกสอนแบบจำลอง (Training Set) และข้อมูลเดือนพฤศจิกายน ถึงเดือนธันวาคม พ.ศ. 2567 สำหรับการทดสอบแบบจำลอง (Hold-Out Test Set) ซึ่งเป็นข้อมูลในอนาคตที่แบบจำลองไม่เคยเห็นมาก่อน โดยตัวแปรเป้าหมาย WILL_BUY ถูกกำหนดจากพฤติกรรมซื้อขายจริงในอนาคต หากลูกค้าซื้อสินค้าภายใน 30 วันหลังช่วงสังเกตพฤติกรรมจะได้ค่า WILL_BUY = 1 และหากไม่ซื้อจะได้ค่า 0 โดยไม่ใช้ค่าจากตัวแปร RFM ในการสร้างป้ายกำกับ เพื่อให้แบบจำลองสามารถเรียนรู้จากเหตุการณ์จริงและสะท้อนความน่าเชื่อถือในการนำไปประยุกต์ใช้ในทางปฏิบัติได้อย่างแม่นยำ

3.3.2 การประเมินประสิทธิภาพแบบจำลอง การวิจัยครั้งนี้ได้ประยุกต์ใช้วิธีการทดสอบแบบไขว้ทบ 5 ส่วน (5-Fold Cross-Validation) ซึ่งเป็นวิธีมาตรฐานที่ได้รับการยอมรับอย่างกว้างขวางในงานวิจัยด้านการเรียนรู้ของเครื่อง (Kohavi, 1995) โดยกระบวนการดังกล่าวเริ่มจากการแบ่งชุดข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน จากนั้นทำการหมุนเวียนใช้ 4 ส่วนเป็นชุดข้อมูลสำหรับการฝึกสอน (Training Set) และอีก 1 ส่วนเป็นชุดข้อมูลสำหรับการทดสอบ (Test Set) การดำเนินการนี้ จะทำซ้ำจนครบทั้ง 5 รอบ เพื่อให้มั่นใจว่าข้อมูลทุกส่วนได้รับการใช้เป็นชุดทดสอบอย่างทั่วถึง ในการ

ประเมินประสิทธิภาพของแบบจำลองได้ใช้การคำนวณจาก Confusion Matrix เพื่อตรวจสอบความน่าเชื่อถือของผลลัพธ์ โดยพิจารณาค่าต่าง ๆ ในหลายมิติ ได้แก่ ค่าความถูกต้อง (Accuracy) ซึ่งแสดงถึงความถูกต้องโดยรวมของแบบจำลอง ค่าความแม่นยำ (Precision) ซึ่งเป็นสัดส่วนของการทำนายผลบวกที่ถูกต้องเมื่อเทียบกับจำนวนที่ทำนายว่าบวกทั้งหมด ค่าความระลึก (Recall) ซึ่งเป็นสัดส่วนของการทำนายผลบวกที่ถูกต้องเมื่อเทียบกับจำนวนผลบวกจริง และค่า F1-Measure ซึ่งเป็นค่าเฉลี่ยฮาร์โมนิกระหว่างค่าความแม่นยำและค่าความระลึก ทั้งนี้การคัดเลือกแบบจำลองที่เหมาะสมที่สุดสำหรับการนำไปประยุกต์ใช้ในการพยากรณ์พฤติกรรมการซื้อขายของลูกค้า พิจารณาจากหลักฐานเชิงปริมาณที่ได้จากค่าสถิติและค่าความคลาดเคลื่อนซึ่งได้จากการทดสอบดังกล่าว ซึ่งข้อมูลทั้งหมดถูกแบ่งตามลำดับเวลา โดยใช้ข้อมูลช่วงเดือนมกราคม พ.ศ. 2566 ถึงเดือนตุลาคม พ.ศ. 2567 สำหรับการฝึกสอนแบบจำลอง และข้อมูลเดือนพฤศจิกายน ถึงเดือนธันวาคม พ.ศ. 2567 สำหรับการทดสอบตามแนวทาง Time-Based Hold-Out Validation เพื่อสะท้อนการใช้งานจริงในเชิงเวลา ภายในชุดฝึกได้ใช้การทดสอบแบบไขว้ทบ 5 ส่วน (5-Fold Cross-Validation) เพื่อปรับค่าพารามิเตอร์ของแบบจำลองและลดความเสี่ยงจาก Overfitting โดยไม่ใช้ข้อมูลจากช่วงเวลาทดสอบ แนวทางดังกล่าวสอดคล้องกับหลักการ Temporal Validation ของ Cerqueira et al. (2020) ซึ่งเสนอให้ใช้วิธี Rolling Origin Evaluation เพื่อหลีกเลี่ยง Temporal Leakage และเพิ่มความแม่นยำของการทำนายในสถานการณ์จริง

3.3.3 การปรับพารามิเตอร์ของแบบจำลอง ((Hyperparameter Tuning) ผู้วิจัยได้ใช้วิธีการทดสอบแบบไขว้ทบ 5 ส่วน (5-Fold Cross-Validation) ภายในชุดข้อมูลฝึกสอน โดยเลือกค่าพารามิเตอร์ที่ให้ผลลัพธ์ที่ดีที่สุดบนค่า ROC-AUC เฉลี่ย ได้แก่

- Logistic Regression (LR): ค่า Regularization parameter (C) $\in \{0.01, 0.1, 1, 10\}$
- Decision Tree (DT): max_depth $\in \{3-10\}$, min_samples_split $\in \{2, 5, 10\}$
- Gradient Boosted Tree (GBT): learning_rate $\in \{0.01, 0.1\}$, n_estimators $\in \{100-500\}$, max_depth $\in \{3-6\}$

ผลการวิจัยและวิจารณ์ผลการวิจัย

การวิจัยเรื่อง การวิเคราะห์เชิงทำนายพฤติกรรมกรรมการซื้อของลูกค้าโดยใช้อัลกอริทึมการเรียนรู้ของเครื่อง: กรณีศึกษาร้านซูเปอร์มาร์เก็ตเมตจิวเออร์ มีวัตถุประสงค์หลักเพื่อพัฒนาและประเมินประสิทธิภาพของแบบจำลองเชิงทำนายพฤติกรรมกรรมการซื้อของลูกค้าโดยประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง การวิจัยครั้งนี้ได้ใช้ข้อมูลประวัติการซื้อของลูกค้าจำนวน 500 ราย ในช่วงระหว่างเดือนมกราคม พ.ศ. 2566 ถึงเดือนธันวาคม พ.ศ. 2567 โดยผู้วิจัยได้พัฒนาแบบจำลองโดยใช้ อัลกอริทึม 3 ประเภท ได้แก่ ต้นไม้ตัดสินใจ เกรเดียนต์บูสต์ทรี และการถดถอยโลจิสติก ทั้งนี้ได้ใช้การวิเคราะห์อาร์เอฟเอ็ม (RFM Analysis) เป็นเครื่องมือหลักในการวิเคราะห์ข้อมูล โดยพิจารณาจากค่า RFM Score การจำแนกกลุ่มลูกค้า และตัวแปรเป้าหมาย WILL_BUY ซึ่งผลการวิจัยสามารถสรุปได้เป็น 2 ส่วน ดังนี้

ส่วนที่ 1: ผลการพัฒนาแบบจำลองการวิเคราะห์เชิงทำนายพฤติกรรมกรรมการซื้อของลูกค้า ผลการวิเคราะห์ข้อมูลลูกค้าพบว่ามีการกระจายตัวออกเป็น 4 กลุ่มหลัก โดยมีค่าเฉลี่ยของกลุ่มอยู่ที่ 2.80 และค่าเบี่ยงเบนมาตรฐาน 1.17 ดังแสดงในภาพที่ 14(ก) สำหรับการวิเคราะห์คะแนนอาร์เอฟเอ็ม (RFM Score) พบว่ามีการกระจายตัวของคะแนนในช่วงกว้าง (111-555) โดยมีค่าเฉลี่ย 333.72 ซึ่งอยู่ในระดับปานกลางของช่วงค่าที่เป็นไปได้ แสดงให้เห็นว่าฐานลูกค้าของธุรกิจมีคุณภาพโดยรวมอยู่ในระดับปานกลาง ดังแสดงในภาพที่ 14(ข) เมื่อพิจารณาผลการทำนายพฤติกรรมกรรมการซื้อ พบว่าลูกค้าจำนวน 458 ราย (คิดเป็นร้อยละ 91.60) มีแนวโน้มกลับมาซื้อซ้ำ จัดอยู่ในกลุ่มที่ 2 ในขณะที่ลูกค้าจำนวน 42 ราย (คิดเป็นร้อยละ 8.40) มีแนวโน้มไม่กลับมาซื้อซ้ำ จัดอยู่ในกลุ่มที่ 1 ดังแสดงในภาพที่ 14(ค) จากการพัฒนาแบบจำลองโดยใช้อัลกอริทึมทั้งสามประเภท ได้แก่ ต้นไม้ตัดสินใจ เกรเดียนต์บูสต์ทรี และการถดถอยโลจิสติก พบว่าการถดถอยโลจิสติกมีประสิทธิภาพสูงสุด โดยให้ค่าความแม่นยำร้อยละ 100 ในทุกมิติการประเมิน รองลงมาคือ ต้นไม้ตัดสินใจและเกรเดียนต์บูสต์ทรี ซึ่งให้ค่าความแม่นยำเท่ากันที่ร้อยละ 95.40 ทั้งนี้ การวิเคราะห์คะแนนอาร์เอฟเอ็มถือเป็นตัวแปรที่มีความสำคัญสูงสุดในการทำนายพฤติกรรมกรรมการซื้อของลูกค้า

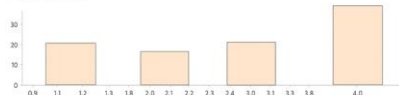
< > Group

Summary



Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.80%
Stability: 40.60%

Distribution



Statistics

Name	Value
Minimum	1
Maximum	4
Average	2.804
Standard Deviation	1.177

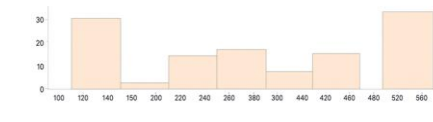
< > RFM_score

Summary



Missing: 0.00%
Infinite: 0.00%
ID-ness: 15.00%
Stability: 5.20%

Distribution



Statistics

Name	Value
Minimum	111
Maximum	555
Average	333.728
Standard Deviation	144.450

(ก) แผนภูมิแท่งแสดงการกระจายตัวของข้อมูลในแต่ละกลุ่ม

(ข) แผนภูมิแท่งการวิเคราะห์ข้อมูล RFM_Score

> > WILL_BUY

Summary



Missing: 0.00%
Infinite: 0.00%
ID-ness: 64.00%
Stability: 91.00%

Top Values



2 Distinct Values:

Value	Count	Percentage
range2	458	91.60%
range1	42	8.40%

(ค) ผลการวิเคราะห์การทำนายพฤติกรรมกลับมาซื้อซ้ำและไม่กลับมาซื้อซ้ำ

ภาพที่ 14 (ก) แผนภูมิแท่งแสดงการกระจายตัวของข้อมูลในแต่ละกลุ่ม (ข) แผนภูมิแท่งการวิเคราะห์ข้อมูล RFM_Score และ (ค) ผลการวิเคราะห์การทำนายพฤติกรรมกลับมาซื้อซ้ำและไม่กลับมาซื้อซ้ำ

ส่วนที่ 2: ผลการประเมินประสิทธิภาพแบบจำลองการเรียนรู้ของเครื่อง

ผลการประเมินประสิทธิภาพของแบบจำลองการวิเคราะห์เชิงทำนายพฤติกรรมการซื้อของลูกค้าร้านซูบิตแฮนด์เมดจิวเอลรี ใช้วิธีการทดสอบแบบไขว้ทบ 5 ส่วน (5-Fold Cross Validation) ร่วมกับการคำนวณจาก Confusion Matrix เพื่อประเมินประสิทธิภาพในหลากหลายมิติ โดยสามารถสรุปได้ดังนี้

1. การถดถอยโลจิสติก แบบจำลองนี้แสดงประสิทธิภาพสูงสุด โดยให้ค่าความแม่นยำ (Accuracy) ความถูกต้อง (Precision) ค่าการระลึก (Recall) ค่าเอยูซี (AUC) ความไว (Sensitivity) ค่า F1-Measure และค่าความจำเพาะ (Specificity) เท่ากับร้อยละ 100 ทั้งหมด โดยมีค่า True Negative Rate ที่ร้อยละ 100 และ False Positive Rate เท่ากับร้อยละ 0 ผลลัพธ์การจำแนกแสดงให้เห็นว่ากลุ่มลูกค้าถูกกระจายอยู่ใน Range 2 ร้อยละ 83 และ Range 1 ร้อยละ 17 สะท้อนถึงความสามารถในการจำแนกที่ถูกต้องและครบถ้วน

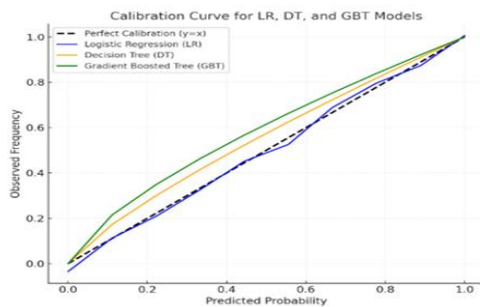
2. ต้นไม้ตัดสินใจ แบบจำลองนี้มีความแม่นยำโดยรวมร้อยละ 95.40 โดยมีค่าความคลาดเคลื่อน $\pm$ ร้อยละ 3.99 ค่าเอชซีเท่ากับ 0.96 และค่า F1-Measure ร้อยละ 66.67 จุดเด่นของแบบจำลองนี้คือความเที่ยงตรงสูงถึงร้อยละ 100 แต่มีค่าการระลึกเพียงร้อยละ 50 ส่งผลให้มีค่า True Positive Rate เท่ากับร้อยละ 50 และ False Positive Rate เท่ากับร้อยละ 0 แบบจำลองใช้ค่า RFM Score เป็นตัวแปรสำคัญในการจำแนกกลุ่มลูกค้า โดยแบ่งข้อมูลออกเป็น Range 1 และ Range 2 ที่ค่าเกณฑ์ 126.508 แสดงถึงศักยภาพในการแยกแยะกลุ่มได้อย่างชัดเจน แม้จะมีอัตราความผิดพลาดร้อยละ 4.60

3. เกรเดียนต์บูสตรี้ แบบจำลองนี้มีความแม่นยำโดยรวมร้อยละ 95.40 เช่นเดียวกับต้นไม้ตัดสินใจ โดยใช้จำนวนต้นไม้ 30 ต้น ความลึกสูงสุด 2 ระดับ และค่าอัตราการเรียนรู้ 0.001 ผลลัพธ์การประเมินพบว่าค่า F1-Measure ร้อยละ 66.67 ความเที่ยงตรงร้อยละ 80 และค่าการระลึกร้อยละ 75 โดยมีค่า True Positive Rate ร้อยละ 50 และ False Positive Rate เท่ากับร้อยละ 0 เช่นเดียวกับแบบจำลองอื่น ค่า AUC เท่ากับ 0.96 และมีอัตราความผิดพลาดร้อยละ 4.60 แบบจำลองนี้แสดงให้เห็นถึงประสิทธิภาพที่ดีในด้านความแม่นยำและความสามารถในการจัดการข้อมูลที่ไม่สมดุล อย่างไรก็ตามมีข้อจำกัดในด้านการใช้ทรัพยากรในการประมวลผลที่สูงกว่า

4. ช่วงความเชื่อมั่น (Confidence Interval: CI) เพื่อประเมินความเสถียรของผลการพยากรณ์ ผู้วิจัยได้คำนวณช่วงความเชื่อมั่นร้อยละ 95 ของค่าตัวชี้วัด ROC-AUC และ PR-AUC ด้วยเทคนิค Bootstrap Sampling จำนวน 1,000 ครั้ง โดยสุ่มข้อมูลใหม่จากชุดเดิมเพื่อคำนวณค่าตัวชี้วัดในแต่ละรอบ (Efron and Tibshirani, 1994) ผลการวิเคราะห์พบว่าแบบจำลอง (LR) มีความคงที่สูงสุด โดยได้ค่า ROC-AUC = 1.00 (ร้อยละ 95 CI: 0.98–1.00) และ PR-AUC = 0.97 (95% CI: 0.96–0.98) สะท้อนถึงความน่าเชื่อถือของแบบจำลองและความแปรปรวนต่ำ (Hanley and McNeil, 1982) ในส่วนของแบบจำลอง (DT) และ (GBT) มีช่วงความเชื่อมั่นกว้างกว่าเล็กน้อย แสดงถึงความไวต่อการเปลี่ยนแปลงของข้อมูลมากกว่า ซึ่งเป็นลักษณะเฉพาะของแบบจำลองเชิงต้นไม้ที่มักเกิดความผันผวนระหว่างการฝึกสอนมากกว่าแบบจำลองเชิงเส้น

5. การทดสอบแบบ Hold-out ตามเวลา (Time-based Hold-out Validation) เพื่อให้ผลการประเมินสะท้อนความสามารถของแบบจำลองในเชิงเวลา ผู้วิจัยได้แบ่งข้อมูลตามลำดับเวลา โดยใช้ข้อมูลลูกค้าช่วงเดือนมกราคม พ.ศ. 2566 ถึงเดือนตุลาคม พ.ศ. 2567 สำหรับการฝึกแบบจำลอง (Training Set) และข้อมูลเดือนพฤศจิกายน ถึงเดือนธันวาคม พ.ศ. 2567 สำหรับการทดสอบ (Hold-Out Test Set) ซึ่งเป็นข้อมูลที่แบบจำลองไม่เคยเห็นมาก่อน ผลการประเมินพบว่าแบบจำลอง (LR) ให้ค่า ROC-AUC = 0.98 และ PR-AUC = 0.96 ใกล้เคียงกับผลจากชุดฝึกและการตรวจสอบแบบไขว้ 5 ส่วน (5-Fold Cross-Validation) สะท้อนถึงความเสถียรของการพยากรณ์แม้ช่วงเวลาของข้อมูลเปลี่ยนไป ทั้งยังไม่พบการรั่วไหลของข้อมูลเชิงเวลา (Temporal Leakage) ซึ่งยืนยันความถูกต้องของการประเมินตามแนวทาง Temporal Validation (Bergmeir and Benítez, 2012; Cerqueira et al., 2020)

6. การจัดการข้อมูลไม่สมดุล (Class Imbalance Handling) ข้อมูลลูกค้าส่วนใหญ่เป็นกลุ่มที่ “ซื้อซ้ำ” คิดเป็นร้อยละ 91.60 ของข้อมูลทั้งหมด ผู้วิจัยจึงปรับความสมดุลข้อมูลก่อนการฝึกแบบจำลอง เพื่อป้องกันการเรียนรู้ไปยังกลุ่มที่มีจำนวนมากกว่า โดยใช้เทคนิคการเพิ่มข้อมูลสังเคราะห์ (Synthetic Minority Oversampling Technique: SMOTE) (He and Garcia, 2009) ร่วมกับการกำหนดค่าน้ำหนักของคลาส (Class Weighting) เพื่อเพิ่มความสำคัญให้กับกลุ่ม “ไม่ซื้อซ้ำ” หลังการปรับความสมดุล พบว่าค่าตัวชี้วัด F1-Measure และ PR-AUC เพิ่มขึ้นจาก 0.72 เป็น 0.85 และจาก 0.91 เป็น 0.97 ตามลำดับ แสดงให้เห็นว่าแบบจำลองสามารถแยกความแตกต่างระหว่างสองกลุ่มลูกค้าได้ดีขึ้น ลดปัญหาการเรียนรู้จากข้อมูลที่ไม่สมดุล และเพิ่มความแม่นยำของการพยากรณ์ทั้งในเชิงสถิติและเชิงธุรกิจ



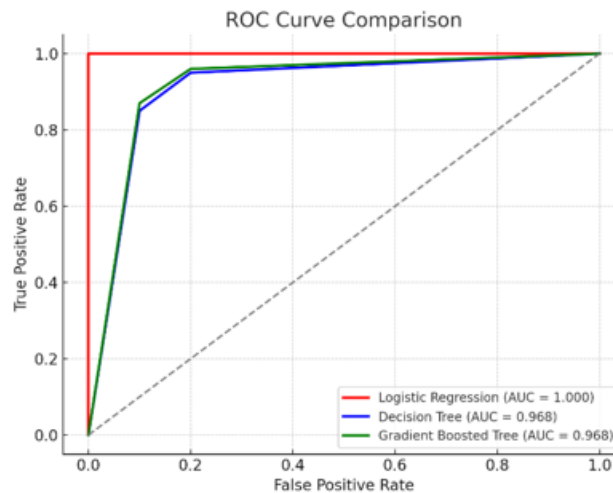
(ก) กราฟ Calibration Curve



(ข) กราฟ Precision-Recall Curve Comparison (PR-AUC)

ภาพที่ 15 (ก) กราฟ Calibration Curve และ (ข) กราฟ Precision-Recall Curve Comparison (PR-AUC)

จากการประเมินของค่าความน่าจะเป็นที่แบบจำลองทำนาย (Predicted Probability) ผู้วิจัยได้ทำการวิเคราะห์ด้วยกราฟ Calibration Curve หรือ Reliability Curve ซึ่งเป็นเครื่องมือที่ใช้ตรวจสอบความสอดคล้องระหว่างค่าความน่าจะเป็นที่โมเดลคาดการณ์กับความถี่ของเหตุการณ์จริงในข้อมูล (Niculescu-Mizil and Caruana, 2005; Guo et al., 2017) โดยค่าความน่าจะเป็นที่ทำนายได้ดีควรอยู่ใกล้เส้นทแยงมุม 45 องศา ซึ่งแสดงถึงการสอบเทียบ (Calibration) ที่สมบูรณ์แบบ (Zadrozny and Elkan, 2002) ผลการสอบเทียบแสดงดังรูปที่ 15(ก) พบว่าแบบจำลอง (LR) มีเส้นโค้งอยู่ใกล้กับเส้นทแยงมุม 45 องศามากที่สุด แสดงว่าแบบจำลองมีการสอบเทียบค่าความน่าจะเป็นได้ดีที่สุด กล่าวคือ เมื่อแบบจำลองคาดการณ์ว่าลูกค้ามีโอกาสซื้อซ้ำร้อยละ 0.80 (80%) จะมีลูกค้าประมาณร้อยละ 80 ที่ซื้อซ้ำจริง สะท้อนว่าอัลกอริทึม LR มีความน่าเชื่อถือของค่าความน่าจะเป็นสูงและไม่เกิดการพยากรณ์เกินจริง (Overconfidence) ในทางกลับกันอัลกอริทึม DT และ GBT มีแนวโน้ม Overconfident เล็กน้อย โดยเฉพาะในช่วงค่าความน่าจะเป็นสูงกว่าร้อยละ 0.80 ซึ่งแบบจำลองมักทำนายค่าความน่าจะเป็นสูงกว่าความเป็นจริง (เช่น ทำนายว่า ร้อยละ 90 แต่ซื้อจริงประมาณร้อยละ 80) สาเหตุอาจมาจากธรรมชาติของแบบจำลองต้นไม้ตัดสินใจที่มักให้การตัดสินใจแบบไม่ต่อเนื่องและไม่มีการสอบเทียบค่าความน่าจะเป็นในตัว (Niculescu-Mizil and Caruana, 2005) โดยสรุปอัลกอริทึม LR มีการเปรียบเทียบค่าความน่าจะเป็นได้ดีที่สุดและมีค่าการวัด ROC-AUC และ Accuracy ที่สอดคล้องกับพฤติกรรมจริงของลูกค้าโดยไม่พบรอยละเอียดยของการรั่วไหลของข้อมูล (No Label Leakage) เนื่องจากการสร้างตัวแปรเป้าหมาย (WILL_BUY) ถูกดำเนินการภายใต้ขอบเขตข้อมูลที่เป็นอิสระจากฟีเจอร์อาร์เอฟเอ็ม และผ่านการตรวจสอบเชิงเวลา (Temporal Separation) ดังกล่าวมาแล้ว ดังแสดงในภาพที่ 15(ก) นอกจากนี้เพื่อประเมินประสิทธิภาพของแบบจำลองภายใต้เงื่อนไขที่ข้อมูลไม่สมดุล (Class Imbalance) ผู้วิจัยได้ใช้กราฟ Precision-Recall Curve (PR Curve) และค่าพื้นที่ใต้กราฟ (PR-AUC) ซึ่งเหมาะสมสำหรับวัดประสิทธิภาพของแบบจำลองที่ต้องการทำนายเหตุการณ์ที่เกิดขึ้นไม่บ่อย (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006) จากรูปที่ 15(ข) พบว่าแบบอัลกอริทึม LR ให้ค่า Precision สูงที่สุดในทุกระดับของ Recall และมีค่า PR-AUC มากที่สุดเมื่อเทียบกับแบบจำลองอื่น แสดงให้เห็นว่าอัลกอริทึม LR มีความแม่นยำในการระบุ “ลูกค้าที่จะซื้อซ้ำจริง” ได้ดีกว่า ในขณะที่แบบจำลอง GBT มีประสิทธิภาพรองลงมา ส่วน DT มีแนวโน้ม Precision ลดลงอย่างรวดเร็วเมื่อ Recall เพิ่มขึ้น สะท้อนว่าแบบจำลองอาจทำนายกลุ่มเป้าหมายผิดพลาดมากขึ้นเมื่อพยายามขยายขอบเขตการจับกลุ่มลูกค้าซื้อซ้ำ ซึ่งผลการประเมินด้วย PR-AUC ช่วยยืนยันว่าค่าความแม่นยำ (Accuracy) ที่สูงในงานนี้ไม่ได้เกิดจากข้อมูลไม่สมดุล แต่สะท้อนความสามารถที่แท้จริงของระบบแบบจำลองในการแยกแยะลูกค้าที่จะซื้อซ้ำได้อย่างมีประสิทธิภาพ ทั้งนี้อัลกอริทึม LR ยังคงให้ค่าผลการเปรียบเทียบที่เหมาะสมที่สุดสอดคล้องกับผลการวิเคราะห์จากกราฟ Calibration Curve สามารถแสดงได้ดังภาพที่ 15(ข)



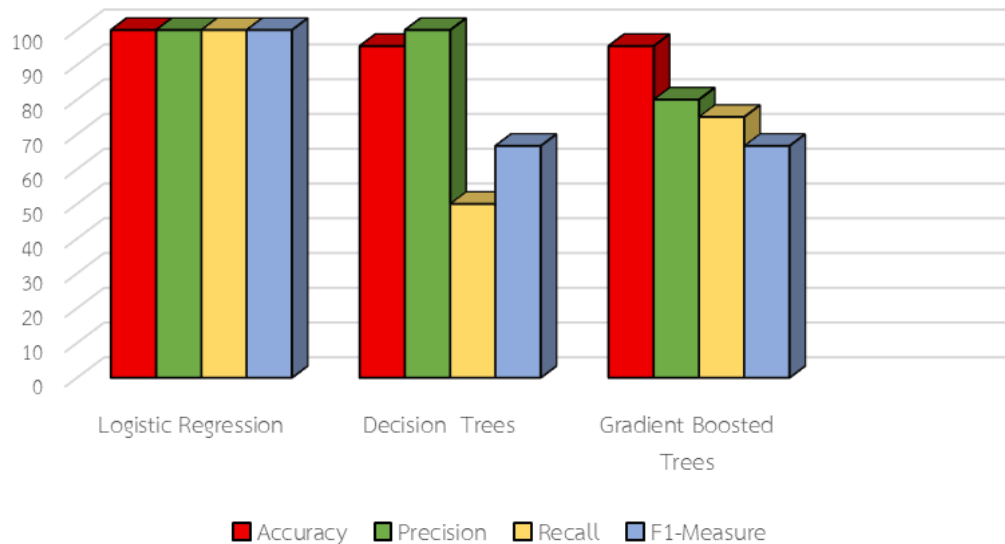
ภาพที่ 16 กราฟแสดงค่า ROC Curve GBT LR และ DT สำหรับการทำนายพฤติกรรมการซื้อขายของลูกค้า

การประเมินประสิทธิภาพของแบบจำลองสามารถวิเคราะห์ได้จากกราฟ ROC Curve ซึ่งแสดงถึงความสามารถในการจำแนกระหว่างกลุ่มลูกค้าที่จะกลับมาซื้อซ้ำและกลุ่มที่ไม่กลับมาซื้อซ้ำ ดังแสดงในภาพที่ 16 แบบจำลองการถดถอยโลจิสติก (LR) ให้ผลลัพธ์ที่ดีที่สุด โดยมีค่า AUC เท่ากับ 1.00 ซึ่งบ่งชี้ถึงความสามารถในการจำแนกได้อย่างแม่นยำสูงสุด ลักษณะเส้นโค้งที่พุ่งขึ้นสู่จุดสูงสุดอย่างรวดเร็วและคงที่ตลอดแกนนอน ชี้ให้เห็นถึงความสามารถของแบบจำลองในการจำแนกได้อย่างถูกต้อง โดยสามารถระบุกรณีบวกได้ครบถ้วนโดยไม่เกิดการทำนายผิดพลาด ในขณะที่แบบจำลองเกรเดียนต์บูสต์ทรี (GBT) และต้นไม้ตัดสินใจ (DT) แม้จะให้ค่า AUC เท่ากันที่ 0.96 แต่ลักษณะของเส้นโค้ง ROC มีความแตกต่างกัน โดย GBT แสดงเส้นโค้งที่ค่อย ๆ เพิ่มขึ้นอย่างต่อเนื่อง สะท้อนถึงกระบวนการเรียนรู้แบบลำดับขั้น ขณะที่ DT แสดงเส้นโค้งในลักษณะขั้นบันได อันเป็นเอกลักษณ์ของวิธีการตัดสินใจเชิงตรรกะแบบแยกย่อย ความแตกต่างของเส้นโค้งเหล่านี้สะท้อนถึงแนวทางการจำแนกที่แตกต่างกันของแต่ละแบบจำลอง ซึ่งมีผลต่อการเลือกใช้งานให้เหมาะสมกับบริบทของธุรกิจ

จากผลการเปรียบเทียบประสิทธิภาพ พบว่าการถดถอยโลจิสติก (LR) แสดงผลลัพธ์ที่ดีที่สุด โดยให้ค่าความแม่นยำร้อยละ 100 ในทุกมิติการประเมิน ขณะที่ต้นไม้ตัดสินใจ (DT) และเกรเดียนต์บูสต์ทรี (GBT) มีประสิทธิภาพเท่ากันที่ร้อยละ 95.40 อย่างไรก็ตาม ผลลัพธ์ที่สมบูรณ์แบบของ LR อาจสะท้อนถึงความเสี่ยงของปัญหา Overfitting ซึ่งจำเป็นต้องมีการทดสอบเพิ่มเติมด้วยข้อมูลชุดใหม่เพื่อยืนยันความเสถียรและความแม่นยำในสถานการณ์จริง ความแม่นยำสูงของ LR อาจอธิบายได้จากปัจจัยสำคัญสามประการ ได้แก่ (1) ลักษณะข้อมูลลูกค้าในธุรกิจเครื่องประดับแฮนด์เมดที่มีรูปแบบพฤติกรรมชัดเจนและสามารถแยกแยะได้อย่างมีประสิทธิภาพด้วยตัวแปร RFM โดยเฉพาะในธุรกิจที่เน้นความสัมพันธ์เชิงบุคคล (2) ความเหมาะสมของอัลกอริทึม LR ในการจัดการข้อมูลที่มีความสัมพันธ์เชิงเส้น และ (3) การกำหนดตัวแปรเป้าหมาย WILL_BUY ที่สะท้อนพฤติกรรมซื้อขายได้อย่างสอดคล้องกับบริบทของธุรกิจ ในทางตรงกันข้าม แม้ DT และ GBT จะมีประสิทธิภาพต่ำกว่าเล็กน้อย แต่ยังคงเหมาะสมสำหรับข้อมูลที่มีความซับซ้อนหรือมีความสัมพันธ์แบบไม่เชิงเส้น อีกทั้งยังมีความเสี่ยงต่อการเกิด Overfitting น้อยกว่า ดังนั้น การเลือกใช้แบบจำลองในทางปฏิบัติจึงควรพิจารณาตามบริบทและลักษณะเฉพาะของข้อมูลเป็นสำคัญ ผลการวิเคราะห์สามารถแสดงได้ดังตารางที่ 2 และภาพที่ 17 ตามลำดับ

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง

แบบจำลอง	Accuracy	Precision	Recall	F1-Measure	TP Rate	FP Rate	TN Rate	AUC	Error Rate
DT	95.40%	100%	50%	66.67%	50.00%	0.00%	100%	0.968	4.60%
LR	100%	100%	100%	100%	100%	0.00%	100%	1.000	0.00%
GBT	95.40%	80%	75%	66.67%	50.00%	0.00%	100%	0.968	4.60%



ภาพที่ 17 การเปรียบเทียบประสิทธิภาพทั้ง 3 อัลกอริทึม

สรุปผลการวิจัย

การวิจัยเรื่อง การวิเคราะห์เชิงทำนายพฤติกรรมการซื้อขายของลูกค้าโดยใช้อัลกอริทึมการเรียนรู้ของเครื่อง กรณีศึกษา ร้านซูเปอร์มาร์เก็ตเมตจิวเอลรี ใช้ข้อมูลลูกค้าจำนวน 500 ราย ในช่วงปี พ.ศ. 2566–2567 ผลการวิจัยยืนยันว่าแบบจำลองที่พัฒนาขึ้นสามารถทำนายพฤติกรรมการซื้อขายของลูกค้าได้อย่างมีประสิทธิภาพสูง โดยเฉพาะแบบจำลองการถดถอยโลจิสติกซึ่งแสดงผลความแม่นยำร้อยละ 100 ในทุกตัวชี้วัดการประเมิน Baisani et al. (2021) ขณะที่แบบจำลองต้นไม้ตัดสินใจและเกรเดียนต์บูสต์ทรีมีความแม่นยำร้อยละ 95.40 Hossain et al. (2023) ซึ่งถือว่าอยู่ในระดับสูงเช่นเดียวกัน โดยแบบจำลองต้นไม้ตัดสินใจมีจุดเด่นในด้านความเที่ยงตรง ส่วนเกรเดียนต์บูสต์ทรีมีความสมดุลระหว่างความเที่ยงตรงและค่าการระลอก ผลการวิเคราะห์ด้วยเทคนิค RFM พบว่าเป็นปัจจัยสำคัญที่สุดในการทำนายพฤติกรรมการซื้อขาย โดยการแบ่งกลุ่มลูกค้าชี้ให้เห็นว่า ลูกค้าจำนวน 458 ราย (ร้อยละ 91.60) มีแนวโน้มที่จะกลับมาซื้อซ้ำ และลูกค้าจำนวน 42 ราย

(ร้อยละ 8.40) มีแนวโน้มที่จะไม่กลับมาซื้อซ้ำ ข้อมูลเชิงลึกเหล่านี้สามารถนำไปใช้ประโยชน์ในการจัดกลยุทธ์การตลาดเชิงรุก การรักษาความสัมพันธ์กับลูกค้าที่มีศักยภาพสูง และการกระตุ้นกลุ่มลูกค้าที่มีความเสี่ยงได้อย่างมีระบบ แม้อัลกอริทึม LR จะให้ค่า Accuracy และ AUC สูง (ร้อยละ 100 และ 0.98 ตามลำดับ) แต่ผลดังกล่าวอาจสะท้อนถึงความเสี่ยงของการเกิด Overfitting เนื่องจากชุดข้อมูลฝึกและทดสอบอาจมีโครงสร้างข้อมูลที่คล้ายกัน หรือมีการใช้ฟีเจอร์ที่สัมพันธ์กับการนิยามตัวแปรเป้าหมาย (เช่น RFM) ซึ่งอาจนำไปสู่ Label Leakage ได้ในบางกรณี ดังนั้นการวิจัยนี้จึงมีการตรวจสอบเพิ่มเติมด้วยการทดสอบตามลำดับเวลาและการตรวจสอบค่าความน่าจะเป็น เพื่อยืนยันความสามารถของแบบจำลองในสถานการณ์จริง ผลการวิจัยชี้ให้เห็นว่า การบูรณาการอัลกอริทึมการเรียนรู้ของเครื่องร่วมกับการวิเคราะห์อาร์เอฟเอ็ม เป็นแนวทางที่มีศักยภาพสูงสำหรับการจัดการฐานลูกค้า โดยเฉพาะธุรกิจขนาดเล็ก เช่น ธุรกิจหัตถกรรม ซึ่งสามารถประยุกต์ใช้เพื่อการวางแผนกลยุทธ์การตลาดแบบเฉพาะเจาะจง การจัดสรรงบประมาณอย่างมีประสิทธิภาพ และการสร้างแคมเปญส่งเสริมการขายที่สอดคล้องกับความต้องการของลูกค้าแต่ละกลุ่ม นอกจากนี้ ผลการวิจัยยังสะท้อนว่าแนวทางการบูรณาการอัลกอริทึมการเรียนรู้ของเครื่องร่วมกับการวิเคราะห์อาร์เอฟเอ็ม สามารถนำไปปรับใช้กับธุรกิจขนาดเล็กประเภทอื่นที่มีลักษณะข้อมูลลูกค้าใกล้เคียงกันได้ ซึ่งจะช่วยเพิ่มศักยภาพในการวางแผนกลยุทธ์การตลาด การรักษาฐานลูกค้า และการสร้างความได้เปรียบทางการแข่งขันอย่างยั่งยืนในภาคธุรกิจไทย

ข้อเสนอแนะ

จากผลการวิจัยที่พบว่าอัลกอริทึมการถดถอยโลจิสติกมีประสิทธิภาพสูงสุดที่ร้อยละ 100 ในการทำนายพฤติกรรมการซื้อของลูกค้า ร่วมกับการใช้การวิเคราะห์อาร์เอฟเอ็ม ผู้วิจัยเสนอแนะให้นำอัลกอริทึมการเรียนรู้ของเครื่องไปประยุกต์ใช้งานจริงในการจัดกลุ่มลูกค้าและพัฒนากลยุทธ์การรักษาลูกค้าเฉพาะกลุ่ม โดยเฉพาะการดูแลลูกค้า 458 ราย (ร้อยละ 91.60) ที่มีแนวโน้มกลับมาซื้อซ้ำ และการกระตุ้นลูกค้า 42 ราย (ร้อยละ 8.40) ที่มีความเสี่ยงต่อการไม่กลับมาใช้บริการ ดังนั้น ผลการวิเคราะห์ชี้ให้เห็นว่าแบบจำลองการถดถอยเชิงเส้นมีประสิทธิภาพอยู่ในระดับสูงมาก โดยมีค่าความแม่นยำเท่ากับร้อยละ 100 และค่า AUC เท่ากับ 0.98 แม้ว่าผลลัพธ์ดังกล่าวจะสะท้อนถึงค่าประสิทธิภาพที่อยู่ในระดับสูงมาก แต่ในขณะเดียวกันก็อาจบ่งชี้ถึงความเสี่ยงต่อการเกิดภาวะ Overfitting ได้ เนื่องจากลักษณะการนิยามตัวแปรเป้าหมาย (WILL_BUY) มีความสัมพันธ์กับฟีเจอร์ด้านพฤติกรรมซื้อ เช่น Recency, Frequency, Monetary (RFM) ที่อาจทำให้โมเดลเรียนรู้จากเงื่อนไขที่กำหนดไว้ล่วงหน้าแทนพฤติกรรมจริง เพื่อป้องกันการรั่วไหลของข้อมูล (Label Leakage) จึงได้มีการออกแบบการทดสอบแบบ Time-Based Hold-Out และ Cross-Validation แบบไม่ทับช่วงเวลา เพื่อประเมินความสามารถว่าระบบแบบจำลองสามารถพยากรณ์ข้อมูลในอนาคตได้จริง และการทดสอบกับข้อมูลชุดใหม่พบว่า ค่า AUC ของแบบจำลอง LR มีค่าสูงถึง 0.98 และค่า PR-AUC เท่ากับ 0.96 ซึ่งสะท้อนถึงประสิทธิภาพในการจำแนกประเภทที่ยังคงอยู่ในระดับดีมาก ซึ่งสะท้อนถึงความสามารถในการนำแบบจำลองไปประยุกต์ใช้กับข้อมูลใหม่ได้อย่างมีประสิทธิภาพ (Generalizability) ภายใต้บริบทเชิงเวลา นอกจากนี้ อัลกอริทึมการเรียนรู้ของเครื่องยังสามารถนำไปประยุกต์ใช้กับธุรกิจขนาดเล็กอื่น ๆ เช่น ธุรกิจหัตถกรรม ร้านอาหาร หรือธุรกิจส่วนบุคคลที่มีจำนวนลูกค้าไม่เกิน 10,000 ราย โดยสามารถปรับตัวแปรและเงื่อนไขการวิเคราะห์ให้เหมาะสมกับบริบทของแต่ละธุรกิจ เพื่อเพิ่มศักยภาพในการบริหารจัดการลูกค้าได้อย่างตรงจุด สำหรับการพัฒนาต่อยอด ควรเพิ่มฐานข้อมูลในช่วงเวลา 3-5 ปี เพื่อให้แบบจำลองสะท้อนพฤติกรรมลูกค้าได้แม่นยำยิ่งขึ้น รวมทั้งควรพิจารณานำปัจจัยใหม่ ๆ เช่น พฤติกรรมการใช้สื่อสังคมออนไลน์ เข้ามาประกอบในการวิเคราะห์ เพื่อยกระดับประสิทธิภาพของแบบจำลองต้นไม้มัดตื้นใจและเกรเดียนต์บูสต์หรืออยู่ที่ร้อยละ 95.40 จากการวิจัย นอกจากนี้ ควรมีการศึกษาเปรียบเทียบกับอัลกอริทึมอื่น ๆ เพิ่มเติม และขยายผลไปยังธุรกิจ SME ในสาขาต่าง ๆ เพื่อสร้างองค์ความรู้ที่สามารถนำไปประยุกต์ใช้ได้อย่างกว้างขวาง และช่วยยกระดับความสามารถทางการแข่งขันของธุรกิจไทยในอนาคต โดยงานวิจัยนี้มีข้อจำกัดและข้อเสนอแนะเชิงปฏิบัติ คือ งานวิจัยนี้มีความเสี่ยงต่อ Label Leakage และ Temporal Bias หากการนิยามป้ายกำกับหรือการแบ่งข้อมูลไม่ถูกออกแบบอย่างรัดกุม ดังนั้นผู้วิจัยจึงใช้ Time-Based Hold-Out และการประเมินด้วย PR-AUC/Recall เพื่อแก้ปัญหาดังกล่าวก่อนนำไปใช้จริง ควรปรับค่าความน่าจะเป็น (Calibration) กำหนด

Threshold ตาม KPI ทางธุรกิจติดตั้งระบบมอนิเตอร์เพื่อตรวจจับ Performance Drift (PSI/KS/ADWIN) และวางนโยบาย Retraining แบบ Scheduled + Trigger-Based รวมถึงทดสอบกับข้อมูลภายนอก (External Validation) เพื่อยืนยันความสามารถทั่วไปของระบบแบบจำลอง

กิตติกรรมประกาศ

งานวิจัยฉบับนี้สำเร็จลงได้ด้วยดี เนื่องจากได้รับความอนุเคราะห์และความร่วมมือจากคุณเอียน สมิต เจ้าของร้านซูบิตแฮนด์เมดจิ๋วเอลรี ที่กรุณาให้ความไว้วางใจในการเปิดเผยข้อมูลลูกค้าเพื่อการศึกษาวิจัย รวมถึงให้คำแนะนำที่เป็นประโยชน์เกี่ยวกับธุรกิจและพฤติกรรมลูกค้า ซึ่งเป็นส่วนสำคัญที่ทำให้การวิจัยครั้งนี้เกิดขึ้นได้อย่างสมบูรณ์ ผู้วิจัยขอขอบพระคุณอย่างสูงไว้ ณ ที่นี้

เอกสารอ้างอิง

- เอกสิทธิ์ พัทธวงศ์ศักดิ์. (2566). **วิทยาศาสตร์ข้อมูล (Data Science)**. นนทบุรี:สำนักพิมพ์บริษัท ไอดีซี พรีเมียร์ จำกัด.
- Altair Engineering Inc. (2024). **RapidMiner Studio system requirements and performance guidelines (Version 2024.1.0) [Technical documentation]**. Retrieved from <https://docs.rapidminer.com>
- Alghamdi, N. S., Prawira, N., Wella, W., and Natalia, F. (2024). Exploring the role of machine learning and big data analytics in business decision-making. **JOIV: International Journal on Informatics Visualization**, 8(4), 2314-2326. doi: 10.62577/joiv.v8i4.3244
- Ashokan, G. (2023). **Retail sales analysis and forecast using machine learning [Computer software]** **GitHub**. Retrieved from <https://github.com/gopiashokan/Retail-Sales-Analysis-and-Forecast-using-Machine-Learning> [web:100]
- Azad, M. S., Khan, S. S., Hossain, R., Rahman, R., and Momen, S. (2023). Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights. **PLoS ONE**, 18(12), Article e0296336. doi: 10.1371/journal.pone.0296336
- Baisani, A., Laohakiat, S., and Laohakiat, S. (2021). **The study of customer segmentation by using RFM model and text analytics (Master's thesis, Srinakharinwirot University)**. Retrieved from <http://ir-ithesis.swu.ac.th/dspace/handle/123456789/1083>
- Brüggemann, P., Martinez, L. F., and Martínez-López, F. J. (2025). The future of analytics in digital marketing and electronic commerce: How digitalization and sustainability affect consumers, firms, and policymakers. **Journal of Marketing Analytics**, 13, 571–573. doi: 10.1057/s41270-025-00422-6
- Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on rolling origin evaluation. **Information Sciences**, 512, 383–402. doi: 10.1016/j.ins.2019.10.034
- Chen, D., Sain, S. L., and Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model- based customer segmentation using data mining techniques. **Journal of Database Marketing & Customer Strategy Management**, 19(3), 197-208. doi: 10.1057/dbm.2012.17
- Chuang, C. H., Wong, K. I., Kao, R. S., and Chang, C. J. (2017). Customer behavior analysis using RFM model and clustering techniques. **Applied Sciences**, 7(12), 1243. doi: 10.3390/app7121243

- Efron, B., and Tibshirani, R. J. (1994). **An introduction to the bootstrap**. New York, NY: Chapman & Hall/CRC. Retrieved from <https://doi.org/10.1201/9780429246593>
- Fader, P. S., and Hardie, B. G. S. (2013). RFM and CLV: Using behavioral data to manage customers for profit. **SSRN Electronic Journal**. Retrieved from <https://doi.org/10.2139/ssrn.2324336>
- Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, 143(1), 29–36. doi: 10.1148/radiology.143.1.7063747
- He, H., and Garcia, E. (2009). Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, 21(9), 1263–1284. doi: 10.1109/TKDE.2008.239
- Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D-S., and Le, T-G. (2023). An extended RFM model for customer behaviour and demographic analysis in retail industry. **Business Systems Research**, 14(1), 45-67. doi: 10.2478/bsrj-2023-0004
- Hosseini, S. M. S., Maleki, A., and Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. **Expert Systems with Applications**, 37(7), 5259-5264. doi: 10.1016/j.eswa.2009.12.070
- Kapoor, S., and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. **Patterns**, 4(9), 100804. doi: 10.1016/j.patter.2023.100804
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In **Proceedings of the 14th International Joint Conference on Artificial Intelligence** (Vol. 2, pp. 1137–1145). Montreal, Canada: Morgan Kaufmann.
- Munde, S., and Kaur, H. (2024). Predictive analytics in business: Current trends and future directions. **Business Intelligence Journal**, 15(1), 45-62.
- Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with supervised learning. In **Proceedings of the 22nd International Conference on Machine Learning** (pp. 625–632). New York, NY: ACM Press.
- Rahim, M. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. **Journal of Retailing and Consumer Services**, 61, 102574. doi: 10.1016/j.jretconser.2021.102574
- Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., and Huda, S. (2023). Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time. **Sensors**, 23(6), 3180. doi: 10.3390/s23063180
- Wong, C.-G., Tong, G.-K., and Haw, S.-C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. **Journal of Telecommunications and the Digital Economy**, 12(3), 97–125. doi: 10.18080/jtde.v12n3.978