การทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

ภูมพัชร พิพัฒศรี[1*], นุวีย์ วิวัฒนวัฒนา[2]

บทคัดย่อ

ในยุคปัจจุบัน ธุรกิจ e-Commerce ต่างแข่งขันกันเพื่อแย่งชิงลูกค้า เนื่องจากประชากรโลกส่วนใหญ่เลือกซื้อสินค้าและบริการผ่านช่องทางออนไลน์กันมากขึ้น ด้วยเหตุนี้ ธุรกิจจึงจำเป็นต้องหาวิธีรักษาลูกค้าไว้ให้ได้ ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาแนวโน้มการเลิกเป็นลูกค้าของเว็บไซต์แห่งหนึ่ง จากข้อมูลสาธารณะในเว็บไซต์ Kaggle.com โดยนำเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน ได้แก่ แบบจำลอง Logistic Regression, Support Vector Machines (SVM) และ Random Forest มาเปรียบเทียบและวัดประสิทธิภาพด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix ร่วมกับการคัดเลือกคุณลักษณะและการจัดการความไม่สมดุลของข้อมูลด้วยวิธี Synthetic Minority Oversampling Technique ผลการทดลองพบว่าแบบจำลอง Random Forest มีประสิทธิภาพดีที่สุดในการทำนาย โดยมีค่า Accuracy 92 เปอร์เซ็นต์, Precision 93 เปอร์เซ็นต์, Recall 92 เปอร์เซ็นต์ และ F1-Score 93 เปอร์เซ็นต์ นอกจากนี้ ผู้วิจัยยังใช้ Local Interpretable Model-Agnostic Explanations มาช่วยอธิบายการทำงานของแบบจำลองเพื่อเพิ่มความน่าเชื่อถือ

คำสำคัญ : การเรียนรู้ของเครื่อง, การทำนายแนวโน้มการเลิกใช้บริการ, LIME

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 065-6194669 E-mail address: phumphatchara.phiphatsri@g.swu.ac.th

CUSTOMER CHURN PREDICTION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES

Phumphatchara Phiphatsri[1*], Nuwee Wiwatwattana[2]

Abstract

In the modern era, e-commerce businesses fight fiercely for customers. As a result, the vast majority of worldwide consumers choose to buy online goods and services and firms must prioritize client retention techniques. In this study, the researchers looked at the pattern of customer attrition for a webpage created using publicly available data from the Kaggle.com website. The performance of three supervised machine learning techniques were logistic regression, support vector machines (SVM), and random forests. The aspect of performance was assessed using criteria such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. The Synthetic Minority Oversampling Technique was used and feature selection to solve unbalanced data and improve data quality. The Random Forest model had the highest predictive performance, with 92% accuracy, 93% precision, 92% recall, and a 93% F1-score. The researchers also used Local Interpretable Model-Agnostic Explanations to explain the model.

Keywords : Machine learning, Churn prediction, LIME

---

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 065-6194669 E-mail address: phumphatchara.phiphatsri@g.swu.ac.th

บทนำ

ในปัจจุบันมีการแข่งขันสำหรับธุรกิจ e-Commerce เป็นจำนวนมาก ประชากรโลกส่วนใหญ่เลือกที่จะบริโภคและอุปโภคในช่องทางออนไลน์กันมากยิ่งขึ้น ซึ่งผลจากการสำรวจนั้นพบว่าโดยเฉลี่ยแล้วร้อยละ 45 ของผู้บริโภคทั่วเอเชียวางแผนที่จะเพิ่มการใช้จ่ายออนไลน์แทนการใช้ช่องทางการค้าปลีกในแบบเดิม [14] จากการที่ผู้บริโภคหันมาพึ่งพาการซื้อของทางออนไลน์กันมากขึ้น ดังนั้นเพื่อให้ธุรกิจอยู่รอด จึงจำเป็นต้องค้นหาวิธีรักษาลูกค้าไว้ให้ได้

อัตราการเลิกซื้อสินค้าของลูกค้าในช่วงระยะเวลาหนึ่ง (Customer Churn) หมายถึงการที่ผู้บริโภคเลิกสนใจ และตัดสินที่จะเลิกเป็นลูกค้า [15] ซึ่งการคาดคะเนการเลิกเป็นลูกค้านั้นเป็นประเด็นที่นักวิจัยสนใจ และหาเทคนิควิธีต่าง ๆ เพื่อทำนายการเลิกใช้งานของลูกค้า ดังนั้นจึงจำเป็นต้องใช้เทคนิคการเรียนรู้ของเครื่องที่สามารถสร้างแบบจำลองการเรียนรู้ของข้อมูลและทำนายผลได้ โดยอาศัยชุดข้อมูลของเว็บไซต์แห่งหนึ่ง เช่น ข้อมูลประเภทประชากร ข้อมูลพฤติกรรมของลูกค้า และประวัติการซื้อสินค้าย้อนหลัง เพื่อหาตัวแปรที่สำคัญที่ส่งผลต่อประสิทธิภาพผลการทำนาย

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาการนำเทคนิคการเรียนรู้ของเครื่องเพื่อใช้ในการทำนายแนวโน้มการเลิกเป็นลูกค้าของเว็บไซต์แห่งหนึ่ง ซึ่งในการทำนายจะใช้เทคนิค Logistic Regression, Support Vector Machines, และ Random Forest ซึ่งเป็นเทคนิคการเรียนรู้แบบมีผู้สอนทำให้สามารถวัดผลประสิทธิภาพการทำงานได้อย่างแม่นยำ อีกทั้งยังเพิ่มความน่าเชื่อถือให้กับแบบจำลอง โดยอาศัยการอธิบายแบบจำลองด้วยค่าความสำคัญของฟีเจอร์ที่ใช้ในการเรียนรู้ (Feature Importance) ซึ่งการตีความการทำงานของแบบจำลองด้วยเครื่องมือที่ชื่อว่า Local Interpretable Model-agnostic Explanations (LIME) และตรวจสอบการใช้งานโดยแสดงค่าสำคัญของฟีเจอร์บนข้อมูล 1 ข้อมูลและตีความการทำงานแบบจำลองออกมาเป็นภาพให้มีความเข้าใจง่ายมากยิ่งขึ้น

งานวิจัยที่เกี่ยวข้อง

บทสรุปงานวิจัย Customer Churn Prediction ได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายผลของลูกค้า ซึ่งมีทั้งหมด 10 บทความ ถูกแบ่งออกเป็น 4 กลุ่มหลัก ดังนี้ โดยกลุ่มที่ 1 แบ่งเป็นการแปรียบเทียบของแบบจำลองซึ่งได้นำเสนอการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) หลากหลายแบบจำลองเพื่อทำนายการเลิกเป็นลูกค้า (Customer Churn Prediction) โดยแบบจำลองที่ได้รับความนิยมสูงสุดที่นำมาใช้ ได้แก่ Random Forest, Logistic Regression, SVM, KNN, XGBoost และ Deep Neural Networks ในงานวิจัยส่วนใหญ่ได้ใช้เมตริก (Metric) เปรียบเทียบประสิทธิภาพของแบบจำลองต่าง ๆ ได้แก่ Accuracy, Precision, Recall, F1-score, และ AUC (Area Under the ROC Curve) ผลการศึกษาพบว่าแบบจำลอง Random Forest มักมีประสิทธิภาพดีที่สุด [1, 3, 6, 7] กลุ่มที่ 2 การปรับความสมดุลของข้อมูล ซึ่งข้อมูล Customer Churn มักมีความไม่สมดุล หมายถึงจำนวนลูกค้าที่เลิกใช้บริการมีน้อยกว่าจำนวนลูกค้าที่ยังใช้งานอยู่ งานวิจัยในกลุ่มนี้เสนอวิธีการปรับสมดุลข้อมูล เช่น Oversampling, Undersampling และ SMOTE ผลการศึกษาพบว่าการปรับสมดุลข้อมูลช่วยเพิ่มประสิทธิภาพให้กับแบบจำลอง [2, 10] กลุ่มที่ 3 การผสมผสานของแบบจำลอง โดยแบบจำลองที่ใช้ในการผสม ได้แก่ Decision Tree, Neural Network, Random Forest โดยงานวิจัยในกลุ่มนี้เสนอการผสมผสานแบบจำลองการเรียนรู้ของเครื่อง (Ensemble Learning) เพื่อเพิ่มประสิทธิภาพการทำนาย ซึ่งผลการศึกษาพบว่าแบบจำลองที่ลผสมผสานมีประสิทธิภาพดีกว่าแบบจำลองเดียว [4, 13] และกลุ่มที่ 4 การเลือกคุณลักษณะ โดยใช้วิธี Feature Selection, Principal Component Analysis (PCA) จากผลการศึกษาพบว่าการเลือกคุณลักษณะช่วยเพิ่มประสิทธิภาพให้กับแบบจำลอง [9, 11]

วิธีดำเนินการ

ขั้นตอนที่ 1 : การเก็บรวบรวมข้อมูลและจัดการกับข้อมูล

ในงานวิจัยนี้ได้ใช้ข้อมูลประชากรของลูกค้าในเว็บไซต์แห่งหนึ่ง ซึ่งเป็นข้อมูลสาธารณะ Kaggle.com จากเว็บไซต์ https://www.kaggle.com/datasets/undersc0re/predict-the-churn-risk-rate ซึ่งประกอบด้วย 23 คุณลักษณะ มีจำนวนข้อมูลทั้งหมด 36,992 ตัวอย่าง โดยตัวแปรเป้าหมายแบ่งออกเป็น 2 กลุ่มคือ 0 หรือลูกค้าที่ยังใช้บริการอยู่ (Exist) และ 1 หรือลูกค้าที่เลิกไม่ใช้บริการแล้ว (Churn) ข้อมูลอยู่ในรูปแบบ CSV หลังจากนั้นทำการสำรวจข้อมูลเบื้องต้นตรวจสอบหาค่าว่างและข้อมูลที่มีผิดปกติพบว่า

สำหรับฟีเจอร์ preferred_offer_types เก็บข้อมูลข้อเสนอของลูกค้ามีจำนวนตัวอย่างข้อมูลทั้งหมด 36,992 ตัวอย่าง พบว่ามี 3 ข้อเสนอคือ Gift Vouchers/Coupons หรือบัตรกำนัล/คูปองมีจำนวน 12,349 ตัวอย่าง และ Credit/Debit Card Offers หรือข้อเสนอบัตรเครดิต/เดบิตมีจำนวน 12,274 ตัวอย่าง และ Without Offers หรือไม่มีข้อเสนอที่ต้องการมีจำนวน 12,081 ตัวอย่าง และพบว่ามีจำนวนค่าว่างที่อยู่ในฟีเจอร์นี้อยู่ 288 ตัวอย่าง ซึ่งหาข้อมูลมาทดแทนไม่ได้ จึงตัดสินใจลบค่าว่างในข้อมูลนี้ออก 288 ตัวอย่าง ซึ่งอยู่ในจำนวนที่ไม่มากเกินไป ซึ่งฟีเจอร์นี้เอง เมื่อลบค่าว่างออกไปแล้วจะส่งผลทำให้ฟีเจอร์อื่น ๆ ที่อยู่ในแถวเดียวกันถูกลบออกไปด้วย ดังนั้นฟีเจอร์ preferred_offer_types จึงเหลือตัวอย่างข้อมูลทั้งหมด 36,704 ตัวอย่าง

สำหรับฟีเจอร์ avg_frequency_login_days เก็บข้อมูลจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ย โดยพบว่าชนิดของข้อมูลมีความผิดจากเดิมคือ object แล้วเปลี่ยนเป็น float64 เพราะว่าจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยควรมีค่าเป็นตัวเลข และยังพบค่าว่างในตัวอย่างข้อมูลเป็นจำนวน 3,522 ตัวอย่าง เมื่อทำการวิเคราะห์ดูข้อมูลแล้วจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์ควรจะเป็นค่าตัวเลข หลังจากนั้นจึงทำการแทนที่ค่าว่างด้วยค่า 0 ให้กับฟีเจอร์นี้ อีกทั้งยังพบว่าในฟีเจอร์นี้มีค่าที่เป็นค่าติดลบอยู่ จึงทำการตัดสินใจเอาค่าติดลบออก เพราะว่าการนับจำนวนครั้งในการเข้าสู่ระบบไม่ควรเป็นค่าติดลบ ดังนั้นจากตอนแรกมีจำนวนข้อมูลทั้งหมด 36,992 ตัวอย่าง จึงเหลือจำนวน 36,028 ตัวอย่าง

สำหรับฟีเจอร์ points_in_wallet เก็บข้อมูลคะแนนสะสมที่ลูกค้าได้รับในแต่ละการทำธุรกรรม พบว่ามีค่าว่างในข้อมูลตัวอย่างอยู่ 3,443 ตัวอย่าง หลังจากนั้นจึงทำการแทนที่ค่าว่างด้วยค่า 0 ให้กับฟีเจอร์นี้ อีกทั้งยังพบว่าในฟีเจอร์นี้มีค่าที่เป็นค่าลบอยู่ จึงทำการเอาค่าติดลบออก เพราะว่าคะแนนสะสมไม่น่าจะมีค่าติดลบได้ จากตอนแรกมีจำนวนข้อมูลทั้งหมด 36,992 ตัวอย่าง เหลือจำนวน 35,896 ตัวอย่าง

สำหรับฟีเจอร์ avg_time_spent เก็บข้อมูลเวลาที่ใช้บนเว็บไซต์โดยเฉลี่ยของลูกค้า โดยพบว่าฟีเจอร์นี้มีค่าที่เป็นค่าติดลบอยู่ จึงทำการเอาค่าติดลบออก จากตอนแรกมีจำนวนข้อมูลทั้งหมด 36,992 ตัวอย่าง เหลือจำนวน 34,250 ตัวอย่าง ซึ่งถ้านำค่าติดลบไปใช้ในแบบจำลองอาจจะทำให้ประสิทธิภาพของแบบจำลองแย่ลงได้

สำหรับฟีเจอร์ medium_of_operation สื่อดำเนินการที่ลูกค้าใช้ทำธุรกรรม โดยภายในฟีเจอร์มีข้อมูลอยู่ 4 ประเภท คือ Desktop หรือเดสก์ท็อปมีจำนวน 13,913 ตัวอย่าง และ Smartphone หรือสมาร์ตโฟนมีจำนวน 13,876 ตัวอย่าง และ Both หรือใช้ทั้งเดสก์ท็อปและสมาร์ตโฟนมีจำนวน 3,810 แถว โดยพบว่ามีข้อมูลที่ไม่รู้ความหมายที่แน่ชัดคือค่า '?' มีจำนวน 5,393 ตัวอย่าง เมื่อเราทำการพิจารณาฟีเจอร์นี้แล้วจึงเลือกทำการแทนที่ค่า '?' ด้วยค่า Unknown แทน เพราะเป็นข้อมูลที่เราไม่รู้หาค่าอื่นมาจัดการไม่ได้ อีกทั้งยังมีเป็นจำนวนมาก จึงไม่สามารถที่จะลบออกไปได้ ดังนั้นเมื่อทำการจัดการกับข้อมูลด้วยการแทนที่ไปแล้ว พบว่า Unknown มีจำนวนทั้งหมด 4,990 ตัวอย่าง ซึ่งข้อมูลที่มีความผิดปกติมีจำนวนลดลงเป็นเพราะว่าบางแถวของฟีเจอร์นี้ตรงกับ avg_time_spent ที่ถูกลบออกไปจากการไม่เอาค่าติดลบ มันจึงส่งผลกับฟีเจอร์ joined_through_referral, region_category และ days_since_last_login อีกด้วย

สำหรับฟีเจอร์ joined_through_referral เก็บข้อมูลประเภทที่ลูกค้าเข้าร่วมเป็นสมาชิกด้วย Code หรือ ID หากลูกค้าไม่ได้ทำการเข้าร่วมเป็นสมาชิกด้วย Code หรือ ID จะเป็นค่า NO พบว่ามีจำนวน 15,839 ตัวอย่าง ถ้าเข้าร่วมการเป็นสมาชิกด้วยด้วย Code หรือ ID จะเป็นค่า Yes พบว่ามีจำนวน 15,715 ตัวอย่าง และยังพบว่ามีข้อมูลแปลกอยู่ในฟีเจอร์นี้ คือค่า '?' มีจำนวน 5,438 ตัวอย่าง เนื่องจากเป็นค่าที่ไม่มีความหมายและอาจทำให้เกิดการวิเคราะห์ที่ผิดต่อการทำงานของแบบจำลองได้ จึงเลือกทำการแทนค่า '?' เป็นค่า Unknown พบว่ามีจำนวน 5,026 ตัวอย่าง หลังจากการแทนที่แล้ว

สำหรับฟีเจอร์ region_category เก็บข้อมูลพื้นที่อยู่อาศัยของลูกค้าพบว่ามี 3 พื้นที่คือ Town หรือเมืองขนาดเล็กมีจำนวน 14,128 ตัวอย่าง และ City หรือเมืองขนาดใหญ่มีจำนวน 12,737 ตัวอย่าง และ Village หรือหมู่บ้านมีจำนวน 4,699 ตัวอย่าง โดยพบค่าว่างมีจำนวน 5,428 ตัวอย่าง ดังนั้นเมื่อลูกค้าไม่ได้กรอกข้อมูลส่วนนี้ อีกทั้งยังพบค่าว่างเป็นจำนวนมาก เพื่อไม่ส่งผลกระทบต่อการทำงานของแบบจำลอง จึงเลือกแทนค่าว่างด้วยค่า Unknown หลังจากได้ทำการแทนที่ของข้อมูลไปแล้วพบว่า Unknown มีจำนวน 5,033 ตัวอย่าง

สำหรับฟีเจอร์ days_since_last_login เก็บข้อมูลจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุด โดยพบข้อมูลที่น่าสงสัยคือค่า -999 มีจำนวน 1,999 ตัวอย่าง เมื่อตรวจสอบดูแล้วจำนวนวันที่ไม่น่าจะมีค่าติดลบได้แล้ว -999 ก็อาจจะหมายถึงลูกค้าแทบจะไม่ได้เข้ามาในระบบเลย ดังนั้นจึงพิจารณาเลือกแทนที่ค่า -999 เป็นค่า 999 แทน เพราะข้อมูลจำนวนวันไม่น่าจะมีค่าติดลบได้แล้ว 999 ซึ่งเป็นค่าที่มีค่ามาก โดยเราให้ความหมายว่า ลูกค้าคนนี้เข้าสู่ระบบเว็บไซต์ครั้งล่าสุดเมื่อ 999 วันที่แล้ว หลังจากทำการแทนที่ไปแล้วพบว่าค่า 999 มีจำนวนทั้งหมด 1,861 ตัวอย่าง

เมื่อจัดการกับค่าว่างและตรวจสอบไม่พบความซ้ำซ้อนของข้อมูลแล้ว เหลือข้อมูลทั้งหมด 34,250 ตัวอย่าง และ 19 คุณลักษณะ ที่สามารถนำไปใช้ในกระบวนการสำรวจข้อมูลและในการเรียนรู้ของแบบจำลอง

ขั้นตอนที่ 2 : กระบวนการสำรวจข้อมูล (Exploratory Data Analysis)

โดยทำการเปรียบเทียบลูกค้าในแต่ละคุณลักษณะ เพื่อแสดงความแตกต่างของลูกค้าที่ยังใช้บริการอยู่ และลูกค้าที่เลิกไม่ใช้บริการแล้ว โดยแสดงเป็นรูปแบบกราฟแท่ง, วงกลม, ไวโอลิน, ฮิสโตแกรม, Stacked

ขั้นตอนที่ 3 : การเตรียมความพร้อมข้อมูล (Data Preprocessing)

ผู้วิจัยแบ่งข้อมูลโดยใช้คำสั่ง train_test_split จากไลบรารี่ scikit-learn ที่สัดส่วน 80% สำหรับข้อมูลในการเรียนรู้ (Training Set) และสัดส่วน 20% สำหรับข้อมูลในการทดสอบ (Test Set) โดยการเตรียมข้อมูลนี้เพื่อใช้กับชุดข้อมูลสำหรับการเรียนรู้เท่านั้น หลังจากนั้นทำการเปลี่ยนแปลงข้อมูลกลุ่มแบบไม่มีลำดับเป็นข้อมูลตัวเลขจะใช้วิธี One-Hot Encoding และปรับเปลี่ยนช่วงของข้อมูลตัวเลขใช้วิธี Standard Scaler ดังนั้นจึงมีการวิธีใช้ make_column_transformer ร่วมกับวิธี make_pipeline เพื่อไม่ให้เกิดชุดทดสอบรั่วไหล (Data Leakage) และคัดเลือกคุณลักษณะ โดยเทคนิค Random Forest เลือกคุณลักษณะที่มีความสำคัญต่ำออก อีกทั้งยังแก้ปัญหาความไม่สมดุลของข้อมูล โดยเลือกใช้วิธี Synthetic Minority Oversampling Technique (SMOTE)

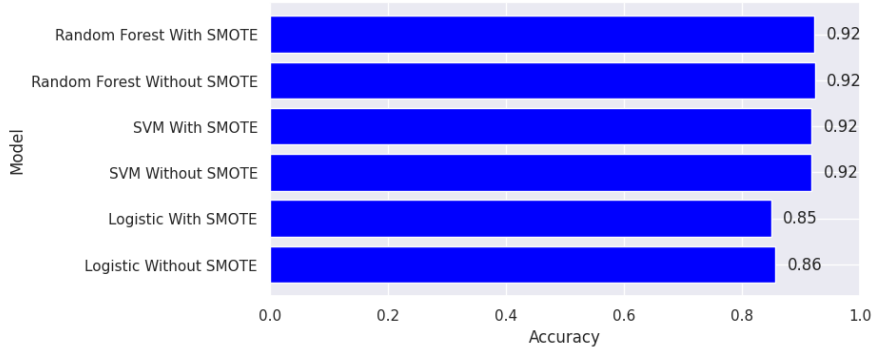
ขั้นตอนที่ 4 : การสร้างแบบจำลองเพื่อทำนายข้อมูล

การหาพารามิเตอร์ที่สุดสำหรับแบบจำลองในการทำนาย โดยใช้ข้อมูลในการเรียนรู้ (Training Set) ทั้งหมดมี 27,400 ตัวอย่าง และข้อมูลที่ใช้ในการทดสอบ (Test Set) ทั้งหมดมี 6,850 ตัวอย่าง อีกทั้งชุดข้อมูลที่ได้ถูกจัดการกับบการแก้ไขปัญหาความไม่สมดุลด้วย SMOTE มีข้อมูลทั้งหมด 29,470 ตัวอย่าง ซึ่งการเลือกหาชุดข้อมูลที่ดีที่สุดทำโดยใช้เทคนิค Grid Search ร่วมกับการทำ Cross Validation ที่ 5 Fold ในขั้นตอนการเรียนรู้ (Training Data) เพื่อทำการปรับจูนพารามิเตอร์ โดยทำการทดลองกับแบบจำลองที่ใช้คือ Logistic Regression, Support Vector Machines (SVM) และ Random Forest หลังจากการทำ Tuning หา Hyperparameter ครบทั้งหมดทุกแบบจำลองที่จะใช้ในงานวิจัยนี้แล้ว เพื่อนำไปทำการเรียนรู้ในข้อมูลชุดเรียนรู้ หลังจากทำการเรียนรู้ทุกแบบจำลองเรียบร้อยแล้ว จึงได้ทำการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง

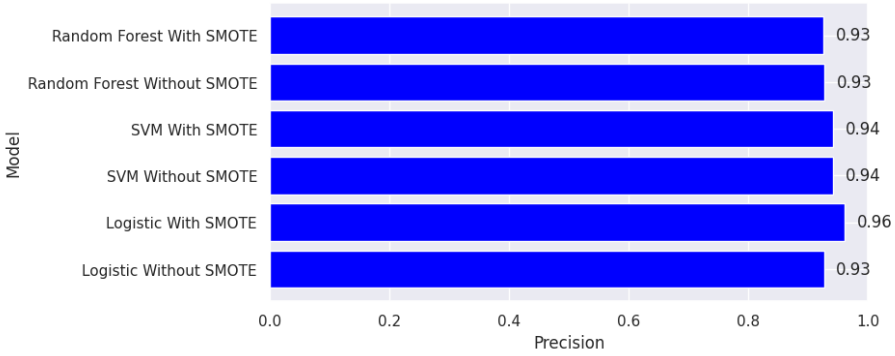
ผลการวิจัยและอภิปรายผลการวิจัย

ชื่อแบบจำลอง	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	เวลาที่ใช้ในการเรียนรู้ (วินาที)
Logistic Regression	0.86	0.93	0.86	0.86	0.09
Logistic Regression with SMOTE	0.85	0.96	0.85	0.85	0.11
SVM	0.92	0.94	0.92	0.92	58.39
SVM with SMOTE	0.92	0.94	0.92	0.92	58.78
Random Forest	0.92	0.93	0.92	0.93	14.31
Random Forest with SMOTE	0.92	0.93	0.92	0.93	4.40

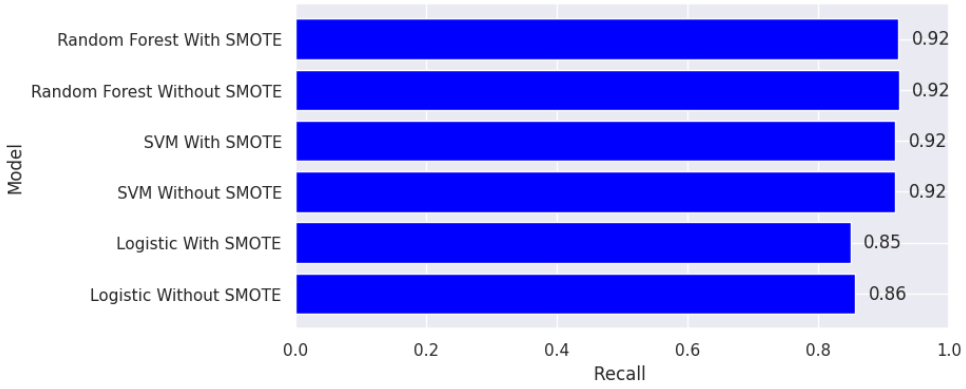
ภาพประกอบที่ 1 แสดงผลลัพธ์ของแบบจำลองทั้งหมด



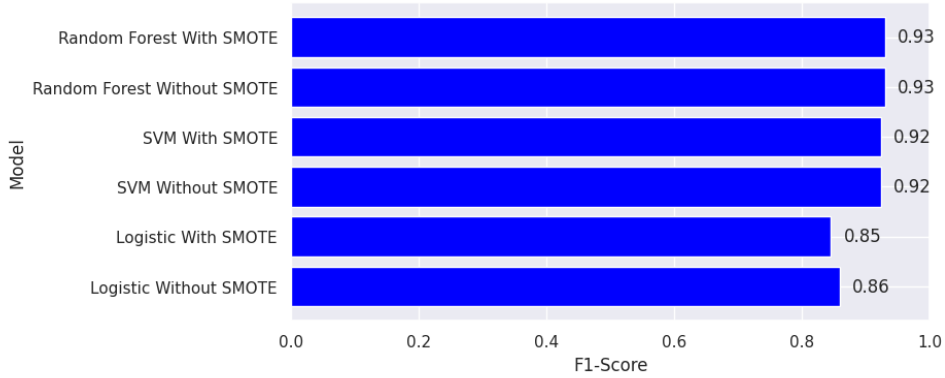
ภาพประกอบ 2 กราฟแท่งแสดงผลลัพธ์ค่า Accuracy ของแบบจำลองทั้งหมด



ภาพประกอบ 3 กราฟแท่งแสดงผลลัพธ์ค่า Precision ของแบบจำลองทั้งหมด

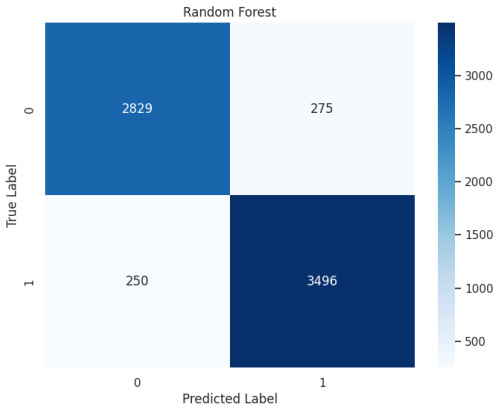


ภาพประกอบ 4 กราฟแท่งแสดงผลลัพธ์ค่า Recall ของแบบจำลองทั้งหมด



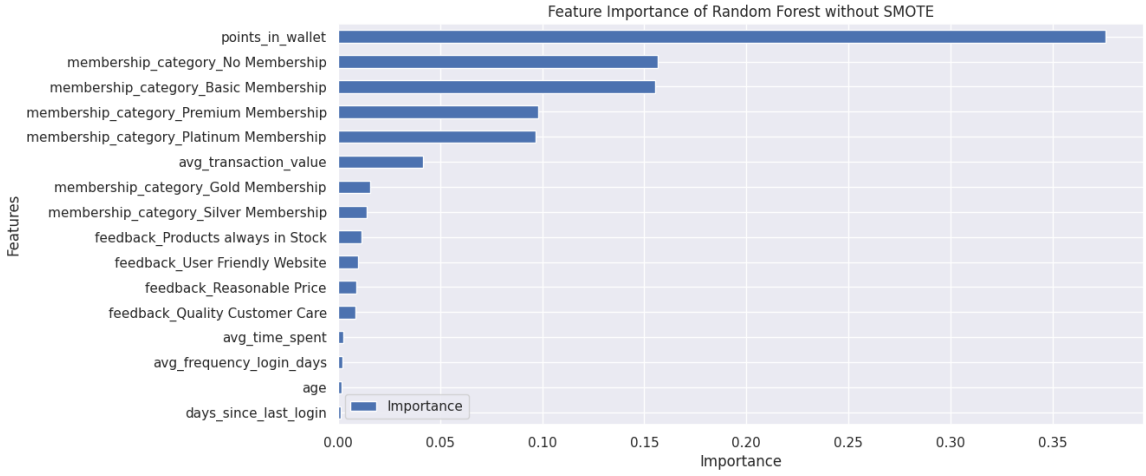
ภาพประกอบ 5 กราฟแท่งแสดงผลลัพธ์ค่า F1-Score ของแบบจำลองทั้งหมด

ในการวิจัยการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง จากการศึกษาผลลัพธ์ของแบบจำลองในภาพประกอบที่ 1 ถึง 5 จะเห็นได้ว่าผลลัพธ์ของแบบจำลอง Random Forest ทั้งแบบที่ใช้และไม่ใช้งาน SMOTE พบว่าแบบจำลองทั้งสองมีประสิทธิภาพโดยรวมดีที่สุด โดยที่ค่า Accuracy 92%, Precision 93%, Recall 93% และ F1-Score 93% ได้ค่าเท่ากันและระยะเวลาที่ใช้ในการเรียนรู้ไม่แตกต่างกันมาก ในขณะที่แบบจำลอง Logistic Regression ซึ่งใช้ระยะเวลาในการเรียนรู้ที่น้อยกว่า แต่มีประสิทธิภาพโดยรวมน้อยกว่าแบบจำลอง Random Forest ทั้งแบบที่ใช้และไม่ใช้งาน SMOTE เพื่อตรวจสังเกตความถูกต้องและความผิดพลาดของตัวอย่างข้อมูลจากชุดทดสอบด้วยการพิจารณา Confusion Matrix พบว่าแบบจำลอง Random Forest แบบไม่ใช้ SMOTE มีการทำนายถูกต้องมากที่สุดและทำนายผิดพลาดน้อยสุด ดังนั้นผู้วิจัยจึงเลือกแบบจำลองที่มีประสิทธิภาพดีที่สุดแสดงผลคุณลักษณะที่สำคัญ (Feature Importance) และ Local Interpretable Model-Agnostic Explanations (LIME)



ภาพประกอบ 6 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง Random Forest

จากภาพประกอบที่ 6 แสดงผล Confusion Matrix ของแบบจำลอง Random Forest พบว่ามีประสิทธิภาพในการทำนายผลได้ถูกต้องในกลุ่มลูกค้า Exist (Label 0) เป็นจำนวน 2,826 ตัวอย่างข้อมูล และทำนายได้ถูกต้องในกลุ่มลูกค้า Churn (Label 1) เป็นจำนวน 3,497 ตัวอย่างข้อมูล ซึ่งแบบจำลองมีการทำนายที่ถูกต้อง 6,323 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 527 ตัวอย่างข้อมูล

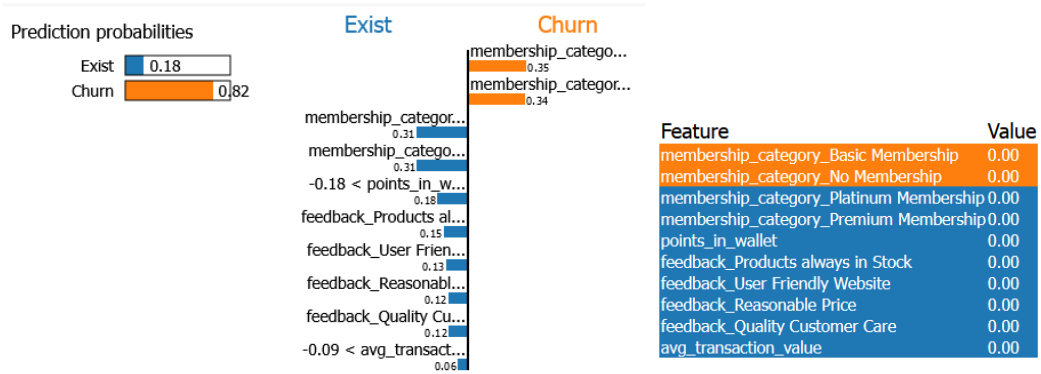


ภาพประกอบที่ 7 แสดงความสำคัญของคุณลักษณะจากแบบจำลอง Random Forest

ในขั้นตอนต่อไปนี้ผู้วิจัยได้ทำการแสดงผลคุณลักษณะที่สำคัญ (Feature Importance) เพื่อเป็นการสำรวจให้เข้าใจว่าคุณลักษณะใดมีอิทธิพลในการกำหนดผลลัพธ์ให้กับกลุ่มลูกค้า Exist และ Churn ตามภาพประกอบที่ 7 สำหรับแบบจำลอง Random Forest แบบไม่ใช้ SMOTE พบว่าคุณลักษณะที่มีความสำคัญต่อการจัดกลุ่มคือ points_in_wallet

ต่อไปพิจารณาสังเกตดูว่าผลลัพธ์ของคุณลักษณะที่สำคัญของแบบจำลองและผลลัพธ์จากการใช้เครื่องมือ LIME มีความสอดคล้องกันหรือไม่ พบว่ากลุ่มลูกค้าเลิกใช้บริการแล้วหรือ Churn มีความสอดคล้องกันของการใช้วิธี LIME ซึ่งพบว่าฟีเจอร์ที่ใช้ใน

การแบ่งกลุ่มเป็นระดับสมาชิก ในส่วนกลุ่มลูกค้าที่ยังใช้บริการอยู่หรือ Exist ซึ่งก็พบว่ามีความสอดคล้องกันจากการใช้วิธี LIME และพบว่าฟีเจอร์ที่ใช้ในการแบ่งกลุ่มเป็นระดับสมาชิกอีกเช่นกัน แต่มีการเปลี่ยนแปลงลำดับความสำคัญของคุณลักษณะกันเล็กน้อย ผู้วิจัยมองว่าการใช้ฟีเจอร์เหล่านี้ในการจัดกลุ่มนั้นสมเหตุสมผลซึ่งระดับของลูกค้านั้นสามารถแบ่งกลุ่มลูกค้าได้ชัดเจน ดังภาพประกอบที่ 8



ภาพประกอบ 8 แสดงผลลัพธ์ด้วยวิธี LIME ของข้อมูลที่ 1784 ในข้อมูลชุดทดสอบ จากแบบจำลอง Random Forest

การวิจัยในครั้งนี้แสดงให้เห็นว่าการสำรวจข้อมูลแต่ละลักษณะของแต่ละกลุ่มนั้น ช่วยให้เข้าใจกลุ่มลูกค้าได้ดีขึ้น โดยพบว่ากลุ่มลูกค้าเลิกใช้บริการแล้วมีความชัดเจนในระดับการเป็นสมาชิกของลูกค้าหรือ membership_category ที่ได้สำรวจพบว่าลูกค้าที่เลิกใช้บริการนั้นเป็นลูกค้าที่ระดับไม่สูงและไม่ได้เป็นสมาชิกด้วย อีกทั้ง points_in_wallet หรือคะแนนสะสมส่วนใหญ่ของลูกค้ากลุ่มนี้ก็ยังมีคะแนนสะสมน้อยกว่ากลุ่มลูกค้าที่ยังใช้บริการอยู่ และการพิจารณาคุณลักษณะอื่นร่วมด้วยเช่น feedback หรือการแสดงความคิดเห็นของลูกค้า พบว่าลูกค้าเลิกใช้บริการนั้นยังแสดงความคิดเห็นของสินค้าและเว็บไซต์ไปในแง่ลบอีกด้วย

สรุปผลการวิจัย

ในการวิจัยนี้ศึกษาการสร้างแบบจำลองการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรของลูกค้าบนเว็บไซต์แห่งหนึ่ง โดยใช้เทคนิคการเรียนรู้ของเครื่องเพื่อเปรียบเทียบประสิทธิภาพในการทำงานของแบบจำลอง ซึ่งมีทั้งหมด 3 แบบจำลอง คือ Logistic Regression, Support Vector Machines (SVM) และ Random Forest แบบจำลองเหล่านี้ใช้การคัดเลือกคุณลักษณะและเทคนิค SMOTE ในการแก้ปัญหาความไม่สมดุลของข้อมูล จากผลการทดลองสรุปได้ว่าแบบจำลองที่ให้ประสิทธิภาพดีที่สุดคือ Random Forest ให้ผลลัพธ์ตามนี้ Accuracy 92%, Precision 93%, Recall 92% และ F1-Score 93% จากนั้นพิจารณา Confusion Matrix พบว่าแบบจำลองมีการทำนายที่ถูกต้อง 6,325 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 525 ตัวอย่าง และยังค้นพบว่าแบบจำลอง Logistic Regression ที่ใช้ SMOTE ได้ใช้ระยะเวลาในการเรียรู้ของแบบจำลองน้อยที่สุด คือ 0.09 วินาที ดังนั้นในการใช้เครื่องมือ LIME มาช่วยในการอธิบายแบบจำลองว่ามีการทำงานหรือตัดสินใจเลือกใช้ฟีเจอร์ใดที่นำไปทำนายนั้น จึงช่วยทำให้นักการตลาดสามารถนำไปตัดสินใจและหาแนวทางวิธีต่าง ๆ เพื่อรักษาฐานลูกค้าไว้ให้ได้ ดังนั้นนักการ

ตลาดสามารถใช้แนวทางนี้เพื่อวางแผนล่วงหน้าและป้องกันไม่ให้ลูกค้าเลิกใช้บริการได้ เพื่อกระตุ้นลูกค้าให้ใช้บริการตลอดไม่หายไปไหน

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] Bhuse, P., Gandhi, A., Meswani, P., Muni, R., & Katre, N. (2020). Machine Learning Based Telecom-Customer Churn Prediction. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 1297–1301. Retrieved from https://doi.org/10.1109/ICISS49785.2020.9315951

[2] Feng, L. (2022). Research on Customer Churn Intelligent Prediction Model based on Borderline-SMOTE and Random Forest. 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS), 803–807. Retrieved from https://doi.org/10.1109/ICPICS55264.2022.9873702

[3] Hassonah, M. A., Rodan, A., Al-Tamimi, A.-K., & Alsakran, J. (2019). Churn Prediction: A Comparative Study Using KNN and Decision Trees. 2019 Sixth HCT Information Technology Trends (ITT), 182–186. Retrieved from https://doi.org/10.1109/ITT48889.2019.9075077

[4] Hu, X., Yang, Y., Chen, L., & Zhu, S. (2020). Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network. 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 129–132. Retrieved from https://doi.org/10.1109/ICCCBDA49378.2020.9095611

[5] Mahesh, B. (2018). Machine Learning Algorithms—A Review. 9(1).

[6] Peddarapu, R. K., Ameena, S., Yashaswini, S., Shreshta, N., & PurnaSahithi, M. (2022). Customer Churn Prediction using Machine Learning. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, 1035–1040. Retrieved from https://doi.org/10.1109/ICECA55336.2022.10009093

[7] Raeisi, S., & Sajedi, H. (2020). E-Commerce Customer Churn Prediction By Gradient Boosted Trees. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), 055–059. Retrieved from https://doi.org/10.1109/ICCKE50421.2020.9303661

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier (arXiv:1602.04938). arXiv. Retrieved from http://arxiv.org/abs/1602.04938

[9] Sharma, A., Gupta, D., Nayak, N., Singh, D., & Verma, A. (2022). Prediction of Customer Retention Rate Employing Machine Learning Techniques. 2022 1st International Conference on Informatics (ICI), 103–107. Retrieved from https://doi.org/10.1109/ICI53355.2022.9786903

[10] Shumaly, S., Neysaryan, P., & Guo, Y. (2020). Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), 082–087. Retrieved from https://doi.org/10.1109/ICCKE50421.2020.9303698

[11] Stehani, S., Karunya, N., Ranjan, D. R. J. B., Sumathipala, S., & Sandanayake, T. C. (2020). Customer Churn Reasoning in Telecommunication Domain. 2020 International Conference on Image Processing and Robotics (ICIP), 1–5. Retrieved from https://doi.org/10.1109/ICIP48927.2020.9367342

[12] Xu, Z., Shen, D., Kou, Y., & Nie, T. (2022). A Synthetic Minority Oversampling Technique Based on Gaussian Mixture Model Filtering for Imbalanced Data Classification. IEEE Transactions on Neural Networks and Learning Systems, 1–14. Retrieved from https://doi.org/10.1109/TNNLS.2022.3197156

[13] Zhang, C., Li, H., Xu, G., & Zhu, X. (2021). Customer churn model based on complementarity measure and random forest. 2021 International Conference on Computer, Blockchain and Financial Development (CBFD), 95–99. Retrieved from https://doi.org/10.1109/CBFD52659.2021.00026

[14] สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์. (2564). E-Commerce ไทย ยุคหลัง COVID-19. สืบค้นจาก https://www.etda.or.th/th/Useful-Resource/Knowledge-Sharing/Perspective-on-Future-of-e-Commerce.aspx

[15] สิรภัทร เกาฏีระ. (ม.ป.ป.). ถึงเวลาแล้วหรือไม่? ที่คำว่า Shopping จะเป็นเพียงอดีต...เมื่อพฤติกรรมผู้บริโภคไม่เหมือนเดิม. สืบค้นจาก https://www.krungsri.com/th/wealth/krungsri-prime/privileges/articles/shopping-will-be-past-consumer-behavior-not-same