การศึกษาเปรียบเทียบประสิทธิภาพอัลกอริทึมเพื่อการจำแนกพฤติกรรมลูกค้าผู้ใช้บริการทางด้านโทรคมนาคม

โดย

ชนาธิป ศรีนวล

โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2567

A COMPARATIVE STUDY OF DIFFERENT ALGORITHMS FOR CLASSIFYING TELECOMMUNICATION CUSTOMER CHURN BEHAVIOR

BY

CHANATIP SRINAUL

A FINAL-YEAR PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE

COMPUTER SCIENCE

FACULTY OF SCIENCE AND TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2024

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

รายงานโครงงานพิเศษ

ของ

ชนาธิป ศรีนวล

เรื่อง

การศึกษาเปรียบเทียบประสิทธิภาพอัลกอริทึมเพื่อการจำแนกพฤติกรรมลูกค้าผู้ใช้บริการทางด้านโทรคมนาคม

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
เมื่อ วันที่ 27 ธันวาคม พ.ศ. 2567

อาจารย์ที่ปรึกษา นวฤกษ์ ชลารักษ์
(ดร.นวฤกษ์ ชลารักษ์)

กรรมการสอบโครงงานพิเศษ
(ผู้ช่วยศาสตราจารย์ ดร.ปกป้อง ส่องเมือง)

กรรมการสอบโครงงานพิเศษ
(รองศาสตราจารย์ ดร.ณัฐธนนท์ หงส์วริทธิ์ธร)

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

รายงานโครงงานพิเศษ

ของ

ชนาธิป ศรีนวล

เรื่อง

การศึกษาเปรียบเทียบประสิทธิภาพอัลกอริทึมเพื่อการจำแนกพฤติกรรมลูกค้าผู้ใช้บริการทางด้านโทรคมนาคม

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
เมื่อ วันที่ 27 ธันวาคม พ.ศ. 2567

อาจารย์ที่ปรึกษา นวฤกษ์ ชลารักษ์
(ดร.นวฤกษ์ ชลารักษ์)

กรรมการสอบโครงงานพิเศษ
(ผู้ช่วยศาสตราจารย์ ดร.ปกป้อง ส่องเมือง)

กรรมการสอบโครงงานพิเศษ
(รองศาสตราจารย์ ดร.ณัฐธนนท์ หงส์วริทธิ์ธร)

หัวข้อโครงงานพิเศษ	การศึกษาเปรียบเทียบประสิทธิภาพอัลกอริทึมเพื่อการจำแนกพฤติกรรมลูกค้าผู้ใช้บริการทางด้านโทรคมนาคม
ชื่อผู้เขียน	ชนาธิป ศรีนวล
ชื่อปริญญา	วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาโครงงานพิเศษ	ดร. นวฤกษ์ ชลารักษ์
ปีการศึกษา	2567

บทคัดย่อ

การสูญเสียลูกค้า (Customer Churn) เป็นปัญหาสำคัญของกลุ่มธุรกิจต่าง ๆรวมไปถึงธุรกิจโทรคมนาคมด้วยเช่นกัน เนื่องจากส่งผลต่อรายได้และความมั่นคงของธุรกิจ การทำนายการสูญเสียลูกค้า (Customer Churn Prediction) จึงเป็นสิ่งสำคัญที่จะช่วยให้กลุ่มธุรกิจสามารถรักษาฐานลูกค้าเอาไว้ได้ โดยการคาดการณ์กลุ่มลูกค้าที่มีแนวโน้มจะเลิกใช้บริการและวางแผนทางการตลาดเพื่อรักษากลุ่มลูกค้ากลุ่มนั้นไว้เช่น การออกโปรโมชั่นต่าง ๆหรือสิทธิพิเศษต่าง ๆ สำหรับที่กลุ่มลูกค้า ซึ่งในงานวิจัยต่าง ๆที่เกี่ยวข้องกับการสร้างโมเดลเพื่อทำนายการสูญเสียลูกค้านั้นได้มีการเลือกใช้อัลกอริทึมที่แตกต่างกันออกไป รวมไปถึงมีขั้นตอนในการจัดเตรียมข้อมูลที่แตกต่างกันตามลักษณะของข้อมูลที่เลือกใช้อีกด้วย

โครงงานนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบโมเดลทำนายการสูญเสียลูกค้าโดยใช้ชุดข้อมูลการสูญเสียลูกค้าในธุรกิจโทรคมนาคมจำนวน 7,043 คนและมีการจัดการกับความไม่สมดุลของข้อมูลด้วยวิธี Random Undersampling, Random Oversampling และ SMOTE โดยอัลกอริทึมที่ใช้ในการสร้างโมเดลได้แก่ Random Forest, XGBoost และ SVM

จากผลการทดลองพบว่าโมเดลที่สร้างด้วยอัลกอริทึม Random Forest และจัดการความไม่สมดุลของข้อมูลด้วยวิธี Random Oversampling ให้ผลลัพธ์ที่ดีที่สุด โดยได้ค่าความถูกต้อง (Accuracy) 88.77% ค่าความแม่นยำ (Precision) 84.83% ค่าความระลึกได้ (Recall) 94.54% และค่าพื้นที่ใต้กราฟ (AUC) 0.95 และจากการหาคุณลักษณะที่สำคัญพบว่าค่าใช้จ่ายรวมของลูกค้า ระยะเวลาที่ใช้บริการ และค่าใช้จ่ายรายเดือน เป็นปัจจัยสำคัญที่ส่งผลต่อการตัดสินใจเลิกใช้บริการของลูกค้า

คำสำคัญ: การทำนายการสูญเสียลูกค้า, การจำแนกประเภท, การสุ่มป่าไม้, เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง, ซัพพอร์ตเวกเตอร์แมชชีน, การจัดการความไม่สมดุลของข้อมูล

Thesis Title	A COMPARATIVE STUDY OF DIFFERENT ALGORITHMS FOR CLASSIFYING TELECOMMUNICATION CUSTOMER CHURN BEHAVIOR
Author	Chanatip Srinaul
Degree	Bachelor of Science
Major Field/Faculty/University	Computer Science Faculty of Science and Technology Thammasat University
Project Advisor	Ph.D. Nawarerk Chalarak
Academic Years	2024

ABSTRACT

Customer churn is a significant problem for various industries, including the telecommunications industry, as it directly impacts revenue and business stability. Customer churn prediction is crucial for businesses to retain their customer base by predicting the customers most likely to leave and implementing marketing strategies to retain them, such as offering promotions or special benefits for these customers. In previous research on building models for customer churn prediction, different algorithms and data preparation steps were utilized depending on the characteristics of the data used.

This research aims to study and compare customer churn prediction models using a dataset of 7,043 customers in the telecommunications industry. Handling imbalanced data by using methods such as Random Undersampling, Random Oversampling, and SMOTE. The algorithms used for build model include Random Forest, XGBoost, and SVM.

The results of the experiments indicate that the model built using the Random Forest algorithm with Random Oversampling for handling imbalanced data achieves the best results, achieving an accuracy of 88.77%, precision of 84.83%, recall of 94.54%, and an AUC of 0.95. Feature importance analysis revealed that total charges, tenure, and monthly charges are the key factors affecting customers' decision to churn.

Keywords: Customer Churn Prediction, Classification, Random Forest, XGBoost, SVM, Handling Imbalanced Data

กิตติกรรมประกาศ

ขอขอบคุณอาจารย์ที่ปรึกษา ดร.นวฤกษ์ ชลารักษ์ที่ให้คำแนะนำและความช่วยเหลือต่างๆ ทั้งในด้านวิชาการ การวางแผน การวิเคราะห์ข้อมูล การนำเสนอโครงงาน และช่วยปรับปรุงแก้ไขตลอดการทำโครงงานนี้จนเสร็จสมบูรณ์และขอขอบคุณคณะกรรมการสอบทุกท่านที่ได้ช่วยให้คำแนะนำและข้อเสนอแนะเพื่อปรับปรุงแก้ไขจนโครงงานนี้สำเร็จลุล่วงไปด้วยดี นอกจากนี้ขอขอบคุณครอบครัวและเพื่อนนิสิตทุกคนที่คอยให้กำลังใจและความช่วยเหลือตลอดระยะเวลาที่ทำโครงงานนี้มาครับ

ชนาธิป ศรีนวล

สารบัญ

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

รายการสัญลักษณ์และคำย่อ

สัญลักษณ์/คำย่อ	คำเต็ม/คำจำกัดความ
1) RF	การสุ่มป่าไม้
2) XGBoost	เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง
3) SVM	ซัพพอร์ตเวกเตอร์แมชชีน
4) SMOTE	การสังเคราะห์ข้อมูลเพิ่ม

บทที่ 1
บทนำ

โครงงานนี้นำเสนอเกี่ยวกับแนวคิดและศึกษาเปรียบเทียบอัลกอริทึมที่ใช้ในการสร้างโมเดลทำนายในการแบ่งประเภทเพื่อทำนายและจำแนกประเภทของลูกค้าที่ยังคงใช้บริการกับลูกค้าที่เลิกใช้บริการไปแล้วของกลุ่มธุรกิจโทรคมนาคม เนื้อหาในบทนำนี้จะนำเสนอที่มาและความสำคัญ วัตถุประสงค์ ขอบเขต ประโยชน์ที่คาดว่าจะได้รับและข้อจำกัดของโครงงาน

1.1 ความเป็นมาและความสำคัญของโครงงาน

ในปัจจุบันที่โลกเจริญก้าวหน้ามากยิ่งขึ้น มีธุรกิจต่าง ๆเพิ่มมากขึ้นในทุก ๆวัน ทำให้เกิดการแข่งขันทางธุรกิจกันมากยิ่งขึ้น ธุรกิจโทรคมนาคมก็เป็นอีกหนึ่งในธุรกิจที่มีความสำคัญต่อผู้คนและธุรกิจด้านอื่นๆและผู้บริโภคยังต้องการการติดต่อสื่อสารที่มีประสิทธิภาพและรวดเร็ว ทำให้ธุรกิจโทรคมนาคมก็เป็นอีกหนึ่งธุรกิจที่มีการแข่งขันทางธุรกิจที่สูง กลุ่มลูกค้าจึงเป็นตัวแปรสำคัญสำหรับธุรกิจ โดยเฉพาะกลุ่มลูกค้าเดิมหรือกลุ่มลูกค้าที่ใช้สินค้าและบริการอย่างสม่ำเสมอเพราะเป็นกลุ่มลูกค้าที่สร้างผลกำไรให้กับบริษัทอย่างสม่ำเสมอทำให้บริษัทมีรายได้คงที่และมีความมั่นคง นอกจากนี้ยังเป็นส่วนสำคัญให้กับบริษัทในเรื่องอื่นๆ เช่น ช่วยลดต้นทุนในการหาลูกค้าใหม่ๆและเป็นสื่อกลางที่ดีสำหรับการแนะนำสินค้าและบริการของบริษัทให้กับกลุ่มลูกค้าใหม่ รวมไปถึงการหาข้อมูลทางการตลาดและการศึกษาความต้องการของลูกค้าโดยใช้ข้อมูลของกลุ่มลูกค้าเดิมเพื่อใช้ข้อมูลดังกล่าวนำมาวางแผนทางการตลาดและใช้ในการปรับปรุงสินค้าและบริการของบริษัทให้ดียิ่งขึ้นได้อีกด้วย

จากที่กล่าวมาข้างต้น จะเห็นได้ว่าการรักษากลุ่มลูกค้าเดิมมีความสำคัญเช่นเดียวกันกับการหากลุ่มลูกค้าใหม่ ดังนั้นการวิเคราะห์หาสาเหตุที่ลูกค้าเลิกใช้สินค้าหรือบริการ (Customer Churn Analysis) และการทำนายการสูญเสียลูกค้า (Customer Churn Prediction) จึงเป็นสิ่งที่ช่วยให้บริษัททราบถึงแนวโน้มหรือโอกาสที่ลูกค้าจะเลิกใช้สินค้าและบริการ เมื่อทราบประเภทของกลุ่มลูกค้าที่มีแนวโน้มดังกล่าวแล้วก็จะทำให้บริษัทสามารถวางแผนทางการตลาดเพื่อรองรับความเสี่ยงนั้นได้ดีมากยิ่งขึ้น เช่น การคิดโปรโมรชั่น สิทธิประโยชน์ต่าง ๆสำหรับกลุ่มลูกค้าเดิม เป็นต้น

การทำนายการสูญเสียลูกค้า (Customer Churn Prediction) เป็นการทำนายแนวโน้มเพื่อจำแนกประเภท (Classification) ของลูกค้าว่าเป็นลูกค้าที่ยังคงใช้สินค้าและบริการอยู่หรือเลิกใช้ไปแล้ว ในการสร้างโมเดลสำหรับทำนายการสูญเสียลูกค้านั้นจำเป็นต้องใช้อัลกอริทึมที่เหมาะสม โดยแต่ละอัลกอริทึมมีหลักการทำงานและการนำไปใช้งานที่แตกต่างกันออกไป การ

เลือกอัลกอริทึมที่เหมาะสมจะช่วยให้โมเดลสำหรับทำนายสามารถทำนายได้อย่างมีประสิทธิภาพและช่วยเพิ่มความแม่นยำในการทำนาย ทำให้สามารถนำไปใช้เพื่อประกอบการตัดสินใจทางธุรกิจได้ดีมากยิ่งขึ้น

โดยทางผู้วิจัยได้เล็งเห็นถึงปัญหานี้ จึงได้จัดทำโครงงานนี้ขึ้นมาเพื่อศึกษาและเปรียบเทียบประสิทธิภาพของอัลกอริทึมที่ใช้ในการสร้างโมเดลสำหรับทำนายว่าอัลกอริทึมใดเหมาะสมสำหรับการสร้างโมเดลทำนายการสูญเสียลูกค้าในกลุ่มธุรกิจโทรคมนาคมเพื่อช่วยในการประกอบการตัดสินใจทางธุรกิจและรักษาฐานลูกค้าเก่าอย่างมีประสิทธิภาพ

1.2 วัตถุประสงค์

1. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการทำงานของโมเดลทำนายการสูญเสียลูกค้าในกลุ่มธุรกิจโทรคมนาคม
2. เพื่อสร้างโมเดลทำนายสำหรับวิเคราะห์หาคุณลักษณะที่สำคัญที่ส่งผลให้ลูกค้าตัดสินใจเลิกใช้บริการในกลุ่มธุรกิจโทรคมนาคม

1.3 ขอบเขตของโครงงาน

1. โครงงานนี้ใช้ชุดข้อมูล การสูญเสียลูกค้าในธุรกิจโทรคมนาคม (Telco Customer Churn) ซึ่งเป็นชุดข้อมูลที่สมมติขึ้นจากข้อมูลการใช้บริการโทรคมนาคมของลูกค้าในแคลิฟอร์เนีย จากเว็บไซต์ Kaggle โดยเป็นข้อมูลที่เปิดเผยแล้ว (Open Data) ซึ่งในชุดข้อมูลประกอบด้วย รหัสสมาชิกของลูกค้า เพศ ความอาวุโส สถานะสมรส สถานะครอบครัว ระยะเวลาที่ใช้บริการ การใช้บริการทางโทรศัพท์ การใช้บริการเบอร์โทรศัพท์หลายสาย การใช้บริการอินเตอร์เน็ต การใช้บริการความปลอดภัยทางออนไลน์ การใช้บริการสำรองข้อมูลออนไลน์ การใช้บริการป้องกันอุปกรณ์ การใช้บริการสนับสนุนทางเทคนิค การใช้บริการสตรีมทีวี การใช้บริการสตรีมภาพยนตร์ ระยะเวลาที่เซ็นสัญญา การเรียกเก็บเงินแบบไม่ใช้กระดาษ วิธีการชำระเงิน ค่าใช้จ่ายรายเดือน ค่าใช้จ่ายทั้งหมดและสถานะการใช้บริการ
2. โครงงานนี้นำเสนอเกี่ยวกับวิธีการและเปรียบเทียบโมเดลทำนายเพื่อแบ่งประเภท (Classification) เพื่อนำไปใช้ในการสร้างโมเดลเพื่อทำนายการสูญเสียลูกค้า (Customer Churn Prediction) แต่ไม่ได้นำเสนอเกี่ยวกับวิธีการแก้ปัญหาเมื่อลูกค้าตัดสินใจจะเลิกใช้บริการหรือนำเสนอวิธีการวางแผนทางการตลาดเพื่อลดจำนวนลูกค้าที่จะเลิกใช้บริการ

1.4 ประโยชน์ของโครงงาน

1. สามารถนำผลการศึกษาเปรียบเทียบอัลกอริทึมไปประยุกต์ใช้ในการเลือกอัลกอริทึมที่เหมาะสมและการพัฒนาโมเดลให้มีประสิทธิภาพมากยิ่งขึ้นทั้งในกลุ่มธุรกิจโทรคมนาคมและกลุ่มธุรกิจอื่นๆ
2. สามารถนำปัจจัยและคุณลักษณะที่สำคัญที่ใช้จำแนกประเภทของลูกค้า ไปประยุกต์ใช้ในการประกอบการตัดสินใจและต่อยอดทางธุรกิจ

1.5 ข้อจำกัดของโครงงาน

1. โครงงานนี้ใช้ชุดข้อมูลเกี่ยวกับ การสูญเสียลูกค้าในธุรกิจโทรคมนาคม (Telco Customer Churn) ซึ่งมีข้อมูลของลูกค้าจำนวน 7,043 คน หากต้องการใช้ชุดข้อมูลชุดอื่นที่มีจำนวนข้อมูลมากขึ้นหรือเป็นชุดข้อมูลในเรื่องอื่น จำเป็นต้องทำการวิเคราะห์ข้อมูลใหม่เสียก่อนเพื่อที่จะได้นำไปใช้ในการสร้างโมเดลได้อย่างมีประสิทธิภาพ
2. อัลกอริทึมที่ใช้ในการสร้างโมเดลการทำนายการสูญเสียลูกค้า (Customer Churn Prediction) ในโครงงานนี้ได้แก่ เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost), การสุ่มป่าไม้ (Random Forest) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) หากใช้อัลกอริทึมประเภทอื่นในการสร้างโมเดลอาจจะทำให้ผลลัพธ์ของการทำนายของโมเดลเปลี่ยนแปลงได้

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดทฤษฎีที่เกี่ยวข้อง

2.1.1 การวิเคราะห์การตัดสินใจเลิกซื้อสินค้าและบริการของลูกค้า (Customer Churn Analysis)

การวิเคราะห์การตัดสินใจเลิกซื้อสินค้าและบริการของลูกค้า (Customer Churn Analysis) คือ การวิเคราะห์และทำนายลักษณะของลูกค้าที่จะเลิกซื้อสินค้าและบริการ เพราะถ้าหากบริษัทไม่สามารถรักษาฐานลูกค้าของตนเองไว้ได้ จะส่งผลกระทบต่อการจัดการต้นทุนและการเติบโตทางการตลาดของบริษัทได้ ดังนั้นเพื่อช่วยให้บริษัทสามารถบริหารและวางแผนการตลาดเพื่อไม่ให้ลูกค้าเลิกซื้อสินค้าและบริการ จึงจำเป็นต้องมีการพัฒนาเครื่องมือเพื่อใช้การวิเคราะห์และทำนายว่าลูกค้าคนใดหรือกลุ่มใดมีแนวโน้มที่เลิกซื้อสินค้าและบริการเพื่อให้บริษัทได้วางแผนทางการตลาดและรับมือได้อย่างเหมาะสม

2.1.2 การทำให้เป็นมาตรฐานของข้อมูล (Data Standardization)

บ่อยครั้งที่ก่อนจะนำข้อมูลของลูกค้ามาใช้ในการสร้างโมเดลทำนายเพื่อทำนายการสูญเสียลูกค้า ข้อมูลของลูกค้าที่มีอยู่มีการกระจายตัวของข้อมูลที่ผิดปกติ มีค่าผิดปกติที่สูงหรือต่ำเกินไป ซึ่งสิ่งเหล่านี้จะทำให้โมเดลทำนายได้ผิดพลาดหรือประสิทธิภาพในการทำนายลดลง ดังนั้นจึงจำเป็นต้องมีการทำให้ข้อมูลเป็นมาตรฐาน (Standardization) เพื่อทำให้ข้อมูลมีความสม่ำเสมอและลดผลกระทบของข้อมูลที่มีค่าผิดปกติ เพื่อช่วยปรับปรุงประสิทธิภาพในการทำนายของโมเดลทำนาย

2.1.2.1 การแปลงข้อมูลให้เป็นมาตรฐาน (Standard Scaling)

การทำให้ข้อมูลเป็นมาตรฐานคือการแปลงข้อมูลให้มีค่าเฉลี่ย (Mean) เท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) เท่ากับ 1 โดยใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานในการคำนวณตามสูตรดังนี้

$$X_{scaled} = \frac{X-\mu}{\sigma}$$

โดยที่ X คือค่าในชุดข้อมูล

μ คือค่าเฉลี่ยของข้อมูล

σ คือค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูล

2.1.3 การจัดการกับความไม่สมดุลของข้อมูล (Handling Imbalanced Data)

ในงานวิจัยที่เกี่ยวข้องกับการสร้างโมเดลทำนายเพื่อทำนายการสูญเสียลูกค้า (Customer Churn Prediction) บ่อยครั้งชุดข้อมูลที่นำมาใช้ในการสร้างโมเดลก็มีสัดส่วนจำนวนของลูกค้าที่เลิกใช้บริการกับลูกค้าที่ยังไม่เลิกใช้บริการอยู่มาก ซึ่งอาจจะทำให้เกิดความไม่สมดุลของข้อมูล (Class Imbalance) ทำให้โมเดลทำนายได้ผลลัพธ์ที่ไม่ดีมากนัก จึงจำเป็นต้องมีการจัดการความไม่สมดุลของข้อมูลเหล่านี้ โดยมีวิธีได้แก่ การลดจำนวนตัวอย่างของข้อมูลแบบสุ่ม, การเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่มและการสังเคราะห์ข้อมูลเพิ่ม

2.1.3.1 การลดจำนวนตัวอย่างข้อมูลแบบสุ่ม (Random Undersampling)

เป็นวิธีการลดตัวอย่างข้อมูลโดยการสุ่มลดลงไป โดยจะลบตัวอย่างข้อมูลในคลาสที่มีจำนวนมากกว่าออกไปเพื่อให้มีจำนวนใกล้เคียงกับคลาสที่มีจำนวนน้อยกว่า ซึ่งเป็นวิธีที่สะดวกรวดเร็วและลดความซ้ำซ้อนของข้อมูล แต่ก็อาจจะไปลบข้อมูลที่มีความสำคัญออกส่งผลให้โมเดลเสียข้อมูลที่สำคัญสำหรับการฝึกไป รวมไปถึงไม่เหมาะกับชุดข้อมูลที่มีจำนวนน้อยเพราะจะทำให้มีจำนวนข้อมูลไม่เพียงพอสำหรับการฝึกและทดสอบโมเดล

2.1.3.2 การเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่ม (Random Oversampling)

เป็นวิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มเพิ่มขึ้นมา โดยจะคัดลอกจำนวนตัวอย่างข้อมูลในคลาสที่มีจำนวนน้อยกว่าและนำมาเพิ่มให้มีจำนวนใกล้เคียงกับคลาสที่มีจำนวนมากกว่า ซึ่งเป็นวิธีที่สะดวกรวดเร็วเหมือนการลดจำนวนตัวอย่างข้อมูลแบบสุ่ม แต่อาจจะทำให้โมเดลเกิด Overfitting ได้ง่ายเนื่องจากการสุ่มเพิ่มตัวอย่างข้อมูลที่สำคัญขึ้นมาจำนวนมากอีก ทั้งยังไม่ทำให้เกิดความหลากหลายของข้อมูล จึงไม่เหมาะสำหรับชุดข้อมูลที่มีสัดส่วนจำนวนข้อมูลของแต่ละคลาสแตกต่างกันมากจนเกินไปเพราะจะทำให้โมเดลไม่เหมาะสำหรับการนำไปใช้งานจริง

2.1.3.3 การสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling: SMOTE)

เป็นวิธีการเพิ่มตัวอย่างข้อมูลโดยการสร้างข้อมูลขึ้นมาใหม่ โดยจะสร้างตัวอย่างข้อมูลใหม่จากตัวอย่างข้อมูลจากคลาสที่มีจำนวนน้อยกว่า โดยสร้างตัวอย่างข้อมูลใหม่จากการคำนวณหาตำแหน่งจุดข้อมูลในคลาสนั้นๆด้วยหลักการเพื่อนบ้านใกล้เคียง (K-Nearest Neighbors) เพื่อหาจุดข้อมูลที่ใกล้เคียงกัน แล้วสร้างตัวอย่างข้อมูลใหม่ขึ้นมาระหว่างจุดข้อมูล

เหล่านั้น ซึ่งเป็นวิธีที่ไม่ทำให้เกิดความซ้ำซ้อนของข้อมูลเหมือนวิธีการลดและเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่ม แต่ทำให้จำเป็นต้องใช้เวลาในการประมวลผลมากขึ้นเพื่อสร้างข้อมูลใหม่ และข้อมูลที่สร้างขึ้นก็อาจจะไม่สะท้อนกับความเป็นจริงซึ่งอาจจะทำให้โมเดลทำนายคาดเคลื่อนได้

2.1.4 อัลกอริทึมที่ใช้ในการสร้างโมเดลเพื่อจำแนกประเภท (Algorithm for Classification Model)

ในงานวิจัยต่างๆที่เกี่ยวกับการวิเคราะห์การเลิกซื้อสินค้าและบริการของลูกค้า (Customer Churn Analysis) และสร้างโมเดลทำนายการสูญเสียลูกค้า (Customer Churn Prediction) มีการใช้อัลกอริทึมที่หลากหลายและจำเป็นต้องเลือกอัลกอริทึมที่เหมาะสมกับชุดข้อมูลที่ใช้เพื่อให้โมเดลทำนายสามารถทำนายได้อย่างมีประสิทธิภาพ แต่จากการทบทวนวรรณกรรม อัลกอริทึมที่ได้ผลลัพธ์ที่ดีคือเอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost) และการสุ่มป่าไม้ (Random Forest) ดังนั้นทางผู้วิจัยจึงได้เลือกอัลกอริทึมทั้ง 2 อัลกอริทึมดังกล่าวมาใช้ในโครงงานนี้และจากปัญหาของโครงงานนี้คือการทำนายการสูญเสียลูกค้า ซึ่งเป็นปัญหาการจำแนกประเภทแบบสองกลุ่ม (Binary Classification Problem) ผู้วิจัยจึงได้เลือกซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) มาใช้ในโครงงานนี้ร่วมด้วย เนื่องจากเป็นอัลกอริทึมที่เน้นสร้างเส้นแบ่ง (Hyperplane) เพื่อจำแนกประเภทของข้อมูล โดยอัลกอริทึมจะพยายามค้นหาเส้นแบ่งที่มีระยะห่าง (Margin) ที่กว้างมากที่สุดเพื่อลดความผิดพลาดในการจำแนกประเภทซึ่งเหมาะสมกับลักษณะของปัญหาของโครงงานนี้

2.1.4.1 เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost)

เป็นอัลกอริทึมที่พัฒนาต่อมาจาก Gradient Boosting โดยเป็นอัลกอริทึมที่จะสร้างต้นไม้ตัดสินใจหลายๆต้นขึ้นมา โดยต้นไม้ตัดสินใจแต่ละต้นจะเรียนรู้จากค่าความผิดพลาดของต้นไม้ตัดสินใจต้นก่อนและปรับปรุงการทำนายให้ดีขึ้นในแต่ละรอบ วนซ้ำจนกว่าค่าความผิดพลาดแทบจะไม่เกิดการเปลี่ยนแปลง

2.1.4.2 การสุ่มป่าไม้ (Random Forest)

เป็นอัลกอริทึมที่สร้างและนำผลลัพธ์ออกมาจากหลายๆต้นไม้ตัดสินใจ (Decision Tree) โดยต้นไม้ตัดสินใจแต่ละต้นจะรับข้อมูลที่แตกต่างกันออกไปโดนสุ่มมาจากชุดข้อมูลเดิม โดยต้นไม้ตัดสินใจแต่ละต้นจะเรียนรู้และทำนายผลจากข้อมูลที่ได้และคำนวณผลการทำนายสุดท้ายด้วยการโหวตจากผลโหวตที่มากที่สุด (Majority Vote) จากต้นไม้ตัดสินใจทุกต้นในกรณีที่เป็นปัญหาแบบจำแนกประเภท (Classification) และในกรณีที่เป็นปัญหาเชิงปริมาณ (Regression) จะใช้ค่าเฉลี่ยของผลลัพธ์จากต้นไม้ตัดสินใจทุกต้น

2.1.4.3 ซัพพอร์ตเวกเตอร์แมชชีน (SVM)

เป็นอัลกอริทึมที่จะพยายามหาเส้นแบ่ง (Hyperplane) เพื่อจำแนกข้อมูลออกเป็นกลุ่มต่างๆให้ได้ชัดเจนมากที่สุด โดยการหาตำแหน่งของเส้นแบ่งที่ทำให้ระยะห่าง (Margin) ระหว่างข้อมูลแต่ละกลุ่มห่างกันมากที่สุด

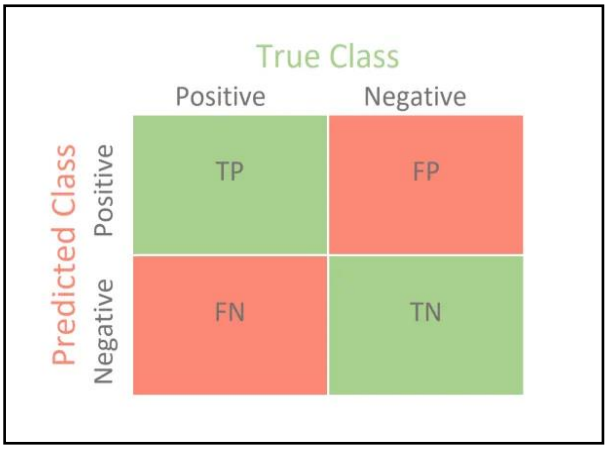
2.1.5 วิธีวัดประสิทธิภาพของตัวทำนาย (Model Evaluation)

วิธีการวัดประสิทธิภาพในการทำนายของโมเดลที่นิยมใช้ในหลายๆงานวิจัย เพื่อวัดผลลัพธ์ที่ได้จากการทำนายของโมเดลมีอยู่หลายวิธี เช่น การวัดค่าความถูกต้อง (Accuracy), การวัดค่าความแม่นยำ (Precision), การวัดค่าความระลึกได้ (Recall) พร้อมทั้งสร้างตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) เพื่อดูภาพรวมของผลการทำนายของโมเดลเทียบกับผลจริงของข้อมูลและการสร้างกราฟ ROC-AUC เพื่อแสดงความสัมพันธ์ระหว่างค่า False Positive Rate และค่า True Positive Rate พร้อมทั้งใช้ค่าพื้นที่ใต้กราฟ (Area Under Curve:AUC) เพื่อประเมินความสามารถในการแยกประเภทคลาสของโมเดลทำนาย

2.1.5.1 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix)

Confusion Matrix คือตารางแสดงผลลัพธ์การทำนายของโมเดลจำแนกประเภท (Classification Model) เพื่อแสดงให้เห็นว่าโมเดลทำนายได้ถูกต้องหรือผิดพลาดเท่าใดโดยภายในตารางจะประกอบไปด้วยค่าทั้งหมด 4 ค่า ได้แก่ True Positive (TP), True Negative (TN), False Positive (FP) และ False Negative (FN)

ภาพที่ 2.1 ตัวอย่างของตาราง Confusion Matrix



จาก Mohajon. J., (2020). Confusion Matrix for Your Multi-Class Machine Learning Model. สืบค้นจาก https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

True Positive (TP) คือข้อมูลที่เป็นจริงและโมเดลสามารถทำนายได้ถูกต้องว่าเป็นจริง

True Negative (TN) คือข้อมูลที่เป็นเท็จและโมเดลสามารถทำนายได้ถูกต้องว่าเป็นเท็จ

False Positive (FP) คือ ข้อมูลที่เป็นเท็จแต่โมเดลทำนายผิดพลาดว่าเป็นจริง

False Negative (FN) คือ ข้อมูลที่เป็นจริงแต่โมเดลทำนายผิดพลาดว่าเป็นเท็จ

2.1.5.2 ค่าความถูกต้อง (Accuracy)

คือค่าความถูกต้องในการทำนายของโมเดลโดยคำนวณจากสัดส่วนของข้อมูลที่โมเดลทำนายได้ถูกต้องต่อจำนวนข้อมูลทั้งหมด ซึ่งคำนวณได้จากสูตรต่อไปนี้

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

2.1.5.3 ค่าความแม่นยำ (Precision)

คือค่าที่วัดความแม่นยำของโมเดลในการทำนายผลลัพธ์ที่เป็นจริง โดยคำนวณจากสัดส่วนของข้อมูลที่เป็นจริงและโมเดลทำนายได้ถูกต้องต่อจำนวนข้อมูลที่โมเดลทำนายว่าเป็นจริงทั้งหมด ซึ่งคำนวณได้จากสูตรต่อไปนี้

$$Precision = \frac{TP}{TP+FP}$$

2.1.5.4 ค่าความระลึกได้ (Recall)

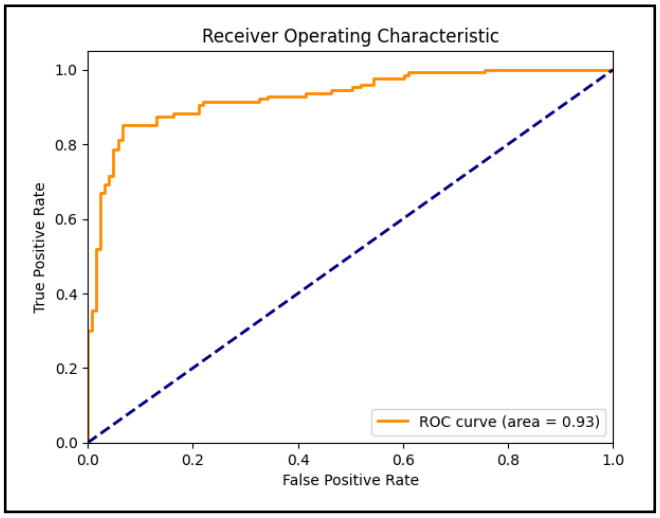
คือค่าที่วัดว่าโมเดลทำนายผลลัพธ์ที่เป็นจริงได้ถูกต้องเท่าใด โดยคำนวณจากสัดส่วนของข้อมูลที่เป็นจริงและโมเดลทำนายได้ถูกต้องต่อจำนวนข้อมูลที่เป็นจริงทั้งหมด ซึ่งคำนวณได้จากสูตรต่อไปนี้

$$Recall = \frac{TP}{TP+FN}$$

2.1.5.5 กราฟ ROC-AUC

กราฟ ROC (Receiver Operating Characteristic) เป็นกราฟที่แสดงให้เห็นถึงความสามารถในการจำแนกประเภทข้อมูลของโมเดลโดยการแสดงความสัมพันธ์ระหว่างค่า True Positive Rate (TPR) หรืออีกชื่อหนึ่งคือค่าความไว (Sensitivity) กับค่า False Positive Rate (FPR) และมีค่าพื้นที่ใต้เส้นโค้ง (Area Under Curve: AUC) เพื่อวัดประสิทธิภาพของโมเดลทำนาย

ภาพที่ 2.2 ตัวอย่างของกราฟ ROC-AUC



จาก Vidhi. C., (2024). AUC and the ROC Curve in Machine Learning. สืบค้นจาก https://www.datacamp.com/tutorial/auc

True Positive Rate (TPR) หรือค่าความไว (Sensitivity) คือค่าที่วัดว่าโมเดลทำนายผลลัพธ์ที่เป็นจริงได้ถูกต้องเท่าใด โดยคำนวณจากสัดส่วนของข้อมูลที่เป็นจริงและโมเดลทำนายได้ถูกต้องต่อจำนวนข้อมูลที่เป็นจริงทั้งหมด ซึ่งคำนวณได้จากสูตรต่อไปนี้

$$TPR = \frac{TP}{TP+FN}$$

False Positive Rate (FPR) คือค่าที่วัดว่าโมเดลทำนายผลลัพธ์ที่เป็นเท็จผิดพลาดเท่าใด โดยคำนวณจากสัดส่วนของข้อมูลที่เป็นเท็จและโมเดลทำนายผิดพลาดว่าเป็นจริงต่อจำนวนข้อมูลที่เป็นเท็จทั้งหมด ซึ่งคำนวณได้จากสูตรต่อไปนี้

$$FPR = \frac{FP}{FP+TN}$$

ค่า Area Under Curve (AUC) หรือค่าพื้นที่ใต้กราฟ คือค่าที่วัดความสามารถในการจำแนกประเภทของโมเดล โดยคำนวณจากพื้นที่ใต้กราฟ ROC มีค่าตั้งแต่ 0 ถึง 1 ถ้าค่า AUC เข้าใกล้ 1 หมายความว่าโมเดลสามารถจำแนกประเภทข้อมูลได้ดี แต่ถ้าค่า AUC น้อยกว่าหรือเท่ากับ 0.5 จะหมายความว่าโมเดลสามารถจำแนกประเภทข้อมูลได้แย่กว่าหรือเทียบเท่ากับการสุ่มเดา

2.2 งานวิจัยที่เกี่ยวข้อง

(Kumar et al., 2023) ได้ทำการวิจัยในเรื่องการวิเคราะห์การเลิกใช้บริการบัตรเครดิตแบบอัตโนมัติของลูกค้า (Credit Card Customer Churn) โดยข้อมูลที่ใช้มีจำนวน 10,127 แถว และมี 23 คุณลักษณะซึ่งเป็นข้อมูลเกี่ยวกับการใช้งานบัตรเครดิตของลูกค้า เช่น ประวัติการทำธุรกรรม ข้อมูลประชากรของลูกค้า เป็นต้น มีเป้าหมายคือการสร้างโมเดลทำนายการสูญเสียลูกค้าได้อย่างมีประสิทธิภาพ หลังจากนั้นจึงทำการจัดเตรียมข้อมูล (Data Preprocessing) โดยประกอบด้วย

1) ทำความสะอาดข้อมูล (Data Cleansing) ด้วยลบข้อมูลในแถวที่มีข้อมูลว่างทิ้ง

2) การแปลงข้อมูล (Data Transformation)

3) การทำให้ข้อมูลเป็นมาตรฐาน (Standardization)

ตามด้วยการสำรวจข้อมูลเชิงสำรวจ (Exploratory Data Analysis) และเลือกอัลกอริทึมเพื่อใช้ในการสร้างโมเดลทำนายได้แก่ การค้นหาเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors), การถดถอยโลจิสติก (Logistic Regression), ต้นไม้ตัดสินใจ (Decision Tree), เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost) และการใช้โมเดลแบบผสมผสานอีก 2 ประเภทคือการค้นหาเพื่อนบ้านใกล้สุด k ตัวกับการถดถอยแบบโลจิสติก (LR-KNN) และ การถดถอยโลจิสติกกับต้นไม้ตัดสินใจ (LR-DT)

1) LR-KNN โดยใช้การถดถอยโลจิสติกทำนายความน่าจะเป็นของผลลัพธ์และใช้วิธีการค้นหาเพื่อนบ้านใกล้สุด k ตัวเพื่อแบ่งกลุ่มของข้อมูล

2) LR-DT โดยใช้การถดถอยโลจิสติกทำนายความน่าจะเป็นของผลลัพธ์และใช้ต้นไม้ตัดสินใจเพื่อแบ่งกลุ่มของข้อมูล

และประเมินผลการทำนายของแต่ละอัลกอริทึมโดยใช้ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึกได้ (Recall) ร่วมกับ Confusion Matrix ซึ่งได้ผลลัพธ์ดังต่อไปนี้

ตารางที่ 2.1 ตารางสรุปผลลัพธ์การทำนายของแต่ละอัลกอริทึม (Kumar et al., 2023)

อัลกอริทึมที่ใช้	ค่าความถูกต้อง (Accuracy)	ค่าความแม่นยำ (Precision)	ค่าความระลึกได้ (Recall)
Logistic Regression	0.85	0.50	0.85
K-Nearest Neighbors	0.85	0.51	0.82
Decision Tree	0.92	0.70	0.80
XGBoost	0.93	0.92	0.88
LR-KNN	0.90	0.80	0.50
LR-DN	0.92	0.73	0.84

ซึ่งจากตารางดังกล่าวจะเห็นได้ว่าอัลกอริทึมที่ได้ค่าความแม่นยำต่างๆที่ดีที่สุดคือ เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost) ตามด้วยการใช้อัลกอริทึมแบบผสมผสานระหว่างการถดถอยโลจิสติกกับต้นไม้ตัดสินใจที่ได้ผลลัพธ์การทำนายที่ดีรองลงมา

(Miao & Wang, 2022) ได้ทำการวิจัยในเรื่องการวิเคราะห์การเลิกใช้บริการบัตรเครดิตแบบอัตโนมัติของลูกค้า (Credit Card Customer Churn) เช่นเดียวกันกับงานวิจัยของ (Kumar et al., 2023) แต่อัลกอริทึมที่นำมาใช้ในการศึกษาได้แก่ การสุ่มป่าไม้ (Random Forest), การถดถอยโลจิสติก (Logistic Regression) และการค้นหาเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) โดยข้อมูลที่ใช้ประกอบด้วยข้อมูลการใช้บัตรเครดิตของลูกค้าจำนวนมากกว่า 10,000 คนและมีคุณลักษณะทั้งหมด 21 คุณลักษณะ เช่น อายุของลูกค้า, รายได้ เป็นต้น หลังจากจัดการเตรียมข้อมูลโดยการจัดเตรียมข้อมูล (Data Preprocessing) แล้ว ผู้วิจัยก็ได้สร้างโมเดลทำนายที่มีการปรับค่าพารามิเตอร์ต่างๆและประเมินประสิทธิภาพของโมเดลทำนายด้วยกราฟ ROC-AUC ร่วมกับ Confusion Matrix โดยได้ค่าพื้นที่ใต้กราฟ (AUC) ของแต่ละอัลกอริทึมดังนี้ Random Forest ได้ค่า AUC เท่ากับ 0.98, Logistic Regression ได้ค่า AUC

เท่ากับ 0.91 และ K-Nearest Neighbors ได้ค่า AUC เท่ากับ 0.90 ซึ่งจะเห็นได้ว่าการสุ่มป่าไม้ (Random Forest) มีประสิทธิภาพในการทำนายมากที่สุด

(Taskin, 2023) ได้ทำการวิจัยเกี่ยวกับการสร้างโมเดลเพื่อทำนายการสูญเสียลูกค้าในกลุ่มธุรกิจโทรคมนาคม (Telco Customer Churn) โดยมีจำนวนข้อมูลของลูกค้า 3,333 คน และมี 16 คุณลักษณะ โดยอัลกอริทึมที่ใช้ ได้แก่ การค้นหาเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors), ซัพพอร์ตเวกเตอร์แมชชีน (SVM), การถดถอยโลจิสติก (Logistic Regression), การสุ่มป่าไม้ (Random Forest), เอดาบูส (AdaBoost), ไลต์เกรเดียนต์บูสติ้งแมนชีน (LGBM), เกรเดียนต์บูสติ้ง (Gradient Boosting) และ เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost) ซึ่งมีวิธีการจัดเตรียมข้อมูล (Data Preprocessing) ดังนี้

1) ลบคอลัมน์เบอร์โทรศัพท์ของลูกค้าออกไป เพราะไม่ได้ใช้ในการวิเคราะห์ข้อมูล
2) แปลงค่าข้อมูลจากข้อมูลประเภทหมวดหมู่ให้เป็นตัวเลข
3) เปลี่ยนค่าตัวแปรเป้าหมาย (Churn) จากค่า True และ False เป็น 1 และ 0

หลังจากการสร้างโมเดลด้วยแต่ละอัลกอริทึมและมีการปรับแต่งค่าพารามิเตอร์ด้วย Grid Search แล้วก็ทำการประเมินผลการทำนายของโมเดลได้ดังต่อไปนี้

ตารางที่ 2.2 ตารางสรุปผลลัพธ์การทำนายของแต่ละอัลกอริทึม (Taskin, 2023)

Models	Accuracy	Precision	Recall	Specificity	G-Mean	Roc-Score	MCC
RF	95.38	**94.86**	72.05	99.33	84.51	85.7	80
KNN	87.96	86.38	20	**99.44**	44.65	59.86	38.08
SVM	91.93	88.88	50.73	98.91	70.68	74.82	63.34
LR	86.05	55.52	21.31	97.02	45.2	59.16	28.1
Adaboost	87.94	64.43	38.12	96.4	60.42	67.25	43.3
LGBM	**95.74**	93.68	75.8	99.12	86.62	87.45	81.93
Grad	94.93	89.76	73.52	98.56	85.08	86.04	78.42
XGboost	**95.74**	92.32	**77.02**	98.91	**87.21**	**87.97**	**81.95**

ซึ่งจะเห็นได้ว่าเมื่อพิจารณาจากค่าความแม่นยำต่าง ๆจะเห็นได้ว่า XGBoost และ LGBM ให้ผลลัพธ์ที่ดีที่สุดโดยมีค่าความแม่นยำ (Accuracy) ที่ 95.74% เท่ากัน และ Random Forest (RF) ให้ผลลัพธ์ดีรองลงมา

(Öztürk, Tunç & Akay, 2023) ได้ศึกษาเกี่ยวกับการวิเคราะห์การเลิกใช้บริการของผู้ขายบนตลาดซื้อขายสินค้าทางออนไลน์ซึ่งมีเป้าหมายคือการพัฒนาโมเดลสำหรับวิเคราะห์การสูญเสียผู้ขาย (Churn) บนตลาดซื้อขายสินค้าออนไลน์ โดยใช้ข้อมูลเกี่ยวกับผู้ขายที่มี 10

คุณลักษณะเช่น เมืองที่ผู้ขายอาศัย, รายได้รวม, ช่องทางในการขาย เป็นต้น อัลกอริทึมที่ผู้วิจัยเลือกใช้คือการถดถอยโลจิสติก (Logistic Regression) และการสุ่มป่าไม้ (Random Forest) และมีการจัดการกับความไม่สมดุลของข้อมูล (Handling Imbalanced Data) ก่อนที่จะนำข้อมูลไปใช้ในการสร้างโมเดลทำนายโดยใช้วิธีการลดจำนวนตัวอย่างข้อมูล (Undersampling) และการเพิ่มจำนวนตัวอย่างข้อมูล (Oversampling) เพื่อปรับสมดุลของข้อมูลโดยมีวิธีในการสร้างโมเดลทำนายแบ่งออกเป็น 3 วิธีดังนี้

1) สร้างโมเดลทำนายโดยไม่มีการจัดเตรียมข้อมูล (Data Preprocessing)
2) สร้างโมเดลทำนายโดยใช้การลดจำนวนตัวอย่างข้อมูลในการปรับสมดุลของข้อมูล
3) สร้างโมเดลทำนายโดยใช้การเพิ่มจำนวนตัวอย่างข้อมูลในการปรับสมดุลของข้อมูล

ซึ่งได้ผลลัพธ์การทำนายของโมเดลเป็นค่า F1-Score ของแต่ละอัลกอริทึมดังต่อไปนี้

ตารางที่ 2.3 ตารางสรุปค่า F1-Score ของการทำนายของแต่ละอัลกอริทึม (Öztürk, Tunç & Akay, 2023)

วิธีที่เลือกใช้	Random Forest	Logistic Regression
วิธีที่ 1	0.76	0.84
วิธีที่ 2	0.71	0.68
วิธีที่ 3	0.92	0.69

ซึ่งจะเห็นได้ว่าวิธีการที่ให้ค่าการทำนายที่ดีที่สุดคือการสร้างโมเดลด้วยอัลกอริทึมการสุ่มป่าไม้ (Random Forest) ที่มีการปรับสมดุลของข้อมูลด้วยการเพิ่มตัวอย่างข้อมูล (Oversampling) โดยได้ค่า F1-Score เท่ากับ 0.92 และรองลงมาคือการสร้างโมเดลด้วยอัลกอริทึมการถดถอยโลจิสติก (Logistic Regression) ที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูลโดยได้ค่า F1-Score เท่ากับ 0.84

จากผลลัพธ์ที่ได้จะเห็นได้ว่าในวิธีการที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูลกับวิธีการที่มีการจัดการกับความไม่สมดุลของข้อมูลนั้นได้ผลของค่า F1-Score ของทั้ง 2 อัลกอริทึมมีความแตกต่างกันอย่างเห็นได้ชัด ทำให้สามารถสรุปได้ว่าวิธีในการจัดการกับความไม่สมดุลของข้อมูล (Handling Imbalanced Data) แต่ละวิธีส่งผลต่อผลลัพธ์การทำนายของอัลกอริทึมแตกต่างกันออกไป ดังนั้นจึงควรพิจารณาเลือกวิธีในการจัดการกับความไม่สมดุลของ

ข้อมูลให้เหมาะสมกับข้อมูลและอัลกอริทึมที่เลือกใช้ในการสร้างโมเดลทำนายเพื่อให้ได้ผลลัพธ์การทำนายที่มีประสิทธิภาพ

สรุปจากการทบทวนวรรณกรรมเกี่ยวกับงานวิจัยทั้งหมดที่กล่าวมาจะเห็นได้ว่าในการสร้างโมเดลเพื่อทำนายการสูญเสียลูกค้าในกลุ่มธุรกิจต่าง ๆ จำเป็นต้องมีการวิเคราะห์และจัดเตรียมข้อมูลอย่างเหมาะสมก่อนที่จะนำมาใช้ในการสร้างโมเดลทำนาย เพื่อให้ได้ผลลัพธ์การทำนายที่มีความถูกต้องแม่นยำและมีประสิทธิภาพต่อการนำไปใช้งานจริงได้ เช่น การสำรวจข้อมูลเชิงสำรวจ (Exploratory Data Analysis) หรือการจัดการกับความไม่สมดุลของข้อมูล (Handling Imbalanced Data) ที่ผู้วิจัยจะนำไปประยุกต์ใช้ในโครงงานนี้ โดยรายละเอียดขั้นตอนวิธีการวิจัยต่างๆจะกล่าวในบทถัดไป

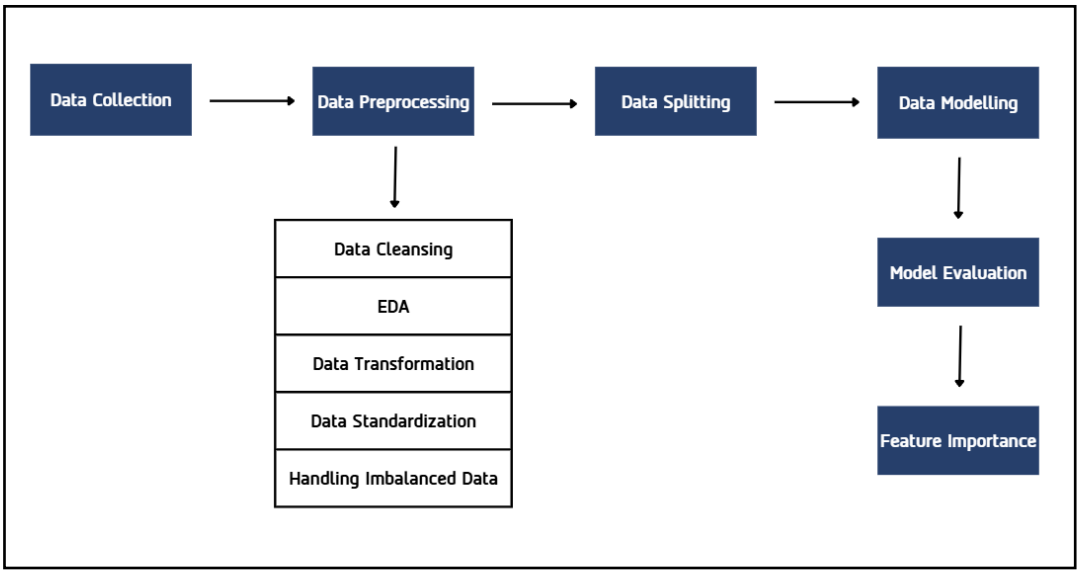
บทที่ 3
วิธีการวิจัย

3.1 ภาพรวมของโครงงาน

วิธีการในการวิจัยครั้งนี้มีขั้นตอนทั้งหมด 6 ขั้นตอนดังนี้

1. การเก็บรวบรวมข้อมูล (Data Collection)
2. การจัดเตรียมข้อมูล (Data Preprocessing)
3. การแบ่งชุดข้อมูลสำหรับการฝึกและการทดสอบ (Data Splitting)
4. การสร้างโมเดล (Data Modelling)
5. การประเมินประสิทธิภาพของโมเดล (Model Evaluation)
6. การหาคุณลักษณะที่สำคัญ (Feature Importance)

ภาพที่ 3.1 แผนภาพแสดงภาพรวมของโครงงาน



3.2 การเก็บรวบรวมข้อมูล (Data Collection)

ชุดข้อมูลที่ใช้ในโครงงานนี้คือชุดข้อมูลการสูญเสียลูกค้าในธุรกิจโทรคมนาคม (Telco Customer Churn) ซึ่งเป็นชุดข้อมูลที่สมมติขึ้นจากข้อมูลการใช้บริการโทรคมนาคมของลูกค้าในแคลิฟอร์เนีย จากเว็บไซต์ Kaggle โดยเป็นข้อมูลที่เปิดเผยแล้ว (Open Data) ซึ่งมีข้อมูลของลูกค้าจำนวน 7,043 คน โดยชุดข้อมูลนี้มีคุณลักษณะทั้งหมด 21 คุณลักษณะโดยแบ่งเป็น

คุณลักษณะเชิงปริมาณ 3 คุณลักษณะกับคุณลักษณะเชิงหมวดหมู่ 17 คุณลักษณะและมีตัวแปรเป้าหมายคือสถานะการใช้บริการ (Churn)

ตารางที่ 3.1 ตารางแสดงตัวอย่างของข้อมูลและคุณลักษณะของชุดข้อมูลที่เลือกนำมาใช้

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes

ตารางที่ 3.2 ตารางแสดงตัวอย่างของข้อมูลและคุณลักษณะของชุดข้อมูลที่เลือกนำมาใช้

OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No

customerID : รหัสสมาชิกของลูกค้า

gender : เพศ

SeniorCitizen : ความอาวุโส (เป็นผู้สูงวัย, ยังไม่เป็นผู้สูงวัย)

Partner : สถานะสมรส

Dependents : สถานะครอบครัว (มีสมาชิกคนอื่นในครอบครัวหรือไม่มี)

tenure : ระยะเวลาที่ใช้บริการ (ต่อเดือน)

PhoneService : การใช้บริการทางโทรศัพท์

MultipleLines : การใช้บริการโทรศัพท์หลายสาย

InternetService	: การใช้บริการทางอินเตอร์เน็ต
OnlineSecurity	: การใช้บริการความปลอดภัยทางออนไลน์
OnlineBackup	: การใช้บริการสำรองข้อมูลออนไลน์
DeviceProtection	: การใช้บริการป้องกันอุปกรณ์
TechSupport	: การใช้บริการสนับสนุนทางเทคนิค
StreamingTV	: การใช้บริการสตรีมทีวี
StreamingMovies	: การใช้บริการสตรีมภาพยนตร์
Contract	: ระยะเวลาที่เซ็นสัญญา (รายเดือน, 1 ปี, 2 ปี)
PaperlessBilling	: การเรียกเก็บเงินแบบไม่ใช้กระดาษ
PaymentMethod	: วิธีการชำระเงิน (เช็คอิเล็กทรอนิกส์, เช็คทางไปรษณีย์ โอนเงินผ่านทางธนาคาร, บัตรเครดิต)
MonthlyCharges	: ค่าใช้จ่ายรายเดือน
TotalCharges	: ค่าใช้จ่ายทั้งหมด
Churn	: สถานะการใช้บริการ

3.3 การจัดเตรียมข้อมูล (Data Preprocessing)

3.3.1 การทำความสะอาดข้อมูล (Data Cleansing)

3.3.1.1 ลบคุณลักษณะ customerID ออกเพราะไม่มีความเกี่ยวข้องในการวิเคราะห์ข้อมูล

3.3.1.2 เปลี่ยนชนิดข้อมูลของ TotalCharges จาก object เป็น float

3.3.1.3 ตรวจหาข้อมูลที่เป็นค่า Null พบว่าทุกคุณลักษณะมียอดรวมเป็น 0 นอกจาก TotalCharges ที่มีค่า Null รวมทั้งหมด 11 จำนวน

3.3.1.4 ลบแถวของข้อมูลที่มีค่า Null จำนวน 11 แถวออกไปเพื่อให้ข้อมูลทุกแถวมีความสมบูรณ์ โดยเหลือข้อมูลของลูกค้า 7,032 คน

3.3.1.5 เปลี่ยนค่าของคุณลักษณะ Senior Citizen จาก 0 เป็น “No” และจาก 1 เป็น

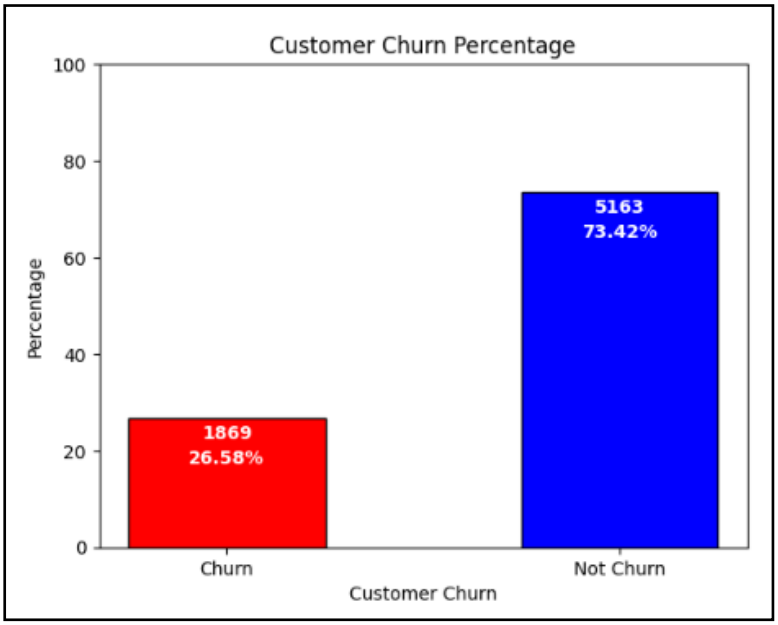
“Yes” และเปลี่ยนชนิดข้อมูลจาก int เป็น object

3.3.1.6 เปลี่ยนค่าของตัวแปรเป้าหมาย Churn จาก “No” เป็น 0 และจาก “Yes” เป็น 1 และเปลี่ยนชนิดข้อมูลจาก object เป็น int

3.3.2 การวิเคราะห์ข้อมูลเชิงสำรวจ (Exploratory Data Analysis)

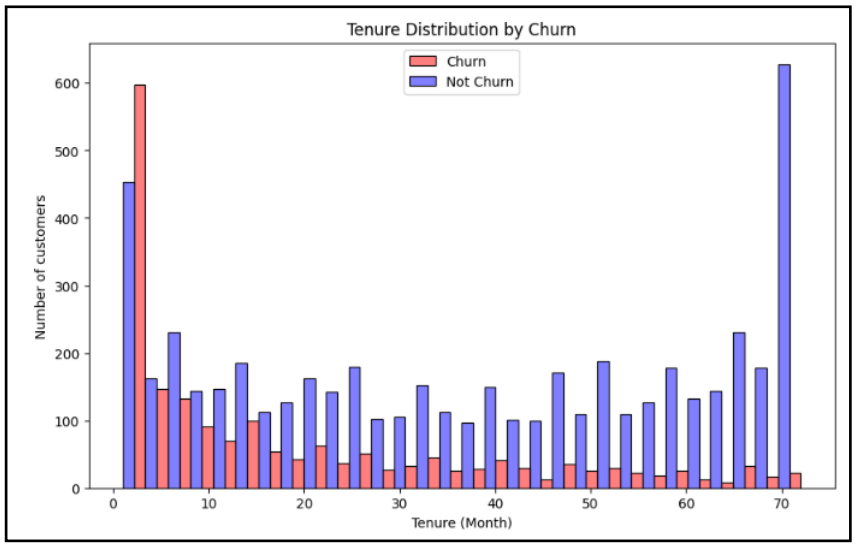
หลังจากทำความสะอาดข้อมูลแล้ว ผู้วิจัยต้องการสำรวจข้อมูลเบื้องต้นเกี่ยวกับความสมดุลของจำนวนตัวแปรเป้าหมาย, การกระจายตัวของข้อมูลในคุณลักษณะเชิงปริมาณทั้ง 3 คุณลักษณะเมื่อเทียบกับตัวแปรเป้าหมายและตรวจสอบเพื่อหาค่าผิดปกติ

ภาพที่ 3.2 กราฟแท่งแสดงสัดส่วนสถานะการใช้บริการของลูกค้า



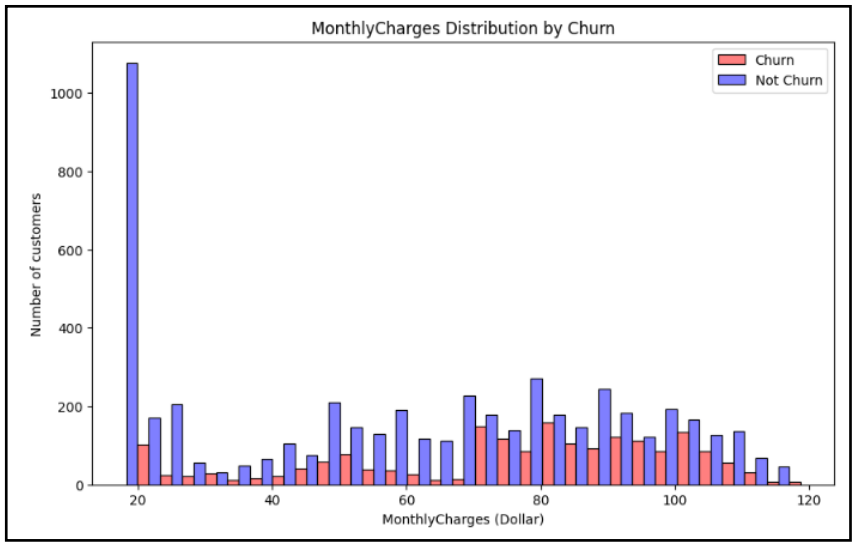
จากกราฟดังกล่าวจะเห็นได้ว่ามีการแบ่งลูกค้าออกเป็น 2 กลุ่มตามสถานะการใช้บริการ โดยลูกค้าที่เลิกใช้บริการไปแล้ว (Churn) คิดเป็น 26.58% หรือ 1,869 คนจากลูกค้าทั้งหมดและลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) คิดเป็น 73.42% หรือ 5,163 จากลูกค้าทั้งหมด ซึ่งเห็นได้ว่าสัดส่วนของตัวแปรเป้าหมายทั้ง 2 คลาสแตกต่างกันค่อนข้างมาก ซึ่งถือว่าเป็นข้อมูลที่ไม่สมดุลกัน (Class Imbalance) จึงจำเป็นต้องจัดการกับความไม่สมดุลนี้ก่อนจะนำไปสร้างโมเดลทำนาย

ภาพที่ 3.3 กราฟฮิสโตแกรมแสดงการกระจายตัวของระยะเวลาที่ใช้บริการตามสถานะการใช้บริการของลูกค้า



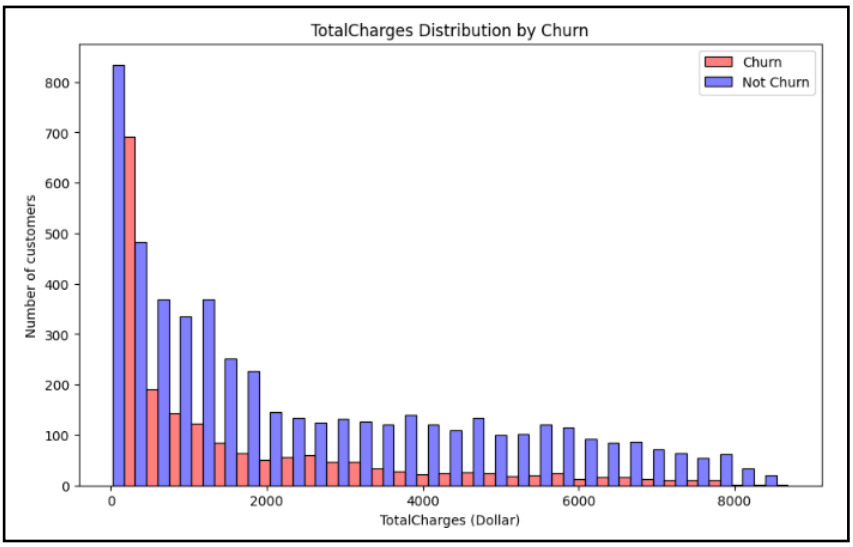
จากกราฟดังกล่าวจะแสดงการกระจายตัวของข้อมูลระยะเวลาที่ใช้บริการเมื่อเปรียบเทียบกับสถานะการใช้บริการของลูกค้า จะเห็นได้ว่าข้อมูลมีการแจกแจงแบบไม่ปกติ โดยพบว่ามีจำนวนลูกค้าที่เลิกใช้บริการ (Churn) มากที่สุดในช่วง 0-10 เดือนแรกของการใช้บริการและมีจำนวนลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) กระจายตัวอยู่ในทุกๆช่วงมากกว่าลูกค้าที่เลิกใช้บริการ โดยมีจำนวนมากที่สุดในช่วง 60-70 เดือนของการใช้บริการ

ภาพที่ 3.4 กราฟฮิสโตรแกรมแสดงการกระจายตัวของค่าใช้จ่ายรายเดือนตามสถานะการใช้บริการของลูกค้า



จากกราฟดังกล่าวจะแสดงการกระจายตัวของข้อมูลค่าใช้จ่ายรายเดือนเมื่อเปรียบเทียบกับสถานะการใช้บริการของลูกค้า จะเห็นได้ว่าข้อมูลมีการแจกแจงแบบไม่ปกติ โดยพบว่ามีจำนวนลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) มากที่สุดในช่วงค่าบริการ 20 ดอลลาร์ต่อเดือน และจำนวนลูกค้าที่เลิกใช้บริการ (Churn) กระจายตัวอยู่มากในช่วงค่าบริการ 70-100 ดอลลาร์ต่อเดือน

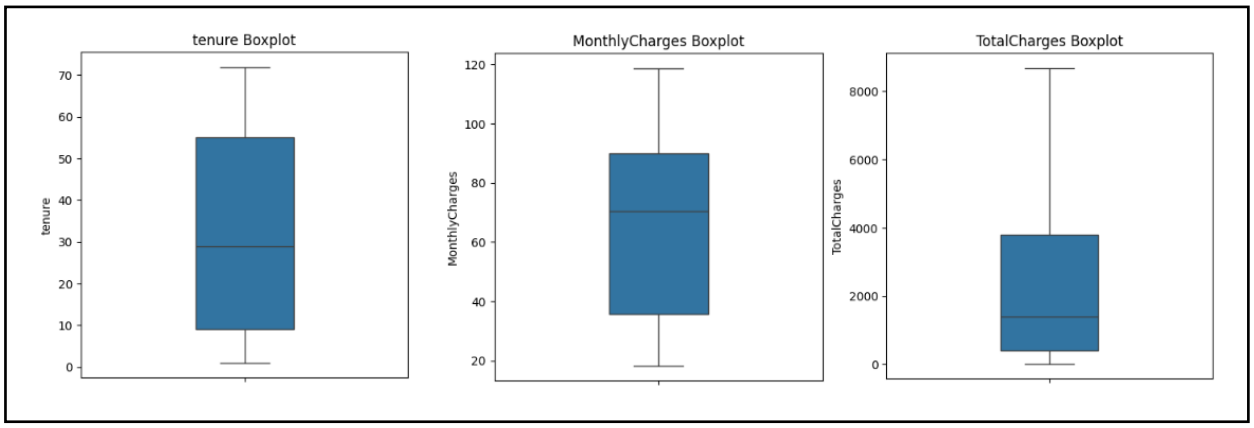
ภาพที่ 3.5 กราฟฮิสโตรแกรมแสดงการกระจายตัวของค่าใช้จ่ายทั้งหมดตามสถานะการใช้บริการของลูกค้า



จากกราฟดังกล่าวจะแสดงการกระจายตัวของข้อมูลค่าใช้จ่ายทั้งหมดเมื่อเปรียบเทียบกับสถานะการใช้บริการของลูกค้า จะเห็นได้ว่าข้อมูลมีการแจกแจงแบบไม่ปกติ โดยพบว่าทั้งจำนวนของลูกค้าที่เลิกใช้บริการ (Churn) และจำนวนลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) มีจำนวนมากในช่วงค่าใช้จ่าย 2,000 ดอลลาร์ของช่วงค่าใช้จ่ายทั้งหมด แต่จำนวนลูกค้าที่ยังไม่เลิกใช้บริการมีจำนวนมากกว่าจำนวนลูกค้าที่เลิกใช้บริการในช่วงค่าใช้จ่ายอื่น ในขณะที่จำนวนลูกค้าที่เลิกใช้บริการในช่วงค่าใช้จ่ายอื่นมีไม่ค่อยมากนัก

จากการพิจารณากราฟฮิสโตแกรมทั้ง 3 กราฟข้างต้นจะพบว่าข้อมูลในคุณลักษณะเชิงปริมาณทั้ง 3 คุณลักษณะมีการกระจายตัวแบบแจกแจงไม่ปกติและเมื่อตรวจสอบการกระจายตัวของข้อมูลในแต่ละคุณลักษณะพบว่าทั้ง 3 คุณลักษณะอาจมีนัยสำคัญต่อการตัดสินใจเลิกใช้บริการของลูกค้า

ภาพที่ 3.6 แผนภาพกล่องแสดงการกระจายตัวของข้อมูลของคุณลักษณะเชิงปริมาณทั้ง 3 คุณลักษณะ



จากการสร้างแผนภาพกล่องเพื่อตรวจสอบการกระจายตัวและตรวจสอบหาค่าผิดปกติในคุณลักษณะเชิงปริมาณได้แก่ ระยะเวลาที่ใช้บริการ, ค่าใช้จ่ายรายเดือนและค่าใช้จ่ายทั้งหมดของลูกค้า พบว่าข้อมูลในแต่ละคุณลักษณะไม่มีการกระจายตัวเกินขอบเขตของเส้นควอไทล์ที่ 1 (Q1) และเส้นควอไทล์ที่ 3 (Q3) เมื่อเทียบกับค่าพิสัยระหว่างควอไทล์ (IQR) ซึ่งสามารถบอกได้ว่าไม่มีข้อมูลที่จัดว่าเป็นค่าผิดปกติในแต่ละคุณลักษณะดังกล่าว

3.3.3 การแปลงข้อมูล (Data Transformation)

ทำการแปลงข้อมูลในคุณลักษณะเชิงหมวดหมู่จากข้อมูลเชิงหมวดหมู่ (Categorical Data) ให้เป็นข้อมูลเชิงคุณภาพ (Numerical Data) โดยใช้ One Hot Encoding ซึ่งเป็นวิธีที่แปลงข้อมูลเชิงหมวดหมู่ให้เป็นตัวเลขที่สามารถนำไปใช้ในการสร้างโมเดลทำนายได้

3.3.4 การทำให้เป็นมาตรฐานของข้อมูล (Data Standardization)

ปรับข้อมูลในคุณลักษณะเชิงปริมาณ ได้แก่ ระยะเวลาที่ใช้บริการ, ค่าบริการรายเดือน และค่าใช้จ่ายทั้งหมด โดยใช้ Standard Scaler ปรับข้อมูลทำให้มีค่าเฉลี่ยเท่ากับ 0 และปรับส่วนเบี่ยงเบนมาตราฐานให้เท่ากับ 1 เพื่อทำให้ข้อมูลอยู่ในมาตราส่วนเดียวกัน ซึ่งทำให้โมเดลสามารถคำนวณและทำงานได้อย่างมีประสิทธิภาพมากขึ้น

3.3.5 การจัดการความไม่สมดุลของข้อมูล (Handling Imbalanced Data)

หลังจากการทำวิเคราะห์ข้อมูลเชิงสำรวจแล้ว จะเห็นได้ว่าสัดส่วนของข้อมูลทั้ง 2 คลาสมีสัดส่วนที่แตกต่างกันค่อนข้างมาก ซึ่งถือได้ว่าชุดข้อมูลนี้เป็นชุดข้อมูลที่มีความไม่สมดุล (Class Imbalance) การนำชุดข้อมูลนี้ไปใช้ในการสร้างโมเดลทำนายโดยไม่มีการจัดการกับความไม่สมดุลของข้อมูล อาจส่งผลให้โมเดลมีอคติต่อคลาสที่มีจำนวนข้อมูลมากกว่า ซึ่งทำให้โมเดลไม่สามารถทำนายกลุ่มลูกค้าที่เลิกใช้บริการได้อย่างมีประสิทธิภาพและไม่เหมาะสมต่อการนำไปใช้งานจริง ดังนั้นจึงจำเป็นต้องมีการจัดการกับความไม่สมดุลของข้อมูลเสียก่อน โดยใช้วิธีการลดตัวอย่างข้อมูลแบบสุ่ม (Random Undersampling), วิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม (Random Oversampling) และการสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling) หรือ SMOTE มาใช้ในการจัดการชุดข้อมูลให้มีความสมดุลมากยิ่งขึ้น ทำให้มีชุดข้อมูลสำหรับการสร้างโมเดลทั้งหมด 4 ชุดคือ ชุดข้อมูลดั้งเดิม, ชุดข้อมูลที่ผ่านการลดตัวอย่างข้อมูลแบบสุ่ม, ชุดข้อมูลที่ผ่านการเพิ่มตัวอย่างข้อมูลแบบสุ่มและชุดข้อมูลที่ผ่านการสังเคราะห์ข้อมูลเพิ่ม

ตารางที่ 3.3 ตารางแสดงจำนวนข้อมูลก่อนและหลังทำการจัดการกับความไม่สมดุลของข้อมูล

เทคนิคที่ใช้	จำนวนข้อมูลก่อนทำการปรับสมดุล		จำนวนข้อมูลหลังทำการปรับสมดุล	
	จำนวนคลาส 0 (not churn)	จำนวนคลาส 1 (churn)	จำนวนคลาส 0 (not churn)	จำนวนคลาส 1 (churn)
ข้อมูลดั้งเดิม	5,163	1,869	5,163	1,869
Random Undersampling	5,163	1,869	1,869	1,869
Random Oversampling	5,163	1,869	5,163	5,163
SMOTE	5,163	1,869	5,163	5,163

3.4 การแบ่งชุดข้อมูลสำหรับฝึกและทดสอบ (Data Splitting)

ในการสร้างโมเดลเพื่อแบ่งประเภทเพื่อทำนายการสูญเสียลูกค้า ผู้วิจัยจะแบ่งชุดข้อมูลทั้ง 4 ชุด โดยแบ่งแต่ละชุดออกเป็น 2 ส่วนคือส่วนข้อมูลสำหรับการฝึกโมเดล (Train set) คิดเป็นสัดส่วน 70% เพื่อใช้สำหรับการเรียนรู้ของโมเดลและส่วนข้อมูลสำหรับการทดสอบความถูกต้องในการทำนายของโมเดล (Test set) คิดเป็นสัดส่วน 30% เพื่อใช้ประเมินประสิทธิภาพ

การทำนายของโมเดล ซึ่งมีจำนวนข้อมูลสำหรับการฝึกและทดสอบของแต่ละชุดข้อมูลดังตารางต่อไปนี้

ตารางที่ 3.4 ตารางแสดงจำนวนข้อมูลของชุดข้อมุลสำหรับการฝึก (Train set)

ชุดข้อมูลที่ใช้	จำนวนคลาส 0 (not churn)	จำนวนคลาส 1 (churn)	จำนวนข้อมูลทั้งหมด
ข้อมูลดั้งเดิม	3,608	1,314	4,922
Random Undersampling	1,308	1,308	2,616
Random Oversampling	3,621	3,607	7,228
SMOTE	3,621	3,607	7,228

ตารางที่ 3.5 ตารางแสดงจำนวนข้อมูลของชุดข้อมูลสำหรับการทดสอบ (Test set)

ชุดข้อมูลที่ใช้	จำนวนคลาส 0 (not churn)	จำนวนคลาส 1 (churn)	จำนวนข้อมูลทั้งหมด
ข้อมูลดั้งเดิม	1,555	555	2,110
Random Undersampling	561	561	1,122
Random Oversampling	1,556	1,542	3,098
SMOTE	1,556	1,542	3,098

3.5 การสร้างโมเดล (Data Modelling)

จากการทบทวนวรรณกรรมจะเห็นได้ว่าอัลกอริทึมที่ใช้ในงานวิจัยเกี่ยวกับการสร้างโมเดลทำนายเพื่อทำนายการสูญเสียลูกค้าแล้วได้ผลลัพธ์ที่ดีคือเอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost) และการสุ่มป่าไม้ (Random Forest) ดังนั้นทางผู้วิจัยจึงได้เลือกอัลกอริทึมทั้ง 2 อัลกอริทึมดังกล่าวมาใช้ในโครงงานนี้และจากปัญหาในโครงงานนี้คือการทำนายการสูญเสีย

ลูกค้า ซึ่งเป็นปัญหาการจำแนกประเภทแบบสองกลุ่ม (Binary Classification Problem) โดยมีผลลัพธ์การทำนายว่าลูกค้าจะเลิกใช้บริการ (Churn) หรือยังคงใช้บริการต่อ (Not Churn) ทางผู้วิจัยจึงได้เลือกซัพพอร์ตเวกเตอร์แมชชีน (SVM) มาใช้ในโครงงานนี้ร่วมด้วย เนื่องจากเป็นอัลกอริทึมที่เน้นสร้างเส้นแบ่ง (Hyperplane) เพื่อจำแนกประเภทของข้อมูล โดยอัลกอริทึมจะพยายามค้นหาเส้นแบ่งที่มีระยะห่าง (Margin) ที่กว้างมากที่สุดเพื่อลดความผิดพลาดในการจำแนกประเภทซึ่งเหมาะสมกับลักษณะของปัญหาของโครงงานนี้และนำผลลัพธ์การทำนายของอัลกอริทึมทั้งหมดที่กล่าวมามาเปรียบเทียบประสิทธิภาพในการแบ่งประเภทของลูกค้า โดยค่าพารามิเตอร์ต่างๆจะใช้ค่าเริ่มต้นของอัลกอริทึมนั้น

3.6 การประเมินประสิทธิภาพของโมเดล (Model Evaluation)

หลังจากที่ได้สร้างโมเดลทำนายแล้ว ผู้วิจัยจะใช้ค่าในชุดข้อมูลสำหรับการทดสอบ (Test set) เปรียบเทียบกับค่าผลการทำนายของโมเดลทำนายและวัดประสิทธิภาพของโมเดลผ่านค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึกได้ (Recall) ประกอบกับการใช้ Confusion Matrix เพื่อแสดงจำนวนผลลัพธ์การทำนายแต่ละประเภท (True Positive, False Positive, True Negative, False Negative) และใช้กราฟ ROC-AUC เพื่อแสดงความสัมพันธ์ระหว่างค่า False Positive Rate และค่า True Positive Rate พร้อมทั้งใช้ค่าพื้นที่ใต้กราฟ (Area Under Curve:AUC) เพื่อประเมินความสามารถในการแยกประเภทคลาสของโมเดลทำนาย

3.7 การหาคุณลักษณะที่สำคัญ (Feature Importance)

หลังจากที่ได้โมเดลทำนายที่มีประสิทธิภาพในการทำนายการสูญเสียลูกค้าแล้ว ผู้วิจัยจะวิเคราะห์หาคุณลักษณะที่เป็นปัจจัยสำคัญที่ส่งผลต่อการทำนายของโมเดลโดยพิจารณาจากค่าคุณลักษณะที่สำคัญ (Feature Importance) จากโมเดลทำนายผล ซึ่งจะช่วยในการหาสาเหตุที่ลูกค้าตัดสินใจเลิกใช้บริการและสามารถนำไปปรับใช้เพื่อลดการสูญเสียลูกค้าในอนาคตได้

บทที่ 4
ผลการดำเนินงาน

เนื้อหาในบทนี้จะนำเสนอผลลัพธ์ของการวิจัยโดยแบ่งออกเป็น 4 ส่วน ได้แก่ ส่วนที่ 1 ผลลัพธ์การทำนายของโมเดลในการทำนายการสูญเสียลูกค้า ส่วนที่ 2 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) ส่วนที่ 3 กราฟ ROC-AUC เพื่อแสดงความสามารถในการจำแนกประเภทของโมเดลและส่วนที่ 4 ผลลัพธ์ของการหาคุณลักษณะที่สำคัญของโมเดลทำนาย

ส่วนที่ 1 ผลลัพธ์การทำนายของโมเดลในการทำนายการสูญเสียลูกค้าของทั้ง 3 อัลกอริทึมที่ผ่านการจัดการความไม่สมดุลของข้อมูลด้วยวิธีต่างๆซึ่งประกอบด้วยค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision) และค่าความระลึกได้ (Recall)

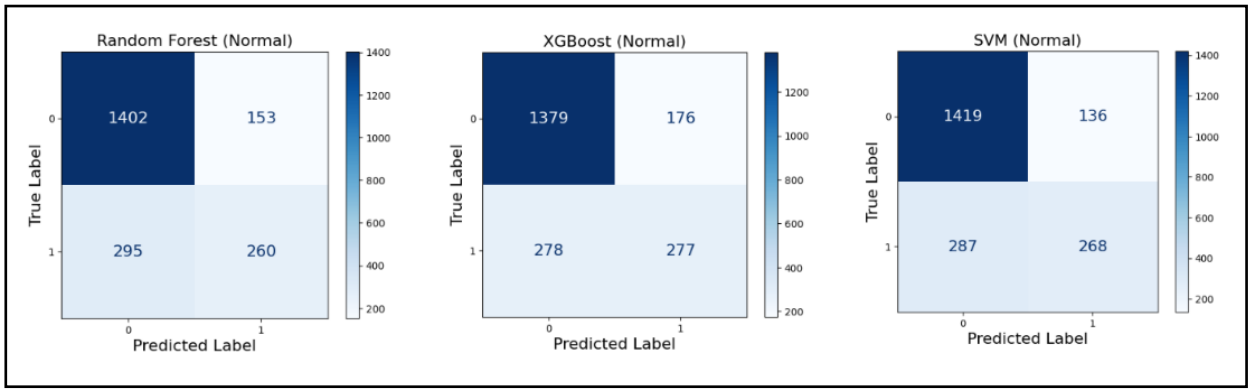
ตารางที่ 4.1 ตารางผลลัพธ์การทำนายของโมเดลในการทำนายการสูญเสียลูกค้า

อัลกอริทึมที่ใช้	ข้อมูลดั้งเดิม			Random Undersampling			Random Oversampling			SMOTE		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Random Forest	78.77%	62.85%	46.85%	72.73%	72.73%	72.73%	88.77%	84.83%	94.54%	85.70%	85.42%	86.25%
XGBoost	78.48%	61.15%	49.91%	73.08%	71.99%	75.58%	85.22%	81.37%	91.52%	85.83%	85.55%	86.38%
SVM	79.95%	66.34%	48.29%	74.33%	72.41%	78.61%	78.15%	76.46%	81.62%	81.18%	78.77%	85.60%

จากตารางดังกล่าวจะเห็นได้ว่าเมื่อพิจารณาจากค่าความถูกต้องแล้ว อัลกอริทึมที่ให้ความถูกต้องสูงที่สุดคือ Random Forest โดยใช้วิธีการเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่มในการจัดการกับความไม่สมดุลของข้อมูล โดยมีค่าความถูกต้องที่ 88.77% และรองลงมาคือ XGBoost โดยใช้วิธีสังเคราะห์ข้อมูลเพิ่ม ซึ่งที่ได้ค่าความถูกต้องที่ 85.83%

ส่วนที่ 2 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) เพื่อตรวจสอบจำนวนผลลัพธ์การทำนายแต่ละประเภทของโมเดลในการทำนายการสูญเสียลูกค้าของทั้ง 3 อัลกอริทึมที่ผ่านการจัดการความไม่สมดุลของข้อมูลด้วยวิธีต่างๆ

ภาพที่ 4.1 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) ของ Random Forest, XGBoost และ SVM ที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล



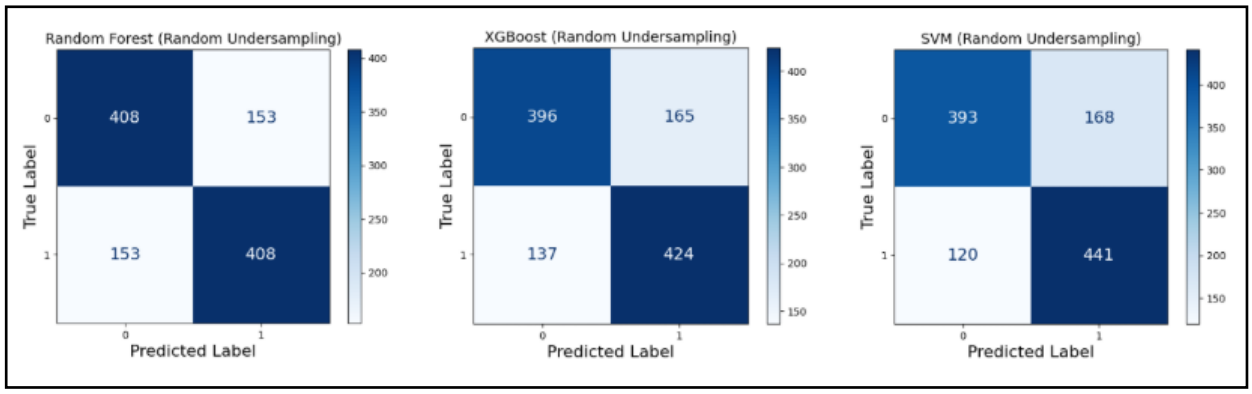
จากตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) นี้ประกอบกับตารางที่ 3.5 จะเห็นได้ว่าโมเดลสามารถทำนายข้อมูลได้อย่างถูกต้องดังนี้

1) Random Forest มีจำนวนลูกค้าที่เลิกใช้บริการ 260 คนและมีลูกค้าที่ยังไม่เลิกใช้บริการ 1,402 คน ซึ่งคิดเป็น 12.32% และ 66.45% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
2) XGBoost มีจำนวนลูกค้าที่เลิกใช้บริการ 277 คน และลูกค้าที่ยังไม่เลิกใช้บริการ 1,379 คน ซึ่งคิดเป็น 13.12% และ 65.36% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
3) SVM มีจำนวนลูกค้าที่เลิกใช้บริการ 268 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 1,419 คน ซึ่งคิดเป็น 12.70% และ 67.25% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ

ประกอบกับผลลัพธ์การทำนายของโมเดลในตารางที่ 4.1 จะเห็นได้ว่าถึงแม้ว่าทั้ง 3 อัลกอริทึมจะสามารถทำนายได้ค่าความถูกต้องได้ค่อนข้างดี ได้แก่ 78.77%, 78.48% และ 79.95% สำหรับ Random Forest, XGBoost และ SVM ตามลำดับ แต่จะเห็นได้ว่าโมเดลทำนายในฝั่งของคลาสที่ลูกค้ายังไม่เลิกใช้บริการ (Not Churn) หรือฝั่ง True Negative ได้ถูกต้องเป็นจำนวนมากเมื่อเทียบกับฝั่งของคลาสที่ลูกค้าเลิกใช้บริการ (Churn) หรือฝั่ง True Positive ซึ่งเป็นคลาสที่เราให้ความสนใจ และจากการพิจารณาค่าความแม่นยำ (Precision) และค่าความระลึกได้ (Recall) ของแต่ละอัลกอริทึมในตารางที่ 4.1 ประกอบด้วย จะพบว่า Random Forest มีค่าความแม่นยำที่ 62.85% และค่าความระลึกได้ที่ 46.85%, XGBoost มีค่าความแม่นยำที่ 61.15% และค่าความระลึกได้ที่ 49.91% และ SVM มีค่าความแม่นยำที่ 66.34% และค่าความระลึกได้

48.29% ซึ่งจะเห็นได้ว่าอัลกอริทึมทั้งสามดังกล่าวมีความโน้มเอียงในการทำนายผลจริง ซึ่งไม่เหมาะสมต่อการนำไปใช้งานจริง

ภาพที่ 4.2 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) ของ Random Forest, XGBoost และ SVM ที่ใช้วิธีลดตัวอย่างข้อมูลแบบสุ่ม



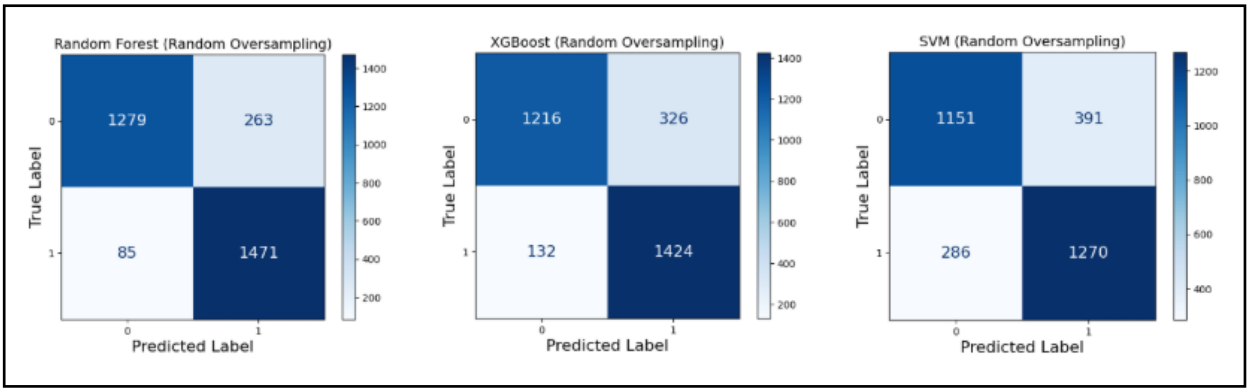
จากตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) นี้ประกอบกับตารางที่ 3.5 จะเห็นได้ว่าโมเดลสามารถทำนายข้อมูลได้อย่างถูกต้องดังนี้

1) Random Forest มีจำนวนลูกค้าที่เลิกใช้บริการ 408 คนและมีลูกค้าที่ยังไม่เลิกใช้บริการ 408 คนเท่ากัน ซึ่งคิดเป็น 36.36% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบ
2) XGBoost มีจำนวนลูกค้าที่เลิกใช้บริการ 424 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 396 คน ซึ่งคิดเป็น 37.79% และ 35.29% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
3) SVM มีจำนวนลูกค้าที่เลิกใช้บริการ 441 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 393 คน ซึ่งคิดเป็น 39.30% และ 35.03% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ

เมื่อพิจารณาผลการทำนายของทั้ง 3 อัลกอริทึมจะเห็นได้ว่าโมเดลสามารถทำนายและแยกแยะคลาสของข้อมูลได้ดีกว่าเมื่อเทียบกับผลการทำนายของโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล เนื่องจากมีจำนวนผลการทำนายของลูกค้าที่เลิกใช้บริการ (Churn) และจำนวนผลการทำนายของลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) ใกล้เคียงกัน เมื่อพิจารณาจากตาราง 4.1 ร่วมด้วยจะเห็นได้ว่าค่าความแม่นยำ (Precision) และค่าความระลึกได้ (Recall) ที่ได้จะมากกว่าผลการทำนายของโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล แต่ได้ค่าความถูกต้อง (Accuracy) น้อยกว่าได้แก่ 72.73%, 73.08% และ 74.33% สำหรับ Random Forest, XGBoost และ SVM ตามลำดับ

จากที่กล่าวมาทำให้สามารถสรุปได้ว่าถึงแม้ว่าโมเดลจะมีการจัดการกับความไม่สมดุลของข้อมูลด้วยวิธีการลดตัวอย่างข้อมูลแบบสุ่มจะช่วยให้โมเดลสามารถทำนายและแยกแยะคลาสของลูกค้าได้ดีมากขึ้น แต่ก็ส่งผลให้ค่าความถูกต้องน้อยลงเพราะจำนวนข้อมูลที่ใช้ในการฝึกและทดสอบมีจำนวนลดน้อยลงแล้วทำให้โมเดลสูญเสียข้อมูลบางส่วนที่อาจจะสำคัญต่อการแยกแยะคลาสของลูกค้าไป

ภาพที่ 4.3 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) ของ Random Forest, XGBoost และ SVM ที่ใช้วิธีเพิ่มตัวอย่างข้อมูลแบบสุ่ม



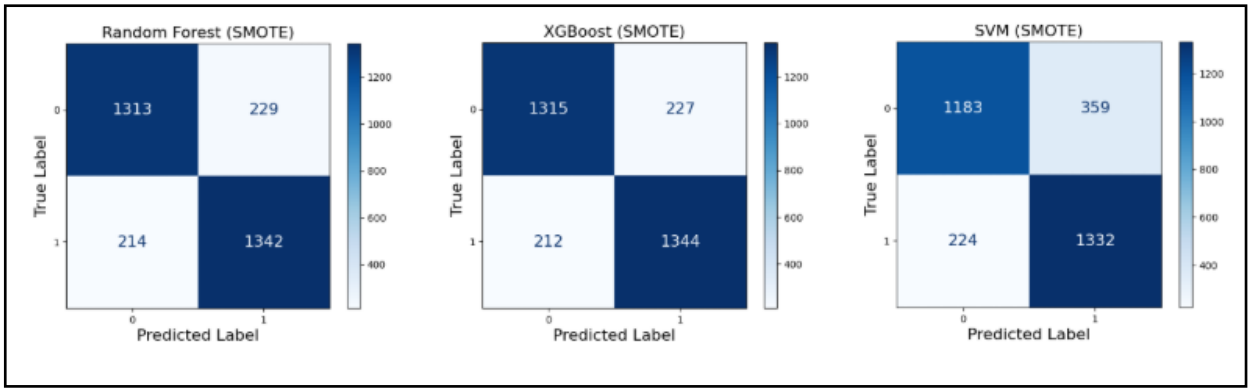
จากตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) นี้ประกอบกับตารางที่ 3.5 จะเห็นได้ว่าอัลกอริทึมสามารถทำนายข้อมูลได้อย่างถูกต้องดังนี้

1) Random Forest มีจำนวนลูกค้าที่เลิกใช้บริการ 1,471 คนและมีลูกค้าที่ยังไม่เลิกใช้บริการ 1,279 คนเท่ากัน ซึ่งคิดเป็น 47.48% และ 41.28% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
2) XGBoost มีจำนวนลูกค้าที่เลิกใช้บริการ 1,424 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 1,216 คน ซึ่งคิดเป็น 45.97% และ 39.25% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
3) SVM มีจำนวนลูกค้าที่เลิกใช้บริการ 1,270 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 1,151 คน ซึ่งคิดเป็น 40.99% และ 37.15% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ

เมื่อพิจารณาผลการทำนายของทั้ง 3 อัลกอริทึมจะเห็นได้ว่าโมเดลสามารถทำนายและแยกแยะคลาสของข้อมูลได้ดีกว่าเมื่อเทียบกับผลการทำนายของโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล เนื่องจากมีจำนวนผลการทำนายของลูกค้าที่เลิกใช้บริการ (Churn) และจำนวน

ผลการทำนายของลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) ใกล้เคียงกัน เมื่อพิจารณาจากตาราง 4.1 ร่วมด้วยจะเห็นว่าได้ค่าความถูกต้อง (Accuracy) ได้แก่ 88.77%, 85.22% และ 78.15% สำหรับ Random Forest, XGBoost และ SVM ตามลำดับ และได้ค่าความแม่นยำ (Precision) กับค่าความระลึกได้ (Recall) สูงกว่าโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูลและโมเดลที่ใช้วิธีการลดจำนวนตัวอย่างข้อมูลแบบสุ่ม

ภาพที่ 4.4 ตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) ของ Random Forest, XGBoost และ SVM ที่ใช้วิธีสังเคราะห์ข้อมูลเพิ่ม



จากตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) นี้ประกอบกับตารางที่ 3.5 จะเห็นได้ว่าอัลกอริทึมสามารถทำนายข้อมูลได้อย่างถูกต้องดังนี้

1) Random Forest มีจำนวนลูกค้าที่เลิกใช้บริการ 1,342 คนและมีลูกค้าที่ยังไม่เลิกใช้บริการ 1,313 คนเท่ากัน ซึ่งคิดเป็น 43.32% และ 42.38% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
2) XGBoost มีจำนวนลูกค้าที่เลิกใช้บริการ 1,344 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 1,315 คน ซึ่งคิดเป็น 43.38% และ 42.45% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ
3) SVM มีจำนวนลูกค้าที่เลิกใช้บริการ 1,332 คนและลูกค้าที่ยังไม่เลิกใช้บริการ 1,183 คน ซึ่งคิดเป็น 43% และ 38.19% ของจำนวนลูกค้าทั้งหมดในชุดข้อมูลสำหรับการทดสอบตามลำดับ

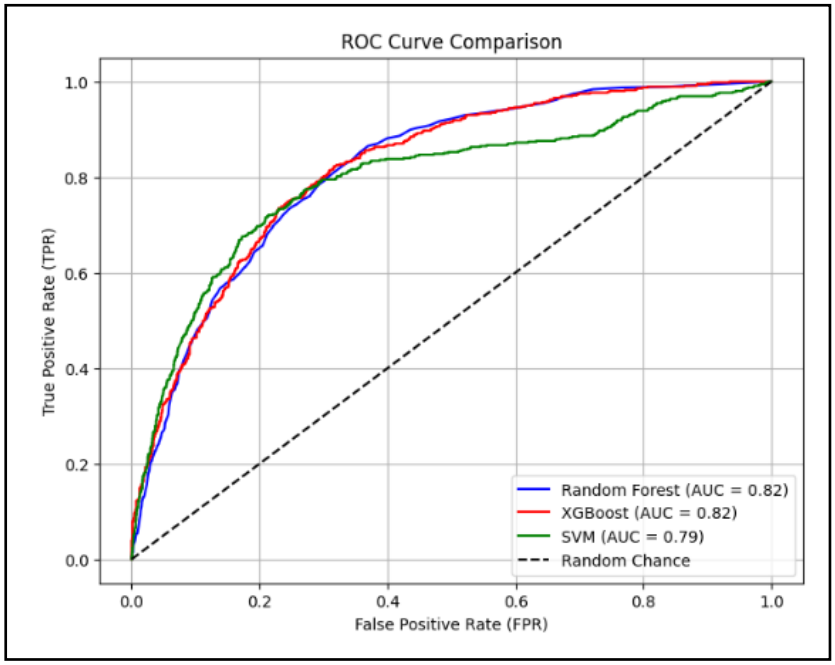
เมื่อพิจารณาผลการทำนายของทั้ง 3 อัลกอริทึมจะเห็นได้ว่าโมเดลสามารถทำนายและแยกแยะคลาสของข้อมูลได้ดีกว่าเมื่อเทียบกับผลการทำนายของโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล เนื่องจากมีจำนวนผลการทำนายของลูกค้าที่เลิกใช้บริการ (Churn) และจำนวน

ผลการทำนายของลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) ใกล้เคียงกัน เมื่อพิจารณาจากตาราง 4.1 ร่วมด้วยจะเห็นว่าได้ค่าความถูกต้อง (Accuracy) ได้แก่ 85.70%, 85.83% และ 81.18% สำหรับ Random Forest, XGBoost และ SVM ตามลำดับ และได้ค่าความแม่นยำ (Precision) กับค่าความระลึกได้ (Recall) สูงกว่าโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูลและโมเดลที่ใช้วิธีการลดจำนวนตัวอย่างข้อมูลแบบสุ่ม ซึ่งมีค่าผลลัพธ์ต่างๆใกล้เคียงกันกับโมเดลที่ใช้วิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม แต่จากตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) จะเห็นได้ว่าโมเดลที่ใช้วิธีสังเคราะห์ข้อมูลเพิ่มทำนายได้ค่า FP และ FN มากกว่าโมเดลที่ใช้วิธีเพิ่มตัวอย่างข้อมูลแบบสุ่มซึ่งแสดงให้เห็นว่าโมเดลที่ใช้วิธีสังเคราะห์ข้อมูลทำนายประเภทของลูกค้าผิดมากกว่าโมเดลที่ใช้วิธีเพิ่มตัวอย่างข้อมูลแบบสุ่มเล็กน้อย

ส่วนที่ 3 กราฟ ROC-AUC เพื่อแสดงความสามารถในการจำแนกประเภทของโมเดลซึ่งกราฟ ROC จะแสดงความสัมพันธ์ระหว่าง True Positive Rate (TPR) กับ False Positive Rate (FPR) และค่าพื้นที่ใต้กราฟ (AUC) จะแสดงความสามารถในการจำแนกโดยรวมของโมเดล โดยจะใช้กราฟนี้เพื่อเปรียบเทียบความสามารถในการจำแนกประเภทของแต่ละอัลกอริทึมโดยแบ่งออกเป็น 4 กราฟตามลักษณะในการจัดการความไม่สมดุลของข้อมูลดังนี้

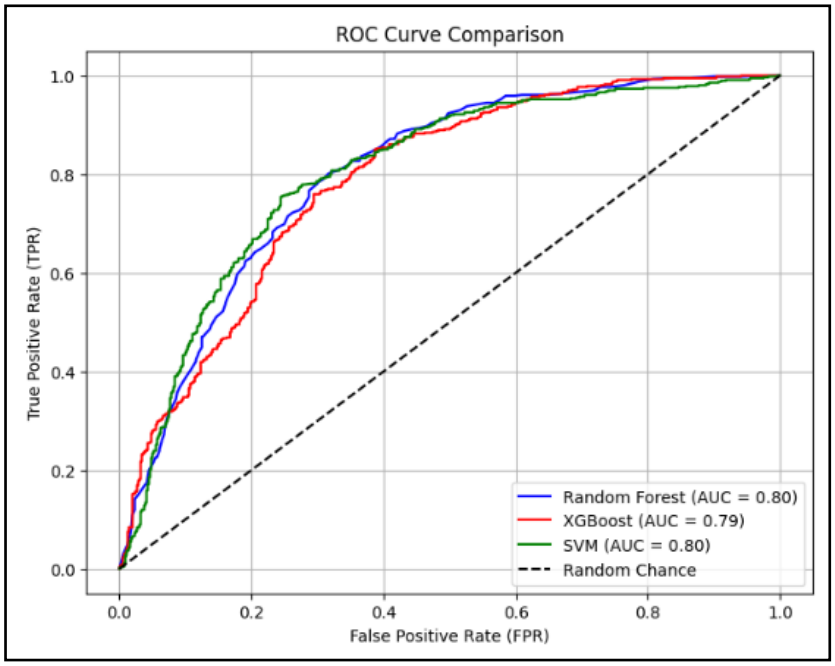
1) กราฟ ROC-AUC เพื่อเปรียบเทียบอัลกอริทึมที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล
2) กราฟ ROC-AUC เพื่อเปรียบเทียบอัลกอริทึมที่ใช้วิธีการลดตัวอย่างข้อมูลแบบสุ่ม (Random Undersampling)
3) กราฟ ROC-AUC เพื่อเปรียบเทียบอัลกอริทึมที่ใช้วิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม (Random Oversampling)
4) กราฟ ROC-AUC เพื่อเปรียบเทียบอัลกอริทึมที่ใช้วิธีการสังเคราะห์ข้อมูลเพิ่ม (SMOTE)

ภาพที่ 4.5 กราฟ ROC-AUC ของ Random Forest, XGBoost และ SVM ที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล



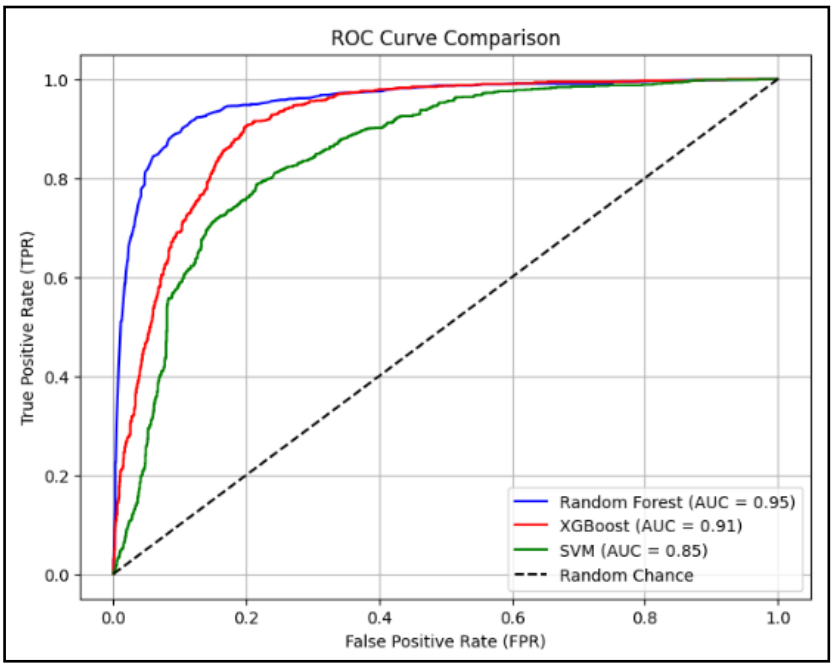
กราฟ ROC-AUC นี้จะเปรียบเทียบความสามารถในการจำแนกประเภทของโมเดลที่ไม่มีการจัดการกับความไม่สมดุลของข้อมูล โดยเส้นสีน้ำเงินแทน Random Forest, เส้นสีแดงแทน XGBoost และเส้นสีเขียวแทน SVM โดยมีเส้นประตรงกลางเป็นเส้นแบ่งเพื่อวัดประสิทธิภาพการทำนายของโมเดลซึ่งมีค่าพื้นที่ใต้กราฟ (AUC) เท่ากับ 0.5 ถ้าเส้นโค้งของโมเดลอยู่ที่เดียวกันกับเส้นประหรืออยู่ต่ำกว่าแสดงว่าโมเดลนั้นมีประสิทธิภาพในการทำนายแย่ไม่ต่างจากการสุ่มเดา ซึ่งจากกราฟนี้ได้เห็นได้ว่า Random Forest มีค่าพื้นที่ใต้กราฟเท่ากับ 0.82, XGBoost มีค่าพื้นที่ใต้กราฟเท่ากับ 0.82 และ SVM มีค่าพื้นที่ใต้กราฟเท่ากับ 0.79 ซึ่งจะเห็นได้ว่า Random Forest และ XGBoost มีค่าพื้นที่ใต้กราฟสูงสุดเท่ากัน

ภาพที่ 4.6 กราฟ ROC-AUC ของ Random Forest, XGBoost และ SVM ที่ใช้วิธีการลดตัวอย่างข้อมูลแบบสุ่ม



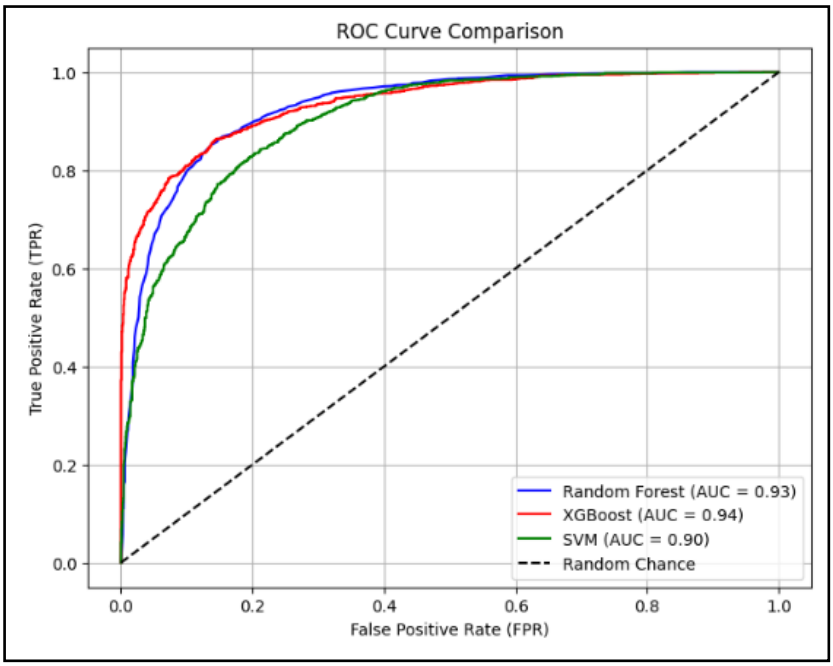
กราฟ ROC-AUC นี้จะเปรียบเทียบความสามารถในการจำแนกประเภทของโมเดลที่ใช้วิธีการลดตัวอย่างข้อมูลแบบสุ่ม โดยเส้นสีน้ำเงินแทน Random Forest, เส้นสีแดงแทน XGBoost และเส้นสีเขียวแทน SVM โดยมีเส้นประตรงกลางเป็นเส้นแบ่งเพื่อวัดประสิทธิภาพการทำนายของโมเดลซึ่งมีค่าพื้นที่ใต้กราฟ (AUC) เท่ากับ 0.5 ถ้าเส้นโค้งของโมเดลอยู่ที่เดียวกันกับเส้นประหรืออยู่ต่ำกว่าแสดงว่าโมเดลนั้นมีประสิทธิภาพในการทำนายแย่ไม่ต่างจากการสุ่มเดา ซึ่งจากกราฟนี้ได้เห็นได้ว่า Random Forest มีค่าพื้นที่ใต้กราฟเท่ากับ 0.80, XGBoost มีค่าพื้นที่ใต้กราฟเท่ากับ 0.79 และ SVM มีค่าพื้นที่ใต้กราฟเท่ากับ 0.80 ซึ่งจะเห็นได้ว่า Random Forest และ SVM มีค่าพื้นที่ใต้กราฟสูงสุดเท่ากัน

ภาพที่ 4.7 กราฟ ROC-AUC ของ Random Forest, XGBoost และ SVM ที่ใช้วิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม



กราฟ ROC-AUC นี้จะเปรียบเทียบความสามารถในการจำแนกประเภทของโมเดลที่ใช้วิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม โดยเส้นสีน้ำเงินแทน Random Forest, เส้นสีแดงแทน XGBoost และเส้นสีเขียวแทน SVM โดยมีเส้นประตรงกลางเป็นเส้นแบ่งเพื่อวัดประสิทธิภาพการทำนายของโมเดลซึ่งมีค่าพื้นที่ใต้กราฟ (AUC) เท่ากับ 0.5 ถ้าเส้นโค้งของโมเดลอยู่ที่เดียวกันกับเส้นประหรืออยู่ต่ำกว่าแสดงว่าโมเดลนั้นมีประสิทธิภาพในการทำนายแย่ไม่ต่างจากการสุ่มเดา ซึ่งจากกราฟนี้ได้เห็นได้ว่า Random Forest มีค่าพื้นที่ใต้กราฟเท่ากับ 0.95, XGBoost มีค่าพื้นที่ใต้กราฟเท่ากับ 0.91 และ SVM มีค่าพื้นที่ใต้กราฟเท่ากับ 0.85 ซึ่งจะเห็นได้ว่า Random Forest มีค่าพื้นที่ใต้กราฟสูงที่สุด

ภาพที่ 4.8 กราฟ ROC-AUC ของ Random Forest, XGBoost และ SVM ที่ใช้วิธีการสังเคราะห์ข้อมูลเพิ่ม

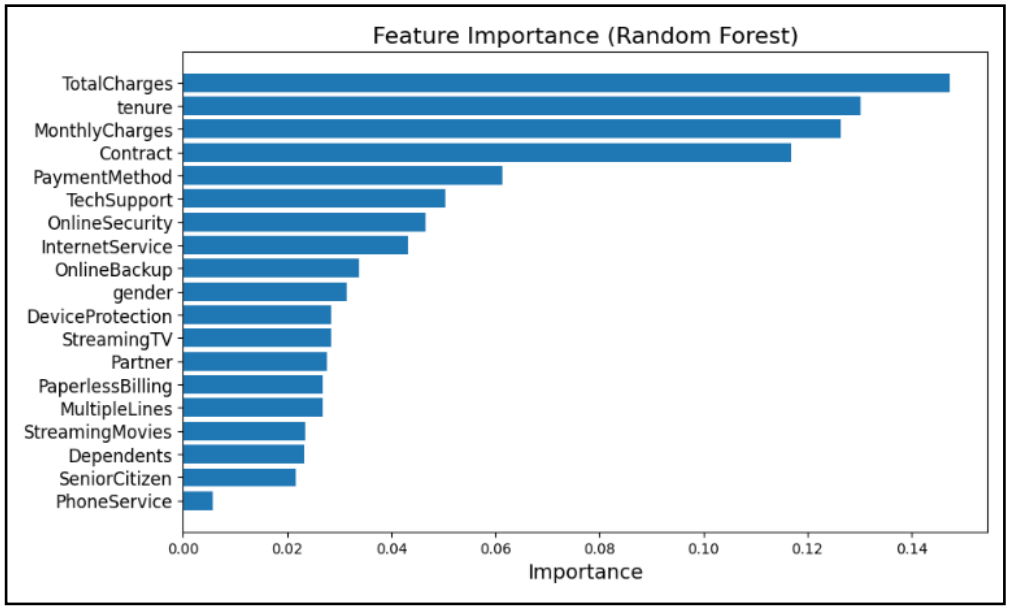


กราฟ ROC-AUC นี้จะเปรียบเทียบความสามารถในการจำแนกประเภทของโมเดลที่ใช้วิธีการสังเคราะห์ข้อมูลเพิ่ม โดยเส้นสีน้ำเงินแทน Random Forest, เส้นสีแดงแทน XGBoost และเส้นสีเขียวแทน SVM โดยมีเส้นประตรงกลางเป็นเส้นแบ่งเพื่อวัดประสิทธิภาพการทำนายของโมเดลซึ่งมีค่าพื้นที่ใต้กราฟ (AUC) เท่ากับ 0.5 ถ้าเส้นโค้งของโมเดลอยู่ที่เดียวกันกับเส้นประหรืออยู่ต่ำกว่าแสดงว่าโมเดลนั้นมีประสิทธิภาพในการทำนายแย่ไม่ต่างจากการสุ่มเดา ซึ่งจากกราฟนี้ได้เห็นได้ว่า Random Forest มีค่าพื้นที่ใต้กราฟเท่ากับ 0.93, XGBoost มีค่าพื้นที่ใต้กราฟเท่ากับ 0.94 และ SVM มีค่าพื้นที่ใต้กราฟเท่ากับ 0.90 ซึ่งจะเห็นได้ว่า XGBoost มีค่าพื้นที่ใต้กราฟสูงที่สุด

ซึ่งเมื่อสรุปจากค่าผลลัพธ์การทำนายทั้ง 3 ค่าของโมเดลในการทำนายการสูญเสียลูกค้าของแต่ละอัลกอริทึมในตารางที่ 4.1 ร่วมกับตารางแสดงผลลัพธ์การทำนาย (Confusion Matrix) และกราฟ ROC-AUC แล้ว อัลกอริทึมที่ให้ผลลัพธ์ที่ดีที่สุดคือ Random Forest ที่ใช้วิธีการจัดการกับความไม่สมดุลด้วยวิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม โดยได้ค่าความถูกต้อง (Accuracy) เท่ากับ 88.77% ค่าความแม่นยำ (Precision) เท่ากับ 84.83% ค่าความระลึกได้เท่ากับ 94.54% และค่าพื้นที่ใต้กราฟ (AUC) เท่ากับ 0.95

ส่วนที่ 4 การหาคุณลักษณะที่สำคัญ (Feature Importance) เพื่อทำความเข้าใจว่าคุณลักษณะใดที่สำคัญและส่งผลต่อการทำนายของโมเดลมากที่สุดและสามารถนำไปใช้ในการวางแผนทางการตลาดต่อได้ โดยได้ผลลัพธ์จากการหาคุณลักษณะที่สำคัญดังนี้

ภาพที่ 4.9 ตารางแสดงผลการหาคุณลักษณะที่สำคัญและส่งผลต่อการทำนายของโมเดล



จากภาพที่ 4.9 จะเห็นได้ว่าคุณลักษณะที่สำคัญและส่งผลต่อการทำนายของโมเดลมากที่สุดคือ TotalCharges หรือค่าใช้จ่ายทั้งหมดที่ลูกค้าจ่ายในการใช้บริการ และรองลงมาคือ tenure หรือระยะเวลาตั้งแต่ที่ลูกค้าใช้บริการมา ซึ่งสามารถสรุปได้ว่าจำนวนค่าใช้จ่ายทั้งหมดและระยะเวลาที่ใช้บริการของลูกค้าส่งผลต่อการเลิกใช้บริการของลูกค้าในกลุ่มธุรกิจโทรคมนาคมมากที่สุด

บทที่ 5
สรุป

5.1 สรุปผลการดำเนินงาน

การทำนายการสูญเสียลูกค้า (Customer Churn Prediction) เป็นสิ่งที่ช่วยให้ธุรกิจโทรคมนาคมสามารถรักษาฐานลูกค้าเก่าไว้ได้ เพราะทำให้ทราบถึงกลุ่มลูกค้าที่มีความเสี่ยงที่จะเลิกบริการล่วงหน้าและทำให้สามารถวางแผนทางการตลาดเพื่อป้องกันการสูญเสียลูกค้าได้ เช่น การวางแผนโปรโมชั่นหรือสิทธิประโยชน์ต่างๆสำหรับลูกค้าที่ใช้บริการ ดังนั้นการทำนายการสูญเสียลูกค้าจึงจำเป็นต้องมีความถูกต้องและแม่นยำเพื่อให้ธุรกิจสามารถรักษากลุ่มลูกค้าเก่าไว้ให้ได้มากที่สุด การที่จะทำเช่นนั้นได้จำเป็นต้องมีการสร้างโมเดลทำนายที่มีประสิทธิภาพและแม่นยำและต้องเลือกใช้อัลกอริทึมที่เหมาะสมและให้ผลลัพธ์การทำนายที่ดี ซึ่งในงานวิจัยต่างๆที่เกี่ยวข้องกับการสร้างโมเดลเพื่อทำนายการสูญเสียลูกค้านั้นได้มีการเลือกใช้อัลกอริทึมที่แตกต่างกันออกไป รวมไปถึงมีขั้นตอนในการจัดเตรียมข้อมูลที่แตกต่างกันตามลักษณะของข้อมูลที่เลือกใช้อีกด้วย โดยในโครงงานนี้ได้มีการศึกษาและเปรียบเทียบโมเดลทำนายการสูญเสียลูกค้าโดยใช้ชุดข้อมูลการสูญเสียลูกค้าในธุรกิจโทรคมนาคมจำนวน 7,043 คน โดยเริ่มจากการทำความสะอาดข้อมูลเพื่อให้ข้อมูลมีความครบถ้วนสมบูรณ์ทำให้เหลือจำนวนข้อมูลลูกค้า 7,032 คนและทำการวิเคราะห์ข้อมูลเชิงสำรวจซึ่งทำให้พบว่ามีจำนวนข้อมูลของลูกค้าที่เลิกใช้บริการคิดเป็น 26.58% หรือ 1,869 คนจากลูกค้าทั้งหมดและลูกค้าที่ยังไม่เลิกใช้บริการ (Not Churn) เป็น 73.42% คิดเป็นจำนวน 5,163 คนจากลูกค้าทั้งหมด ซึ่งถือว่ามีความไม่สมดุลของข้อมูลจึงมีการจัดการกับความไม่สมดุลของข้อมูลด้วยวิธีลดตัวอย่างข้อมูลแบบสุ่ม (Random Undersampling), วิธีการเพิ่มตัวอย่างข้อมูล (Random Oversampling) และการสังเคราะห์ข้อมูลเพิ่ม (SMOTE) เพื่อปรับสมดุลของข้อมูล จากนั้นได้ทำการสร้างโมเดลเพื่อทำนายการสูญเสียลูกค้าโดยอัลกอริทึมที่ใช้ได้แก่ การสุ่มป่าไม้ (Random Forest), เอ็กซ์ตรีมเกรเดียนต์บูสติ้ง (XGBoost) และซัพพอร์ตเวกเตอร์แมชชีน (SVM) ในการประเมินประสิทธิภาพในการทำนายของโมเดลได้ประเมินจากค่าความแม่นยำต่างๆประกอบกับ Confusion Matrix และกราฟ ROC-AUC โดยผลจากการทดลอง อัลกอริทึมที่ได้ผลลัพธ์ที่ดีที่สุดคือการสุ่มป่าไม้ (Random Forest) ที่ใช้วิธีการจัดการกับความไม่สมดุลด้วยวิธีการเพิ่มตัวอย่างข้อมูลแบบสุ่ม โดยได้ค่าความถูกต้อง (Accuracy) เท่ากับ 88.77% ค่าความแม่นยำ (Precision) เท่ากับ 84.83% ค่าความระลึกได้เท่ากับ 94.54% และค่าพื้นที่ใต้กราฟ (AUC) เท่ากับ 0.95 และจากการหาคุณลักษณะที่สำคัญพบว่าค่าใช้จ่ายทั้งหมดของลูกค้ามีความสำคัญและส่งผลต่อการจำแนกว่าลูกค้าจะเลิกใช้บริการหรือยังไม่เลิกใช้บริการมากที่สุด ตามด้วยระยะเวลาที่ลูกค้าใช้

บริการและค่าใช้จ่ายรายเดือนของลูกค้าตามลำดับ ซึ่งภาคธุรกิจสามารถนำผลลัพธ์ดังกล่าวไปประยุกต์ใช้ในการวางแผนทางธุรกิจเพื่อรักษาฐานลูกค้าเก่าให้มีประสิทธิภาพมากยิ่งขึ้นได้

5.2 ข้อเสนอแนะ

1) สำหรับโครงงานนี้ ในขั้นตอนการสร้างโมเดลสำหรับทำนายการสูญเสียลูกค้าได้ใช้ค่าพารามิเตอร์ต่าง ๆของแต่ละอัลกอริทึมเป็นค่าเริ่มต้นโดยที่ไม่ได้มีการปรับค่าพารามิเตอร์เพิ่มเติม หากต้องการปรับปรุงประสิทธิภาพของโมเดลทำนายให้ดีมากขึ้นสามารถปรับค่าพารามิเตอร์ด้วยวิธีต่าง ๆได้ เช่น Grid Search เพื่อหาค่าพารามิเตอร์ที่เหมาะสมสำหรับอัลกอริทึม

2) สำหรับโครงงานนี้มีสัดส่วนระหว่างจำนวนข้อมูลของทั้งสองคลาสค่อนข้างมากโดยคิดเป็น 26.58% และ 73.42% ตามลำดับ จึงมีการจัดการกับความไม่สมดุลของข้อมูล 3 วิธี ได้แก่ การลดตัวอย่างข้อมูลแบบสุ่ม (Random Undersampling), การเพิ่มตัวอย่างข้อมูลแบบสุ่ม (Random Oversampling) และการสังเคราะห์ข้อมูลเพิ่ม (SMOTE) ซึ่งจากตารางที่ 4.1 จะเห็นได้ว่าวิธีในการจัดการความไม่สมดุลของข้อมูลที่ส่งผลให้โมเดลสามารถทำนายการสูญเสียลูกค้าได้อย่างมีประสิทธิภาพคือ Random Oversampling และ SMOTE โดยมีค่าเฉลี่ยค่าความถูกต้อง (Accuracy) ของทั้งสามอัลกอริทึมคือ 84.05% และ 84.24% ตามลำดับซึ่งจะเห็นได้ว่าวิธี SMOTE ทำให้ได้ค่าเฉลี่ยค่าความถูกต้องของทั้งสามอัลกอริทึมมากกว่าเล็กน้อย แต่เมื่อพิจารณาที่ค่าเฉลี่ยของค่าความระลึกได้ (Recall) ของทั้งสามอัลกอริทึมที่ได้จากวิธี Random Oversampling และ SMOTE คิดเป็น 89.27% และ 86.08% ตามลำดับ จะเห็นได้ว่าวิธี Random Oversampling ทำให้ได้ค่าเฉลี่ยของค่าความระลึกได้มากกว่า ซึ่งค่า Recall เป็นค่าที่บ่งบอกว่าโมเดลทำนายค่า FN หรือทำนายลูกค้าที่เลิกใช้บริการ (Churn) ผิดว่ายังไม่เลิกใช้บริการ (Not Churn) ได้มากน้อยแค่ไหนจึงเป็นค่าที่สำคัญไม่แพ้กันกับค่า Accuracy สำหรับการรักษากลุ่มลูกค้าเดิมไว้สำหรับกลุ่มธุรกิจโทรคมนาคม จึงสามารถสรุปได้ว่าวิธีการจัดการกับความไม่สมดุลของข้อมูลที่เหมาะสมกับโครงงานนี้คือการเพิ่มจำนวนตัวอย่างข้อมูลแบบสุ่ม (Random Oversampling) แต่หากเป็นในโครงงานอื่นที่ชุดข้อมูลมีสัดส่วนระหว่างจำนวนข้อมูลของทั้งสองคลาสแตกต่างกันอย่างมากเช่นจำนวนข้อมูลของคลาสแรกคิดเป็น 5% ของจำนวนข้อมูลทั้งหมดและจำนวนข้อมูลของคลาสที่สองคิดเป็น 95% ของจำนวนข้อมูลทั้งหมด ซึ่งถ้าเป็นกรณีนี้วิธี Random Oversampling จะไม่เหมาะสมมากนักเพราะวิธีนี้จะสุ่มคัดลอกและเพิ่มข้อมูลจากฝั่งที่น้อยกว่าให้เท่ากับฝั่งที่มากกว่า ซึ่งจะทำให้ชุดข้อมูลที่ผ่านการปรับสมดุลมีความซ้ำซ้อนของข้อมูลเป็นจำนวนมากและทำให้โมเดลเกิด Overfitting และไม่เหมาะกับการนำไปใช้จริง วิธี

SMOTE จึงเหมาะสมกับการจัดการความไม่สมดุลของข้อมูลในกรณีนี้มากกว่าเนื่องจากเป็นการสร้างข้อมูลขึ้นมากใหม่ทำให้ช่วยเพิ่มความหลากหลายให้กับชุดข้อมูลและช่วยลดการเกิด Overfitting ได้มากกว่า Random Oversampling แต่ก็มีข้อเสียคือจำเป็นต้องใช้เวลาในการประมวลผลที่นานกว่า

รายการอ้างอิง

Chugh, V. (2024). AUC and the ROC Curve in Machine Learning. Datacamp. https://www.datacamp.com/tutorial/auc

Grus, J. (2022). เรียนรู้หลักการ Data Science ด้วย Python. สำนักพิมพ์คอร์ฟังก์ชั่น

Kumar, R., Sahithi, B., Neeharika, K., Shivaleela, M., Singh, D. & Reddy, K. (2023). Automation of Credit Card Customer Churn Analysis using Hybrid Machine Learning Models. Doi: 10.1051/e3sconf/202343001034

Miao, X. & Wang, H. (2022). Customer Churn Prediction on Credit Card Services using Random Forest Method. Doi: 10.2991/aebmr.k.220307.104

Mohajon, J. (2020). Confusion Matrix for Your Multi-Class Machine Learning Model. Medium. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Öztürk, M., Tunç, A. & Akay, F. (2023). Machine Learning based churn analysis for sellers on the e-commerce marketplace. Doi: 10.2478/ijmce-2023-0013

Taskin, N. (2023). Customer Churn Prediction Model in Telecommunication Sector Using Machine Learning Technique. Doi: 10.1016/j.rico.2023.100342