การวิเคราะห์การทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมโดยวิธีการเรียนรู้ของเครื่อง

ANALYSIS OF THE CHURN PREDICTION FOR TELECOM CUSTOMERS USING MACHINE LEARNING

อัญชิสา สิทธิวิริยะชัย

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

การวิเคราะห์การทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมโดยวิธีการเรียนรู้ของเครื่อง

อัญชิสา สิทธิวิริยะชัย

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566

ANALYSIS OF THE CHURN PREDICTION FOR TELECOM CUSTOMERS USING MACHINE LEARNING

ANCHISA SITTIVIRIYAHCAI

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2023

สารนิพนธ์

เรื่อง

การวิเคราะห์การทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมโดยวิธีการเรียนรู้ของเครื่อง

ของ

อัญชิสา สิทธิวิริยะชัย

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... กรรมการ

(อาจารย์ ดร.โสภณ มงคลลักษมี)

ชื่อเรื่อง	การวิเคราะห์การทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมโดยวิธีการเรียนรู้ของเครื่อง
ผู้วิจัย	อัญชิสา สิทธิวิริยะชัย
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2566
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาแนวโน้มการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยประกอบไปด้วยแบบจำลอง 6 แบบดังนี้ 1. แบบจำลอง Logistic Regression 2. แบบจำลอง Naive Bayes 3. แบบจำลอง KNN 4. แบบจำลอง Decision Tree 5. แบบจำลอง Random Forest และ 6. แบบจำลอง XGBoost โดยใช้ชุดข้อมูลที่เก็บรวบรวมเกี่ยวกับพฤติกรรมของลูกค้าบริษัทโทรคมนาคมแห่งหนึ่ง เพื่อใช้ในการทำนายการยกเลิกใช้บริการของลูกค้า ซึ่งประกอบด้วยข้อมูลทั้งหมด 7,043 แถว จากฐานข้อมูลสาธารณะแบบเปิด https://www.kaggle.com ผู้วิจัยสนใจที่จะศึกษาปัจจัยหรือคุณลักษณะที่บ่งชี้ว่าลูกค้าจะเลิกใช้บริการของบริษัท และศึกษาหลักการการทำงานของการเรียนรู้ด้วยเครื่องสำหรับการนำมาประยุกต์ใช้ในขั้นตอนการคัดเลือกคุณลักษณะที่ส่งผลต่อการทำนายสูง ผลที่ได้คือปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 3อันดับ ได้แก่ Tenure, Total Charges และ Contract และแบบจำลอง Logistic Regression ให้ผลลัพธ์ที่ดีที่สุดในแง่ของ Accuracy แบบจำลอง XGBoost มีประสิทธิภาพรองลงมา และแบบจำลอง Decision Tree มีประสิทธิภาพต่ำสุดผู้วิจัยจะนำข้อมูลไปประยุกต์ใช้ในการบริหารจัดการทรัพยากร สร้างกลยุทธ์การตลาด ปรับปรุงการบริการและการสร้างสินค้าใหม่ เพื่อตอบสนองความต้องการของลูกค้าและเพิ่มประสิทธิภาพในการแข่งขันในตลาดต่อไป

คำสำคัญ : การขอยกเลิกใช้บริการ, บริษัทโทรคมนาคม, การเรียนรู้ของเครื่อง, แบบจำลอง Logistic Regression

Title	ANALYSIS OF THE CHURN PREDICTION FOR TELECOM CUSTOMERS USING MACHINE LEARNING
Author	ANCHISA SITTIVIRIYAHCAI
Degree	MASTER OF SCIENCE
Academic Year	2023
Thesis Advisor	Assistant Professor Dr. Sirisup Laohakiat

The objective of this research is to study the trends of customer churn for a telecommunications company using machine learning techniques, comprising six models: (1) Logistic Regression; (2) Naive Bayes; (3) KNN; (4) Decision Tree; (5) Random Forest, and (6) XGBoost. These models are applied using a dataset collected on the behavior of customers from a telecommunications company, totaling 7,043 rows, sourced from an open dataset on https://www.kaggle.com. The researchers aimed to investigate the factors or characteristics indicating customer churn and understand the principles of machine learning for practical application in feature selection to enhance predictive accuracy. The results reveal the top three most influential factors leading to customer churn are Tenure, Total Charges, and Contract. Logistic Regression model yields the highest accuracy, followed by XGBoost, while Decision Tree model performed the least effectively. Researchers intend to utilize the data for resource management, devising marketing strategies, improving services, and developing new products to meet customer demands and enhance competitiveness in the market.

Keyword : Churn Prediction, Telecom, Machine Learning, Logistic Regression

กิตติกรรมประกาศ

การจัดทำวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากการสนับสนุน ความรู้ ความช่วยเหลือ คำแนะนำ ตลอดจนแนวทางในการทำวิจัยและจัดทำสารนิพนธ์ของ ผศ.ดร.ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษา และคณาจารย์ทุกท่านในภาควิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ การสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

อัญชิสา สิทธิวิริยะชัย

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและความเป็นมาของการวิจัย

ในปัจจุบันมีธุรกิจด้านโทรคมนาคมที่ให้บริการสัญญาณทางการสื่อสาร เช่น สัญญาณโทรศัพท์เคลื่อนที่ สัญญาณอินเทอร์เนต และสัญญาณโทรทัศน์ดิจิตอลเป็นต้น ซึ่งเป็นธุรกิจที่มูลค่าทางการตลาดสูงมาก โดยมีมูลค่าราว 6.1 แสนล้านบาท โดยสามารถแบ่งออกเป็น ตลาดบริการด้านการสื่อสารมูลค่า 3.6 แสนล้านบาท และตลาดอุปกรณ์สื่อสารมูลค่า 2.6 แสนล้านบาท โดยการขยายตัวของอุตสาหกรรมสื่อสารส่วนใหญ่มีผลมาจากธุรกิจบริการระบบโทรศัพท์แบบเคลื่อนที่ ซึ่งแบ่งได้เป็นตลาดบริการและตลาดอุปกรณ์ อันเป็นผลจากเทคโนโลยีที่เปลี่ยนแปลงอย่างรวดเร็วทั้งระบบ และเครื่องรับสัญญาณ รวมไปจนถึงอุปกรณ์เสริมต่าง ๆ ผนวกกับททางด้านผู้ประกอบการมีการลงทุนเพิ่มเติมเพื่อขยายโครงข่ายอย่างต่อเนื่องในการขยายพื้นที่ให้บริการ ส่งผลให้คุณภาพการสื่อสารแบบไร้สายเป็นไปได้อย่างดียิ่งขึ้น และสามารถใช้งานได้อย่างกว้างขวางมากยิ่งขึ้น เมื่อประกอบกับพฤติกรรมของผู้บริโภคที่ต้องการความคล่องตัวและสะดวกสบายในทุกที่ทุกเวลา จึงล้วนเป็นปัจจัยสำคัญที่ช่วยหนุนการขยายตัวของธุรกิจ

โครงสร้างผู้ให้บริการระบบโทรศัพท์เคลื่อนที่ ปัจจุบันแบ่งเป็น 2 กลุ่ม ได้แก่

- กลุ่มที่ 1 ผู้ให้บริการที่มีสิทธิ์ในโครงข่าย (Mobile Network Operators-MNOs) ซึ่งได้รับสิทธิ์หรือใบอนุญาตให้ใช้คลื่นความถี่และมีโครงสร้างพื้นฐานหรือมีโครงข่ายของตนที่สามารถให้บริการโทรศัพท์เคลื่อนที่ได้โดยอิสระ แบ่งเป็น
    - ผู้ประกอบการที่เป็นองค์กรรัฐวิสาหกิจ ได้แก่ บมจ. ทีโอที (TOT) และ บมจ. กสท โทรคมนาคม (CAT)
    - ผู้ประกอบการภาคเอกชน ได้แก่ กลุ่มบริษัท AIS (บริษัท แอดวานซ์ ไวร์เลส เน็ทเวอร์ค: AWN) กลุ่มบริษัท DTAC (บริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น: DTAC และบริษัท ดีแทค ไตรเน็ต: DTN) และกลุ่มบริษัท TRUE (บริษัท ทรู มูฟ เอช ยูนิเวอร์แซล คอมมิวนิเคชั่น: TUC)

ผู้ประกอบการภาคเอกชนในกลุ่มที่ 1 มีส่วนแบ่งตลาดด้านจำนวนผู้ใช้บริการรวมกัน 97.8% ของผู้ใช้บริการทั้งหมด และมีการให้บริการที่ครอบคลุมมากมายหลายประเภท อาทิ การ

สื่อสารทางเสียง (Voice) และการสื่อสารทางข้อมูล (Non-voice/Data) ซึ่งรวมถึงการให้บริการการเชื่อมต่ออินเตอร์เน็ต บริการ Content และบริการเสริมที่มีความเกี่ยวกับการสื่อสาร ขณะที่รูปแบบการให้บริการนั้นมีทั้งแบบรายเดือนและเติมเงิน ทั้งนี้ หลังจากที่มีการนำเทคโนโลยี 3G มาเปิดให้บริการในประเทศไทย ได้ทำให้ปริมาณการใช้บริการด้านข้อมูลเพิ่มสูงขึ้นอย่างมาก ส่งผลให้รายได้จากการให้บริการข้อมูลกลายเป็นรายได้หลักของผู้ให้บริการแทนรายได้จากการให้บริการเสียง

- กลุ่มที่ 2 ผู้ประกอบการที่ให้บริการโทรศัพท์เคลื่อนที่บนโครงข่ายเสมือน (Mobile Virtual Network Operators: MVNOs) ซึ่งสามารถให้บริการระบบโทรศัพท์เคลื่อนที่ได้โดยไม่จำเป็นต้องได้รับใบอนุญาตใช้คลื่นความถี่ และไม่จำเป็นต้องมีโครงข่ายของตนเอง แต่สามารถดำเนินการประกอบกิจการโดยเช่าใช้ความจุโครงข่ายของรัฐวิสาหกิจ (TOT และ CAT) และบางรายเป็นตัวแทนขายส่งขายต่อ (Wholesale-resale) บริการสู่ผู้บริโภค ผู้ประกอบการในกลุ่มนี้ ได้แก่ 1) ผู้ประกอบการที่ให้บริการบนโครงข่ายของ TOT ได้แก่ บริษัท โมบาย เอท เทลโค (ไทยแลนด์) (Buzzme) และบริษัท ล็อกซเล่ย์ จำกัด (TuneTalk) และ 2) ผู้ประกอบการที่ให้บริการบนโครงข่ายของ CAT ได้แก่ บริษัทเรียลมูฟ (Real Move) บริษัท 168 คอมมูนิเคชั่น จำกัด บริษัทเดอะไวท์สเปซ จำกัด (Penguin Sim) และบริษัท ดาต้า ซีดีเอ็มเอ คอมมูนิเคชั่น (My World)

ธุรกิจการให้บริการระบบโทรศัพท์เคลื่อนที่มีลักษณะของตลาดกึ่งผูกขาดที่มีผู้ประกอบการน้อยราย ถือได้ว่าเป็นธุรกิจที่ต้องการเงินลงทุนที่สูง ทั้งในการวางโครงสร้างโครงข่ายและการลงทุนทางด้านเทคโนโลยีที่มีการเปลี่ยนแปลงอย่างรวดเร็ว ผู้ประกอบการรายใหญ่ที่มีฐานะทางด้านการเงินแข็งแกร่งจึงมีความได้เปรียบและมีอำนาจผูกขาดในตลาดเหนือผู้ประกอบการรายย่อย การเข้าสู่ตลาดของผู้ประกอบการรายใหม่จึงนับได้ว่ามีอุปสรรคอยู่มาก อย่างไรก็ตาม ตลาดนี้ก็มีการแข่งขันในระหว่างผู้ให้บริการต่าง ๆ อยู่สูงมาก ผู้ให้บริการจึงมีความจำเป็นที่จะต้องรักษาฐานลูกค้าของบริษัทของตน พร้อมๆไปกับการสำรวจและค้นหากลุ่มลูกค้าใหม่ ๆ เนื่องจากธุรกิจโทรคมนาคมมีข้อมูลลูกค้าอยู่เป็นจำนวนมาก การรักษาฐานลูกค้าเดิมที่มีอยู่จึงมีต้นทุนที่น้อยกว่าการหาลูกค้าใหม่

เพื่อให้บริษัทเพิ่มความรู้ความเข้าใจในพฤติกรรมของลูกค้าได้เพิ่มมากยิ่งขึ้น จึงได้มีการนำระบบการเรียนรู้ของเครื่อง (Machine Learning) มาช่วยในกระบวนการวิเคราะห์ข้อมูลการใช้งานของลูกค้าอย่างหลากหลาย หนึ่งในบรรดาการใช้งานการเรียนรู้ของเครื่องที่ได้รับความนิยมอย่างสูงในการวิเคราะห์พฤติกรรมของผู้บริโภค ได้แก่ การทำนายความเป็นไปได้ที่ลูกค้าจะยกเลิกการใช้บริการของบริษัท (churn prediction) ทั้งนี้เนื่องจาก หากเราสามารถที่จะทำนายกลุ่มลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการของบริษัท เราจะสามารถวางแผนเพื่อนำเสนอโปรโมชั่นที่มีความเหมาะสม เพื่ออาจสามารถรักษาฐานลูกค้าในกลุ่มนี้ไว้ได้ นอกจากนั้น ยังมีส่วนสำคัญในการออกแบบและพัฒนานโยบายทางการตลาดของบริษัท เพื่อช่วยเพิ่มประสิทธิภาพในการแข่งขันได้

นอกจากนี้ การสร้างแบบจำลองขึ้นมาเพื่อทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าร่วมกับการวิเคราะห์คุณลักษณะ ยังเป็นส่วนสำคัญในการช่วยให้บริษัทเกิดความรู้ความเข้าใจในพฤติกรรมของลูกค้าของตนอย่างลึกซึ้งขึ้น อันอาจจะนำไปสู่การพัฒนาต่อยอด เพื่อสร้างแบบจำลองในการทำนายพฤติกรรมโดยรวมของลูกค้า ที่มีขอบเขตกว้างขวางมากยิ่งขึ้น เช่น การทำนายความชอบของลูกค้าต่อข้อเสนอทางการตลาดแบบใหม่ ๆ เป็นต้น

ในการสร้างแบบจำลอง เราจะใช้ข้อมูล ‘Churn Prediction: Telco Customer Churn’ โดยข้อมูลดังกล่าวนั้นประกอบไปด้วย ข้อมูลส่วนตัวของลูกค้า เช่น ข้อมูลด้านเพศและอายุ กับข้อมูลในการรับบริการของลูกค้า ว่ามีการซื้อบริการอะไรของทางบริษัทบ้าง โดยข้อมูลเหล่านี้อาจนำมาเป็นส่วนหนึ่งในการวิเคราะห์เพื่อบ่งชี้ ประสิทธิภาพในการทำนายของคุณลักษณะที่มีอยู่ในข้อมูลได้ว่า คุณลักษณะเหล่านี้เพียงพอต่อการทำนายการรับบริการต่อ หรือ เลิกใช้บริการของลูกค้าได้หรือไม่ และทำนายได้อย่างดีเพียงไร โดยการวิเคราะห์คุณลักษณะด้วยแบบจำลอง จะช่วยให้เราสามารถนำไปพัฒนารูปแบบของชุดข้อมูลในอนาคตได้อีกว่า ยังขาดคุณลักษณะประการใด และลักษณะใดไม่มีประโยชน์ในการทำนายการขอยกเลิกใช้บริการได้อีกด้วย ในการวิจัยนี้ได้มุ่งเน้นไปยังการใช้งานการเรียนรู้ด้วยเครื่อง (Machine Learning)

นอกจากนี้ยังมุ่งเน้นในส่วนของการใช้ประโยชน์จากการกระบวนการคัดเลือกคุณลักษณะ (Feature Selection) โดยมีวัตถุประสงค์เพื่อค้นหาคุณลักษณะที่มีความสำคัญต่อการทำนายการยกเลิกการใช้บริการ อันจะนำไปสู่การพัฒนาการออกแบบการจัดเก็บข้อมูลของบริษัท รวมไปถึง

การออกแบบนโยบายและข้อเสนอต่าง ๆ ทางการตลาด เพื่อให้เพิ่มประสิทธิภาพในการแข่งขันได้อย่างสูงสุด

1.2 จุดประสงค์ของการวิจัย

ในการทดลองวิจัยครั้งนี้ได้ตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อศึกษาปัจจัยหรือคุณลักษณะต่างๆที่บ่งชี้ว่าลูกค้าจะเลิกใช้บริการของบริษัท
2. เพื่อศึกษาหลักการการทำงานของการเรียนรู้ด้วยเครื่อง สำหรับการนำมาประยุกต์ใช้ในขั้นตอนการคัดเลือกคุณลักษณะที่ส่งผลต่อการทำนายสูง
3. เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ด้วยเครื่องแบบต่างๆ เพื่อค้นหาแบบจำลองที่เหมาะสมต่อการนำมาใช้ทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม

1.3 ขอบเขตของการวิจัย

ในการดำเนินการวิจัยมีการใช้งานชุดข้อมูล ‘Churn Prediction: Telco Customer Churn’ ซึ่งเป็นชุดข้อมูลสาธารณะ https://www.kaggle.com/code/danwheble/churn-prediction-telco-customer-churn/notebook ซึ่งเป็นชุดข้อมูลที่เก็บรวบรวมเกี่ยวกับพฤติกรรมของลูกค้าบริษัทโทรคมนาคมแห่งหนึ่ง เพื่อใช้ในการทำนายการยกเลิกใช้บริการของลูกค้า ซึ่งประกอบด้วยข้อมูลจำนวนทั้งหมด 7,043 แถว และ 21 คอลัมน์ โดยตัวแปรที่ศึกษามีดังนี้

- CustomerID (รหัสลูกค้า)
- Gender (เพศ)
- SeniorCitizen (บัตรผู้สูงอายุ)
- Partner (แต่งงาน)
- Dependents (อยู่กับครอบครัว)
- Tenure (ระยะเวลาการใช้งานมีหน่วยเป็นเดือน)
- PhoneService (การใช้งานโทรศัพท์)
- MultipleLines (มีการเปิดใช้งานหลายเบอร์)
- InternetService (การใช้งานอินเตอร์เน็ต)

- OnlineSecurity (ระบบรักษาความปลอดภัย)
- OnlineBackup (ระบบสำรองข้อมูล)
- DeviceProtection (ระบบป้องกันอุปกรณ์)
- TechSupport (มีการบริการทางเทคนิค)
- StreamingTV (มีการ Streaming TV)
- StreamingMovies (มีการ Streaming ภาพยนตร์)
- Contract (ระยะเวลาในสัญญา)
- PaperlessBilling (มีการวางบิลโดยไม่ใช้กระดาษ)
- PaymentMethod (รูปแบบการชำระเงิน)
- MonthlyCharges (ค่าใช้บริการรายเดือน)
- TotalCharges (ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญา)
- Churn (ยกเลิกบริการหรือไม่)

1.4 สมมุติฐานในการวิจัย

1. TotalCharges หรือค่าใช้จ่ายทั้งหมดตลอดระยะสัญญา อาจมีผลต่อการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม มากกว่าปัจจัยตัวอื่น ๆ โดยคาดว่าค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาที่มากขึ้น โอกาสที่ลูกค้าจะยกเลิกการใช้บริการจะยิ่งน้อยลง

2. จากแบบจำลองการเรียนรู้ด้วยเครื่องแบบต่าง ๆ ที่นำมาทดลอง จะมีบางแบบจำลองที่มีความเหมาะสมกับการทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมมากกว่ายิ่งกว่าแบบจำลองอื่น ๆ โดยเราจะทำการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง เพื่อค้นหาแบบจำลองที่มีความเหมาะสมต่อการนำมาใช้กับชุดข้อมูลนี้มากที่สุด

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สามารถนำแบบจำลองไปใช้ประโยชน์ในการทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม

2. ทำให้ทราบถึงคุณลักษณะหรือปัจจัยที่มีอิทธิพลหรือส่งผลต่อการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม เพื่อให้สามารถระบุกลุ่มของลูกค้าที่มีแนวโน้มในการขอยกเลิกใช้บริการได้

3. ข้อมูลผลลัพธ์จากการทำนายสามารถนำไปใช้ปรับปรุงประสิทธิภาพนโยบายการตลาดของบริษัทให้มีความสอดคล้องกับสภาวะของกลุ่มลูกค้าได้

4. สามารถนำไปพัฒนาเพื่อต่อยอดในการสร้างแบบจำลอง สำหรับการวิเคราะห์พฤติกรรมส่วนอื่นๆ ของลูกค้าในกิจการได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่มีความเกี่ยวข้องกับการดำเนินการวิจัยและได้มีการนำเสนอตามหัวข้อต่อไปนี้

1. แบบจำลอง Logistic Regression
2. แบบจำลอง Naive Bayes
3. แบบจำลอง KNN
4. แบบจำลอง Decision Tree
5. แบบจำลอง Random Forest
6. แบบจำลอง XGBoost
7. งานวิจัยที่เกี่ยวข้อง

แบบจำลอง Logistic Regression เป็นแบบจำลองที่จัดอยู่ในประเภท Supervise Learning สำหรับการทำนายผลแบบ Classification โดยข้อมูลตัวอย่างแต่ละรายการจะมีคอลัมน์ที่เป็น Target หรือตัวแปรตาม (y) เช่น Yes/No, True/False, 1/0 เป็นต้น ส่วนคอลัมน์ที่เป็น Feature หรือตัวแปรอิสระ (x) อาจมีตั้งแต่ 1 คอลัมน์ขึ้นไป โดยถ้าค่าตัวแปรหนึ่งเพิ่มจะส่งผลให้ค่าตัวแปรเปลี่ยนแปลงไปด้วย ซึ่งจะมีความสัมพันธ์แบบเชิงเส้นตรงหรือเส้นโค้งในลักษณ์ที่เพิ่มขึ้นหรือลดลงก็ได้ เขียนเป็นสมการได้ดังนี้

$$\mathrm{P(x)} = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}}$$

จึงเป็นที่มาของคำว่า Logistic Regression เพราะ e ยกกำลังสมการแบบ Linear Regression นั่นเอง หรือกล่าวได้ว่า Logistic Regression ก็คือการหาความน่าจะเป็นด้วย Sigmoid Function ทั้งนี้หากเราทราบค่า intercept ($\beta 0$) และ coefficient ($\beta 1$, $\beta 2$, ...) ก็สามารถนำตัวแปรอิสระ x1, x2, ... มาแทนค่าในฟังก์ชัน ก็จะได้ค่าความน่าจะเป็นออกมา แต่อย่างไรก็ตาม ความน่าจะเป็นต้องมีค่าระหว่าง 0 - 1 ในขณะที่ผลลัพธ์ของ Logistic Regression จะต้องอยู่ในรูปแบบ Classification ที่จำแนกได้ 2 กลุ่ม เช่น Yes/No ซึ่งไม่สอดคล้องกัน ดังนั้น จึงต้องนำค่าความน่าจะเป็นที่ได้ไปทำการเปรียบเทียบต่อ ถ้าความน่าจะเป็นมีค่าระหว่าง 0-0.5 จะ

ทำนายผลเป็นคลาสกลุ่มที่ 1 และ ถ้าความน่าจะเป็นมีค่าระหว่าง 0.5 – 1 จะทำนายผลเป็นคลาสกลุ่มที่ 2 (บัญชา ปะสีละตัง, 2564)

แบบจำลอง Naive Bayes เป็นแบบจำลองที่จัดอยู่ในประเภท Supervise Learning โดยจะทำนายผลแบบ Classification ซึ่งเป็นการสร้างแบบจำลองเพื่อทำนายผลโดยอาศัยทฤษฎีการคำนวณความน่าจะเป็น หลักการพื้นฐานคือการนับจำนวนความถี่ของรายการ จากนั้นคำนวณความน่าจะเป็น แล้วนำมาเปรียบเทียบกัน (บัญชา ปะสีละตัง, 2563) ดังนั้น ลักษณะข้อมูลตัวอย่างที่จะใช้ Train Model ชนิด Naive Bayes ทั้งคอลัมน์ Features ตัวแปรอิสระ (x) และ Target หรือตัวแปรตาม (y) ควรเป็นแบบ Classification หรือสามารถจัดกลุ่มได้ เนื่องจากแบบจำลอง Naïve Bayes จะทำนายผลโดยใช้วิธีการคำนวณความน่าจะเป็นคล้ายกับ Logistic Regression แต่ Logistic Regression จะใช้ฟังก์ชัน Sigmoid ซึ่งให้ผลลัพธ์ออกมาเป็นค่าระหว่าง 0-1 ในขณะที่แบบจำลอง Naïve Bayes จะคำนวณโดยใช้หลักการหรือทฤษฎีของความน่าจะเป็น (Provability) ซึ่งความน่าจะเป็น คือโอกาสที่จะเกิดเหตุการณ์ (Event) อย่างใดอย่างหนึ่ง เมื่อเทียบกับความเป็นไปได้ทั้งหมด ความน่าจะเป็นต้องมีค่าระหว่าง 0-1 เท่านั้น โดยส่วนใหญ่เราจะแทนด้วยตัว P ซึ่งสามารถเขียนเป็นสูตรได้ดังนี้

$$\text{Probability} = \frac{\text{จำนวนเหตุการณ์ที่สิ่งนั้นจะเกิดขึ้น}}{\text{จำนวนความเป็นไปได้ทั้งหมด}}$$

ส่วนทฤษฎีของเบย์ (Bayes Theorem) สามารถเขียนเป็นกฎได้ดังนี้

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

P(A|B) คือความน่าจะเป็นของเหตุการณ์ A ภายใต้เงื่อนไขของ B ซึ่งเป็นค่าที่เราต้องการหา

P(BIA) คือความน่าจะเป็นของเหตุการณ์ B ภายใต้เงื่อนไขของ A ซึ่งเป็นข้อมูลที่เราทราบอยู่แล้ว

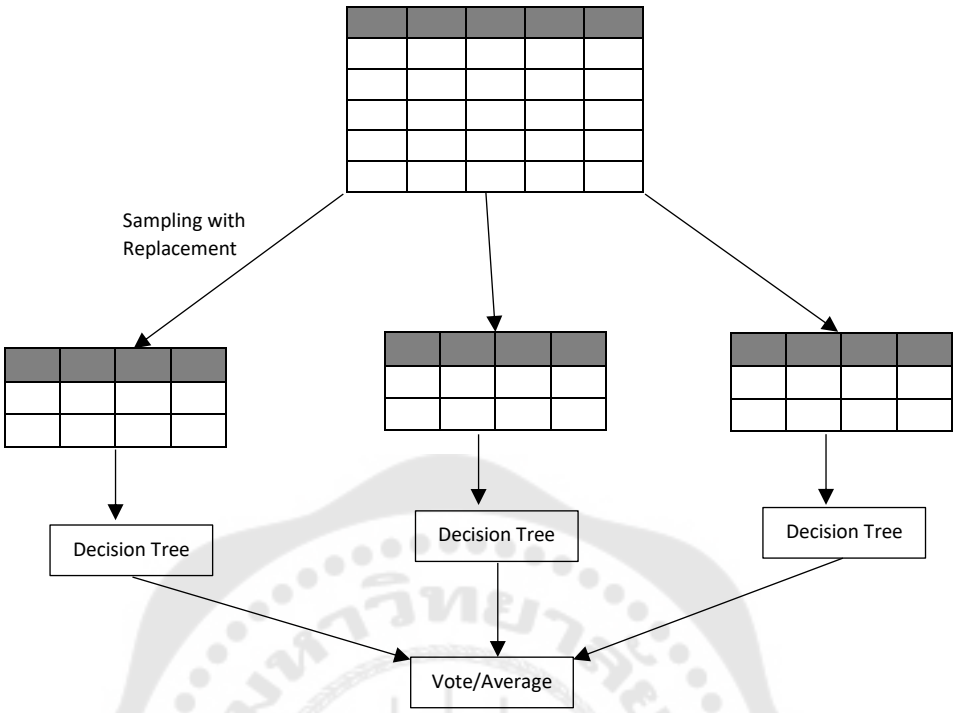
P(A) คือความน่าจะเป็นของเหตุการณ์ A ซึ่งเกิดขึ้นแล้ว

P(B) คือความน่าจะเป็นของเหตุการณ์ B ซึ่งเกิดขึ้นแล้ว

แบบจำลอง KNN หรือ K-Nearest Neighbors เป็นอีกแบบจำลองที่ง่ายที่สุดอันหนึ่ง เนื่องจากใช้เพียงแค่หลักการคำนวณคณิตศาสตร์เบื้องต้นเท่านั้นแต่ให้ผลลัพธ์ที่มีความถูกต้อง แม่นยำ และเชื่อถือได้ แบบจำลองนี้รองรับการทำนายผลทั้งแบบ Classification และแบบ Regression โดยใช้วิธีเปรียบเทียบระยะห่างของตำแหน่งข้อมูล หลักการทำงานคือการเทียบหาจุดใหม่ ถ้าพบว่าอยู่ใกล้กับกลุ่มใดก็จะจัดให้ข้อมูลใหม่อยู่ในกลุ่มนั้น ซึ่งค่า k จะเป็นตัวกำหนดว่าจะเทียบหาจุดเพื่อนบ้านกี่จุด แต่หากมีข้อมูลตัวอย่างเป็นจำนวนมาก การทำนายผลอาจช้ากว่าแบบจำลองอื่นๆ

แบบจำลอง Decision Tree หรือการตัดสินใจด้วยแผนภูมิต้นไม้ โดยสร้างลำดับขั้นการตัดสินใจ เป็นแผนภูมิที่มีลักษณะโครงสร้างคล้ายกับต้นไม้แบบกลับหัว จัดอยู่ในกลุ่มของ Supervise Learning เหมาะสำหรับการจำแนกหรือแยกแยะว่าข้อมูลอยู่ในกลุ่มไหน มีลักษณะการทำงานคล้ายกับการถามและตอบ และยังมีกราฟแสดงผลที่ได้จากแบบจำลองอีกด้วย โดยผลลัพธ์จะมีเพียงสองกลุ่มหรือมากกว่าก็ได้ สามารถทำนายผลได้ทั้งแบบ Classification และแบบ Regression ทั้งนี้ข้อมูลที่จะทำนายผลด้วย Decision Tree นั้น คอลัมน์ที่เป็น Features หรือ ตัวแปรอิสระ (x) ควรเป็นข้อมูลที่จัดกลุ่มได้ ส่วนคอลัมน์ Target หรือตัวแปรตาม (y) อาจเป็นแบบ Classification หรือแบบ Regression ก็ได้ หลักการเบื้องต้นของการเขียนแผนภูมิต้นไม้โดยทั่วไปคือการนำเงื่อนไขมาแบ่งเป็นกรณีย่อยๆ ซ้อนกันลงไปเรื่อยๆ จนกว่าจะไม่มีเงื่อนไขที่จะแบ่งต่อไปอีก (กอบเกียรติ สระอุบล, 2563)

แบบจำลอง Random Forest จะนำแบบจำลอง Decision Tree หลายๆอันมาใช้ร่วมกัน ซึ่งก็เปรียบได้กับต้นไม้ (Tree) หลาย ๆ ต้น เมื่อรวมกันก็จะ กลายเป็นป่า (Forest) สามารถทำนายผลได้ทั้งแบบ Classification และแบบ Regression เช่นเดียวกับ Decision Tree ซึ่งโดยทั่วไปแล้ว Random Forest มักจะมีความแม่นยำมากกว่า สำหรับหลักการทำงานของ Random Forest มีแนวทางโดยใช้วิธีแบบ Bagging คือแบ่งข้อมูลตัวอย่างออกเป็นหลายๆชุดแล้วนำแต่ละชุดไป train ด้วยแบบจำลอง จากนั้นก็นำผลลัพธ์จากแต่ละแบบจำลองไปเปรียบเทียบโดยการโหวตหรือหาค่าเฉลี่ย หากเป็นการทำนายผลแบบ Classification ก็จะทำนายผลโดยวิธีการโหวต แต่หากเป็นการทำนายผลแบบ Regression ก็จะทำนายผลโดยวิธีการหาค่าเฉลี่ย



ภาพประกอบ 1 หลักการทำงานของแบบจำลอง Random Forest

แบบจำลอง XGBoost (EXtreme Gradient Boosting) การเรียนรู้ในกลุ่ม Ensemble ถือเป็น แบบจำลองที่ล้ำสมัย พัฒนาขึ้นจากแนวคิดพื้นฐานเดียวกับ Gradient Boosting แต่ว่าใช้วิธีคำนวณที่แตกต่างกัน สามารถทำนายผลได้ทั้งแบบ Classification และแบบ Regression โดยใช้วิธีแบบ Boosting คือนำข้อมูลทั้งหมดไป train ด้วยแบบจำลองอันแรกก่อน หากมีรายการที่ทำนายผิดพลาดก็จะกำหนดน้ำหนักให้กับแต่ละรายการ แล้วสุ่มเลือกรายการข้อมูลใหม่ จากนั้นนำไป train ด้วยแบบจำลองตัวที่สอง และทำเช่นนี้ไปเรื่อยๆ จนกว่าการทำนายผลจะถูกต้องทั้งหมด แล้วก็นำผลการทำนายจากแต่ละแบบจำลองไปทำการโหวตหรือหาค่าเฉลี่ย โดยแบบจำลองนี้มีจุดเด่นที่น่าสนใจคือ ทำงานได้เร็ว สามารถจัดการกับข้อมูลที่หายไปได้โดยอัตโนมัติ และผู้ใช้สามารถกำหนดเงื่อนไขการทำงานเพิ่มเติมในแบบ Custom ได้ และยังสามารถประมวลผลข้อมูลหลายมิติได้ เช่น ข้อมูลแบบช่วงเวลา (Time Series) และข้อมูลแบบที่มีการแบ่งกลุ่ม (Clustering)

ในการวิจัยครั้งนี้ได้มีการศึกษาค้นคว้าและทบทวนวรรณกรรมงานวิจัยอื่นๆ ที่มีความเกี่ยวข้องและเป็นประโยชน์กับงานวิจัย ดังต่อไปนี้

บทความวิจัยเรื่อง Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression **โดย** Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade (Dalvi et al., 2016) ในงานวิจัยนี้ต้องการทำนายผลการยกเลิกบริการของบริษัทโทรคมนาคมแห่งหนึ่ง งานวิจัยนี้ใช้ชุดข้อมูลจริงจากผู้ให้บริการโทรคมนาคมซึ่งเพื่อสร้างแบบจำลองสำหรับคาดการณ์ว่าลูกค้ามีแนวโน้มที่จะยกเลิกบริการหรือไม่ การสร้างระบบมี 3 ขั้นตอนหลัก ได้แก่ การพัฒนาเว็บอินเตอร์เฟส การสกัดหา Feature ที่สำคัญ และการทำนาย เว็บอินเตอร์เฟสจะแสดงภาพรวมของผลลัพธ์ที่ได้ ส่วนการสกัดหา Feature ที่สำคัญจะประกอบด้วยการประมาณค่าพารามิเตอร์ใน Logistic Regression และการสร้างกฎสำหรับ Decision Tree การทำนายจะขึ้นอยู่กับการประมาณค่าพารามิเตอร์ ที่ได้จาก Logistic Regression และกฎที่ได้จาก Decision Tree ชุดข้อมูลนำเข้าของระบบประกอบด้วย Training Data และ Test Data โดยคุณลักษณะของชุดข้อมูลการฝึก เช่น รัฐ, รหัสพื้นที่, เบอร์โทรศัพท์, นาทีโทรวันธรรมดา, นาทีโทรเย็น, นาทีโทรกลางคืน, นาทีโทรต่างประเทศ, จำนวนครั้งที่ติดต่อฝ่ายบริการลูกค้า, แพ็กเกจอินเตอร์เน็ต, แพ็กเกจ Voicemail, จำนวนสายวันธรรมดา, ค่าบริการวันธรรมดา, จำนวนสายวันเย็น, ค่าบริการวันเย็น, จำนวนสายกลางคืน, ค่าบริการกลางคืน, จำนวนสายต่างประเทศ, ค่าบริการต่างประเทศ และการยกเลิกบริการ โดยสรุปแล้วงานวิจัยชี้ให้เห็นว่าการใช้แบบจำลอง Logistic Regression และ Decision Trees ประกอบกันสามารถให้ผลที่ดีกว่าการใช้แบบจำลองเดียว

บทความวิจัยเรื่อง Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques **เขียนโดย** Abhishek Gaur (Gaur & Dubey, 2018) บทความนี้มุ่งเน้นไปที่เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ต่างๆ สำหรับการทำนายการย้ายค่ายของลูกค้า เทคนิคเหล่านี้ใช้ในการสร้างแบบจำลองการจำแนก (classification model) เพื่อระบุลูกค้าที่มีแนวโน้มจะย้ายค่าย ตัวอย่างของแบบจำลองที่กล่าวถึง ได้แก่ Logistic Regression, SVM, Random Forest และ Gradient Boosted Tree บทความเปรียบเทียบประสิทธิภาพของแบบจำลองเหล่านี้สำหรับการทำนายการย้ายค่าย งานวิจัยส่วนใหญ่เน้นแค่การระบุคนที่ จะย้ายค่ายแต่ไม่ค่อยสนใจกรณีที่ ระบุผิด คือคนที่ไม่ย้ายค่าย กลับถูกจัดว่าจะย้ายค่าย

การรณรงค์ให้ลูกค้าอยู่ต่อ (retention campaign) มักใช้การลดราคาหรือโปรโมชั่นซึ่งสิ้นเปลืองงบประมาณ ถ้าหากทำกับคนที่ไม่คิดจะย้ายค่ายอยู่แล้ว ยิ่งทำให้เปลืองเงินไปโดยไม่เกิดประโยชน์ บทวิเคราะห์แบบเดิมมักให้ผลเป็นแค่ 0 (ไม่ย้ายค่าย) หรือ 1 (ย้ายค่าย) ซึ่งเป็นข้อจำกัดในการวิเคราะห์ข้อมูล ในบทความวิจัยนี้จะเสนอวิธีวิเคราะห์โดยแบ่งลูกค้าเป็นกลุ่ม (profile-based analysis) วิธีนี้คาดว่าจะช่วยให้การทำนายแม่นยำขึ้น โดยการจัดกลุ่มลูกค้าที่มีพฤติกรรมคล้าย ๆ กัน เพื่อดูว่ากลุ่มไหนมีแนวโน้มจะย้ายค่ายมากที่สุด นอกจากนั้น วิธีนี้ยังช่วยลดการระบุผิด ด้วยการแยกกลุ่มลูกค้าที่ผลการทำนายไม่แม่นยำออกไป ผลการศึกษาพบว่า Gradient Boosted Tree คือแบบจำลองที่มีประสิทธิภาพดีที่สุด ในขณะที่ Logistic Regression และ Random Forest มีประสิทธิภาพปานกลาง ส่วน SVM มีประสิทธิภาพต่ำสุดในการทำนายการย้ายค่าย

บทความวิจัยเรื่อง A Comparison of Machine Learning Algorithms for Customer Churn Prediction **เขียนโดย** Parth Pulkundwar (Pulkundwar et al., 2023) งานวิจัยนี้กล่าวถึงบทบาทของแมชชีนเลิร์นนิง ในการทำนายการย้ายค่าย เริ่มต้นด้วยการเน้นย้ำถึงความสำคัญของการย้ายค่ายในสภาวะการแข่งขัน พร้อมกับชี้ให้เห็นถึงผลกระทบของการใช้ข้อมูลในการแก้ไขปัญหานี้ ต่อไปจะกล่าวถึงอัลกอริทึม ML ต่างๆ ที่เหมาะสำหรับการทำนายการย้ายค่ายและเปรียบเทียบผลลัพธ์เพื่อหาอัลกอริทึมที่ดีที่สุด งานวิจัยพบว่าอัลกอริทึม Decision Tree Classification, Random Forest Classification, AdaBoost และ XGBoost Classification เหมาะสำหรับการทำนายการย้ายค่าย นอกจากนี้ บทความยังกล่าวถึงการนำผลลัพธ์ไปใช้ในแอปพลิเคชันทำนายการย้ายค่ายด้วย ผลลัพธ์ที่ได้คืออัลกอริทึมที่ใช้เวลาน้อยที่สุดในการฝึก (train) คือ Naïve Bayes และ KNN Classification แต่ความแม่นยำในการทดสอบ (testing) ต่ำ ส่วน Logistic Regression ก็ใช้เวลาน้อยในการฝึก แต่ความแม่นยำก็ไม่สูงมาก อัลกอริทึม Gradient Boosting (AdaBoost และ XGBoost) ใช้เวลานานกว่าในการฝึก แต่มีความแม่นยำสูง สรุปได้ว่า Random Forest คือ อัลกอริทึมที่มีประสิทธิภาพดีที่สุด

บทความวิจัยเรื่อง Churn Prediction Estimation Based on Machine Learning Methods **เขียนโดย** Mykola Malyar Mykola Robotyshyn M.V. **และ** Maryana Sharkadi (Malyar et al., 2020) บทความวิเคราะห์แนวทางที่มีอยู่สำหรับการทำนายการย้ายค่ายของลูกค้าในหลายธุรกิจ และเสนอวิธีการกำหนดช่วงเวลาที่ลูกค้าอาจจะย้ายค่าย และเลือกวิธีการติดป้ายข้อมูล (data labeling) ที่เหมาะสม เพื่อเปลี่ยนปัญหาให้กลายเป็นปัญหาการจัดกลุ่มแบบแบ่งเป็น

สองกลุ่ม (Binary Classification) โดยเลือกใช้วิธีการ Ensemble Tree (Random Forest, XGBoost, LightGBM) เป็นอัลกอริทึมการเรียนรู้ สรุปบทความกล่าวถึงเทคนิค Bagging และ Boosting ว่าเป็นเทคนิคยอดนิยม ในการทำงานกับข้อมูลแบบตาราง (tabular data) ในปัจจุบัน เทคนิคเหล่านี้ถูกนำมาใช้แก้ปัญหาการทำนายการย้ายค่ายของลูกค้า อัลกอริทึมที่ใช้ทดสอบ 4 แบบ ได้แก่ Decision Tree, Random Forest, LightGBM และ XGBoost โดย Decision Tree ถูกใช้เป็นอัลกอริทึมพื้นฐานเบื้องต้น LightGBM และ XGBoost ใช้แนวคิดแบบ Boosting ส่วน Random Forest ใช้แนวคิดแบบ Bagging โดยสรุปว่าอัลกอริทึมที่แสดงผลลัพธ์ดีที่สุดคือ XGBoost ด้วยคะแนน AUC 0.8322 ส่วนผลลัพธ์ที่ต่ำสุดคือ Decision Tree แบบเดิมที่ได้คะแนน AUC 0.65 ผลลัพธ์ของ LightGBM อยู่ที่ 0.818 (AUC score) ซึ่งน้อยกว่า XGBoost อยู่พอสมควร ผลการทดสอบแสดงให้เห็นว่า การใช้กลุ่มของ Ensemble Trees มีประสิทธิภาพมากกว่าการใช้ Tree เพียงต้นเดียว นอกจากนี้ บทความยังชี้ให้เห็นว่า เทคนิค Boosting มีประสิทธิภาพ ดีกว่าเทคนิค Bagging เมื่อใช้กับชุดข้อมูลนี้ และคุณสมบัติสำคัญที่มีผลต่อการย้ายค่ายของลูกค้า มีดังนี้ จำนวนครั้งที่ลูกค้าใช้บริการ, ราคาเฉลี่ยของการใช้บริการ, จำนวนวันที่ผ่านไปนับตั้งแต่ลูกค้าเริ่มใช้บริการครั้งแรก และจำนวนลูกค้ารายอื่นที่เคยใช้บริการสินค้าประเภทเดียวกัน

บทความวิจัยเรื่อง "Customer Churn Prediction Using Machine Learning Approaches" โดย R. Srinivasan (Srinivasan et al., 2023) เป็นการศึกษาเกี่ยวกับการใช้เทคนิคและวิธีการทางด้าน (Machine Learning) เพื่อทำนายการลูกค้าที่จะเลิกใช้บริการหรือ "Customer Churn" ในองค์กร การศึกษานี้มุ่งเน้นการวิเคราะห์และทำนายพฤติกรรมของลูกค้าที่มีแนวโน้มที่จะตัดสินใจยุติการใช้บริการขององค์กรนั้น ๆ บทความวิจัยนี้กล่าวถึงการใช้แมชชีนเลิร์นนิง (ML) เพื่อสร้างแบบจำลองทำนายการย้ายค่าย ซึ่งจะช่วยให้ผู้ให้บริการโทรคมนาคมสามารถคาดการณ์ ลูกค้าที่มีแนวโน้มจะย้ายค่ายได้ เปรียบเทียบประสิทธิภาพของอัลกอริทึม ML ต่างๆ เพื่อหาแบบจำลองที่ดีที่สุด ผลการทดลองพบว่า การใช้Random Forest ร่วมกับ SMOTE-ENN มีประสิทธิภาพดีที่สุดในการทำนายการย้ายค่าย โดยวัดจากค่า F1-score จากการวิเคราะห์แบบจำลองนี้สามารถทำนายการย้ายค่ายได้สูงสุดถึง 95%

บทที่ 3
การดำเนินการวิจัย

ในงานวิจัย ผู้วิจัยได้ดำเนินตามขั้นตอนดังนี้

3.1 การเก็บรวบรวมข้อมูล

3.2 การนำเข้าข้อมูล ตรวจสอบข้อมูล และทำความสะอาดข้อมูล

3.3 การสำรวจข้อมูล (Exploratory Data Analysis)

3.4 การเตรียมข้อมูล (Data Preprocessing)

3.5 การสร้างแบบจำลองเพื่อทำการทำนายและหา Feature Importance

3.1 การเก็บรวบรวมข้อมูล

ข้อมูลลูกค้าบริษัทให้บริการสัญญาณโทรศัพท์ ประกอบด้วยข้อมูลจำนวนทั้งหมด 7,043 แถว และ 21 คอลัมน์ จากข้อมูลสาธารณะบนเว็บไซต์ www.kaggle.com โดยมีการแบ่งข้อมูลออกเป็น 2ชุด คือ กลุ่มลูกค้าที่ยกเลิกบริการ และลูกค้าที่ใช้บริการต่อ

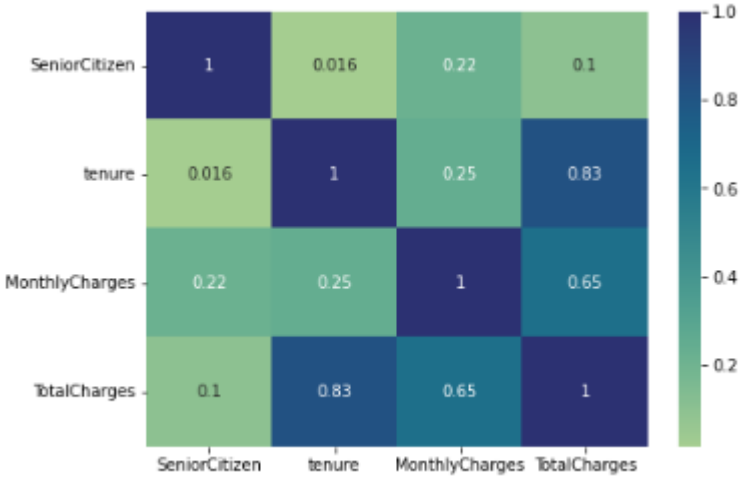
3.2 การนำเข้าข้อมูล ตรวจสอบข้อมูล และทำความสะอาดข้อมูล

ผู้วิจัยใช้ภาษาไพทอน (Python) ในการวิเคราะห์ข้อมูลและการเรียนรู้ของเครื่อง เริ่มต้นด้วยการนำเข้าโมดูลสำคัญสำหรับการสร้างแบบจำลอง ต่อมานำเข้าไฟล์ข้อมูลและข้อมูลที่ใช้สำหรับสร้างแบบจำลอง เริ่มกระบวนการตรวจสอบและสำรวจข้อมูลเบื้องต้น เพื่อหาข้อมูลเชิงลึกโดยใช้ไลบรารี่ Pandas, Numpy จากนั้นทำความสะอาดข้อมูล ตรวจสอบดูค่าที่หายไป พบว่าคอลัมน์ TotalCharges มีค่าว่างอยู่ 11ค่า จึงได้ทำการลบออกไป และได้เปลี่ยนชนิดของข้อมูลของ TotalCharges จาก Object เป็น Float64

3.3 การสำรวจข้อมูล (Exploratory Data Analysis)

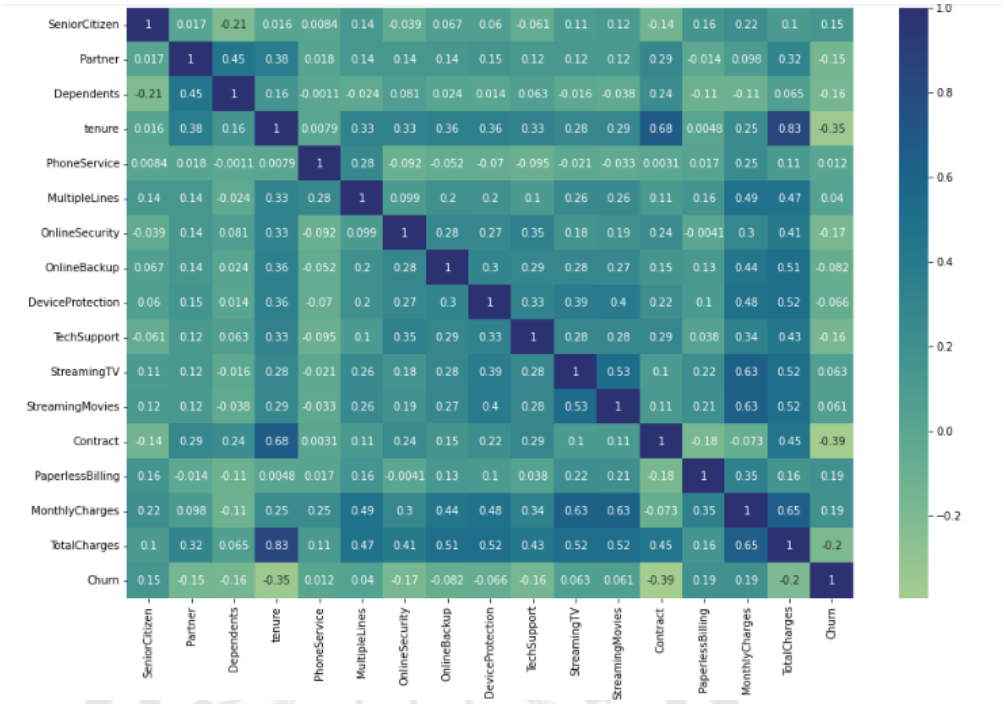
วิเคราะห์ระดับความสัมพันธ์ของตัวแปรของแต่ละคอลัมน์ โดยการใช้เทคนิคที่เรียกว่า Correlation Coefficient หรือ ค่าสัมประสิทธิ์สหสัมพันธ์ เป็นค่าที่บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว คือเป็นค่าที่บ่งบอกถึงความสัมพันธ์ของตัวแปร 2 ตัว ว่ามีความสัมพันธ์กันมากน้อย

เพียงใด และมีความสัมพันธ์ในเชิงบวกหรือเชิงลบ ค่าสัมประสิทธิ์สหสัมพันธ์จะมีค่าอยู่ระหว่าง -1.0 จนถึง +1.0 โดยหากพบว่าค่าเข้าใกล้ -1.0 หมายความว่า ตัวแปรทั้ง 2 ตัวมีความสัมพันธ์เชิงลบ หรือแปรผกผันกัน เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้น อีกตัวแปรจะลดลง หากพบว่าค่าเท่ากับ 0.0 หมายความว่า ตัวแปรทั้ง 2 ตัวไม่มีความสัมพันธ์กัน และ หากพบว่าค่าเข้าใกล้ +1.0 หมายความว่า ตัวแปรทั้ง 2 ตัวมีความสัมพันธ์เชิงบวก หรือแปรผันตามกัน เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้น อีกตัวแปรจะเพิ่มขึ้นตาม



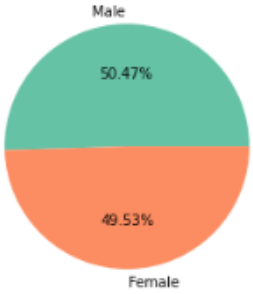
ภาพประกอบ 2 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร 4 ตัวที่คอลัมน์มีค่าเป็นตัวเลข

จากภาพประกอบ 2 จะเห็นได้ว่าตัวแปร Tenure มีความสัมพันธ์กับ TotalCharges สูงสุด โดยมีค่า สัมประสิทธิ์สหสัมพันธ์อยู่ที่ 0.83 โดยตัวแปรทั้งสองนี้ จะมีความสัมพันธ์ตามกัน นั่นคือ ลูกค้าที่ใช้บริการมานาน ก็จะมีค่าใช้จ่ายตลอดระยะสัญญาสูงตามไปด้วย ในขณะที่ลูกค้าใหม่ซึ่งมีค่า Tenure ต่ำ ก็จะมีจะมีค่าใช้จ่ายตลอดระยะสัญญาต่ำกว่าลูกค้าที่ใช้บริการมานานแล้ว และ TotalCharges ความสัมพันธ์กับ MontlyCharges โดยค่า correlation อยู่ที่ 0.65 เนื่องจากลูกค้าที่จ่ายรายเดือนสูงก็จะมีแนวโน้มที่จะมีค่าใช้จ่ายตลอดระยะสัญญาสูงตามไปด้วย



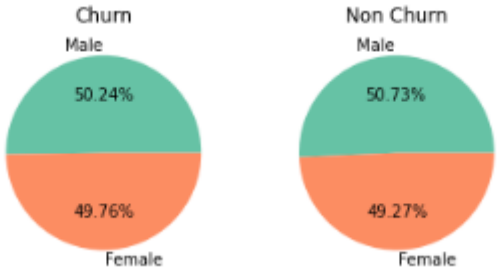
ภาพประกอบ 3 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรทุกคอลัมน์

จากภาพประกอบ 3 จะเห็นได้ว่าตัวแปร Tenure มีความสัมพันธ์กับ Contract โดยมีค่าสัมประสิทธิ์สหสัมพันธ์อยู่ที่ 0.68 โดยตัวแปรทั้งสองนี้ จะมีความสัมพันธ์ตามกัน นั่นคือ ลูกค้าที่ใช้บริการมานานก็จะมีแนวโน้มใช้บริการผูกสัญญาระยะยาวมากกว่า ลำดับถัดมาค่าสัมประสิทธิ์สหสัมพันธ์ที่ 0.63 ซึ่งก็ถือว่าค่อนข้างสูง เป็นค่าความสัมพันธ์ระหว่าง MonthlyCharges กับ StreamingTV และ MonthlyCharges กับ StreamingMovies แสดงให้เห็นถึงว่าหากมีการใช้บริการ StreamingTV และ StreamingMovies ก็จะทำให้ MonthlyCharges สูงตามไปด้วย



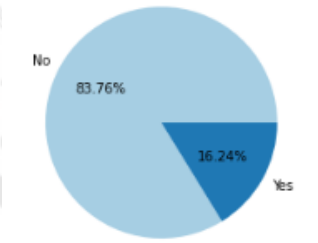
ภาพประกอบ 4 สัดส่วนระหว่างผู้ใช้บริการที่เป็นเพศชายกับเพศหญิง

จากภาพประกอบ 4 จะพบว่าสัดส่วนของเพศลูกค้าไม่ได้แตกต่างกันมากนัก จำนวนลูกค้าที่เป็นผู้ชายมี 3,549 คน คิดเป็นสัดส่วนร้อยละ 50.47 และจำนวนลูกค้าที่เป็นผู้หญิงมี 3,483 คน คิดเป็นสัดส่วนร้อยละ 49.53



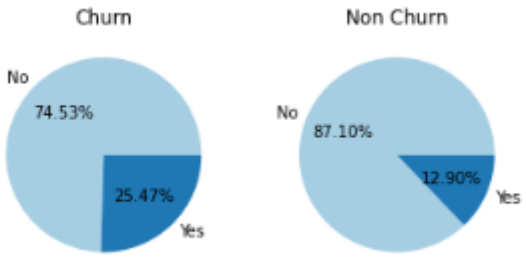
ภาพประกอบ 5 สัดส่วนระหว่างผู้ใช้บริการที่เป็นเพศชายกับเพศหญิงและแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ

จากภาพประกอบที่ 5 จะเห็นได้ว่าสัดส่วนระหว่างผู้ใช้บริการที่เป็นเพศชายกับเพศหญิงและแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อไม่ได้ต่างกันมากนัก มีสัดส่วนที่เท่าๆกัน



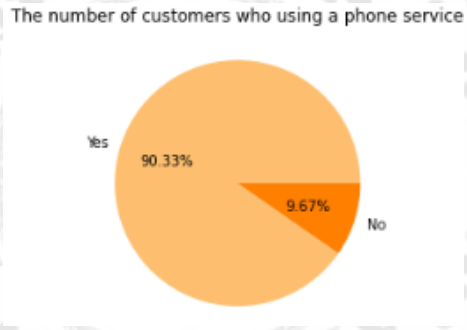
ภาพประกอบ 6 สัดส่วนของลูกค้าวัยผู้สูงอายุ

จากภาพประกอบที่ 6 ลูกค้าวัยผู้สูงอายุมีจำนวน 1,142 คน คิดเป็นร้อยละ 16.24 และจำนวนลูกค้าที่ไม่ใช่ผู้สูงอายุมีจำนวน 5,890 คิดเป็นร้อยละ 83.76



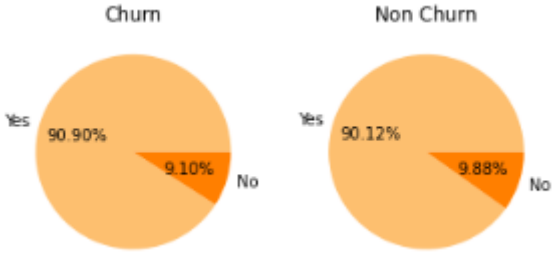
ภาพประกอบ 7 สัดส่วนของลูกค้าวัยผู้สูงอายุและแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ

จากภาพประกอบที่ 7 จะเห็นได้ว่าจะเห็นได้ว่าสัดส่วนของลูกค้าวัยผู้สูงอายุและแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อไม่ได้ต่างกันอย่างมีนัยสำคัญเท่าไหร่



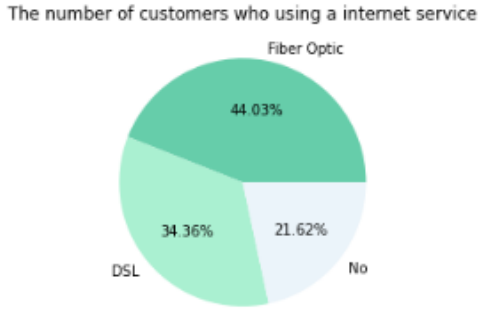
ภาพประกอบ 8 สัดส่วนของผู้ใช้บริการโทรศัพท์

จากภาพประกอบ 8 แสดงสัดส่วนของผู้ใช้บริการโทรศัพท์ที่มีจำนวน 6,352 คน คิดเป็นร้อยละ 90.33



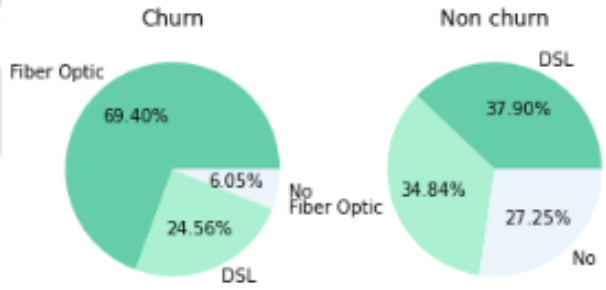
ภาพประกอบ 9 สัดส่วนของผู้ใช้บริการโทรศัพท์ และแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ

จากภาพประกอบ 9 แสดงสัดส่วนของผู้ใช้บริการโทรศัพท์ และแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ จะเห็นได้ว่าสัดส่วนไม่ได้ต่างกันมากนัก



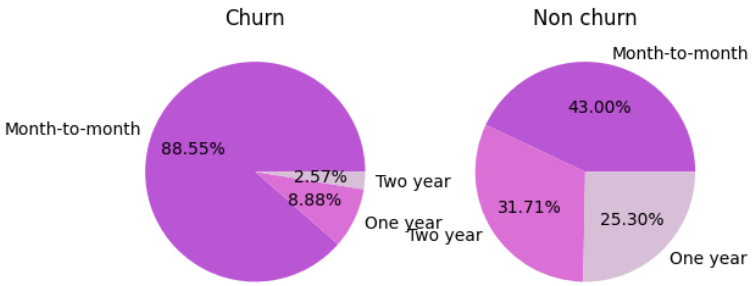
ภาพประกอบ 10 สัดส่วนของผู้ใช้บริการอินเตอร์เน็ต

จากภาพประกอบ 10 บริษัทมีอินเตอร์เน็ตให้บริการอยู่ 2 รูปแบบ คือ DSL และ Fiber optic ซึ่ง Fiber optic เป็นอินเตอร์เน็ตรูปแบบที่ลูกค้าใช้บริการมากที่สุด คิดเป็นร้อยละ 44.03 รองลงมาเป็น DSL คิดเป็นร้อยละ 34.36 และที่เหลือที่ไม่ได้ใช้งานอินเตอร์เน็ตคิดเป็นร้อยละ 21.62



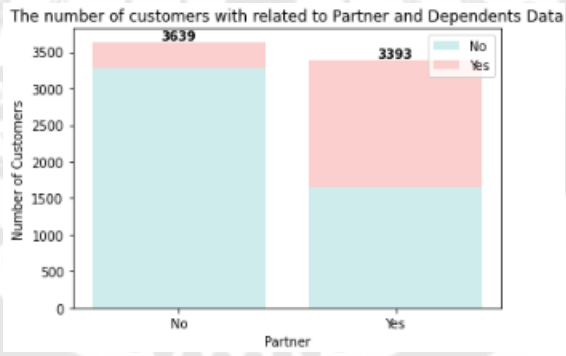
ภาพประกอบ 11 สัดส่วนของผู้ใช้บริการอินเตอร์เน็ต และแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ

จากภาพประกอบ 11 จะเห็นได้ว่าลูกค้าที่ยกเลิกบริการ เป็นลูกค้าที่ใช้งานอินเตอร์เน็ตแบบ Fiber optic ถึงร้อยละ 69.40 แต่ลูกค้าที่ไม่ได้ใช้งานอินเตอร์เน็ตกลับยกเลิกบริการเพียงแค่ร้อยละ 6.05 และสำหรับกลุ่มที่ใช้บริการต่อ มีสัดส่วนลูกค้าอินเตอร์เน็ตแบบ Fiber Optic กับ DSL ไม่ได้ต่างกันมากนัก



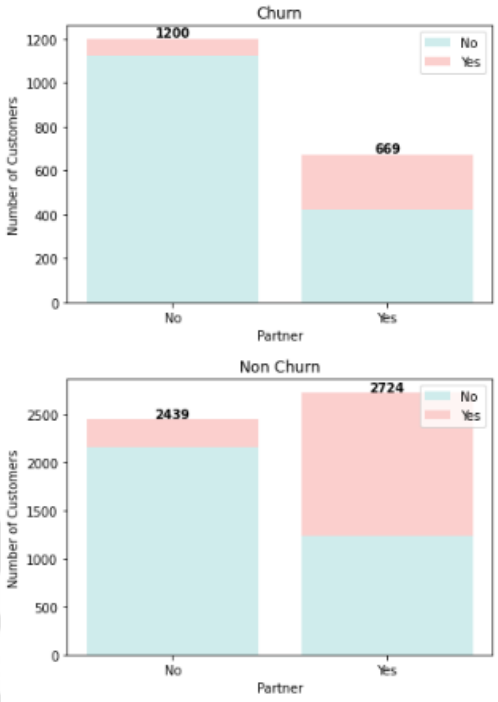
ภาพประกอบ 12 สัดส่วนของผู้ใช้บริการอินเตอร์เน็ต และแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ

จากภาพประกอบ 12 จะเห็นได้ว่าส่วนใหญ่กลุ่มลูกค้าที่ยกเลิกบริการจะเป็นกลุ่มที่เป็นสัญญาระยะสั้น ร้อยละ 88.55 มีแค่ร้อยละ 11.45 เท่านั้นที่เป็นสัญญาระยะยาว ส่วนกลุ่มลูกค้าที่ใช้บริการต่อมีสัดส่วนของสัญญาต่างๆเท่าๆกัน



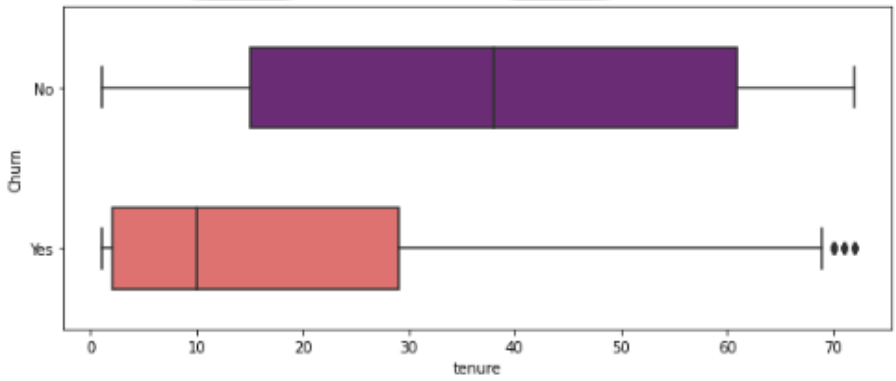
ภาพประกอบ 13 จำนวนลูกค้าแบ่งโดยสถานะการแต่งงานและการอยู่อาศัยร่วมกันกับผู้อื่น

จากภาพประกอบ 13 ลูกค้าที่มีสถานะโสดมีจำนวนเยอะกว่าลูกค้าที่แต่งงานแล้วเพียงเล็กน้อย ในส่วนของลูกค้าที่แต่งงานแล้ว มีสัดส่วนระหว่างการอยู่ร่วมกันกับอยู่คนเดียวใกล้เคียงกัน แต่ในกลุ่มลูกค้าที่โสดนั้นมีสัดส่วนการอยู่ร่วมกันกับผู้อื่นมากกว่าอย่างมาก จากกราฟนี้จะแบ่งลูกค้าออกได้เป็น 4กลุ่มด้วยกันคือ โสดและอยู่คนเดียว โสดแต่อยู่ร่วมกับผู้อื่น แต่งงานแล้วและอยู่ร่วมกันผู้อื่น และแต่งงานแล้วแต่อยู่คนเดียว



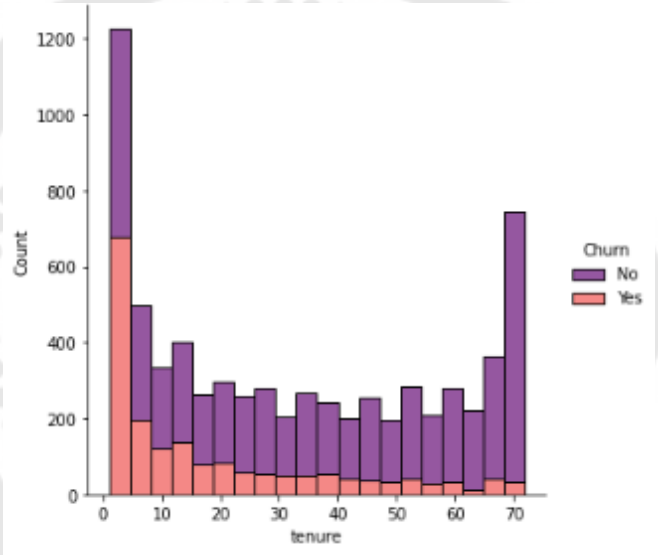
ภาพประกอบ 14 จำนวนลูกค้าแบ่งโดยสถานะการแต่งงานและการอยู่อาศัยร่วมกันกับผู้อื่น และแบ่งเป็นกลุ่มลูกค้าที่ยกเลิกบริการกับกลุ่มลูกค้าที่ใช้บริการต่อ

จากภาพประกอบ 14 จะเห็นได้ว่ากลุ่มลูกค้าที่ยกเลิกการใช้บริการส่วนใหญ่เป็นกลุ่มลูกค้าโสดที่อยู่ร่วมกับผู้อื่น ส่วนกลุ่มลูกค้าที่ใช้บริการมีสัดส่วนระหว่างคนที่แต่งงานงานแล้วกับคนโสดไม่ได้แตกต่างกันมากนัก



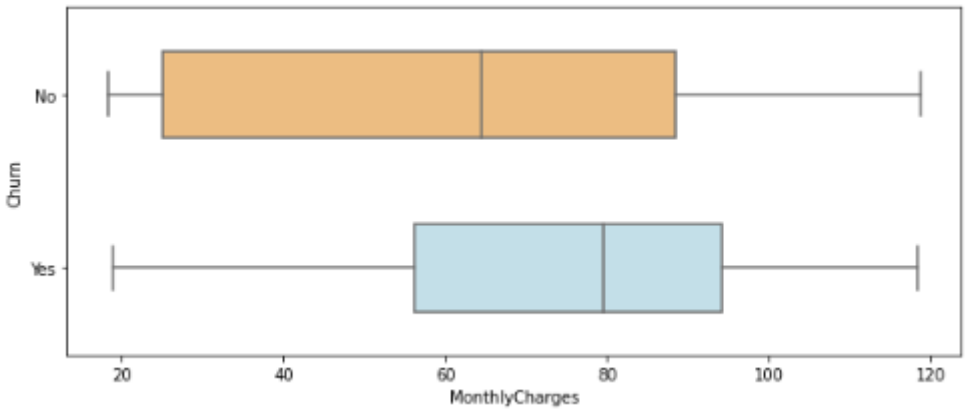
ภาพประกอบ 15 แสดงกราฟ Box Plot ของข้อมูล Tenure หรือระยะเวลาการใช้บริการของลูกค้า

จากภาพประกอบ 15 แสดงให้เห็นว่าค่าเฉลี่ยของระยะเวลาการใช้บริการของลูกค้าที่ยกเลิกการใช้บริการอยู่ที่เพียงแค่ 10 เดือนเท่านั้น และค่อนข้างกระจุกตัวในบริเวณที่ค่าไม่สูงมากนัก ส่วนค่าเฉลี่ยของระยะเวลาการใช้บริการของลูกค้าที่ใช้งานต่ออยู่ที่ 40 เดือนและที่การกระจายตัวของข้อมูลที่กว้างกว่า พอที่จะอธิบายได้ว่า ลูกค้าที่ยกเลิกการใช้บริการส่วนใหญ่เป็นลูกค้าที่มีระยะเวลาการใช้บริการมาไม่นาน แต่ว่ามีข้อมูล Outliers หรือ ค่าผิดปกติ หมายถึง จุดข้อมูลที่แตกต่างอย่างมากจากจุดข้อมูลอื่นๆ ในชุดข้อมูลกลุ่มลูกค้าที่ยกเลิกการใช้บริการ ซึ่งกลุ่มนี้มีความสำคัญเนื่องจากเป็นกลุ่มลูกค้าที่มีระยะเวลาในการใช้บริการมาเป็นเวลานาน อาจต้องมีการสอบถามลูกค้าว่าเหตุใดถึงยกเลิกการใช้บริการเพื่อเก็บข้อมูลมาพัฒนาสินค้าและบริการต่อไป



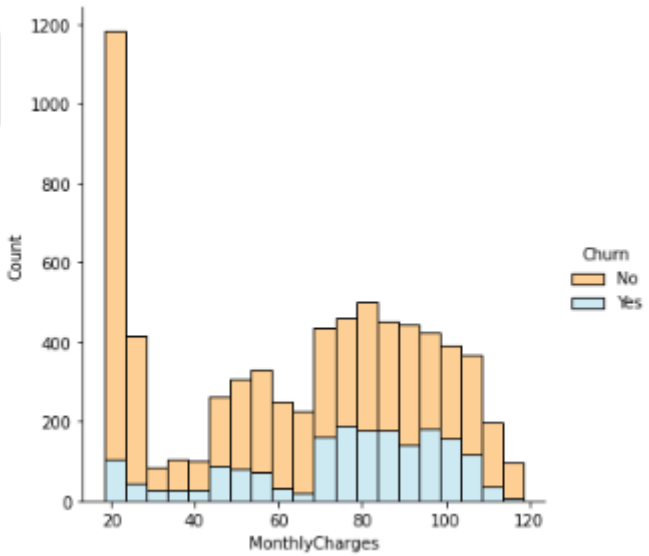
ภาพประกอบ 16 กราฟ Histogram ของข้อมูล Tenure หรือระยะเวลาการใช้บริการของลูกค้า

จากภาพประกอบ 16 จะเห็นได้ว่าระยะเวลาการใช้บริการของลูกค้าที่ยกเลิกการใช้บริการค่อนข้างกระจุกตัวในบริเวณที่ค่าต่ำกว่า 10 เดือน แสดงว่ายิ่งระยะเวลาที่ลูกค้าใช้บริการนานก็มีโอกาสที่จะใช้บริการต่อมากกว่า



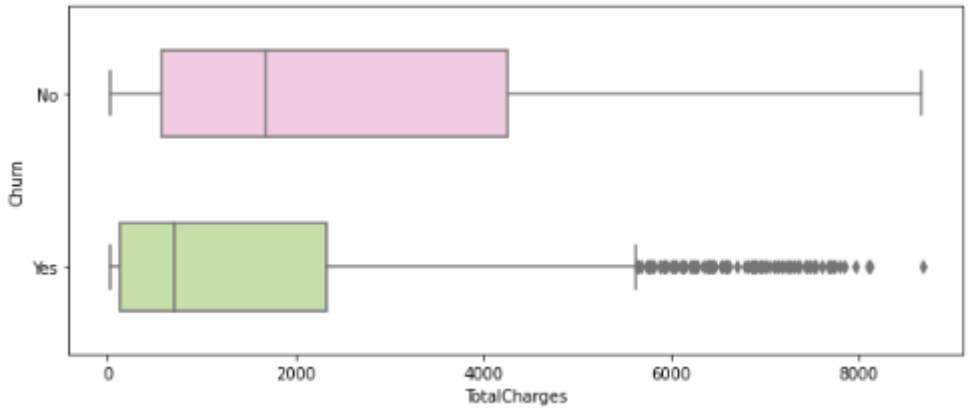
ภาพประกอบ 17 กราฟ Box Plot ของข้อมูล MonthlyCharges หรือข้อมูลค่าใช้จ่ายต่อเดือนของลูกค้า

จากภาพประกอบ 17 จะเห็นได้ว่าค่าเฉลี่ยของค่าใช้จ่ายต่อเดือนของลูกค้าที่ยกเลิกการใช้บริการอยู่ที่ประมาณ 80 USD ซึ่งสูงกว่าค่าเฉลี่ยของค่าใช้จ่ายต่อเดือนของลูกค้าที่ใช้บริการต่อที่จะอยู่ที่ประมาณ 60 USD กว่าๆ อาจสรุปว่าลูกค้าที่มีค่าใช้จ่ายต่อเดือนเยอะมีแนวโน้มที่จะยกเลิกการใช้บริการมากกว่ากลุ่มที่มีค่าใช้จ่ายต่อเดือนน้อย ซึ่งก็มีประเด็นที่น่าสนใจที่ควรนำไปศึกษาต่อว่าทำไมลูกค้าที่มีค่าใช้จ่ายต่อเดือนเยอะจึงยกเลิกการใช้บริการมากกว่า



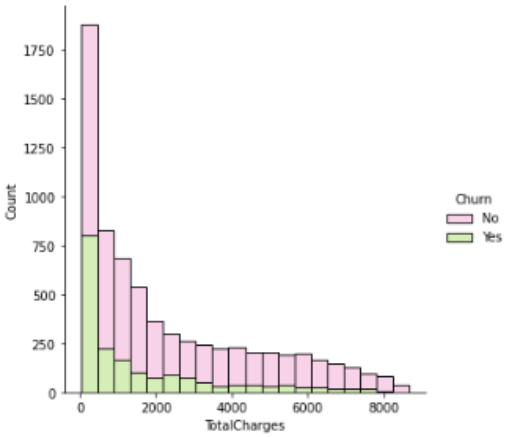
ภาพประกอบ 18 กราฟ Histogram ของข้อมูล MonthlyCharges หรือข้อมูลค่าใช้จ่ายต่อเดือนของลูกค้า

จากภาพประกอบ 18 จะเห็นได้ว่าข้อมูลไม่ได้กระจายตัวอย่างเท่ากัน ส่วนใหญ่ค่าใช้จ่ายรายเดือนของลูกค้าที่ใช้บริการต่อจ่ายน้อยกว่าหรือเท่ากับ 20USD



ภาพประกอบ 19 กราฟ Box Plot ของข้อมูล TotalCharges หรือ ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้า

จากภาพประกอบ 19 จะเห็นได้ว่าค่าเฉลี่ยของค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้าที่ยกเลิกการใช้บริการอยู่ที่ 500 USD ซึ่งน้อยกว่าค่าเฉลี่ยของค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้าที่ใช้บริการต่ออยู่ที่ 1,800 USD และมีการกระจายตัวของข้อมูลแคบกว่า แต่ชุดข้อมูลของลูกค้าที่ยกเลิกการใช้บริการมี Outliers หรือ ค่าผิดปกติ หมายถึง จุดข้อมูลที่แตกต่างอย่างมากจากจุดข้อมูลอื่นๆ เป็นจำนวนมาก ซึ่งมีค่าใช้จ่ายรวมทั้งหมดมากกว่า 6,000 USD ซึ่งกลุ่มนี้มีความสำคัญเนื่องจากเป็นกลุ่มลูกค้าที่มีการใช้จ่ายค่อนข้างเยอะ อาจต้องมีการสอบถามลูกค้าว่าเหตุใดถึงยกเลิกการใช้บริการเพื่อเก็บข้อมูลมาพัฒนาสินค้าและบริการต่อไป



ภาพประกอบ 20 กราฟ Histogram ของข้อมูล TotalCharges หรือ ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้า

จากภาพประกอบ 20 จะเห็นได้ว่ายิ่งค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้ายิ่งน้อยเท่าไหร่ จำนวนคนที่ยกเลิกใช้บริการยิ่งสูงมากขึ้น ซึ่งจะสอดคล้องกับระยะเวลาการใช้งาน เนื่องจากว่า TotalCharges คือ ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้า

3.4 การเตรียมข้อมูล (Data Pre-processing)

ผู้วิจัยได้ทำการเข้ารหัสตัวอักษรให้เป็นตัวเลขโดยใช้เทคนิค One-Hot Encoding คือ เทคนิคที่ใช้แปลงข้อมูลประเภท Categorical ให้เป็นรูปแบบ Binary Vector ที่สามารถใช้งานกับแบบจำลองได้ง่ายขึ้น เนื่องจากแบบจำลองมักทำงานกับข้อมูลเชิงตัวเลข ช่วยให้แบบจำลองเข้าใจและประมวลผลข้อมูลประเภท Categorical ได้อย่างมีประสิทธิภาพ และยังได้ทำการปรับสเกลของข้อมูลโดยใช้เทคนิค Standard Scaler เป็นเทคนิคที่ใช้ปรับขนาดข้อมูลเชิงตัวเลข (Numerical data) ในแบบจำลอง ให้มีค่าเฉลี่ย (Mean) อยู่ที่ 0 และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) อยู่ที่ 1 เพื่อช่วยให้แบบจำลองเรียนรู้ได้เร็วขึ้นและมีประสิทธิภาพดีขึ้น

3.5 การสร้างแบบจำลองเพื่อทำการทำนายและหา Feature Importance

ผู้วิจัยได้สร้างแบบจำลองทั้งหมด 6 แบบ ได้แก่ 1. แบบจำลอง Logistic Regression 2. แบบจำลอง Naive Bayes 3. แบบจำลอง KNN 4. แบบจำลอง Decision Tree 5. แบบจำลอง Random Forest และ 6. แบบจำลอง XGBoost ต่อจากนั้นได้ทำการหา Feature Importance ของแบบจำลองทั้ง 3 ได้แก่ แบบจำลอง Logistic Regression 2. แบบจำลอง Random Forest และ 3. แบบจำลอง XGBoost

บทที่ 4
ผลการดำเนินงานวิจัย

การวิจัยนี้ เป็นงานวิจัยในการศึกษาการพัฒนาแบบจำลองเพื่อวิเคราะห์แนวโน้มการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม โดยอาศัยการเรียนรู้ของเครื่องมือ ที่ช่วยในการวิเคราะห์และตัดสินใจในการวางแผนกลยุทธ์ที่จะช่วยแก้ปัญหาได้อย่างตรงประเด็น รวมถึง ศึกษาปัจจัยในการหรือคุณลักษณะที่บ่งชี้ว่า ลูกค้าจะเลิกใช้บริการของบริษัท และ ค้นหาแบบจำลองที่เหมาะสมต่อการนำมาใช้ทำนายการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามขบวนการและขั้นตอนต่างๆ รวมไปถึงการประเมินผลการทดลองเพื่อใช้ในการพิจารณา แนวโน้มการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม

1. ผลลัพธ์ของการเตรียมข้อมูล
2. ผลลัพธ์ของการพัฒนาแบบจำลอง

4.1 ผลลัพธ์ของการเตรียมข้อมูล

การตรวจสอบและทำความสะอาดข้อมูล ข้อมูลลูกค้าบริษัทให้บริการสัญญาณโทรศัพท์ ประกอบด้วยข้อมูลจำนวนทั้งหมด 7,043 แถว และ 21 คอลัมน์ จากข้อมูลสาธารณะบนเว็บไซต์ www.kaggle.com โดยมีการแบ่งข้อมูลออกเป็น 2ชุด คือ กลุ่มลูกค้าที่ยกเลิกบริการ และลูกค้าที่ใช้บริการต่อ การตรวจสอบข้อมูลเพื่อหาข้อมูลที่หายไป และทำความสะอาดข้อมูลโดยการลบหรือแทนค่าข้อมูลที่หายไป เช่น ในกรณีของคอลัมน์ "TotalCharges" ที่มีค่าว่าง การลบข้อมูลที่หายไปนี้จะทำให้สะอาดและเตรียมข้อมูลได้ดีขึ้น

โดยผู้วิจัยได้ทำการแปลงข้อมูล โดยการแปลงชนิดข้อมูลให้เหมาะสมกับการวิเคราะห์ ในกรณีนี้การเปลี่ยนชนิดของข้อมูลในคอลัมน์ "TotalCharges" จาก Object เป็น Float64 ทำให้สามารถนำไปใช้ในการคำนวณและวิเคราะห์ได้ถูกต้อง และวิเคราะห์ความสัมพันธ์ของข้อมูลด้วยการใช้เทคนิค Correlation Coefficient พบว่า Tenure มีความสัมพันธ์กับ TotalCharges สูงสุด นั่นคือลูกค้าที่ใช้บริการมานาน ก็จะมีแนวโน้ม ใช้บริการมากทำให้ค่าใช้จ่ายมาก และอีกหนึ่งความสัมพันธ์ที่พบคือ Tenure มีความสัมพันธ์กับ Contract ลูกค้าที่ใช้บริการมานานก็จะมีแนวโน้มใช้บริการผูกสัญญาระยะยาวมากกว่า การระบุความสัมพันธ์นี้ช่วยให้บริษัทสามารถดำเนินการปรับแต่งราคาและบริการให้เหมาะสมกับความต้องการของลูกค้าได้

จากนั้นผู้วิจัยได้แสดงสัดส่วนของข้อมูลในรูปแบบกราฟ หรือตารางสรุป ช่วยให้ผู้ใช้งานสามารถเข้าใจและวิเคราะห์ข้อมูลได้ง่ายขึ้น โดยในที่นี้การแสดงสัดส่วนของเพศและอายุของ

ลูกค้าเป็นต้น ช่วยให้ผู้ใช้งานเห็นภาพรวมของลูกค้าที่ให้บริการ การวิเคราะห์จำนวนผู้ใช้บริการโดยนับจำนวนผู้ใช้บริการในแต่ละกลุ่ม เช่น ผู้ใช้บริการโทรศัพท์และอินเตอร์เน็ต ช่วยให้บริษัททราบถึงกลุ่มลูกค้าที่ให้บริการมากที่สุด และให้มีการวางแผนการบริหารจัดการทรัพยากรอย่างเหมาะสม และพบข้อสังเกตที่น่าสนใจดังนี้

ลูกค้าที่ยกเลิกการใช้บริการส่วนใหญ่เป็นลูกค้าที่มีระยะเวลาการใช้บริการมาไม่นาน

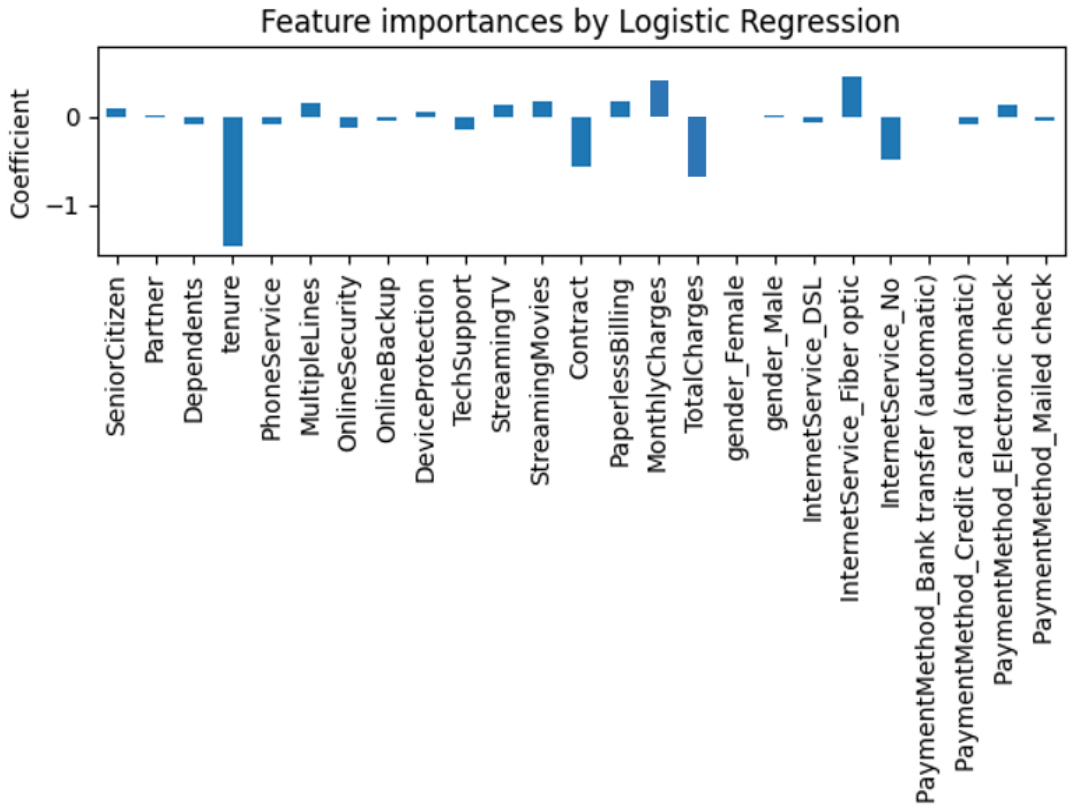
ยิ่งระยะเวลาที่ลูกค้าใช้บริการนานก็มีโอกาสที่จะใช้บริการต่อมากกว่า

ค่าเฉลี่ยของค่าใช้จ่ายต่อเดือนของลูกค้าที่ยกเลิกการใช้บริการสูงกว่าค่าเฉลี่ยของค่าใช้จ่ายต่อเดือนของลูกค้าที่ใช้บริการต่อ

ลูกค้าที่มีค่าใช้จ่ายต่อเดือนเยอะมีแนวโน้มที่จะยกเลิกการใช้บริการมากกว่า

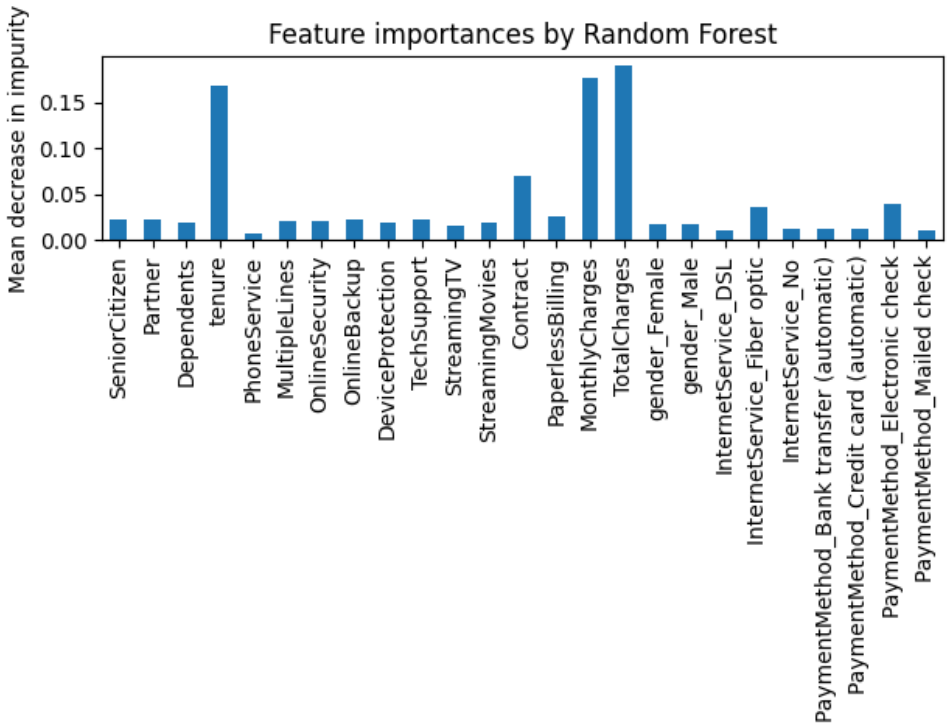
ค่าเฉลี่ยของค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้าที่ยกเลิกการใช้บริการน้อยกว่าค่าเฉลี่ยของค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้าที่ใช้บริการต่อ

ยิ่งค่าใช้จ่ายน้อยเท่าไหร่ จำนวนคนที่ยกเลิกใช้บริการยิ่งสูงมากขึ้น



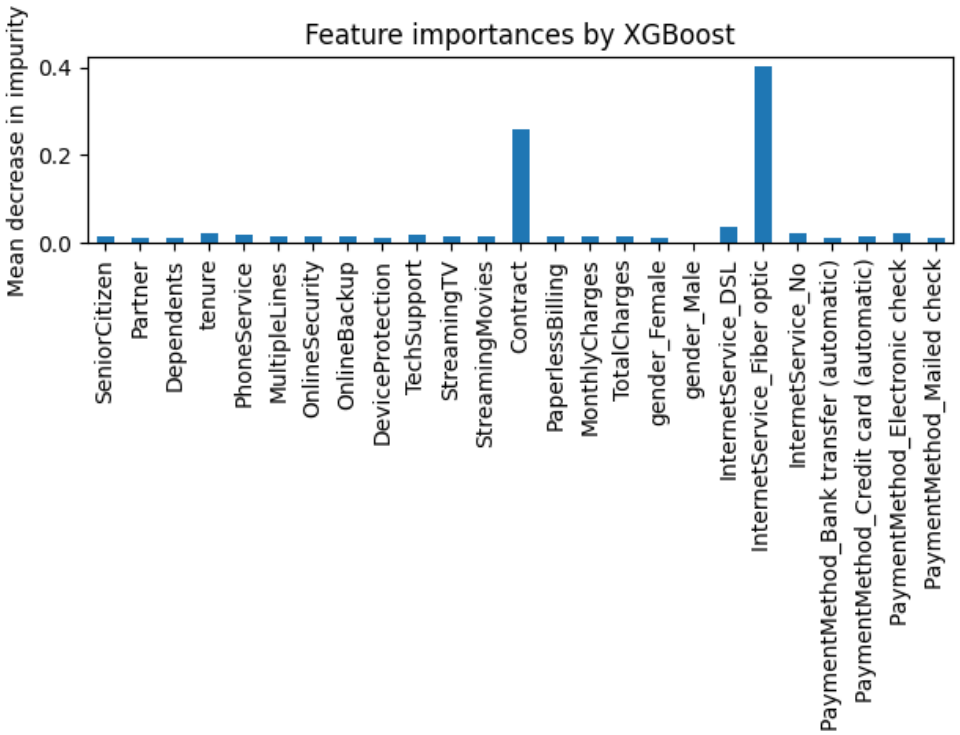
ภาพประกอบ 21 Best feature ของแบบจำลอง Logistic Regression

จากภาพประกอบ 21 แสดงให้เห็นถึงปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 3อันดับ ได้แก่ Tenure, TotalCharges และ Contract โดยทั้งหมดอยู่ฝั่งติดลบจะแสดงถึงความสัมพันธ์ที่แปรผกผันกัน กล่าวคือ ยิ่งค่าของ Tenure มากเท่าไหร่ อัตราการยกเลิกใช้บริการก็จะยิ่งน้อยลง และ TotalCharges ที่มากขึ้น ก็จะทำให้อัตราการยกเลิกการใช้บริการน้อยลง เช่นเดียวกันกับ Contract ยิ่งอยู่นานก็จะยิ่งมีอัตราการยกเลิกการใช้บริการน้อยลง



ภาพประกอบ 22 Best feature ของแบบจำลอง Random Forest

จากภาพประกอบ 22 แสดงให้เห็นถึงปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 3อันดับ ได้แก่ Tenure, MonthlyCharges และ TotalCharges ซึ่งแบบจำลอง Random Forest จะไม่ได้บอกถึงทิศทางความสัมพันธ์ว่าเป็นแปรผันตามหรือแปรผกผัน เหมือนกับการหา Best Feature ของแบบจำลอง Logistic Regression โดยสรุปคือถ้า Tenure, MonthlyCharges และ TotalCharges มีค่าสูง ก็จะส่งผลถึงการยกเลิกการบริการสูงแต่บอกไม่ได้ว่าส่งผลไปในทิศทางไหน อาจจะต้องดูการกระจายตัวของข้อมูลเพิ่มเติมเพื่อช่วยในการวิเคราะห์ต่อไป

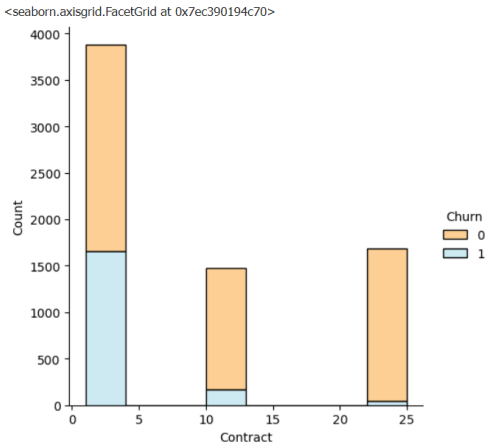


ภาพประกอบ 23 Best feature ของแบบจำลอง XGBoost

จากภาพประกอบ 23 แสดงให้เห็นถึงปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 2อันดับ ได้แก่ Internet Service Fiber Optic และ Contract ซึ่งจะไม่ได้บอกถึงทิศทางความสัมพันธ์ว่าเป็นแปรผันตามหรือแปรผกผัน อาจเป็นไปได้ว่าผู้ที่ใช้บริการ Internet Service Fiber Optic มากจะยกเลิกการใช้บริการสูง หรือ ผู้ที่ใช้บริการ Internet Service Fiber Optic น้อยจะยกเลิกการใช้บริการสูงก็ได้

สรุปจากการหา Best Feature ของทั้ง3วิธี มีปัจจัยสำคัญที่ซ้ำกันทั้ง3วิธี คือ Contract และซ้ำกัน 2 วิธีขึ้นไปคือ Tenure และ TotalCharge แสดงว่า ทั้ง Contract, Tenure และ TotalCharges เป็นปัจจัยหลักที่สำคัญในการยกเลิกการใช้บริการของลูกค้า

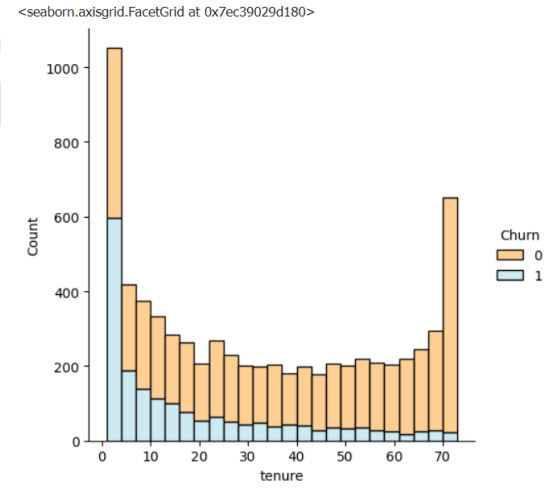
ต่อมาเราจะมาดูการกระจายตัวของลูกค้าที่ยกเลิกบริการจากทั้ง 3 Feature นี้



ภาพประกอบ 24 กราฟ Histogram ของข้อมูล Contract หรือสัญญาการให้บริการ

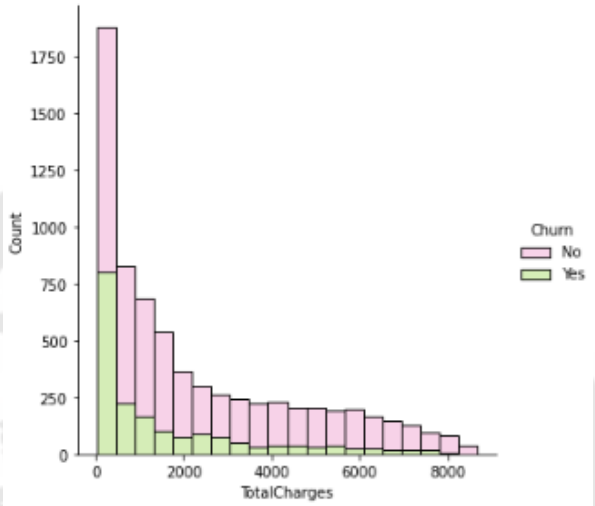
จากภาพประกอบ 24 จะเห็นได้ว่าข้อมูลที่อยู่ในช่วง 0-5 หรือสัญญาระยะสั้น มีการยกเลิกบริการค่อนข้างสูง แต่ในขณะเดียวกันคนที่ไม่ยกเลิกบริการก็มีปริมาณที่ไม่ต่างจากกันมาก และในช่วง 10-15 หรือสัญญาระยะกลาง ในช่วงนี้ลูกค้าส่วนใหญ่จะไม่ยกเลิกการใช้บริการ และสุดท้ายในช่วง 20-25 หรือสัญญาระยะยาว ในช่วงนี้แทบจะไม่มีการยกเลิกสัญญาเลย

โดยสรุป ยิ่งระยะเวลาสัญญามาก อัตราลูกค้าที่ยกเลิกบริการก็จะยิ่งต่ำ ส่วนถ้าระยะเวลาสัญญาสั้น อัตราที่ลูกค้าจะยกเลิกบริการจะยิ่งสูง



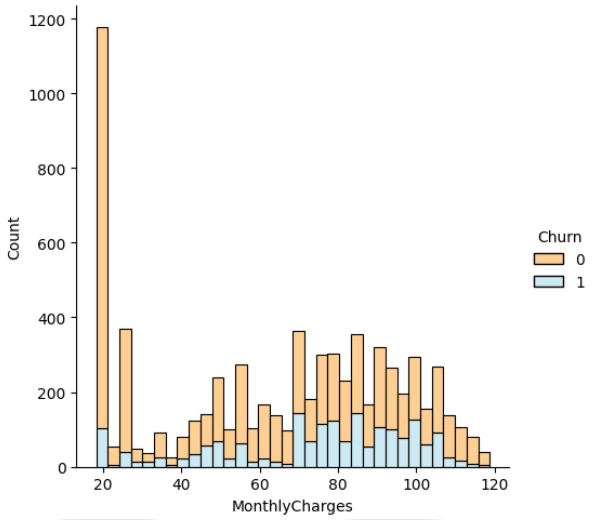
ภาพประกอบ 25 กราฟ Histogram ของข้อมูล Tenure หรือระยะเวลาการใช้บริการของลูกค้า

จากภาพประกอบ 25 จะเห็นได้ว่าจำนวนที่ผู้ใช้บริการยกเลิกสูงที่สุดคือช่วงที่ผู้ใช้งานใช้บริการมายังไม่นานมากคืออยู่ในช่วงระหว่าง 0-10 เมื่อสังเกตว่ายิ่งค่า Tenure เยอะ จำนวนผู้ใช้บริการที่ยกเลิกก็จะยิ่งลดลง โดยสรุปคือ จำนวนระยะเวลาที่ผู้ใช้งานมีผลในทางตรงกันข้ามกับอัตราการยกเลิกการใช้บริการ กล่าวคือยิ่งจำนวนเวลาที่ใช้งานน้อยเท่าไหร่ ผู้ใช้งานก็จะยิ่งมีโอกาสยกเลิกสูงเท่านั้น และยิ่งระยะเวลาที่ลูกค้าใช้บริการนานก็มีโอกาสที่จะใช้บริการต่อมากกว่า



ภาพประกอบ 26 กราฟ Histogram ของข้อมูล TotalCharges หรือ ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้า

จากภาพประกอบ 26 จะเห็นได้ว่าช่วงที่ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาน้อยๆ มีจำนวนคนที่ยกเลิกใช้บริการสูงมาก ยิ่งค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้ายิ่งน้อยเท่าไหร่ จำนวนคนที่ยกเลิกใช้บริการยิ่งสูงมากขึ้น ซึ่งจะสอดคล้องกับระยะเวลาการใช้งาน เนื่องจากว่า TotalCharges คือ ค่าใช้จ่ายทั้งหมดตลอดระยะสัญญาของลูกค้า



ภาพประกอบ 27 กราฟ Histogram ของข้อมูล MonthlyCharges หรือ ค่าใช้รายเดือนของลูกค้า

จากภาพประกอบ 27 จะเห็นได้ว่ากลุ่มลูกค้าที่มีค่าใช้จ่ายต่อเดือนน้อยกว่าหรือเท่ากับ 20 USD ต่อเดือน มีการใช้บริการต่อในจำนวนที่ค่อนข้างสูง

	Churn	Contract	tenure	MonthlyCharges	TotalCharges
Churn	1.000000	-0.393888	-0.354049	0.192858	-0.199484
Contract	-0.393888	1.000000	0.675048	-0.073124	0.448418
tenure	-0.354049	0.675048	1.000000	0.246862	0.825880
MonthlyCharges	0.192858	-0.073124	0.246862	1.000000	0.651065
TotalCharges	-0.199484	0.448418	0.825880	0.651065	1.000000

ภาพประกอบ 28 ค่าสัมประสิทธิ์สหสัมพันธ์ของ Best Feature กับ Target

จากภาพประกอบ 28 จะเห็นได้ว่าตัวแปร MonthlyCharges มีความสัมพันธ์กับ Churn สูงสุดโดยมีค่า สัมประสิทธิ์สหสัมพันธ์อยู่ที่ 0.19 โดยตัวแปรทั้งสองนี้ จะมีความสัมพันธ์ตามกัน นั่นคือ ลูกค้าที่มีค่าใช้จ่ายต่อเดือนน้อยก็จะมีการยกเลิกบริการที่น้อยตามไปด้วย ในขณะที่ Contract มีความสัมพันธ์กับ Churn สูงสุดในด้านตรงกันข้ามโดยมีค่าสัมประสิทธิ์สหสัมพันธ์อยู่ที่ -0.39 คือลูกค้าที่ใช้บริการสัญญาระยะสั้นจะมีการยกเลิกการใช้บริการสูง

ผู้วิจัยนำข้อมูลไปประยุกต์ใช้ในการบริหารจัดการ โดยหลังจากที่ทำความเข้าใจข้อมูลและมีความเสร็จสมบูรณ์ในกระบวนการเตรียมข้อมูล บริษัทสามารถนำข้อมูลที่ผ่านการวิเคราะห์ไปใช้ในการบริหารจัดการทรัพยากร สร้างกลยุทธ์การตลาด ปรับปรุงการบริการและการสร้างสินค้าใหม่เพื่อตอบสนองความต้องการของลูกค้าและเพิ่มประสิทธิภาพในการแข่งขันในตลาด ผู้วิจัยแนะนำว่าหลังจากนำข้อมูลไปใช้งานในการดำเนินธุรกิจ บริษัทควรติดตามและประเมินผลการดำเนินงานที่มาจากการนำข้อมูลไปใช้งาน นำผลตอบรับจากลูกค้าและข้อมูลการดำเนินงานเข้ามาวิเคราะห์เพื่อให้สามารถปรับปรุงและพัฒนากลยุทธ์การบริหารจัดการให้มีประสิทธิภาพและเหมาะสมกับความต้องการของลูกค้าต่อไป

4.2 ผลลัพธ์ของการพัฒนาแบบจำลอง

ผลการทำงานเบื้องต้นของแบบจำลอง

ประสิทธิภาพของแบบจำลองต่างๆที่จะถูกนำมาใช้ในการศึกษานี้ แสดงไว้ดังต่อไปนี้

จากการทดสอบประสิทธิภาพของการพัฒนาแบบจำลอง เพื่อใช้ในการทำนายการยกเลิกใช้บริการของลูกค้าจะได้ผลลัพธ์ตามตาราง โดยผู้วิจัยพบว่า แบบจำลอง Logistic Regression ให้ประสิทธิภาพการทำงานสูงสุด

	Model	Accuracy(%)
0	Logistic Regression	80.739161
1	Naive Bayes	75.692964
2	KNN	77.540867
3	Decision Tree	71.357498
4	Random Forest	78.820185
5	XGBoost	78.891258

ภาพประกอบ 29 การทดสอบประสิทธิภาพของการพัฒนาแบบจำลอง

จากภาพประกอบ 29 จะเห็นได้ว่าแบบจำลอง Logistic Regression ให้ผลลัพธ์ที่ดีที่สุดในแง่ของ Accuracy แบบจำลอง XGBoost มีประสิทธิภาพรองลงมา และแบบจำลอง Decision Tree มีประสิทธิภาพต่ำสุด

การทดสอบประสิทธิภาพของแบบจำลองเป็นขั้นตอนสำคัญในการพัฒนาแบบจำลอง เพื่อประเมินความแม่นยำและประสิทธิภาพของแบบจำลองก่อนนำไปใช้งานจริง ในการทดสอบนี้ ผู้วิจัยได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองต่างๆในการทำนายการยกเลิกการใช้บริการ

ผู้วิจัยใช้ชุดข้อมูลที่มีข้อมูลเกี่ยวกับลูกค้าและพฤติกรรมการใช้บริการ เพื่อฝึกแบบจำลองทั้ง 3 ประเภท ผู้วิจัยแบ่งชุดข้อมูลออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกอบรม (Training Set) : 80% ของชุดข้อมูลทั้งหมด และชุดข้อมูลทดสอบ (Test Set) : 20% ของชุดข้อมูลทั้งหมด

ผู้วิจัยใช้ตัวชี้วัดประสิทธิภาพดังต่อไปนี้เพื่อเปรียบเทียบแบบจำลอง

Accuracy เป็นอัตราความแม่นยำโดยรวม โดยคิดจากจำนวนที่ทำนายถูกทั้งหมดการด้วยจำนวนรายการทั้งหมด

Precision เป็นอัตราความแม่นยำ เป็นการวัดว่าแบบจำลองทำนายผลบวกได้ถูกต้องกี่เปอร์เซ็นต์ จากทั้งหมดที่แบบจำลองทำนายว่าเป็นผลบวก

Recall เป็นการวัดว่าแบบจำลองทำนายผลบวกได้ครบถ้วนกี่เปอร์เซ็นต์ จากทั้งหมดที่มีผลบวกจริง

F1-Score เป็นค่าที่แสดงถึงระดับความสอดคล้องกันระหว่าง Precision และ Recall ซึ่งถ้าค่า F1-score สูง แสดงว่าแบบจำลองมีทั้ง Precision และ Recall สูง

สรุปแบบจำลอง Logistic Regression ให้ประสิทธิภาพสูงสุดในการทำนายการยกเลิกการใช้บริการ แบบจำลอง XGBoost มีประสิทธิภาพรองลงมา และแบบจำลอง Decision Tree มีประสิทธิภาพต่ำสุด ผลลัพธ์จากการทดสอบนี้ ช่วยให้เลือกแบบจำลองที่เหมาะสมสำหรับการใช้งานจริง

Accuracy of logistic regression classifier on test set: 0.81

	precision	recall	f1-score	support
0	0.85	0.90	0.87	1038
1	0.66	0.55	0.60	369
accuracy			0.81	1407
macro avg	0.75	0.72	0.74	1407
weighted avg	0.80	0.81	0.80	1407

ภาพประกอบ 30 ตัวชี้วัดประสิทธิภาพของแบบจำลอง Logistic Regression

จากภาพประกอบ 30 ค่าชี้วัดประสิทธิภาพที่ให้มา แสดงว่าแบบจำลอง Logistic Regression นี้มีประสิทธิภาพค่อนข้างดี โดยสามารถทำนายผลได้ถูกต้อง 81%

Class 0 เป็นกลุ่มลูกค้าที่ใช้บริการต่อ

Precision = 0.85 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ใช้บริการต่อได้ถูกต้อง 85% จากทั้งหมดที่แบบจำลองทำนายว่าเป็นกลุ่มลูกค้าที่ใช้บริการต่อ

Recall = 0.9 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ใช้บริการต่อได้ครบถ้วน 90% จากทั้งหมดที่มีกลุ่มลูกค้าที่ใช้บริการต่อจริง

F1-score = 0.87 หมายความว่าแบบจำลอง มีประสิทธิภาพโดยรวมดี (ค่า F1-score สูง แสดงว่าแบบจำลองมีทั้ง Precision และ Recall สูง)

Class 1 เป็นกลุ่มลูกค้าที่ยกเลิกการใช้บริการ

Precision = 0.66 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้ถูกต้อง 66% จากทั้งหมดที่แบบจำลองทำนายว่าเป็นกลุ่มลูกค้าที่ยกเลิกการใช้บริการ

Recall = 0.55 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้ครบถ้วน 55% จากทั้งหมดที่มีกลุ่มลูกค้าที่ยกเลิกการใช้บริการจริง

F1-score = 0.60 หมายความว่าแบบจำลองมีประสิทธิภาพโดยรวมปานกลาง (ค่า F1-score ต่ำ แสดงว่าแบบจำลองมี Precision หรือ Recall ต่ำ)

Accuracy of Random Forest classifier on test set: 0.79

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1038
1	0.63	0.48	0.54	369
accuracy			0.79	1407
macro avg	0.73	0.69	0.70	1407
weighted avg	0.78	0.79	0.78	1407

ภาพประกอบ 31 ตัวชี้วัดประสิทธิภาพของแบบจำลอง Random Forest

จากภาพประกอบ 31 ค่าชี้วัดประสิทธิภาพที่ให้มา แสดงว่าแบบจำลอง Random Forest นี้มีประสิทธิภาพค่อนข้างดี โดยสามารถทำนายผลได้ถูกต้อง 79%

Class 0 เป็นกลุ่มลูกค้าที่ใช้บริการต่อ

Precision = 0.83 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ใช้บริการต่อได้ถูกต้อง 83% จากทั้งหมดที่แบบจำลองทำนายว่าเป็นกลุ่มลูกค้าที่ใช้บริการต่อ

Recall = 0.9 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ใช้บริการต่อได้ครบถ้วน 90% จากทั้งหมดที่มีกลุ่มลูกค้าที่ใช้บริการต่อจริง

F1-score = 0.86 หมายความว่าแบบจำลอง มีประสิทธิภาพโดยรวมดี (ค่า F1-score สูง แสดงว่าแบบจำลองมีทั้ง Precision และ Recall สูง)

Class 1 เป็นกลุ่มลูกค้าที่ยกเลิกการใช้บริการ

Precision = 0.63 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้ถูกต้อง 63% จากทั้งหมดที่แบบจำลองทำนายว่าเป็นกลุ่มลูกค้าที่ยกเลิกการใช้บริการ

Recall = 0.48 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้ครบถ้วน 48% จากทั้งหมดที่มีกลุ่มลูกค้าที่ยกเลิกการใช้บริการจริง

F1-score = 0.54 หมายความว่าแบบจำลองมีประสิทธิภาพโดยรวมค่อนข้างต่ำ (ค่า F1-score ต่ำ แสดงว่าแบบจำลองมี Precision หรือ Recall ต่ำ)

Accuracy of XGBoost classifier on test set: 0.79

	precision	recall	f1-score	support
0	0.84	0.88	0.86	1038
1	0.61	0.53	0.57	369
accuracy			0.79	1407
macro avg	0.73	0.70	0.71	1407
weighted avg	0.78	0.79	0.78	1407

ภาพประกอบ 32 ตัวชี้วัดประสิทธิภาพของแบบจำลอง XGBoost

จากภาพประกอบ 32 ค่าชี้วัดประสิทธิภาพที่ให้มา แสดงว่าแบบจำลอง XGBoost นี้มีประสิทธิภาพค่อนข้างดี โดยสามารถทำนายผลได้ถูกต้อง 79%

Class 0 เป็นกลุ่มลูกค้าที่ใช้บริการต่อ

Precision = 0.84 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ใช้บริการต่อได้ถูกต้อง 84% จากทั้งหมดที่แบบจำลองทำนายว่าเป็นกลุ่มลูกค้าที่ใช้บริการต่อ

Recall = 0.88 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ใช้บริการต่อได้ครบถ้วน 88% จากทั้งหมดที่มีกลุ่มลูกค้าที่ใช้บริการต่อจริง

F1-score = 0.86 หมายความว่าแบบจำลอง มีประสิทธิภาพโดยรวมดี (ค่า F1-score สูง แสดงว่าแบบจำลองมีทั้ง Precision และ Recall สูง)

Class 1 เป็นกลุ่มลูกค้าที่ยกเลิกการใช้บริการ

Precision = 0.61 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้ถูกต้อง 61% จากทั้งหมดที่แบบจำลองทำนายว่าเป็นกลุ่มลูกค้าที่ยกเลิกการใช้บริการ

Recall = 0.53 หมายความว่าแบบจำลองทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้ครบถ้วน 53% จากทั้งหมดที่มีกลุ่มลูกค้าที่ยกเลิกการใช้บริการจริง

F1-score = 0.57 หมายความว่าแบบจำลองมีประสิทธิภาพโดยรวมค่อนข้างต่ำ (ค่า F1-score ต่ำ แสดงว่าแบบจำลองมี Precision หรือ Recall ต่ำ)

Class 0 (Non-Churn)	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.81	0.85	0.90	0.87
Random Forest	0.79	0.83	0.90	0.86
XGBoost	0.79	0.84	0.88	0.86

ตาราง *1* ผลตัวชี้วัดประสิทธิภาพของ Class 0 หรือผู้ที่ใช้บริการต่อ
ของแบบจำลองทั้ง 3แบบ

Class 1 (Churn)	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.81	0.66	0.55	0.60
Random Forest	0.79	0.63	0.48	0.54
XGBoost	0.79	0.61	0.53	0.57

ตาราง *2* ผลตัวชี้วัดประสิทธิภาพของ Class 1 หรือผู้ที่ยกเลิกการใช้บริการ
ของแบบจำลองทั้ง 3แบบ

จากตาราง 1 และตามราง 2 จะเห็นได้ว่า ทั้ง 3 แบบจำลองนี้มีประสิทธิภาพดีในการทำนาย Class 0 เนื่องจากมีค่า Precision, Recall และ F1-score สูง แต่แบบจำลองนี้มีประสิทธิภาพปานกลางในการทำนาย Class 1 เนื่องจากมีค่า Precision, Recall และ F1-score ต่ำ ส่วนสาเหตุที่ประสิทธิภาพในการทำนายของ Class 1 ต่ำกว่า Class 0 เนื่องมาจากความไม่สมดุลของข้อมูล หรือ Data Imbalance มักพบในกรณีที่มีข้อมูล Class 0 มากกว่า Class 1 มาก ส่งผลให้แบบจำลองมีแนวโน้มที่จะเรียนรู้ Class 0 มากกว่า Class 1 แบบจำลองจึงมีประสิทธิภาพในการทำนาย Class 0 มากกว่า Class 1

ดังนั้นผู้วิจัยจึงได้มีการเพิ่มการทำ Class Weight เข้าไป การตั้ง Class Weight = balance เป็นเทคนิคที่ใช้ในแบบจำลองการเรียนรู้ของเครื่อง เพื่อจัดการกับปัญหาข้อมูลไม่สมดุล หรือ Data Imbalance ซึ่งปัญหาข้อมูลไม่สมดุล หมายถึง สถานการณ์ที่ชุดข้อมูลการ Train มีจำนวนตัวอย่างในแต่ละคลาสไม่เท่ากัน การตั้ง class weight = balance ช่วยแก้ปัญหานี้โดยกำหนดค่า weight ให้กับแต่ละคลาสคลาสที่มีจำนวนน้อยจะได้รับ weight มากกว่า และคลาสที่มีจำนวนมากจะได้รับ weight น้อยลง ผลลัพธ์คือ แบบจำลองจะให้ความสำคัญกับคลาสที่มีจำนวนน้อยมากขึ้น ได้ผลลัพธ์ดังนี้

Accuracy of logistic regression classifier on test set: 0.75

	precision	recall	f1-score	support
0	0.91	0.73	0.81	1038
1	0.51	0.81	0.63	369
accuracy			0.75	1407
macro avg	0.71	0.77	0.72	1407
weighted avg	0.81	0.75	0.76	1407

ภาพประกอบ *33* ตัวชี้วัดประสิทธิภาพของแบบจำลอง Logistic Regression หลังจากทำ Class Weight

Accuracy of Random Forest classifier on test set: 0.79

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1038
1	0.62	0.47	0.53	369
accuracy			0.79	1407
macro avg	0.72	0.68	0.70	1407
weighted avg	0.77	0.79	0.78	1407

ภาพประกอบ 34 ตัวชี้วัดประสิทธิภาพของแบบจำลอง Random Forest หลังจากทำ Class Weight

	Logistic Regression			
Class 1 (Churn)	Accuracy	Precision	Recall	F1-Score
ก่อนทำ Class Weight	0.81	0.66	0.55	0.60
หลังทำ Class Weight	0.75	0.51	0.81	0.63

ตาราง *3* ตารางเปรียบเทียบประสิทธิภาพของแบบจำลอง *logistic regression* ก่อนและหลังการปรับน้ำหนักข้อมูล

จากตาราง 3 จะเห็นได้ว่า หลังจากการทำ Class Weight แล้ว ค่า Recall เพิ่มขึ้นเป็น 0.81 จากเดิม 0.55 ซึ่งเป็นการแสดงให้เห็นว่า การปรับน้ำหนักข้อมูล เพื่อปรับความสมดุลย์ข้อมูล มีผลทำให้ค่า Recall เพิ่มขึ้น เนื่องจากแบบจำลอง ทำนายกลุ่มลูกค้าเลิกใช้บริการเพิ่มขึ้น หากแต่

จะเป็นการลดความแม่นยำของการทำนายลง ซึ่งเห็นได้จากการที่ ค่า ความแม่นยำของแบบจำลอง ลดลงจาก 0.81 เหลือ 0.75 โดยการวิเคราะห์การทำนายการขอยกเลิกใช้บริการนี้ ควรจะเน้นการพิจารณาที่ค่า Recall เนื่องจากเน้นการทำนายกลุ่มลูกค้าที่ยกเลิกการใช้บริการได้แค่ไหน จากทั้งหมดที่มีกลุ่มลูกค้าที่ยกเลิกการใช้บริการจริง

	Random Forest			
Class 1 (Churn)	Accuracy	Precision	Recall	F1-Score
ก่อนทำ Class Weight	0.79	0.63	0.48	0.54
หลังทำ Class Weight	0.79	0.62	0.47	0.53

ตาราง 4 ตารางเปรียบเทียบประสิทธิภาพของแบบจำลอง *Random forest* ก่อนและหลังการปรับน้ำหนักข้อมูล

จากตาราง 4 จะเห็นได้ว่า หลังจากการทำ Class Weight แล้ว ค่า Recall ไม่ได้เพิ่มขึ้นมากเท่าไหร่ เนื่องจากการทำ Class Weight นี้ไม่ได้มีผลกับแบบจำลอง Random Forest มากนัก ซึ่งแตกต่าง จากแบบจำลอง logistic regression ที่เพิ่มประสิทธิภาพในค่า recall ขึ้นมาอย่างมาก

บทที่5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการทำวิจัย การศึกษาการพัฒนาแบบจำลองเพื่อวิเคราะห์แนวโน้มการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคม ผู้วิจัยได้ทำการประเมินประสิทธิภาพของการพัฒนาแบบจำลองเพื่อวิเคราะห์และตัดสินใจในการวางแผนกลยุทธ์และสรุปผล โดยแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

1. อภิปรายผลการวิจัย
2. ข้อเสนอแนะ

5.1 อภิปรายผลการวิจัย

จากการทดสอบประสิทธิภาพของการพัฒนาแบบจำลอง เพื่อใช้ในการทำนายการยกเลิกใช้บริการของลูกค้าจะได้ผลลัพธ์ตามตาราง โดยผู้วิจัยพบว่า แบบจำลอง random forest ให้ประสิทธิภาพการทำงานสูงสุดในทุกด้านของดรรชนีตัววัด

	Model	Accuracy(%)
0	Logistic Regression	80.739161
1	Naive Bayes	75.692964
2	KNN	77.540867
3	Decision Tree	71.357498
4	Random Forest	78.820185
5	XGBoost	78.891258

ภาพประกอบ 35 การทดสอบประสิทธิภาพของการพัฒนาแบบจำลอง

จากภาพประกอบ 35 การเปรียบเทียบและอภิปรายผลของ Feature Importance ในบรรดาแบบจำลองแบบจำแนกประเภทที่ผู้วิจัยกล่าวถึง ได้แก่ Logistic Regression, Naïve Bayes, KNN, Decision Tree, Random Forest, XGBoost อภิปรายผลร่วมกับค่า Accuracy ในการประเมินประสิทธิภาพของแบบจำลองต่อกัน ควรพิจารณาคุณลักษณะเฉพาะของแต่ละแบบจำลองตลอดจากค่า Accuracy และ Feature Importance ดังนี้

1. Logistic Regression ให้ Feature Importance เป็นค่าจำนวนจริงที่บ่งบอกถึงความสำคัญของคุณลักษณะในการทำนาย โดยแบบจำลองนี้มีความแม่นยำสูงที่สุด ที่ 80.74% และสามารถจำแนกข้อมูลได้ โดยเฉพาะถ้าคุณลักษณะที่สำคัญถูกใช้ในแบบจำลองนี้
2. Naïve Bayes ไม่มีค่า Feature Importance แบบเดียวกันกับแบบจำลองอื่น ๆ มีเนื่องจากวิธีการทำงานของ Naïve Bayes ไม่ได้มีการใช้ค่า Feature Importance แต่ได้ใช้ความน่าจะเป็นของคุณลักษณะที่มีต่อคลาสเป้าหมาย โดยความแม่นยำของ Naïve Bayes นี้อยู่ในระดับปานกลาง โดยมีค่า Accuracy อยู่ที่ 75.69%
3. K-Nearest Neighbors (KNN) เป็นแบบจำลองที่ไม่มีค่า Feature Importance แบบเดียวกันกับแบบจำลองอื่น ๆ เนื่องมาจากการทำนายในแบบจำลอง KNN นี้ได้มีการใช้การค้นหาค่าใกล้เคียงจากข้อมูล Training set โดยที่ค่า Accuracy อยู่ที่ 77.54% แบบจำลอง KNN ได้ใช้หลักการของความคล้ายกันระหว่างข้อมูลในการทำนาย แต่ค่าความแม่นยำก็อาจมีความแปรปรวนเพิ่มมากขึ้นได้ในบางกรณีต่าง ๆ
4. Decision Tree ไม่มีค่า Feature Importance แต่มีความสามารถในการประมาณความสำคัญของคุณลักษณะและจัดลำดับแต่ละคุณลักษณะตามความสำคัญ ความแม่นยำของ Decision Tree โดยจากงานวิจัยนี้มี Accuracy 71.36%
5. Random Forest เป็นแบบจำลองที่มีการใช้ค่า Feature Importance มาจากแต่ละต้นไม้ (Tree) เพื่อนำมาประมาณความสำคัญของคุณลักษณะ โดยมีค่า Accuracy อยู่ที่ 78.82% ซึ่งถือว่ามีระดับค่าความแม่นยำสูงที่สุดในกลุ่มของแบบจำลองทั้งหมดที่ผู้วิจัยกล่าวถึง
6. XGBoost เป็นอัลกอริทึม Gradient Boosting ที่มีความสามารถในการปรับปรุงความแม่นยำของแบบจำลอง มีความแม่นยำสูงและสามารถใช้ Feature Importance เพื่อระบุคุณลักษณะที่สำคัญในการทำนาย โดยมี Accuracy 78.89%

การเลือกใช้แบบจำลองหรือการพิจารณาคุณลักษณะที่สำคัญต้องพิจารณาวัตถุประสงค์ของงานและลักษณะของข้อมูลของคุณ ความแม่นยำและค่า Feature Importance มีความสำคัญแต่องค์ประกอบอื่น ๆ อาจมีผลในการตัดสินใจในการเลือกแบบจำลองที่เหมาะสมสำหรับงาน

จากผลการทดลองที่ได้รับ แบบจำลอง Logistic Regression คือแบบจำลองที่มีประสิทธิภาพในการทำนายที่ดีที่สุด มีความแม่นยำ (Accuracy) สูงสุดที่ประมาณ 80.74% ซึ่งเป็นแบบจำลองที่เหมาะสำหรับใช้ในการทำนายลูกค้าที่อาจจะยกเลิกบริการของบริษัท อันดับต่อมาคือ XGBoost และ Random Forest ซึ่งมีประสิทธิภาพในการทำนายความถูกต้องอยู่ที่ประมาณ 78.89% และ 78.82 ตามลำดับ ส่วนแบบจำลอง Decision Tree มีประสิทธิภาพในการทำนายความถูกต้องอยู่ที่ประมาณ 71.36% ซึ่งต่ำที่สุด

ปัจจัยที่มีผลต่อการยกเลิกใช้บริการมากที่สุดคือ Contract, Tenure และ Total Charge ตามลำดับ โดยตัวแปรทั้ง 3ตัวนี้ จะมีความสัมพันธ์แบบผกผันกับการยกเลิกการใช้บริการ กล่าวคือ หากลูกค้าผูกสัญญาระยะยาว หรือมีระยะเวลาการใช้งานทั้งหมดมาก หรือมีค่าใช้จ่ายตลอดระยะสัญญามาก จะมีการยกเลิกการใช้บริการที่น้อย แต่หากลูกค้าผูกสัญญาระยะสั้น หรือมีระยะเวลาการใช้งานทั้งหมดน้อย หรือมีค่าใช้จ่ายตลอดระยะสัญญาน้อย จะมีการยกเลิกการใช้บริการที่มากขึ้น

5.2 ข้อเสนอแนะ

ผลลัพธ์ดังกล่าวจะขึ้นอยู่กับวัตถุประสงค์และความต้องการของงานที่กำลังดำเนินการ ความสำคัญของวัตถุประสงค์ , ความสำคัญของความแม่นยำ (Accuracy) และ Feature Importance จะขึ้นอยู่กับวัตถุประสงค์ของงานของผู้วิจัย หากความแม่นยำสำคัญมากกว่าให้กำหนดความสำคัญต่อแบบจำลองที่มีความแม่นยำสูง ๆ อย่าง Random Forest หรือ Decision Tree เป็นต้น

โดยผู้วิจัยได้วิเคราะห์ความแม่นยำ (Accuracy) และความคงที่ ในการจำแนกประเภทและทำนาย, ความแม่นยำและความคงที่ของแบบจำลองมีความสำคัญ ผู้วิจัยแนะนำให้พิจารณาความคงที่และการทดสอบแบบจำลองเพื่อควบคุมการ Overfitting และให้ความสำคัญต่อความแม่นยำในข้อมูลทดสอบและทดสอบแบบจำลอง

ซึ่งความซับซ้อนของแบบจำลอง เป็นปัจจัยสำคัญ ที่อาจทำให้เกิดความล่าช้าของงานได้ โดยผู้วิจัยขอแนะนำแบบจำลองที่มีความเร็วและความง่ายในการใช้งานเป็นสิ่งสำคัญ ได้แก่

แบบจำลอง Logistic Regression หรือ Naïve Bayes ซึ่งเป็นตัวเลือกที่ดี เนื่องจากมีความเร็วและเรียบง่ายในการใช้งาน

Feature Importance ทำให้เข้าใจว่าคุณลักษณะใดมีผลสำคัญที่สุดในการทำนายผล โดยค่า Feature Importance จะช่วยในการคัดเลือกคุณลักษณะที่สำคัญที่สุดสำหรับการทำนาย ค่า Feature Importance สูงสุดสามารถช่วยในการคัดเลือกคุณลักษณะที่สำคัญสำหรับการทำนายและลดคุณลักษณะที่ไม่สำคัญ

การปรับปรุงและการทดสอบ ควรพิจารณาการปรับปรุงแบบจำลองตามผลลัพธ์และความคาดหวัง ทดสอบความแม่นยำของแบบจำลองและพิจารณาการใช้ค่า Feature Importance เพื่อปรับปรุงแบบจำลอง

การเปรียบเทียบ ควรทำการทดสอบและเปรียบเทียบหลายแบบจำลองก่อนตัดสินใจเพื่อที่จะเลือกแบบจำลองที่เหมาะสมรวมทั้ง การทดสอบและปรับปรุงเป็นขั้นตอนสำคัญในการพัฒนาแบบจำลอง

การทำนายการย้ายค่ายของลูกค้ามีความสำคัญอย่างยิ่งในยุคที่ตลาดเปลี่ยนแปลงรวดเร็ว ด้วยการมาของแบบจำลอง Machine Learning (ML) ควบคู่กับการตัดสินใจตามข้อมูล (data-driven decision making) ทำให้การทำนายการย้ายค่ายมีแนวโน้มที่จะยิ่งมีความสำคัญมากขึ้น เทคโนโลยีคอมพิวเตอร์และ Deep Learningของโลกยุคใหม่กำลังมีการเติบโตอย่างมหาศาล เทคโนโลยีคอมพิวเตอร์ มีชิ้นส่วนใหม่ๆที่มีประสิทธิภาพมากขึ้นวางจำหน่ายในตลาดอยู่ทุกวัน ซึ่งหมายความว่า พลังการประมวลผลจะเพิ่มขึ้นอย่างมาก ภายในระยะเวลาและพลังงานที่เท่าเดิม คาดการณ์ว่า ในอนาคต แบบจำลอง Deep Learning (DL) จะได้รับการยอมรับและใช้งาน เหมือนกับที่แบบจำลอง Machine Learning (ML) ถูกใช้ในปัจจุบัน สิ่งนี้จะช่วยให้การทำนายต่างๆ แม่นยำและรวดเร็วยิ่งขึ้น ไม่เฉพาะแค่การทำนายการย้ายค่าย แต่ยังรวมไปถึงวัตถุประสงค์อื่นๆ อีกมากมาย แบบจำลอง DL จะช่วยให้สามารถนำข้อมูลต่างๆ มาคำนวณได้มากขึ้น ทำให้การทำนายมีความแม่นยำใกล้เคียงความเป็นจริงมากที่สุด ในอนาคตอาจมีการศึกษาเพื่อปรับปรุงประสิทธิภาพของแบบจำลอง DL ให้ตอบสนองทั้งข้อจำกัดด้านเวลาและมาตรฐานประสิทธิภาพ

บรรณานุกรม

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, 18-19 March 2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. 2016 Symposium on Colossal Data Analysis and Networking (CDAN),

Gaur, A., & Dubey, R. (2018, 28-29 Dec. 2018). Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques. 2018 International Conference on Advanced Computation and Telecommunication (ICACAT),

Malyar, M., Robotyshyn, M. V. M., & Sharkadi, M. (2020, 5-9 Oct. 2020). Churn Prediction Estimation Based on Machine Learning Methods. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC),

Pulkundwar, P., Rudani, K., Rane, O., Shah, C., & Virnodkar, S. (2023, 8-9 Dec. 2023). A Comparison of Machine Learning Algorithms for Customer Churn Prediction. 2023 6th International Conference on Advances in Science and Technology (ICAST),

Srinivasan, R., Rajeswari, D., & Elangovan, G. (2023, 5-7 Jan. 2023). Customer Churn Prediction Using Machine Learning Approaches. 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF),

กอบเกียรติ สระอุบล. (2563). เรียนรู้ *Data Science* และ *AI:Machine Learning* ด้วย *Python*. มีเดียเนทเวิร์ค.

บัญชา ปะสีละตัง. (2563). จัดการและวิเคราะห์ข้อมูลด้วย *Python Data Science*. ซีเอ็ดยูเคชั่น.

บัญชา ปะสีละตัง. (2564). สร้างการเรียนรู้สำหรับ *AI* ด้วย *Python Machine Learning*. ซีเอ็ดยูเคชั่น.

ประวัติผู้เขียน

ชื่อ-สกุล	อัญชิสา สิทธิวิริยะชัย
วัน เดือน ปี เกิด	16 กรกฎาคม 2535
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	พ.ศ. 2556 บัญชีบัณฑิต สาขาวิชาการบัญชี จาก มหาวิทยาลัยธรรมศาสตร์
ที่อยู่ปัจจุบัน	กรุงเทพมหานคร