# การวิเคราะห์แนวโน้มการขอยกเลิกใช้บริการสำหรับลูกค้าบริษัทโทรคมนาคมโดยวิธีการเรียนรู้ของเครื่อง

อัญชิสา สิทธิวิริยะชัย[1*], ศิริสรรพ เหล่าหะเกียรติ[2]

## บทคัดย่อ

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 และศึกษาหลักการการทำงานของการเรียนรู้ด้วยเครื่อง สำหรับการนำมาประยุกต์ใช้ในขั้นตอนการคัดเลือกคุณลักษณะที่ส่งผลต่อการทำนายสูง ผลที่ได้คือปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 3อันดับ ได้แก่ Tenure, Total Charges และ Contract และแบบจำลอง Logistic Regression ให้ผลลัพธ์ที่ดีที่สุดในแง่ของ Accuracy แบบจำลอง XGBoost มีประสิทธิภาพรองลงมา และแบบจำลอง Decision Tree มีประสิทธิภาพต่ำสุด ผู้วิจัยจะนำข้อมูลไปประยุกต์ใช้ในการบริหารจัดการทรัพยากร สร้างกลยุทธ์การตลาด ปรับปรุงการบริการและการสร้างสินค้าใหม่เพื่อตอบสนองความต้องการของลูกค้าและเพิ่มประสิทธิภาพในการแข่งขันในตลาดต่อไป

**คำสำคัญ** : การขอยกเลิกใช้บริการ, บริษัทโทรคมนาคม, การเรียนรู้ของเครื่อง, แบบจำลอง Logistic Regression

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 099-4549628 E-mail address: anchisa.sit@g.swu.ac.th

# Analysis of The Churn Prediction for Telecom Customers Using Machine learning

Anchisa Sittiviriyachai[1*], Sirisup Laohakiat[2]

## Abstract

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. Researchers aim 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 performs 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.

**Keywords:** Churn Prediction, Telecom, Machine Learning, Logistic Regression

---

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

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

* Corresponding author:Tel.: 099-4549628 E-mail address: anchisa.sit@g.swu.ac.th

## บทนำ

ปัจจุบัน มีธุรกิจโทรคมนาคม ที่ให้บริการสัญญาณการสื่อสาร เช่นสัญญาณโทรศัพท์เคลื่อนที่ สัญญาณอินเทอร์เนต และสัญญาณโทรทัศน์ดิจิตอลเป็นต้น เป็นธุรกิจที่มูลค่าการตลาดสูงมาก มีมูลค่าราว 6.1 แสนล้านบาท ธุรกิจบริการระบบโทรศัพท์เคลื่อนที่มีลักษณะของตลาดกึ่งผูกขาดที่มีผู้ประกอบการน้อยราย เป็นธุรกิจที่ต้องใช้เงินลงทุนสูงทั้งการวางโครงข่าย และการลงทุนด้านเทคโนโลยีที่เปลี่ยนแปลงรวดเร็ว ผู้ประกอบการรายใหญ่ที่มีฐานะเงินทุนแข็งแกร่งจึงมีความได้เปรียบและมีอำนาจผูกขาดในตลาด การเข้าสู่ตลาดของผู้ประกอบการรายใหม่จึงนับว่ามีอุปสรรคอยู่มาก อย่างไรก็ตาม ตลาดนี้ก็มีการแข่งขันระหว่างผู้ให้บริการอยู่สูงมา ผู้ให้บริการจึงมีความจำเป็นต้องรักษาฐานลูกค้าของบริษัท พร้อมๆไปกับการค้นหาลูกค้าใหม่ เนื่องจากธุรกิจโทรคมนาคมมีข้อมูลลูกค้าจำนวนมาก การรักษาฐานลูกค้าเดิมมีต้นทุนน้อยกว่าการหาลูกค้าใหม่

เพื่อให้บริษัทเพิ่มความเข้าใจในพฤติกรรมของลูกค้าเพิ่มขึ้น จึงได้มีการนำระบบ การเรียนรู้ของเครื่องมาช่วยในการวิเคราะห์ข้อมูลการใช้งานของลูกค้าอย่างหลากหลาย หนึ่งในบรรดาการใช้งานการเรียนรู้ของเครื่องที่ได้รับความนิยมอย่างสูง ในการวิเคราะห์พฤติกรรมของผู้บริโภคได้แก่ การทำนายความเป็นไปได้ ที่ลูกค้าจะยกเลิกการใช้บริการของบริษัท (churn prediction) ทั้งนี้เนื่องจาก หากเราสามารถทำนายกลุ่มลูกค้า ที่มีแนวโน้มจะยกเลิกการใช้บริการของบริษัท เราสามารถนำเสนอโปรโมชั่นที่เหมาะสม เพื่ออาจรักษาฐานลูกค้ากลุ่มนี้ไว้ได้ นอกจากนั้น ยังช่วยในการออกแบบและพัฒนานโยบายการตลาดของบริษัท เพื่อช่วยเพิ่มประสิทธิภาพในการแข่งขันได้ และการสร้างแบบจำลองเพื่อทำนายการขอยกเลิกใช้บริการสำหรับลูกค้า ร่วมกับการวิเคราะห์คุณลักษณะ ยังเป็นการช่วยให้บริษัทเกิดความเข้าใจในพฤติกรรมของลูกค้าอย่างลึกซึ้งขึ้น อันอาจจะนำไปพัฒนาต่อ เพื่อสร้างแบบจำลองในการทำนายพฤติกรรมโดยรวมของลูกค้า ที่มีเขบเขตกว้างขวางขึ้น เช่น การทำนายความชอบของลูกค้าต่อข้อเสนอทางการตลาดแบบใหม่ๆ เป็นต้น

## วิธีดำเนินการ

**ขั้นตอนที่ 1 :** การเก็บรวบรวมข้อมูล

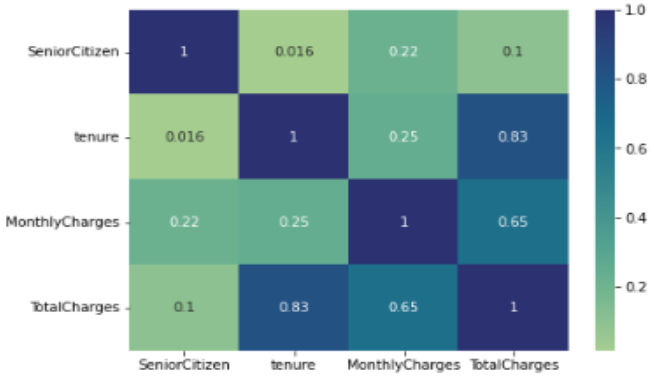
ผู้วิจัยนำข้อมูลลูกค้าบริษัทให้บริการสัญญาณโทรศัพท์ มาวิเคราะห์ในการทำวิจัยครั้งนี้ ประกอบด้วยข้อมูลจำนวนทั้งหมด 7,043 แถว และ 21 คอลัมน์ จากข้อมูลสาธารณะบนเว็บไซต์ www.kaggle.com โดยมีการแบ่งข้อมูลออกเป็น 2ชุด คือ กลุ่มลูกค้าที่ยกเลิกบริการ และลูกค้าที่ใช้บริการต่อ

**ขั้นตอนที่ 2 :** การนำเข้าข้อมูล ตรวจสอบข้อมูล และพิจารณาข้อมูล

ผู้วิจัยใช้ภาษาไพทอนในการวิเคราะห์ข้อมูลและการเรียนรู้ของเครื่อง เริ่มต้นด้วยการนำเข้าโมดูลสำคัญสำหรับการสร้างแบบจำลอง ต่อมานำเข้าไฟล์ข้อมูลและข้อมูลที่ใช้สำหรับสร้างแบบจำลอง เริ่มกระบวนการตรวจสอบและสำรวจข้อมูลเบื้องต้น เพื่อหาข้อมูลเชิงลึกโดยใช้ไลบรารี่ Pandas, Numpy จากนั้นทำความสะอาดข้อมูล ตรวจสอบดูค่าที่หายไป พบว่า คอลัมน์ Total Charge มีค่าว่างอยู่ 11ค่า จึงได้ทำการลบออกไป และได้เปลี่ยนชนิดของข้อมูลของ Total Charge จาก Object เป็น Float64

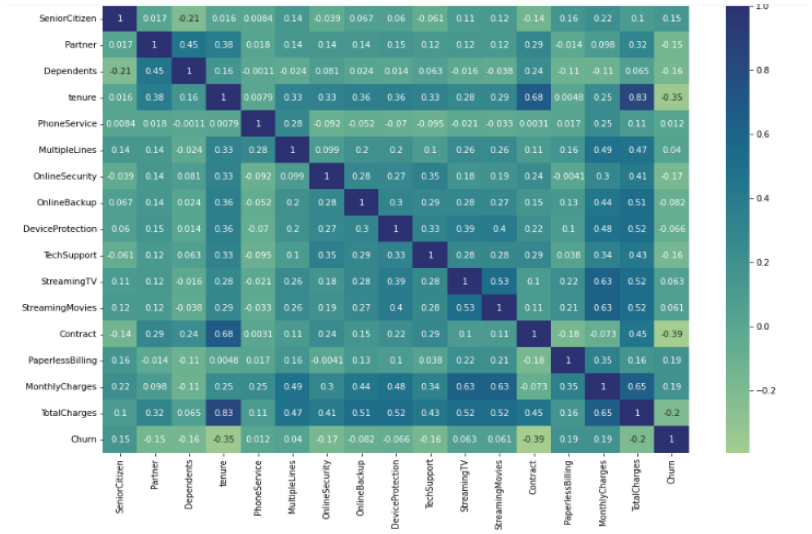
**ขั้นตอนที่ 3 :** การสำรวจข้อมูล

วิเคราะห์ระดับความสัมพันธ์ของตัวแปรของแต่ละคอลัมน์ โดยการใช้เทคนิคที่เรียกว่า Correlation Coefficient หรือ ค่าสัมประสิทธิ์สหสัมพันธ์ เป็นค่าที่บ่งชี้ถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว คือเป็นค่าที่บ่งบอกถึงความสัมพันธ์ของตัวแปร 2 ตัว ว่ามีความสัมพันธ์กันมากน้อยเพียงใด 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 1 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร 4 ตัวที่คอลัมน์มีค่าเป็นตัวเลข*

จากภาพประกอบ 1 จะเห็นได้ว่าตัวแปร Tenure มีความสัมพันธ์กับ TotalCharges สูงสุดโดยมีค่า สัมประสิทธิ์สหสัมพันธ์อยู่ที่ 0.83 โดยตัวแปรทั้งสองนี้ จะมีความสัมพันธ์ตามกัน นั่นคือ ลูกค้าที่ใช้บริการมานาน ก็จะมีแนวโน้ม ใช้บริการมากทำให้ค่าใช้จ่ายมาก ในขณะที่ ลูกค้าใหม่ซึ่งมีค่า Tenure ต่ำ ก็จะมีจะมีตัวแปรค่าบริการ TotalCharges ต่ำกว่าลูกค้าที่ใช้บริการมานานแล้วด้วย และ TotalCharges ความสัมพันธ์กับ MontlyCharges โดยค่า correlation อยู่ที่ 0.65 เนื่องจากลูกค้าที่ใช้งานมาเป็นระยะเวลานานมีแนวโน้มที่จะจ่ายค่าบริการสูงด้วย และลูกค้าที่จ่ายรายเดือนแพงก็จะต้องจ่ายค่าบริการทั้งหมดสูงตามไปด้วย




*ภาพประกอบ 2 ค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรทุกคอลัมน์*

จากภาพประกอบ 2 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 MonthlyCharges ก็จะทำให้ MonthlyCharges สูงตามไปด้วย

**ขั้นตอนที่ 4 :** การเตรียมข้อมูล

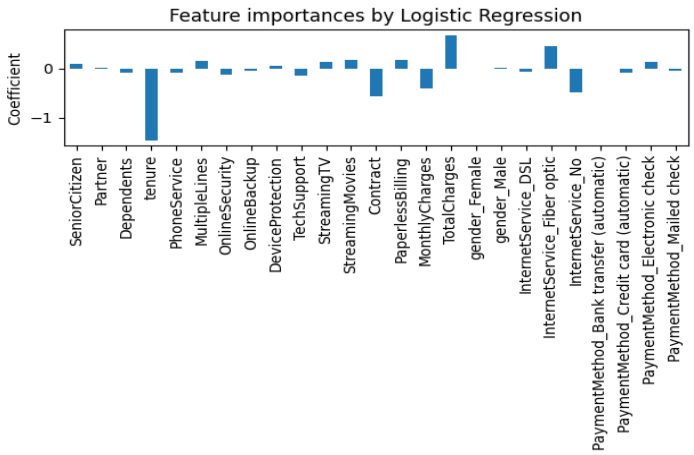
ผู้วิจัยได้ทำการ One-Hot Encoding คือเทคนิคที่ใช้แปลงข้อมูลประเภท Categorical ให้เป็นรูปแบบ Binary Vector ที่สามารถใช้งานกับแบบจำลองได้ง่ายขึ้น เนื่องจากแบบจำลองมักทำงานกับข้อมูลเชิงตัวเลข ช่วยให้แบบจำลองเข้าใจและประมวลผลข้อมูลประเภท Categorical ได้อย่างมีประสิทธิภาพและยังได้ทำการปรับสเกลของข้อมูลโดยใช้เทคนิค Standard Scaler เป็นเทคนิคที่ใช้ปรับขนาดข้อมูลเชิงตัวเลข (Numerical data) ในแบบจำลอง ให้มีค่าเฉลี่ย (Mean) อยู่ที่ 0 และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) อยู่ที่ 1 เพื่อช่วยให้แบบจำลองเรียนรู้ได้เร็วขึ้นและมีประสิทธิภาพดีขึ้น

**ขั้นตอนที่ 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

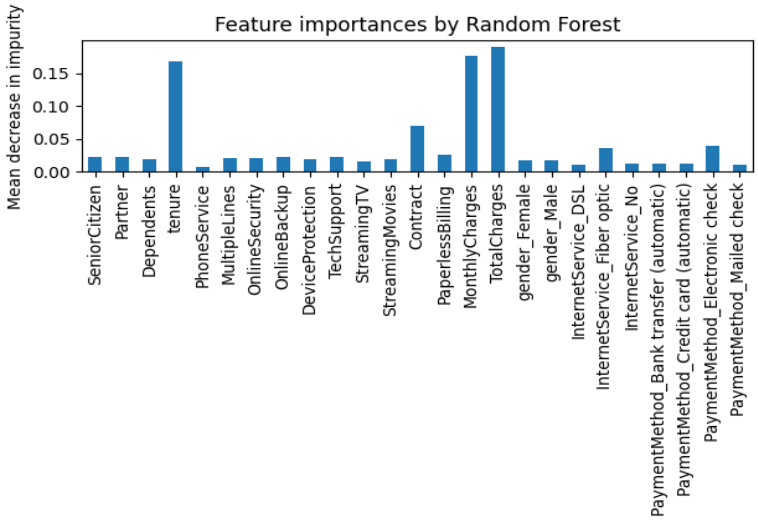
## ผลการวิจัยและอภิปรายผลการวิจัย

จากการวิจัยผู้วิจัยพบว่าปัจจัยที่มีผลต่อการการยกเลิกการใช้บริการของลูกค้า ของแบบจำลองทั้ง 3 ได้แก่ แบบจำลอง Logistic Regression 2. แบบจำลอง Random Forest และ 3. แบบจำลอง XGBoost ได้ผลออกมาดังนี้



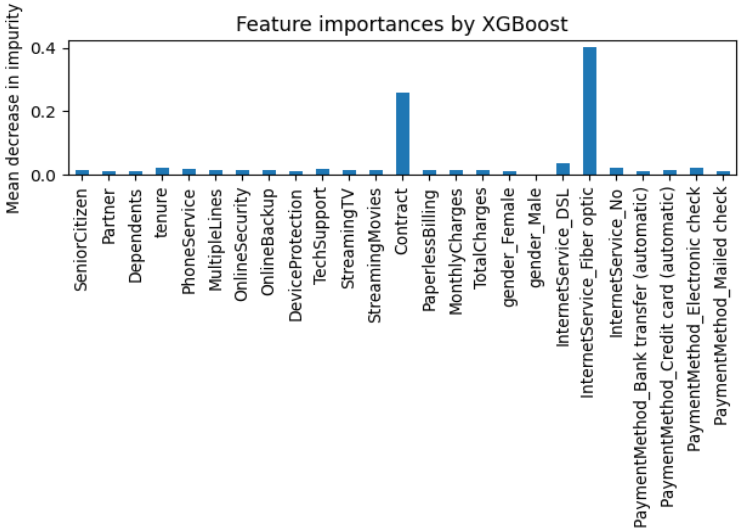

*ภาพประกอบ 3 Best feature ของแบบจำลอง Logistic Regression*

จากภาพประกอบ 3 แสดงให้เห็นถึงปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 3อันดับ ได้แก่ Tenure, Total Charges และ Contract โดย Tenure และ Contract ที่อยู่ฝั่งติดลบ จะแสดงถึงความสัมพันธ์ที่แปรผกผันกัน กล่าวคือ ยิ่งค่าของ Tenure มากเท่าไหร่ อัตราการยกเลิกใช้บริการก็จะยิ่งน้อยลง เช่นเดียวกันกับ Contract ยิ่งอยู่นานก็จะยิ่งมีอัตราการยกเลิกการใช้บริการน้อยลง แต่สำหรับ Total Charge จะมีความสัมพันธ์ที่แปรผันตาม คือถ้าค่า Total Charge มาก อัตราการยกเลิกใช้บริการก็จะมากตาม




*ภาพประกอบ 4 Best feature ของแบบจำลอง Random Forest*

จากภาพประกอบ 4 แสดงให้เห็นถึงปัจจัยที่มีความสำคัญส่งผลถึงการยกเลิกการใช้บริการมากที่สุด 3อันดับ ได้แก่ Tenure, Monthly Charges และ Total Charges ซึ่งแบบจำลอง Random Forest จะไม่ได้บอกถึงทิศทางความสัมพันธ์ว่าเป็นแปรผันตามหรือแปรผกผัน เหมือนกับการหา Best Feature ของแบบจำลอง Logistic Regression โดยสรุปคือถ้า Tenure, Monthly Charges และ Total Charges มีค่าสูง ก็จะส่งผลถึงการยกเลิกการบริการสูงแต่บอกไม่ได้ว่าส่งผลไปในทิศทางไหน อาจจะต้องดูการกระจายตัวของข้อมูลเพิ่มเติมเพื่อช่วยในการวิเคราะห์ต่อไป




*ภาพประกอบ 5 Best feature ของแบบจำลอง XGBoost*

จากภาพประกอบ 5 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 Total Charge แสดงว่า ทั้ง Contract, Tenure และ Total Charge เป็นปัจจัยหลักที่สำคัญในการยกเลิกการใช้บริการของลูกค้า

| | 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 6 การทดสอบประสิทธิภาพของการพัฒนาแบบจำลอง*

จากภาพประกอบ 6 จะเห็นได้ว่าแบบจำลอง Logistic Regression ให้ผลลัพธ์ที่ดีที่สุดในแง่ของ Accuracy แบบจำลอง XGBoost มีประสิทธิภาพรองลงมา และแบบจำลอง Decision Tree มีประสิทธิภาพต่ำสุด

## สรุปผลการวิจัย

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 และปัจจัยที่มีผลต่อการยกเลิกใช้บริการมากที่สุดคือ Contract, Tenure และ Total Charge ตามลำดับ

## กิตติกรรมประกาศ

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

## เอกสารอ้างอิง

[1] 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),

[2] 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),

[3] 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),

[4] 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),

[5] 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),