

การทำนายพฤติกรรมการยกเลิกบริการในธุรกิจท่องเที่ยวด้วยเทคนิคการเรียนรู้ของเครื่อง

วาสิการ จิตติศักดิ์^{1*} สุพจน์ เสงี่ยมพรม² และ ไกรรุ่ง เสงี่ยมพรม²

¹สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

²สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม

*wasikabkk@gmail.com

บทคัดย่อ

การสูญเสียลูกค้าหรือการยกเลิกบริการเป็นความท้าทายสำคัญที่ส่งผลกระทบต่อรายได้และการเติบโตของธุรกิจท่องเที่ยว งานวิจัยนี้จึงมุ่งเน้นการพัฒนาระบบเพื่อพยากรณ์ความเสี่ยงในการยกเลิกบริการของลูกค้าล่วงหน้า โดยประยุกต์ใช้วิทยาการข้อมูลและการเรียนรู้ของเครื่อง ผ่านการวิเคราะห์ชุดข้อมูลพฤติกรรมนักท่องเที่ยวจำนวน 954 รายการ ผู้วิจัยได้ทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึม 3 รูปแบบ ได้แก่ การถดถอยโลจิสติก (Logistic Regression) ต้นไม้ตัดสินใจ (Decision Tree) และ ป่าสุ่ม (Random Forest) ผลการศึกษาพบว่า เทคนิคต้นไม้ตัดสินใจ ให้ประสิทธิภาพโดดเด่นที่สุด โดยให้ความแม่นยำ (Accuracy) ในการทำนายสูงถึง 87% ผลลัพธ์จากการศึกษานี้เปรียบเสมือนเครื่องมือช่วยตัดสินใจที่ช่วยให้ธุรกิจท่องเที่ยวสามารถระบุกลุ่มลูกค้าที่มีความเสี่ยงสูง นำไปสู่การวางแผนกลยุทธ์การตลาดเชิงรุก การรักษาสถานลูกค้า และการเพิ่มขีดความสามารถในการแข่งขันได้อย่างตรงจุด

คำสำคัญ: การทำนายการยกเลิกบริการ ธุรกิจท่องเที่ยว การเรียนรู้ของเครื่อง การรักษาสถานลูกค้า

Predicting Customer Churn in the Tourism Industry Using Machine Learning

Wasika Jittisak^{1*}, Supoj Hengprapohm² and Kairung Hengprapohm²

¹Program in Data Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

²Program in Computer Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

*wasikabkk@gmail.com

Abstract

Customer churn, or service cancellation, poses a critical challenge that directly impacts the revenue and long-term sustainability of the tourism industry. This research aims to develop a proactive predictive system to assess customer cancellation risks by applying data analytics and Machine Learning techniques. Utilizing a travel behavior dataset comprising 954 records, the study evaluated and compared the performance of three predictive algorithms: Logistic Regression, Decision Tree, and Random Forest. The findings reveal that the Decision Tree model outperformed the others, achieving a robust prediction accuracy of 87%. This research provides a strategic advantage for tourism enterprises, enabling them to identify at-risk customers in advance, tailor proactive retention strategies, and effectively optimize their marketing resources.

Keywords: Customer Churn Prediction, Tourism Industry, Machine Learning, Customer Retention

1. บทนำ

อุตสาหกรรมการท่องเที่ยวเปรียบเสมือนฟันเฟืองสำคัญในการขับเคลื่อนเศรษฐกิจ แต่ในขณะเดียวกันก็เป็นหนึ่งในธุรกิจที่ต้องเผชิญกับความผันผวนและสถานการณ์การแข่งขันที่สูงมาก ในปัจจุบัน ปัญหาสำคัญประการหนึ่งที่ผู้ประกอบการธุรกิจท่องเที่ยว (Tour & Travels) ทุกรายต้องเผชิญคือ “อัตราการยกเลิกบริการของลูกค้า” (Customer Churn) ซึ่งหมายถึงการที่ลูกค้าตัดสินใจยกเลิกการจองหรือเลิกใช้บริการกลางคัน ปรากฏการณ์นี้ไม่เพียงแต่ส่งผลกระทบต่อรายได้และผลกำไรของธุรกิจเท่านั้น แต่การสูญเสียลูกค้าไปโดยไม่สามารถคาดการณ์ได้ล่วงหน้า ยังส่งผลให้องค์กรเสียโอกาสทางธุรกิจ และสูญเสียประสิทธิภาพในการวางแผนจัดสรรทรัพยากรในระยะยาวอีกด้วย ในอดีตองค์กรท่องเที่ยวมักพยายามทำความเข้าใจพฤติกรรมยกเลิกของลูกค้าผ่านกระบวนการวิเคราะห์แบบดั้งเดิม เช่น การทำแบบสำรวจความพึงพอใจ หรือการสังเกตพฤติกรรมโดยพนักงาน ซึ่งมีข้อจำกัดอย่างมากเมื่อต้องเผชิญกับข้อมูลประวัติการใช้บริการและพฤติกรรมของผู้บริโภคในยุคดิจิทัลที่มีปริมาณมหาศาลและความซับซ้อนสูง ส่งผลให้ไม่สามารถพยากรณ์ความน่าจะเป็นที่ลูกค้าแต่ละรายจะยกเลิกบริการได้อย่างแม่นยำและทัน่วงที

เพื่อก้าวข้ามข้อจำกัดดังกล่าว การประยุกต์ใช้เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) จึงกลายเป็นเครื่องมือสำคัญในการเปลี่ยนผ่านองค์กรจากการทำงานแบบดั้งเดิม ไปสู่การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics) โดยอาศัยการเรียนรู้จากข้อมูลในอดีตเพื่อระบุความสัมพันธ์ของตัวแปรแฝงที่ส่งผลต่อการยกเลิกบริการ ดังที่ได้มีงานวิจัยและทฤษฎีการจำแนกประเภทข้อมูลนำตัวแบบทางคณิตศาสตร์และสถิติมาช่วยทำนายพฤติกรรมในมิติต่าง ๆ [1] ไม่ว่าจะเป็นการ

ใช้สมการถดถอยเชิงเส้นและโลจิสติก (Linear/Logistic Regression) เพื่อหาความสัมพันธ์ของตัวแปร การใช้โครงสร้างแบบต้นไม้ตัดสินใจ (Decision Tree) เพื่อสร้างเงื่อนไขคัดแยกที่อธิบายผลลัพธ์ได้ [2] หรือการใช้เทคนิคขั้นสูงอย่างป่าสุ่ม (Random Forest) เพื่อลดความเอนเอียงและเพิ่มความเสถียรให้แก่ผลการทำนาย [3] ซึ่งการเลือกใช้โมเดลที่เหมาะสมจะช่วยให้ระบบสามารถจำแนกข้อมูลและวิเคราะห์แนวโน้มได้อย่างมีประสิทธิภาพ

ด้วยเหตุนี้ งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาระบบพยากรณ์พฤติกรรมการยกเลิกบริการของลูกค้าในธุรกิจท่องเที่ยว โดยทำการวิเคราะห์ชุดข้อมูลพฤติกรรมนักท่องเที่ยวจำนวน 954 รายการ เพื่อเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึม Logistic Regression, Decision Tree และ Random Forest ในการค้นหาตัวแปรที่ให้ความแม่นยำสูงสุด ยิ่งไปกว่านั้น เพื่อให้งานวิจัยเกิดประโยชน์ในเชิงประจักษ์และสามารถนำไปใช้งานได้จริง (Practical Deployment) ผู้วิจัยจึงได้ต่อยอดนำโมเดลที่ดีที่สุดไปพัฒนาเป็นเว็บแอปพลิเคชันผ่าน Streamlit API เพื่อให้ผู้ประกอบการ ตลอดจนฝ่ายการตลาดและฝ่ายบริการลูกค้า สามารถใช้เป็นเครื่องมือช่วยตัดสินใจในการตรวจสอบแนวโน้มของลูกค้าแต่ละรายได้ล่วงหน้า นำไปสู่การวางแผนกลยุทธ์การรักษาฐานลูกค้า (Customer Retention) การนำเสนอโปรโมชั่นจูงใจที่ตรงจุด และการบริหารจัดการทรัพยากรขององค์กรได้อย่างมีประสิทธิภาพสูงสุด

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดเกี่ยวกับการยกเลิกใช้บริการของลูกค้า

ในสภาวะการแข่งขันทางธุรกิจที่สูงขึ้นในปัจจุบัน หลักการทางกลยุทธ์การตลาดที่สำคัญประการหนึ่งระบุว่า ต้นทุนในการจัดหาลูกค้าใหม่ (Customer Acquisition Cost) มีมูลค่าสูงกว่าต้นทุนในการรักษาลูกค้าเดิม (Customer Retention Cost) ถึง 5 – 25 เท่า ดังนั้น การบริหารจัดการความสัมพันธ์กับลูกค้า (CRM) จึงเปลี่ยนผ่านจากการพยายามหาลูกค้าใหม่เพียงอย่างเดียว ไปสู่การป้องกันไม่ให้ลูกค้าเดิมเลิกใช้บริการ หรือที่เรียกว่า "การทำนายการยกเลิกใช้บริการ" (Customer Churn Prediction) ในบริบทของอุตสาหกรรมท่องเที่ยว การยกเลิกบริการมักอยู่ในรูปแบบของการยกเลิกการจอง (Booking Cancellation) ซึ่งการทราบแนวโน้มล่วงหน้าจะช่วยให้ผู้ประกอบการสามารถเสนอสิทธิพิเศษหรือโปรโมชั่นเพื่อดึงดูดใจลูกค้ากลับมาได้ทันที [4]

2.2 การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (Machine Learning) [5] เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (AI) ที่มุ่งเน้นการพัฒนาระบบให้สามารถเรียนรู้และหาความสัมพันธ์จากข้อมูลในอดีตได้ด้วยตนเองโดยไม่ต้องเขียนโปรแกรมสั่งการในทุกเงื่อนไข สำหรับปัญหาการยกเลิกบริการของลูกค้า นั้น จัดเป็นปัญหาประเภทการจำแนกประเภทแบบมีผู้สอน (Supervised Classification) ซึ่งชุดข้อมูลฝึกสอนจะต้องประกอบด้วย "ตัวแปรต้น" (Features เช่น อายุ, รายได้, ประวัติการจอง) และ "ตัวแปรตาม" (Target Label คือ 0: ไม่ยกเลิก และ 1: ยกเลิก) อัลกอริทึมที่ใช้ในการวิจัย ประกอบด้วย

2.2.1 การถดถอยโลจิสติก (Logistic Regression) [6] เป็นตัวแบบทางสถิติพื้นฐานที่ปรับใช้กับงานจำแนกประเภท โดยอาศัยฟังก์ชันซิกมอยด์ (Sigmoid function) ในการแปลงผลรวมเชิงเส้นของตัวแปรต้นให้อยู่ในรูปของค่าความน่าจะเป็นที่มีค่าระหว่าง 0 ถึง 1 ข้อดีของโมเดลนี้คือมีความเรียบง่ายและสามารถอธิบายทิศทางความสัมพันธ์ของตัวแปร (Coefficient) ได้อย่างชัดเจนทางสถิติ

2.2.2 ต้นไม้ตัดสินใจ (Decision Tree Classifier) [7] เป็นอัลกอริทึมที่สร้างเงื่อนไขการตัดสินใจในรูปแบบโครงสร้างต้นไม้ (Tree structure) โดยทำการเลือกตัวแปรที่สามารถแบ่งแยกกลุ่มข้อมูลให้มีความบริสุทธิ์สูงสุด (เช่น Information Gain หรือ Gini Impurity) ในแต่ละโหนดย่อย ข้อดีของต้นไม้ตัดสินใจคือเป็น "ตัวแบบกล่องขาว" (White-box Model) ที่มนุษย์สามารถอ่านและเข้าใจเงื่อนไข (If-Else Rules) ได้ว่าเหตุใดลูกค้าจึงถูกตัดสินว่าเสี่ยงยกเลิก

2.2.3 ป่าสุ่ม (Random Forest Classifier) เป็นเทคนิคขั้นสูงในกลุ่มการเรียนรู้แบบกลุ่มก้อน (Ensemble Learning) ที่ถูกนำเสนอโดย L. Breiman [3] โดยมีหลักการทำงานคือการสร้างต้นไม้ตัดสินใจจำนวนมากหลายร้อยต้นขึ้นมาพร้อมกัน ผ่านเทคนิคการสุ่มตัวอย่างข้อมูลแบบใส่คืน (Bootstrap Aggregating หรือ Bagging) ร่วมกับการสุ่มเลือกตัวแปรในแต่ละโหนด จากนั้นใช้หลักการโหวตเสียงข้างมาก (Majority Vote) เพื่อตัดสินผลลัพธ์สุดท้าย ซึ่งช่วยแก้ปัญหาการเรียนรู้มากเกินไป (Overfitting) ของต้นไม้ตัดสินใจเดี่ยวได้อย่างมีประสิทธิภาพ

2.3 การประเมินประสิทธิภาพของโมเดล

การประเมินประสิทธิภาพของโมเดลในงานจำแนกประเภทข้อมูล นิยมอาศัยตารางเมทริกซ์ความสับสน (Confusion Matrix) เพื่อคำนวณมาตรวัดมาตรฐาน 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) เพื่อดูภาพรวม, ค่าความแม่นยำ (Precision) เพื่อพิจารณาความผิดพลาดในการเตือนภัยปลอม, ค่าความครบถ้วน (Recall) ซึ่งมีความสำคัญสูง เนื่องจากการวัดว่าโมเดลสามารถตรวจจับลูกค้าที่คิดจะยกเลิกได้มากน้อยเพียงใด และค่า F1-Score ซึ่งเป็นค่าเฉลี่ยฮาร์มอนิกระหว่าง Precision และ Recall [8]

2.4 งานวิจัยที่เกี่ยวข้อง

การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อทำนายการยกเลิกบริการ ได้รับความสนใจในหลากหลายอุตสาหกรรม ดังที่ จิรฤต บุญหมื่นไวย และคณะ [1] ได้ทำการศึกษาการวิเคราะห์แนวโน้มการยกเลิกใช้บริการของลูกค้าธนาคาร โดยเปรียบเทียบอัลกอริทึม Logistic Regression, Decision Tree, Naïve Bayes และ k-NN ผลการศึกษพบว่าโมเดลกลุ่มต้นไม้ตัดสินใจสามารถระบุปัจจัยความเสี่ยงของลูกค้าที่ต้องการปิดบัญชีได้อย่างแม่นยำ

ในการเปรียบเทียบเชิงลึก A. De Caigny และคณะ [2] ได้นำเสนอตัวแบบลูกผสมระหว่าง Logistic Regression และ Decision Tree ในการทำนาย Customer Churn ซึ่งชี้ให้เห็นประเด็นสำคัญว่า แม้สมการโลจิสติกจะสามารถจับแนวโน้มความสัมพันธ์เชิงเส้นในภาพรวมได้ดี แต่โครงสร้างแบบ Decision Tree มีความจำเป็นอย่างยิ่งในการดักจับพฤติกรรมเฉพาะกลุ่ม (Local patterns) ของลูกค้าในเซกเมนต์ย่อย

ในบริบทของอุตสาหกรรมการท่องเที่ยวและบริการ A. Antonio และคณะ ได้ทำการศึกษาวิจัยขนาดใหญ่เพื่อทำนายพฤติกรรมการยกเลิกการจองโรงแรมของนักท่องเที่ยวจำนวนกว่าแสนรายการ ผลการวิจัยยืนยันสอดคล้องกันว่า กลุ่มโมเดลต้นไม้ตัดสินใจให้ประสิทธิภาพในการพยากรณ์การยกเลิกบริการที่สูงที่สุดเมื่อเทียบกับโมเดลพื้นฐานอื่น ๆ [9]

3. วิธีดำเนินการวิจัย

การวิจัยเรื่องนี้มุ่งศึกษาการนำเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) มาประยุกต์ใช้ในการคาดการณ์การยกเลิกการสมัครสมาชิกหรือการยกเลิกใช้บริการของลูกค้าในธุรกิจทัวร์และการท่องเที่ยว โดยพิจารณาจากข้อมูลส่วนบุคคลและข้อมูลการใช้งานของลูกค้า เช่น อายุ เพศ ประวัติการเดินทาง และพฤติกรรมการจอง เพื่อนำไปสู่การเปรียบเทียบและคัดเลือกโมเดลที่ดีที่สุด โดยมีขั้นตอนวิธีดำเนินการวิจัยแบ่งออกเป็น 4 ขั้นตอนหลัก ดังนี้

3.1 การรวบรวมข้อมูล (Data Set)

ผู้วิจัยได้ดำเนินการรวบรวมข้อมูลทุติยภูมิจากเว็บไซต์ www.kaggle.com ซึ่งเป็นชุดข้อมูลเกี่ยวกับพฤติกรรมการใช้บริการของลูกค้าเพื่อนำมาวิเคราะห์การยกเลิกใช้บริการ (Churn Prediction) โดยมีรายละเอียดและขั้นตอนการรวบรวมข้อมูล ดังนี้

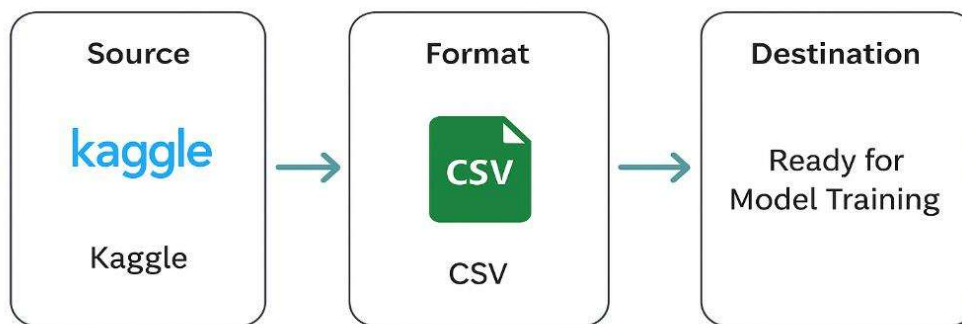
- 3.1.1 ปริมาณข้อมูล: ชุดข้อมูลที่นำมาใช้มีจำนวนทั้งสิ้น 954 รายการ แบ่งเป็น ลูกค้าที่ยังไม่เลิกใช้บริการ จำนวน 730 รายการ และ ลูกค้าที่เลิกใช้บริการ จำนวน 224 รายการ

- 3.1.2 รูปแบบการจัดเก็บ: ข้อมูลถูกจัดเก็บอยู่ในรูปแบบไฟล์นามสกุล .csv

3.1.3 คุณลักษณะของข้อมูล (Attributes): ประกอบด้วยตัวแปรที่ระบุคุณลักษณะของลูกค้า จำนวน 7 คุณลักษณะ ได้แก่

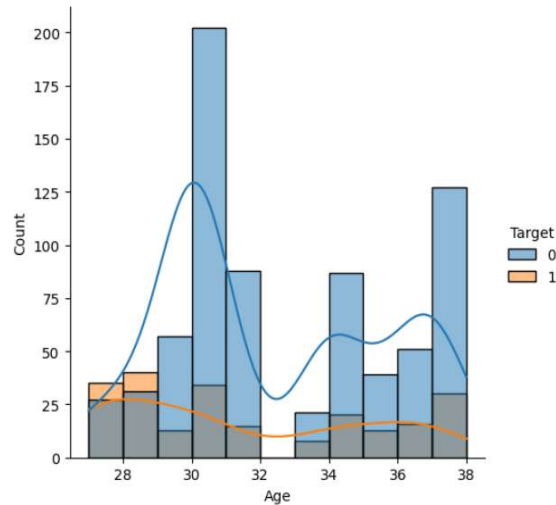
- 1) อายุ (Age)
- 2) การเดินทางด้วยเครื่องบินบ่อยหรือไม่ (FrequentFlyer)
- 3) ระดับรายได้ต่อปี (AnnualIncomeClass)
- 4) จำนวนครั้งที่ลูกค้าเลือกใช้บริการต่าง ๆ ของบริษัทในช่วงหลายปีที่ผ่านมา (ServicesOpted)
- 5) บัญชีลูกค้าที่ใช้กับบริษัทเชื่อมโยงกับบัญชีสื่อสังคมออนไลน์หรือไม่ (AccountSyncedToSocialMedia)
- 6) ลูกค้าเคยจองโรงแรมหรือที่พักผ่านบริการของบริษัทหรือไม่ (BookedHotelOrNot)
- 7) คำตอบที่ต้องการให้โมเดลทำนายว่า ลูกค้าเลิกใช้บริการหรือไม่ (Target)

3.1.4 ขั้นตอนการได้มาซึ่งข้อมูล: เริ่มต้นจากการดาวน์โหลดไฟล์ข้อมูลดิบ (Raw Data) จากเว็บไซต์ Kaggle ในรูปแบบไฟล์ CSV จากนั้นจึงนำเข้าสู่ระบบเพื่อเตรียมความพร้อมสำหรับการฝึกสอนโมเดล (ตามที่ปรากฏในภาพที่ 1 ขั้นตอนการดาวน์โหลดข้อมูล)



ภาพที่ 1 ขั้นตอนการดาวน์โหลดข้อมูล

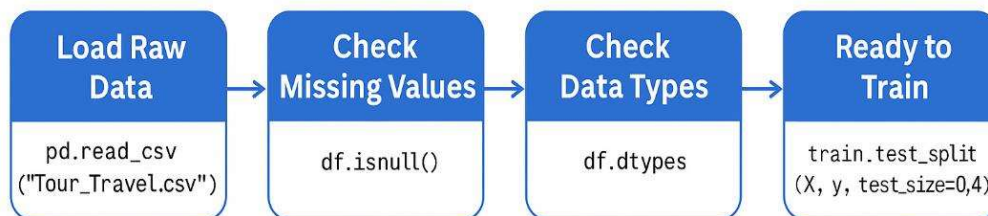
3.1.5 นอกจากนี้ จากการวิเคราะห์ความสัมพันธ์เบื้องต้นของชุดข้อมูลพบว่า คุณลักษณะด้าน "อายุ" มีสัดส่วนความสัมพันธ์ต่อการตัดสินใจยกเลิกใช้บริการของลูกค้าอย่างเห็นได้ชัด (ตามที่ปรากฏในภาพที่ 2 อัตราส่วนการยกเลิกใช้ตามสัดส่วนอายุ)



ภาพที่ 2 อัตราส่วนการยกเลิกใช้ตามสัดส่วนอายุ

3.2 การเตรียมข้อมูล (Data Preparation)

การจัดรูปแบบข้อมูลเป็นขั้นตอนการแปลงข้อมูลที่รวบรวมมาให้กลายเป็นข้อมูลที่ถูกต้องและพร้อมต่อการนำไปวิเคราะห์ในอัลกอริทึมขั้นถัดไป โดยมีกระบวนการทำงาน (ดังภาพที่ 3) มีดังนี้



ภาพที่ 3 การเปลี่ยนแปลงข้อมูลจากตัวอักษรเป็นเลขใน Attribute Education City Gender

3.2.1 การโหลดข้อมูลดิบ (Load Raw Data): ทำการอ่านไฟล์ข้อมูลเข้าสู่ระบบผ่านคำสั่ง
pd.read_csv("Tour_Travel.csv")

3.2.2 การตรวจสอบข้อมูลสูญหาย (Check Missing Values): ทำการตรวจสอบหาค่าว่างหรือข้อมูลที่ตกหล่นในแต่ละคุณลักษณะผ่านคำสั่ง df.isnull()

3.2.3 การตรวจสอบชนิดข้อมูล (Check Data Types): ตรวจสอบประเภทของตัวแปรแต่ละชุดผ่านคำสั่ง df.dtypes

3.2.4 การแปลงข้อมูล (Data Transformation): ดำเนินการแปลงข้อมูลประเภทตัวอักษร (Text) ให้เป็นข้อมูลประเภทตัวเลข (Numeric) เพื่อให้อยู่ในช่วงข้อมูลเดียวกัน โดยทำการแปลงในกลุ่มคุณลักษณะ ได้แก่ Education, City และ Gender

3.2.5 การแบ่งชุดข้อมูลเตรียมฝึกสอน (Ready to Train): ดำเนินการแบ่งกลุ่มตัวแปรออกเป็นตัวแปรอิสระ (X) และตัวแปรตาม (y) เพื่อส่งต่อไปยังกระบวนการทดสอบ

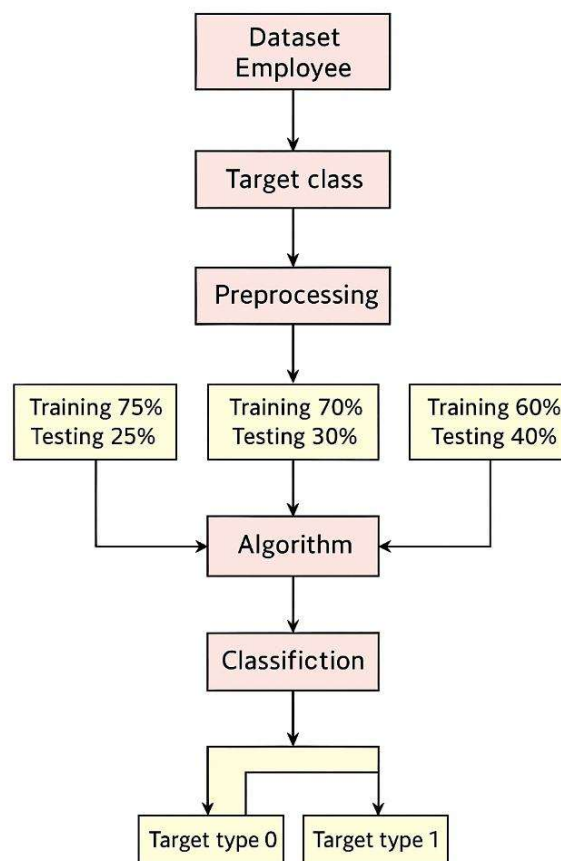
3.3 การสร้างโมเดล (Modeling)

ในกระบวนการนี้ ผู้วิจัยได้ใช้โปรแกรม Google Colab ในการสร้างตัวแบบข้อมูลสำหรับการเรียนรู้ และทดสอบเพื่อเปรียบเทียบโมเดลที่ดีที่สุด ก่อนที่จะนำไปพัฒนาเป็นเว็บแอปพลิเคชัน โดยมีกรอบการวิเคราะห์คัดแยก (ดังภาพที่ 4) ดังนี้

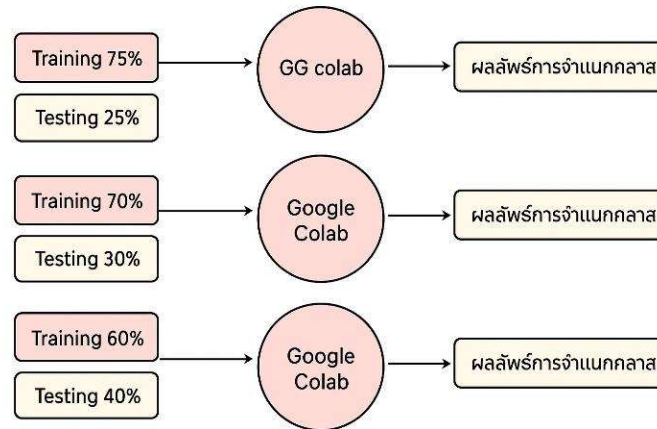
3.3.1 การกำหนดคลาสเป้าหมาย (Target Class): กำหนดตัวแปรผลลัพธ์การจำแนกออกเป็น 2 กลุ่ม คือ Target type 0 (ลูกค้าที่ไม่ยกเลิกบริการ) และ Target type 1 (ลูกค้าที่ยกเลิกบริการ)

3.3.2 การแบ่งสัดส่วนชุดข้อมูล (Data Splitting): ทำการแบ่งชุดข้อมูลผ่านการประมวลผลเบื้องต้น (Preprocessing) ออกเป็น ข้อมูลฝึกสอน (Training Set) และ ข้อมูลทดสอบ (Testing Set) เพื่อประเมินประสิทธิภาพของโมเดล ดังแสดงในภาพที่ 5 ออกเป็น 3 อัตราส่วน ได้แก่

- อัตราส่วน 75 : 25 (Training 75% / Testing 25%)
- อัตราส่วน 70 : 30 (Training 70% / Testing 30%)
- อัตราส่วน 60 : 40 (Training 60% / Testing 40%)



ภาพที่ 4 โมเดลการวิเคราะห์คัดแยกการยกเลิกใช้บริการ



ภาพที่ 5 ความสัมพันธ์ที่ใช้ประเมินประสิทธิภาพของโมเดล

3.3.3 การเรียนรู้ของอัลกอริทึม (Algorithm): นำชุดข้อมูลฝึกสอนทั้ง 3 อัตราส่วนไปป้อนเข้าสู่โมเดล Machine Learning 3 รูปแบบ ได้แก่ Logistic Regression, Decision Tree Classifier และ Random Forest Classifier เพื่อทำการจำแนกคลาส (Classification)

3.4 การวัดประสิทธิภาพ (Evaluation)

หลังจากทำการแบ่งข้อมูลออกเป็นชุดฝึกสอนและชุดทดสอบในอัตราส่วน 75:25, 70:30 และ 60:40 เรียบร้อยแล้ว ในแต่ละชุดสัดส่วนจะทำการทดสอบซ้ำทั้งหมด 3 ครั้ง เพื่อประเมินประสิทธิภาพของโมเดล โดยใช้ ตารางเมทริกซ์ความสับสน (Confusion Matrix) เป็นเครื่องมือในการวัดค่าความถูกต้องของการทำนาย

ผู้วิจัยได้ใช้สมการทางสถิติในการคำนวณตัววัดประสิทธิภาพทั้ง 6 ค่า ดังต่อไปนี้ (สมการที่ 1 - 6)

3.4.1 ค่าความแม่นยำในการจำแนกประเภทข้อมูลของ Positive (True Positive Rate: *TPR*) คำนวณได้จากสมการ:

$$TPR = \frac{TP}{Actual\ Positive} = \frac{TP}{TP + FN} \quad (1)$$

3.4.2 ค่าความแม่นยำในการจำแนกประเภทของ Negative (False Negative Rate: *FNR*) คำนวณได้จากสมการ:

$$FNR = \frac{FN}{Actual\ Positive} = \frac{FN}{TP + FN} \quad (2)$$

3.4.3 ค่า Precision เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส คำนวณได้จากสมการ:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

3.4.4 ค่า Recall (เป็นสูตรเดียวกันกับการหาค่า True Positive Rate) เป็นการวัดค่าความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส คำนวณได้จากสมการ:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

3.4.5 ค่า F1-Score เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดลโดยพิจารณาแยกทีละคลาส คำนวณได้จากสมการ:

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (5)$$

3.4.6 ค่า Accuracy เป็นการวัดค่าความถูกต้องของโมเดลโดยพิจารณารวมทุกคลาส คำนวณได้จากสมการ:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (6)$$

โดยกำหนดให้ TP = True Positive, TN = True Negative, FP = False Positive และ FN = False Negative

เมื่อเสร็จสิ้นกระบวนการคำนวณ ผู้วิจัยจะทำการเปรียบเทียบผลลัพธ์และเลือกโมเดลที่ให้ผลการทำนาย (Accuracy) ที่ดีที่สุด เพื่อนำไปใช้งานในขั้นตอนการพัฒนาระบบจริงต่อไป

4. ผลการทดลอง

จากการนำชุดข้อมูลพฤติกรรมการใช้บริการของลูกค้าธุรกิจท่องเที่ยวจำนวน 954 รายการ เข้าสู่กระบวนการเตรียมข้อมูลและฝึกสอนผ่านอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning) 3 รูปแบบ ได้แก่ Logistic Regression, Decision Tree และ Random Forest ผู้วิจัยได้ทำการทดสอบประสิทธิภาพของโมเดลโดยแบ่งสัดส่วนข้อมูลฝึกสอนและทดสอบ (Train-Test Split) ออกเป็น 3 อัตราส่วน ได้แก่ 75:25, 70:30 และ 60:40 ซึ่งสามารถแสดงผลการเปรียบเทียบตามอัตราส่วนต่าง ๆ ได้ดังนี้

4.1 ผลการเปรียบเทียบประสิทธิภาพในอัตราส่วน 75 : 25

ในการทดสอบชุดแรก ผู้วิจัยได้แบ่งชุดข้อมูลสำหรับฝึกสอน (Training Set) ร้อยละ 75 และข้อมูลสำหรับทดสอบ (Testing Set) ร้อยละ 25 โดยผลการวัดประสิทธิภาพผ่านค่าความถูกต้อง (Accuracy) (สมการ 6) ความแม่นยำ (Precision) (สมการ 3) ความครบถ้วน (Recall) (สมการ 4) และค่า F1-Score (สมการ 5) แสดงได้ดังตารางที่ 1

ตารางที่ 1 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล โดยการแบ่งข้อมูลฝึกสอนและทดสอบในอัตราส่วน 75:25

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.8703	0.7077	0.7931	0.7479
Random Forest	0.8536	0.6769	0.7586	0.7154
Logistic Regression	0.7782	0.5556	0.4310	0.4854

จากตารางที่ 1 พบว่า เมื่อแบ่งชุดข้อมูลในอัตราส่วน 75:25 อัลกอริทึม Decision Tree ให้ประสิทธิภาพโดยรวมสูงที่สุด โดยมีค่าความถูกต้อง (Accuracy) อยู่ที่ 0.8703 รองลงมาคือ Random Forest มีค่าความถูกต้องอยู่ที่ 0.8536 และ Logistic Regression มีค่าความถูกต้องต่ำที่สุดอยู่ที่ 0.7782 นอกจากนี้ ในมิติของความครบถ้วน (Recall) ซึ่งเป็นค่าที่ชี้วัดความสามารถในการตรวจจับลูกค้าที่ต้องการยกเลิกบริการได้จริง พบว่า Decision Tree สามารถตรวจจับได้ดีที่สุดถึง 0.7931 ซึ่งห่างจาก Logistic Regression ที่ทำได้เพียง 0.4310 อย่างมีนัยสำคัญ

4.2 ผลการเปรียบเทียบประสิทธิภาพในอัตราส่วน 70 : 30

ในการทดสอบชุดที่สอง ผู้วิจัยได้ปรับสัดส่วนข้อมูลทดสอบเพิ่มขึ้นเป็นร้อยละ 30 และข้อมูลฝึกสอนร้อยละ 70 เพื่อทดสอบความเสถียรของตัวแบบเมื่อต้องเผชิญกับชุดข้อมูลทดสอบที่ใหญ่ขึ้น โดยได้ผลลัพธ์ดังตารางที่ 2

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล โดยการแบ่งข้อมูลฝึกสอนและทดสอบในอัตราส่วน 70:30

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.8746	0.6914	0.8358	0.7567
Random Forest	0.8537	0.6623	0.7612	0.7083
Logistic Regression	0.7979	0.5818	0.4776	0.5245

จากตารางที่ 2 แสดงให้เห็นว่า เมื่อปรับสัดส่วนข้อมูลทดสอบเพิ่มขึ้นเป็นร้อยละ 30 อัลกอริทึม Decision Tree ยังคงรักษามาตรฐานการทำนายไว้ได้ดีที่สุด โดยมีค่าความถูกต้อง (Accuracy) ขยับเพิ่มขึ้นเป็น 0.8746 และมีค่าความครบถ้วน (Recall) สูงถึง 0.8358 ในขณะที่ Random Forest ทำความถูกต้องได้ 0.8537 และ Logistic Regression ทำความถูกต้องได้ 0.7979 ซึ่งทั้งสองโมเดลยังคงมีค่า Precision และ Recall ที่ต่ำกว่ากลุ่มต้นไม้ตัดสินใจ

4.3 ผลการเปรียบเทียบประสิทธิภาพในอัตราส่วน 60 : 40

ในการทดสอบชุดที่สาม ผู้วิจัยได้ลดสัดส่วนข้อมูลฝึกสอนลงเหลือร้อยละ 60 และใช้เป็นข้อมูลทดสอบถึงร้อยละ 40 เพื่อพิจารณาพฤติกรรมของโมเดลเมื่อได้รับข้อมูลสำหรับเรียนรู้ร้อยละ 60 โดยแสดงผลที่ได้ดังตารางที่ 3

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล โดยการแบ่งข้อมูลฝึกสอนและทดสอบในอัตราส่วน 60:40

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.8770	0.7326	0.7241	0.7283
Random Forest	0.8665	0.7143	0.6897	0.7017
Logistic Regression	0.7984	0.5714	0.4598	0.5095

จากตารางที่ 3 พบประเด็นที่น่าสนใจว่า แม้โมเดลจะได้รับข้อมูลสำหรับฝึกสอนลดลงเหลือเพียงร้อยละ 60 แต่อัลกอริทึม Decision Tree กลับสามารถทำค่าความถูกต้อง (Accuracy) สูงสุดในการวิจัยนี้ที่ 0.8770 พร้อมทั้งมีค่าความแม่นยำ (Precision) ขยับขึ้นมาสูงสุดที่ 0.7326 ส่วน Random Forest ขยับความถูกต้องตามขึ้นมาเป็น 0.8665 และ Logistic Regression ทำความถูกต้องได้ 0.7984

อย่างไรก็ตาม เป็นที่น่าสังเกตว่าในสัดส่วน 60:40 นี้ ค่า Recall ของทุกโมเดลมีแนวโน้มลดลง (Decision Tree ลดลงเหลือ 0.7241) ซึ่งเป็นผลกระทบทางสถิติจากการที่ชุดข้อมูลฝึกสอนมีขนาดเล็กจนสูญเสียตัวอย่างของกลุ่มเป้าหมาย (ผู้ยกเลิกบริการ) บางส่วนไปในการเรียนรู้

5. สรุปผลและอภิปรายผลการทดลอง

การศึกษาวิจัยเรื่อง "การทำนายพฤติกรรมการยกเลิกบริการในธุรกิจท่องเที่ยวด้วยเทคนิคการเรียนรู้ของเครื่อง" มีวัตถุประสงค์เพื่อพัฒนาระบบพยากรณ์ความเสี่ยงในการยกเลิกบริการ (Customer Churn Prediction) ของลูกค้าธุรกิจท่องเที่ยว โดยนำชุดข้อมูลพฤติกรรมนักท่องเที่ยวจำนวน 954 รายการ มาผ่านกระบวนการวิทยาการข้อมูล และเปรียบเทียบประสิทธิภาพของอัลกอริทึม 3 รูปแบบ ได้แก่ Logistic Regression, Decision Tree และ Random Forest ซึ่งสามารถสรุปผล อภิปรายผล และมีข้อเสนอแนะดังต่อไปนี้

5.1 สรุปผลการทดลอง

จากการดำเนินการวิจัยและทดสอบประสิทธิภาพของตัวแบบผ่านการแบ่งข้อมูลฝึกสอนและทดสอบ (Train-Test Split) ใน 3 อัตราส่วน ได้แก่ 75:25, 70:30 และ 60:40 สามารถสรุปผลได้ดังนี้

5.1.1 ผลการเปรียบเทียบประสิทธิภาพ: อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree Classifier) เป็นโมเดลที่มีประสิทธิภาพสูงที่สุดในการวิจัยครั้งนี้ โดยให้ค่าความถูกต้อง (Accuracy) เฉลี่ยสูงถึงร้อยละ 87 ในทุกอัตราส่วน (87.03%, 87.46% และ 87.70% ตามลำดับ)

5.1.2 ความสามารถในการตรวจจับการยกเลิก: นอกเหนือจากค่าความถูกต้องแล้ว Decision Tree ยังให้ค่าความครบถ้วน (Recall) สูงที่สุดถึงร้อยละ 83.58 (ในสัดส่วน 70:30) ซึ่งพิสูจน์ได้ว่าโมเดลมีความสามารถในการตรวจจับและระบุตัวลูกค้าที่มีความเสี่ยงจะยกเลิกบริการได้อย่างมีประสิทธิภาพเหนือกว่า Random Forest และ Logistic Regression อย่างมีนัยสำคัญ

5.2 อภิปรายผลการทดลอง

จากผลการทดลองข้างต้น มีประเด็นที่น่าสนใจซึ่งสามารถนำมาอภิปรายผลโดยเชื่อมโยงกับทฤษฎีและงานวิจัยที่เกี่ยวข้องได้ดังนี้

5.2.1 ความเหมาะสมของโครงสร้างต้นไม้ตัดสินใจ (Decision Tree) ต่อข้อมูลพฤติกรรมลูกค้า: ผลการทดลองสอดคล้องกับสมมติฐานและงานวิจัยที่เกี่ยวข้อง (เช่น งานวิจัยของ A. Antonio และคณะ [9]) ที่ว่าพฤติกรรมการตัดสินใจของนักท่องเที่ยวมักมีลักษณะเป็นเงื่อนไขแบบขั้นบันได (If-Else logic) เช่น "ถ้าอายุอยู่ในช่วงนี้ และมีรายได้ระดับนี้ จะมีโอกาสยกเลิกสูง" อัลกอริทึม Decision Tree ซึ่งมีกลไกการสร้างกฎเกณฑ์การตัดสินใจแบบแบ่งแยกข้อมูลตามเกณฑ์ (Information Gain) จึงสามารถเรียนรู้และดักจับรูปแบบความสัมพันธ์เฉพาะกลุ่ม (Local patterns) ของลูกค้าได้อย่างแม่นยำ

5.2.2 ปริมาณข้อมูลกับการทำงานของ Random Forest: ตามทฤษฎีการเรียนรู้แบบกลุ่มก้อน (Ensemble Learning) อัลกอริทึม Random Forest ควรจะให้ประสิทธิภาพที่สูงและเสถียรกว่า Decision Tree แบบเดี่ยว แต่จากผลการทดลองกลับพบว่าทำคะแนนได้ต่ำกว่าเล็กน้อย (Accuracy ~85-86%) ปรากฏการณ์นี้สามารถอธิบายได้ว่า เนื่องจากชุดข้อมูลที่ใช้นั้นมีขนาดค่อนข้างเล็ก (954 รายการ) เมื่อ Random Forest ทำการสุ่มตัวอย่างข้อมูลแบบใส่คืน (Bootstrap) เพื่อสร้างต้นไม้จำนวนมาก อาจทำให้เกิดการเรียนรู้ข้อมูลซ้ำซ้อนกันมากเกินไป (Redundancy) จนความสามารถในการแบ่งแยกคลาสลดลงเมื่อเทียบกับการสร้างต้นไม้เพียงต้นเดียวที่ครอบคลุมข้อมูลทั้งหมด

5.3.3 ข้อจำกัดของตัวแบบ Logistic Regression ที่ให้ผลลัพธ์ต่ำที่สุดในทุกตัวชี้วัด (Accuracy ~77-79% และ Recall ต่ำกว่า 50%) สิ่งนี้เป็นการยืนยันว่า ปัจจัยที่ส่งผลต่อการยกเลิกบริการของลูกค้าในธุรกิจท่องเที่ยว ไม่ได้มีความสัมพันธ์กันเป็นลักษณะเส้นตรง (Non-linear relationship) การใช้สมการเชิงเส้นจึงไม่สามารถดักจับความซับซ้อนของพฤติกรรมมนุษย์ที่มีความแปรปรวนสูงได้



6. เอกสารอ้างอิง (References)

- [1] Boonmuenwai, J., Tantanuch, J., & Rojanadit, B. (2020). Analysis of the churn prediction for bank customers by using machine learning. Proceedings of the 5th National Conference on Quality Management and Technology Innovation, May 23, 2020. Available at: https://math.sut.ac.th/~jessada/cv/Full-Paper_TAM2020-Toon.pdf (In Thai)
- [2] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research*, vol. 269, no. 2, pp. 760–772, 2018.
- [3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [4] Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson Education.
- [5] T. M. Mitchell, *Machine Learning*, New York: McGraw-Hill, 1997.
- [6] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. John Wiley & Sons, 2013.
- [7] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
- [9] A. Antonio, G. de Almeida, and L. Nunes, "Predicting hotel booking cancellations to decrease uncertainty and increase revenue," *Tourism Management*, vol. 13 No. 2, pp. 25-39, 2017.