# การทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

# CUSTOMER CHURN PREDICTION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES

ภูมพัชร พิพัฒศรี

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

# การทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

ภูมพัชร พิพัฒศรี

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566

# CUSTOMER CHURN PREDICTION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES

PHUMPHATCHARA PHIPHATSRI

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

การทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

ของ

ภูมพัชร พิพัฒศรี

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

---

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

---

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... กรรมการ

(ดร.เรืองศักดิ์ ตระกูลพุทธิรักษ์)

| | |
|---|---|
| ชื่อเรื่อง | การทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง |
| ผู้วิจัย | ภูมพัชร พิพัฒศรี |
| ปริญญา | วิทยาศาสตรมหาบัณฑิต |
| ปีการศึกษา | 2566 |
| อาจารย์ที่ปรึกษา | ผู้ช่วยศาสตราจารย์ ดร. นุวีย์ วิวัฒนวัฒนา |

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

คำสำคัญ : การเรียนรู้ของเครื่อง, การทำนายแนวโน้มการเลิกใช้บริการ, LIME

| | |
|---|---|
| Title | CUSTOMER CHURN PREDICTION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES |
| Author | PHUMPHATCHARA PHIPHATSRI |
| Degree | MASTER OF SCIENCE |
| Academic Year | 2023 |
| Thesis Advisor | Assistant Professor Dr. Nuwee Wiwatwattana |

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

Keyword : Machine learning, Churn prediction, LIME

# กิตติกรรมประกาศ

การจัดทำวิจัยฉบับนี้สำเร็จไปได้ด้วยดี ด้วยการสนับสนุนด้านความรู้ แนวทางการดำเนินการวิจัย คำแนะนำในการจัดการทำสารนิพนธ์ ขอขอบคุณ ผศ.ดร.นุวีย์ วิวัฒนวัฒนา อาจารย์ที่ปรึกษา และ ผศ.ดร. จันตรี ผลประเสริฐ รวมทั้งคณะกรรมการสอบสารนิพนธ์ทุกท่านที่ได้ชี้แนะแนวทาง ขอขอบพระคุณอาจารย์ในภาควิชาวิทยาการข้อมูลทุกท่านที่ให้ความรู้ การสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย และขอขอบคุณพี่ ๆ เพื่อน ๆ ที่คอยช่วยเหลือในช่วงเวลาเรียน สุดท้ายขอขอบพระคุณครอบครัวของผู้วิจัยที่ให้โอกาสในการศึกษาและเป็นกำลังใจให้จนสำเร็จการศึกษา ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

ภูมพัชร พิพัฒศรี

# สารบัญ

หน้า

# สารบัญตาราง

หน้า

# สารบัญรูปภาพ

**หน้า**

# บทที่ 1

# บทนำ

## 1.1 ที่มาและความสำคัญ

ในช่วงไม่กี่ปีที่ผ่านพฤติกรรมการซื้อสินค้าของผู้บริโภคมีการเปลี่ยนแปลงจากการซื้อผ่านช่องทางออฟไลน์ไปเป็นออนไลน์มากขึ้น ในขณะเดียวกันร้านค้าออนไลน์ก็เพิ่มจำนวนขึ้นตามอุปสงค์ที่เพิ่มมากขึ้น ส่งผลให้ตลาดอีคอมเมิร์ซเติบโตอย่างต่อเนื่องทางออนไลน์กันมากขึ้น มูลค่าทางการตลาดเติบโตอย่างต่อเนื่อง ซึ่งการศึกษารูปแบบธุรกิจดิจิทัลและวิธีการรับเอาเทคโนโลยีต่าง ๆ มาปรับใช้ในทางธุรกิจนั้นเป็นสิ่งที่สำคัญที่จะต้องทำความเข้าใจและนำมาใช้ประโยชน์ในการประกอบธุรกิจ e-Commerce (Electronic Commerce) หรือการพาณิชย์อิเล็กทรอนิกส์ จนกระทั่งเกิดการแพร่ระบาดของ COVID-19 ขึ้นมากลายเป็นตัวการสำคัญที่เร่งให้ตลาดค้าปลีกในภูมิภาคเอเชีย รวมถึงประเทศไทยเกิดผลทำให้เป็นแรงผลักดันที่ทำให้ผู้บริโภคนำไปสู่การปรับเปลี่ยนพฤติกรรมไปใช้อีคอมเมิร์ซอย่างเต็มรูปแบบในช่วงหลังโควิด จึงส่งผลให้มูลค่าอีคอมเมิร์ซเติบโตแบบก้าวกระโดด การซื้อสินค้าและบริการผ่านทางออนไลน์นั้นได้กลายเป็นส่วนหนึ่งของการใช้ชีวิตแบบใหม่ไปแล้ว ซึ่งผลจากการสำรวจนั้นพบว่าโดยเฉลี่ยแล้วร้อยละ 45 ของผู้บริโภคทั่วเอเชียวางแผนที่จะเพิ่มการใช้จ่ายออนไลน์แทนการใช้ช่องทางการค้าปลีกในแบบเดิม (สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์, 2564) จากการที่ผู้บริโภคหันมาพึ่งพาการซื้อของทางออนไลน์กันมากขึ้น ทำให้เกิดการแข่งขันทางการตลาดที่รุนแรงและแย่งชิงลูกค้า ดังนั้นการดำเนินธุรกิจอีคอมเมิร์ซจึงไม่ใช่เรื่องง่าย การป้องกันลูกค้าไม่ให้กระโดดข้ามไปใช้บริการของคู่แข่งอื่น นั้นจึงเป็นเรื่องที่ท้าทายและสำคัญมากที่จะทำให้ธุรกิจอยู่รอด

อัตราการเลิกซื้อสินค้าของลูกค้าในช่วงระยะเวลาหนึ่ง (Customer Churn) หมายถึงการที่ผู้บริโภคเลิกสนใจ และตัดสินที่จะเลิกเป็นลูกค้า (สิรภัทร เกาฏีระ, ม.ป.ป.) ซึ่งการคาดคะเนการเลิกเป็นลูกค้านั้นเป็นประเด็นที่นักวิจัยสนใจ และหาเทคนิควิธีต่าง ๆ เพื่อทำนายการเลิกใช้งานของลูกค้า เพราะเชื่อว่ากลยุทธ์ทางการตลาดที่ดี คือการรักษาลูกค้าเก่าที่มีอยู่หรือหลีกเลี่ยงการเปลี่ยนใจเลิกใช้ของลูกค้า เนื่องจากการคาดการณ์แนวโน้มที่ลูกค้าจะเลิกซื้อสินค้าจึงไม่ใช่เรื่องง่าย อีกทั้งการเลิกเป็นลูกค้าก็ไม่ได้มาจากปัจจัยหนึ่งเพียงอย่างเดียว ทั้งนี้มีหลายปัจจัยมากที่เกี่ยวข้อง เช่น ปัจจัยความไม่พอใจในสินค้าหรือบริการ เนื่องจากลูกค้าไม่พอใจในคุณภาพประสิทธิภาพ หรือคุณค่าของสินค้า รวมถึงการบริการลูกค้าที่ไม่ดี จนทำให้สูญเสียความรู้สึกไว้ใจ

และปัจจัยความต้องการของลูกค้าที่เปลี่ยนแปลง พบว่าความต้องการและความชอบของลูกค้าอาจเปลี่ยนแปลงไปตามกาลเวลาหากสินค้าหรือบริการไม่ปรับตัวหรือไม่ตอบสนองความต้องการที่เปลี่ยนแปลงไป ซึ่งพวกเขาอาจมองหาทางเลือกที่สอดคล้องกับความต้องการในปัจจุบันของพวกเขามากขึ้น และปัจจัยปัญหาหรือข้อร้องเรียนที่ยังไม่ได้รับการแก้ไข หากปัญหาหรือข้อร้องเรียนของลูกค้าไม่ได้รับการแก้ไขอย่างเต็มที่หรือได้รับการแก้ไขอย่างทันท่วงทีอาจทำให้เกิดความไม่พอใจและสูญเสียลูกค้าในที่สุด สิ่งสำคัญที่ธุรกิจต้องเข้าใจปัจจัยเฉพาะที่ผลักดันให้เกิดการสูญเสียในภาคอุตสาหกรรมและฐานลูกค้า ด้วยการระบุและแก้ไขปัจจัยเหล่านี้ธุรกิจสามารถพัฒนากลยุทธ์การรักษาเป้าหมายลูกค้าเพื่อลดการสูญเสียและเพิ่มความพึงพอใจและความภักดีของลูกค้า จากเหตุผลดังกล่าวทำผู้วิจัยเล็งเห็นความสำคัญของการใช้การเรียนรู้ของเครื่องในการทำนายการเลิกเป็นลูกค้า เนื่องจากมีความสามารถในการวิเคราะห์ข้อมูลจำนวนมากและระบุรูปแบบที่อาจบ่งบอกถึงการเลิกเป็นลูกค้าที่อาจเกิดขึ้น และจับความสัมพันธ์ที่ซับซ้อนระหว่างตัวแปร ดังนั้นจึงจำเป็นต้องใช้เทคนิคการเรียนรู้ของเครื่องที่สามารถสร้างแบบจำลองการเรียนรู้ของข้อมูลและทำนายผลได้ โดยอาศัยชุดข้อมูลของเว็บไซต์แห่งหนึ่ง เช่น ข้อมูลประเภทประชากร ข้อมูลพฤติกรรมของลูกค้า และประวัติการซื้อสินค้าย้อนหลัง เพื่อหาตัวแปรที่สำคัญที่ส่งผลต่อประสิทธิภาพผลการทำนาย

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

### 1.2 **จุดประสงค์ของงานวิจัย**

1. เพื่อคาดการณ์ลูกค้าที่มีแนวโน้มที่จะเลิกเป็นลูกค้าโดยใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน เพื่อให้สามารถรักษาลูกค้าเอาไว้ได้

2. เพื่อศึกษาว่าคุณลักษณะใดที่มีความสำคัญและการตีความแบบจำลองที่ได้ด้วยการอธิบายแบบจำลอง

### 1.3 ขอบเขตของการวิจัย

งานวิจัยนี้ใช้ข้อมูลของลูกค้าบนเว็บไซต์แห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com ซึ่งข้อมูลถูกรวบรวมตั้งแต่ปี ค.ศ.2015-2017 ซึ่งประกอบด้วย 23 แอททริบิวต์หรือตัวแปร ซึ่งรวมเลเบลแล้วตามตาราง 1 ในการทำการทดสอบแบบจำลองข้อมูลที่ใช้ในการทำนายจะมี 2 ค่าเท่านั้น คือ 0 หมายถึง ลูกค้ายังใช้บริการอยู่หรือ Exist และ 1 หมายถึง ลูกค้าที่เลิกใช้บริการแล้วหรือ Churn โดยมีข้อมูลของลูกค้าทั้งหมด 36,992 ตัวอย่าง ซึ่งเป้าหมายคือต้องการคาดการณ์ลูกค้าที่มีแนวโน้มที่จะเลิกเป็นลูกค้าโดยใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน ซึ่งแบบจำลองที่เลือกใช้คือ Logistic Regression, Support Vector Machines, และ Random Forest อีกทั้งยังจัดการกับปัญหาหาความไม่สมดุลของข้อมูลด้วยวิธี Synthetic Minority Oversampling Technique (SMOTE) เนื่องจากงานวิจัยนี้เป็นการศึกษาปัญหาการทำนายผลในรูปแบบการจำแนกกลุ่มข้อมูล (Classification) จึงใช้การวัดผลประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score รวมถึงการพิจารณาด้วย Confusion Matrix อีกทั้งยังแสดงคุณลักษณะที่สำคัญหรือ Feature Importance และ Local Interpretable Model-agnostic Explanations (LIME)

ตาราง 1 ข้อมูลตัวแปรของชุดข้อมูลในการดำเนินงานวิจัย

| ลำดับ | ชื่อตัวแปร | ข้อมูลภายในตัวแปร | คำอธิบายตัวแปรของข้อมูล |
|---|---|---|---|
| 1 | age | ปี | อายุของลูกค้า |
| 2 | gender | ‘F’,<br>‘M’,<br>‘Unknown’ | เพศของลูกค้า |
| 3 | security_no | รหัสที่ไม่ซ้ำกัน | รหัสหมายเลขที่ไม่ซ้ำกันของลูกค้าเพื่อระบุตัวบุคคล |
| 4 | region_category | ‘City’,<br>‘Village’,<br>‘Town’ | พื้นที่อยู่อาศัยของลูกค้า |
| 5 | membership_category | ‘Platinum Membership’,<br>‘Premium Membership’,<br>‘No Membership’,<br>‘Silver Membership’,<br>‘Gold Membership’,<br>‘Basic Membership’ | ระดับการเป็นสมาชิกของลูกค้า |

ตาราง 1 (ต่อ)

| ลำดับ | ชื่อตัวแปร | ข้อมูลภายในตัวแปร | คำอธิบายตัวแปรของข้อมูล |
|---|---|---|---|
| 6 | joining_date | เดือน / วัน / ปี | วันที่ลูกค้าเข้าร่วมเป็นสมาชิก |
| 7 | joined_through_referral | 'Yes', 'No' | ลูกค้าเข้าร่วมเป็นสมาชิกโดย Code หรือ ID |
| 8 | referral_id | รหัส | การระบุตัวตนที่อ้างอิงการเป็นสมาชิกของลูกค้าด้วย Code หรือ ID |
| 9 | preferred_offer_types | 'Gift Vouchers/Coupons', 'Credit/Debit Card Offers', 'Without Offers' | ข้อเสนอของลูกค้า |
| 10 | medium_of_operation | 'Desktop', 'Smartphone', 'Both' | สื่อดำเนินการที่ลูกค้าใช้ทำธุรกรรม |
| 11 | Internet_option | 'Wi-Fi', 'Fiber_Optic', 'Mobile_Data' | ประเภทบริการอินเทอร์เน็ตที่ลูกค้าใช้ |
| 12 | last_visit_time | ชั่วโมง / นาที / วินาที | เวลาที่ลูกค้าเข้ามาเว็บไซต์ครั้งล่าสุด |
| 13 | days_since_last_login | จำนวนวัน | จำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุด |
| 14 | avg_time_spent | นาที | เวลาที่ใช้เวลาบนเว็บไซต์โดยเฉลี่ยของลูกค้า |
| 15 | avg_transaction_value | มูลค่า | มูลค่าการซื้อโดยเฉลี่ยของลูกค้า |
| 16 | avg_frequency_login_days | จำนวนนับครั้ง | จำนวนนับครั้งการเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยของลูกค้า |
| 17 | points_in_wallet | ค่าคะแนน | คะแนนสะสมที่ลูกค้าได้รับในแต่ละการทำธุรกรรม |
| 18 | used_special_discount | 'Yes', 'No' | ลูกค้าต้องการใช้ส่วนลดพิเศษ |
| 19 | offer_application_preference | 'Yes', 'No' | ลูกค้าต้องการใบเสร็จ |
| 20 | past_complaint | 'Yes', 'No' | ลูกค้าต้องการร้องเรียน |
| 21 | complaint_status | 'Not Applicable', 'Solved', 'Solved in Follow-up', 'Unsolved', 'No Information Available' | สถานะการร้องเรียนจากลูกค้า |

ตาราง 1 (ต่อ)

| ลำดับ | ชื่อตัวแปร | ข้อมูลภายในตัวแปร | คำอธิบายตัวแปรของข้อมูล |
|---|---|---|---|
| 22 | feedback | 'Products always in Stock',<br>'Quality Customer Care',<br>'User Friendly Website',<br>'Reasonable Price',<br>'Poor Website',<br>'No reason specified',<br>'Poor Product Quality',<br>'Poor Customer Service',<br>'Too many ads' | การแสดงความคิดเห็นของลูกค้า |
| 23 | churn_risk_score<br>(Target) | '0', '1' | ตัวแปรเป้าหมายหรือเลเบล<br>(0 หรือลูกค้ายังใช้บริการอยู่,<br>1 หรือลูกค้าที่เลิกใช้บริการแล้ว) |

### 1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถเป็นเครื่องมือสำหรับใช้ในการตรวจสอบพฤติกรรมของลูกค้าที่มีแนวโน้มที่จะเลิกใช้บริการ เพื่อสร้างแคมเปญหรือโปรโมชั่นดึงดูดลูกค้าตลอดจนใช้ในการวางแผนกลยุทธทางการตลาด (Marketing Strategy) เพื่อสร้างความจงรักภักดีในแบรนด์ให้แก่ลูกค้า (Customer Loyalty)

2. สามารถเข้าใจถึงคุณลักษณะที่ส่งผลต่อการเลิกใช้บริการ เพื่อป้องกันและจัดหากลยุทธ์ได้อย่างเหมาะสม

# บทที่ 2

# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง โดยนำเสนอตามหัวข้อต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.2 งานวิจัยที่เกี่ยวข้อง

## 2.1 ทฤษฎีที่เกี่ยวข้อง

### 2.1.1 การเรียนรู้ของเครื่อง (Machine Learning)

Machine Learning เป็นสาขาหนึ่งของปัญญาประดิษฐ์ที่เกี่ยวข้องกับการใช้อัลกอริทึมเพื่อให้คอมพิวเตอร์สามารถเรียนรู้จากข้อมูลและทำการคาดการณ์หรือการตัดสินใจได้โดยไม่ต้องมีการตั้งโปรแกรมที่เน้นการพัฒนาขั้นตอนวิธีที่สามารถเรียนรู้และคาดการณ์ข้อมูลได้ สิ่งที่ได้ออกมาจากการเรียนรู้คือ ผลลัพธ์หรือ Output ซึ่งขั้นตอนวิธีการเรียนรู้ของเครื่องสามารถแบ่งได้เป็น 3 ประเภทอย่างกว้าง ๆ คือ การเรียนรู้แบบมีผู้สอน การเรียนรู้แบบไม่มีผู้สอน และการเรียนรู้แบบเสริมกำลัง (Mahesh, 2018)

2.1.1.1 Supervised Learning หรือการเรียนรู้แบบมีผู้สอน คือการฝึกอบรมรูปแบบข้อมูลที่มีป้ายกำกับที่ Input และ Output อัลกอริทึมการเรียนรู้แบบมีผู้สอนที่ใช้สำหรับงาน เช่น การจำแนกประเภท และการถดถอย

2.1.1.2 Unsupervised Learning หรือการเรียนรู้แบบไม่มีผู้สอน คือการฝึกอบรมรูปแบบข้อมูลที่ไม่มีป้ายกำกับที่ตัวแปรอินพุต อัลกอริทึมการเรียนรู้แบบไม่มีผู้สอนใช้สำหรับงาน เช่น การจัดกลุ่ม และการลดมิติ

2.1.1.3 Reinforcement Learning หรือการเรียนรู้แบบเสริมกำลัง คือการเรียนรู้ของเอเจนต์ในการตัดสินใจในสภาพแวดล้อมโดยรับข้อเสนอแนะในรูปแบบของรางวัลหรือการลงโทษ อัลกอริทึมการเรียนรู้แบบเสริมกำลังถูกนำมาใช้สำหรับงาน เช่น การเล่นเกม และหุ่นยนต์

### 2.1.1 ทฤษฎีอัลกอริทึม Logistic Regression

Logistic Regression เป็นวิธีการทางสถิติที่ใช้ในการวิเคราะห์ชุดข้อมูลซึ่งมีตัวแปรอิสระหนึ่งตัวหรือมากกว่าที่เป็นไปได้ในผลลัพธ์ เป็นการวิเคราะห์การถดถอยชนิดหนึ่งที่ใช้กันทั่วไปในการทำนายผลลัพธ์ของตัวแปรตามที่เป็นหมวดหมู่ตามตัวแปรทำนายหนึ่งตัวหรือมากกว่า

กล่าวอีกนัยหนึ่ง มันถูกใช้เพื่อจำลองความน่าจะเป็นของคลาสหรือเหตุการณ์บางอย่างที่มีอยู่ เช่น ผ่าน/ไม่ผ่าน ชนะ/แพ้ หรือสุขภาพดี/ป่วย ในการสร้างแบบจำลองความสัมพันธ์ระหว่างตัวแปรอิสระและความน่าจะเป็นของผลลัพธ์ Logistic Function จะแปลงค่าของตัวแปรอิสระให้เป็นค่าความน่าจะเป็นระหว่าง 0 ถึง 1 แสดงดังสมการที่ (1)

$$p(y = k|x) = \frac{1}{1+e^{-kw^T x}} \quad (1)$$

โดยที่

$p(y = k|x)$ คือความน่าจะเป็นที่ $x$ อยู่ในคลาส $k$

$x$ คือข้อมูลที่ต้องการทำนายคลาส

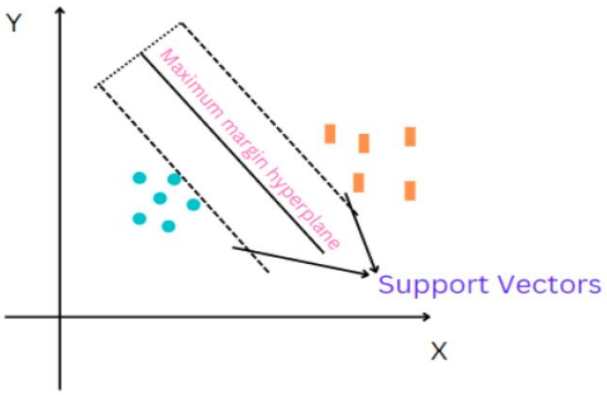
$y$ คือคลาสของข้อมูล $x$ แต่ละตัว

$k$ คือคลาสโดยที่ $k \in \{0,1, \dots\}$

$w^T$ คือ Normal Vector

**2.1.2 ทฤษฎีอัลกอริทึม Support Vector Machines (SVM)**

Support Vector Machine (SVM) เป็นอัลกอริทึมที่ใช้สำหรับการจำแนกประเภทและการวิเคราะห์การถดถอย มันทำงานโดยการค้นหาไฮเปอร์เพลนที่แยกข้อมูลออกเป็นคลาสต่าง ๆ ซึ่งไฮเปอร์เพลนที่ดีจะถูกเลือกในลักษณะที่เพิ่มระยะขอบระหว่างสองคลาสให้สูงสุด เรียกว่า Maximum Margin ระยะขอบคือระยะห่างระหว่างไฮเปอร์เพลนและจุดข้อมูลที่ใกล้เคียงที่สุดแต่ละคลาส เรียกว่า Support Vector สามารถจัดการได้ทั้งข้อมูลเชิงเส้นและไม่ใช่เชิงเส้นโดยใช้ฟังก์ชัน Kernel ที่แตกต่างกัน แสดงตามภาพประกอบ 1

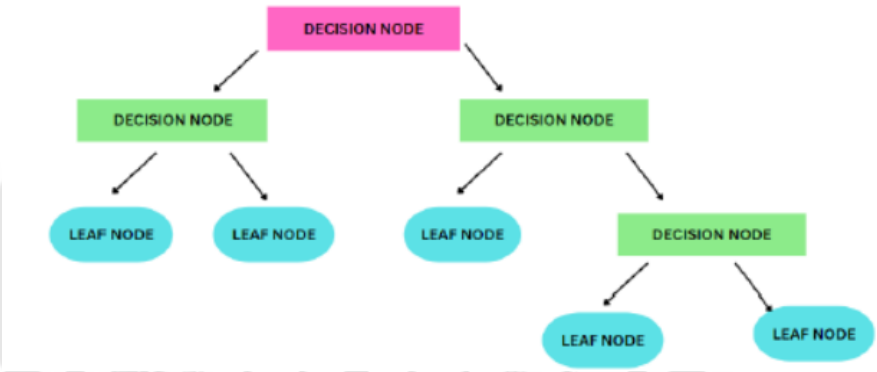



ภาพประกอบ 1 แสดง hyperplane ที่ดีที่สุดในการจำแนกข้อมูลชุดหนึ่ง

ที่มา: (Peddarapu et al., 2022)

### 2.1.3 ทฤษฎีอัลกอริทึม Random Forest

Random Forest เป็นขั้นตอนวิธีเรียนรู้ของเครื่อง (Machine Learning Algorithm) เป็นประเภทการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่สร้างต้นไม้ตัดสินใจ (Decision Tree) หลายต้นเป็นจำนวนมาก ซึ่งแต่ละอันจะได้รับการฝึกฝนบนเซตย่อยของข้อมูลแบบสุ่มในระหว่างกระบวนการฝึกอบรม ขั้นตอนวิธีจะเลือกเซตย่อยของฟีเจอร์แบบสุ่มเพื่อแยกข้อมูลแต่ละโหนดของต้นไม้ซึ่งจะช่วยลดการ Overfitting และปรับปรุงลักษณะทั่วไปของแบบจำลองการทำนายขั้นสุดท้ายจะทำโดยการรวมการคาดการณ์ของต้นไม้ทั้งหมดในอัลกอริทึม Random Forest ได้รับการแสดงให้เห็นว่ามีประสิทธิภาพในการจัดการข้อมูลมิติสูงและจัดการกับชุดข้อมูลที่ไม่สมดุล อีกทั้งยังเป็นขั้นตอนวิธีที่นิยมสำหรับการจัดหมวดหมู่ ดังภาพประกอบ 2




ภาพประกอบ 2 แสดงการทำงานของแบบจำลอง Random Forest

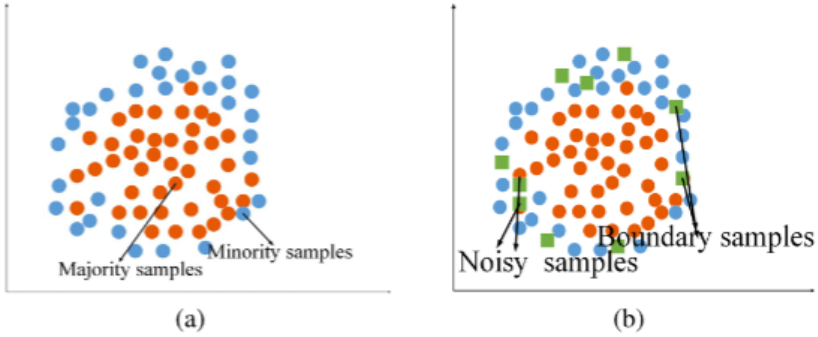
ที่มา: (Peddarapu et al., 2022)

### 2.1.4 การสังเคราะห์ข้อมูลใหม่ (Synthetic Minority Oversampling TEchnique: SMOTE)

SMOTE เป็นเทคนิค Oversampling การสุ่มตัวอย่างเป็นวิธีการปรับสมดุลข้อมูลที่ใช้ในการเรียนรู้ของเครื่องเพื่อแก้ไขปัญหาของข้อมูลที่ไม่สมดุล ซึ่งข้อมูลที่ใช้ในการจัดการกับข้อมูลที่ไม่สมดุลมันสร้างตัวอย่างสังเคราะห์สำหรับชนกลุ่มน้อยโดยการสอดแทรกระหว่างตัวอย่างที่มีอยู่ วิธีการนี้จะช่วยให้ความสมดุลของข้อมูลและปรับปรุงประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องโดยปรับสมดุลข้อมูลและลดความลำเอียงต่อชนชั้นข้างมาก

ตามภาพประกอบ 3 แสดงการกระจายของตัวอย่างหลังจากใช้เทคนิคการสังเคราะห์ข้อมูล SMOTE ประกอบด้วยภาพย่อย 2 ภาพคือ (a) Original และ (b) SMOTE โดยในภาพประกอบ (a) คือการแสดงการกระจายตัวอย่างข้อมูล Original ในชุดข้อมูลก่อนใช้ SMOTE

แสดงภาพการกระจายเบื้องต้นของตัวอย่าง เน้นย้ำถึงการมีอยู่ของ Boundary samples และความไม่สมดุลของคลาส ซึ่งมีความแตกต่างชัดเจนระหว่างสองคลาส อย่างไรก็ตามหลังจากใช้อัลกอริทึม SMOTE และในภาพประกอบ (b) ชุดข้อมูลจะมีความสมดุลมากขึ้น แสดงผลของเทคนิคการ Oversampling ด้วยวิธี SMOTE สร้างตัวอย่างสังเคราะห์เพื่อสร้างความสมดุลให้กับคลาสในชุดข้อมูล โดยการสังเคราะห์ตัวอย่างใหม่ เพื่อแก้ไขปัญหาความไม่สมดุลของคลาส ซึ่งคลาสกลุ่มน้อยมีตัวอย่างน้อยกว่าคลาสกลุ่มมากอย่างเห็นได้ชัด แต่ในบางกรณี SMOTE อาจนำไปสู่การสร้าง Noise samples และ Boundary samples ซึ่งส่งผลต่อประสิทธิภาพของอัลกอริทึมการจำแนกได้




ภาพประกอบ 3 การกระจายตัวอย่างหลังจากสังเคราะห์ด้วย SMOTE (a) Original (b) SMOTE

ที่มา: (Xu et al., 2022)

### 2.1.5 การวัดประสิทธิภาพการทำงานของอัลกอริทึม

ในงานวิจัยนี้เป็นการจัดการกับปัญหา Classification จึงใช้การวัดประสิทธิภาพด้วยค่า Accuracy, Precision, Recall, และ F1-Score ซึ่งคำนวณได้จาก Confusion Matrix ดังตาราง 2

ตาราง 2 แสดง Confusion Matrix

| | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Predict | Positive | TP | FP |
| | Negative | FN | TN |

โดยที่

True positive (TP) คือ การทำนายว่าลูกค้ามีโอกาสเลิกใช้บริการ “ถูก”

True negative (TN) คือ การทำนายว่ายังเป็นลูกค้าที่ยังใช้บริการอยู่ “ถูก”

False positive (FP) คือ การทำนายว่าลูกค้ามีโอกาสเลิกใช้บริการ “ผิด”

False negative (FN) คือ การทำนายว่ายังเป็นลูกค้าที่ยังใช้บริการอยู่ “ผิด”

**2.1.5.1 Accuracy**

คืออัตราส่วนของการทำนายถูกกับจำนวนข้อมูลทั้งหมด แสดงได้ดังสมการ (2)

$$\text{Accuracy}= \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \quad (2)$$

**2.1.5.2 Precision**

คือการเปรียบเทียบการทำนายลูกค้ามีโอกาสเลิกใช้บริการว่าจริง แล้วเกิดขึ้นจริง (TP) เทียบกับการทำนายลูกค้ามีโอกาสเลิกใช้บริการว่าจริง แต่สิ่งที่เกิดไม่จริง (FP) แสดงได้ดังสมการ (3)

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

**2.1.5.3 Recall**

คือค่าความถูกต้องของการทำนายลูกค้ามีโอกาสเลิกใช้บริการว่าจริง เทียบกับจำนวนครั้งของเหตุการณ์ ทั้งการทำนายและการเกิดขึ้นจริง ว่าเป็นจริง แสดงได้ดังสมการ (4)

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

**2.1.5.4 F1-Score**

คือค่าเฉลี่ยถ่วงน้ำหนักระหว่าง Precision และ Recall แสดงได้ดังสมการ (5)

$$\text{F1–Score} = 2 \times \left(\frac{\mathbf{Recall \times Precison}}{\mathbf{Recall+Precison}}\right) \quad (5)$$

**2.1.6 Local Interpretable Model-agnostic Explanations**

Local Interpretable Model-agnostic Explanations หรือ LIME เป็นเทคนิคที่สามารถช่วยในการอธิบายการคาดการณ์ของตัวจำแนกประเภทใด ๆ ในลักษณะที่ตีความและซื่อสัตย์ต่อแบบจำลอง โดยช่วยในการทำความเข้าใจเหตุผลที่อยู่เบื้องหลังการทำนายซึ่งเป็นสิ่งสำคัญในการประเมินความน่าเชื่อถือของแบบจำลอง เพื่อใช้เปรียบเทียบคำอธิบายต่าง ๆ และเลือกแบบที่น่าเชื่อถือที่สุด ซึ่งจะเป็นประโยชน์เมื่อตัดสินใจเลือกรูปแบบที่จะปรับใช้ในการประยุกต์ได้จริง (Ribeiro et al., 2016)

## 2.2 งานวิจัยที่เกี่ยวข้อง

การทบทวนวรรณกรรมของงานวิจัยนี้ได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายผลของลูกค้า มีรายละเอียดและสรุปผลของงานวิจัยตามตาราง 3 ดังต่อไปนี้

### 2.2.1 บทความวิจัยเรื่อง Customer Churn Prediction using Machine Learning

โดย Rama Krishna Peddarapu, Sofia Ameena, Surepally Yashaswini, Nadipelli Shreshta, และ Muppidi PurnaSahithi (Peddarapu et al., 2022)

งานวิจัยนี้นำสนอการนี้ใช้เทคนิคการเรียนรู้ของเครื่องด้วยแบบจำลอง Logistic Regression, Random Forest, SVM และ XGboost เพื่อพัฒนารูปแบบการทำนายการเลิกสมัครใช้งานของลูกค้าธนาคาร โดยข้อมูลที่ใช้ในการสร้างแบบจำลองมี 14 คุณลักษณะ และ 1,000 แถว ในงานวิจัยนี้ได้เลือกวิธี Feature Selection โดยการใช้แบบจำลอง Random Forest เพื่อเลือกคุณลักษณะที่มีความสำคัญมากที่สุด จากนั้นได้ศึกษาเปรียบเทียบวัดประสิทธิภาพของแต่ละแบบจำลองได้แก่ Accuracy, Precision, Recall และ F1-score และได้ใช้เส้นโค้ง ROC มาช่วยในการพิจารณาอีกด้วย จากผลการทดลองพบว่าแบบจำลองที่ดีที่สุดคือ Random Forest มีผลการทำนายสำหรับค่า Accuracy ที่ 86%

### 2.2.2 บทความวิจัยเรื่อง Customer Churn Reasoning in Telecommunication Domain

โดย S. Stehani, N. Karunya, D. R. J. B. Ranjan, Sagara Sumathipala, และ T. C. Sandanayake (Stehani et al., 2020)

งานวิจัยนี้นำสนอการนี้ใช้เทคนิคการเรียนรู้ของเครื่องด้วยแบบจำลอง Random Forest, Naïve Bayes, SVM, Ada Boost, Logistic Regression และ K-Nearest Neighbor เพื่อการคาดการณ์การเลิกใช้งานของลูกค้าในอุตสาหกรรมโทรคมนาคม โดยข้อมูลที่ใช้ในการสร้างแบบจำลองมีทั้งหมด 3,333 แถว และ 20 คุณลักษณะ เนื่องจากในงานนี้มีจำนวนคุณลักษณะจำนวนมากจึงทำการเลือกคุณลักษณะที่โดดเด่นออกมา อีกทั้งยังใช้เทคนิค principal component analysis (PCA) เพื่อสร้างคุณลักษณะใหม่จากข้อมูลชุดเดิม ในงานวิจัยนี้ได้ศึกษาเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง โดยใช้การวัดประสิทธิภาพในการทำงาน คือ Accuracy, Precision, Recall และ F1-score พบว่าแบบจำลอง Random Forest ให้ค่า Accuracy ที่ 89%, Naïve Bayes ให้ค่า Accuracy ที่ 0.85%, SVM ให้ค่า Accuracy ที่ 78%, Ada Boost ค่า Accuracy ที่ 84%, Logistic Regression ให้ค่า Accuracy ที่ 76%, K-Nearest Neighbor ให้ค่า Accuracy ที่ 78% จากผลการทดลองพบว่า Random Forest เป็นเทคนิคที่ Accuracy ดีที่สุด

**2.2.3 บทความวิจัยเรื่อง Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees**

โดย Sajjad Shumaly, Pedram Neysaryan, และ Yanhui Guo (Shumaly et al., 2020)

งานวิจัยนี้นำสนอการคาดการณ์การเลิกเป็นลูกค้าในอุตสาหกรรมโทรคมนาคมโดยใช้เทคนิคการเรียนรู้ของเครื่องด้วยแบบจำลอง Decision Tree, Support Vector Machine, Multi-Layer Perceptron, Random Forest และ Gradient Boosting อีกทั้งยังใช้วิธีการปรับสมดุลข้อมูลคือ Over-Sampling, Under-Sampling และ SMOTE โดยเปรียบเทียบประสิทธิภาพทุกแบบวิธีทั้งข้อมูลที่ไม่สมดุล และข้อมูลที่ปรับความสมดุลแล้ว ซึ่งงานวิจัยนี้เป็นการศึกษาข้อมูลไม่สมดุล จึงไม่ได้สนใจค่า Accuracy แต่ให้ความสนใจค่า AUC มากกว่า จากผลการทดลองพบว่าประสิทธิภาพของแบบจำลอง Random Forest กับแบบจำลอง Gradient Boosting มีประสิทธิภาพที่ใกล้เคียงกัน แต่แบบจำลองที่ดีที่สุดคือแบบจำลอง Gradient Boosting ร่วมกับการใช้วิธีปรับความสมดุลของข้อมูล Over-Sampling มีค่า Accuracy ที่ 93.5% และค่า AUC ที่ 90.1%

**2.2.4 บทความวิจัยเรื่อง Research on Customer Churn Intelligent Prediction Model based on Borderline-SMOTE and Random Forest**

โดย Linmao Feng (Feng, 2022)

งานวิจัยนี้นำเสนอการทำนายการเปลี่ยนใจของลูกค้าของธนาคาร โดยใช้เทคนิคการเรียนรู้ของเครื่องด้วยแบบจำลอง Random Forest ร่วมกับใช้วิธีการแก้ไขปัญหาข้อมูลไม่สมดุลด้วย Borderline SMOTE เพื่อนำเปรียบเทียบกับแบบจำลองอื่น ๆ ได้แก่ KNN, Decision tree และ Naïve Bayes ซึ่งได้รับการประเมินและการวัดประสิทธิภาพการทำงานของอัลกอริทึมด้วยค่า OOB Error Rate, AUC, Precision, Recall, และ F-mean จากผลการทดลองพบว่าแบบจำลอง Random Forest ร่วมกับใช้วิธี Borderline-SMOTE มีประสิทธิภาพดีที่สุดมีค่า OOB Error Rate ที่ 92.3%, AUC ที่ 92.1%, Precision ที่ 90.3%, Recall ที่ 94.4%, และ F-mean ที่ 92.3%

**2.2.5 บทความวิจัยเรื่อง E-Commerce Customer Churn Prediction By Gradient Boosted Trees**

โดย Shamim Raeisi และ Hedieh Sajedi (Raeisi & Sajedi, 2020)

งานวิจัยนี้นำเสนอการใช้เทคนิคการเรียนรู้ของเครื่องด้วยแบบจำลอง Gradient Boosted Trees, KNN Decision Trees, Naïve Bayes Random Forest, และ Rule Induction เพื่อทำนายลูกค้าที่เลิกซื้อในบริการสั่งซื้ออาหารออนไลน์ในกรุงเตหะราน ประเทศอิหร่าน จากผล

การทดลองพบว่าแบบจำลอง Gradient Boosted Trees มีค่า Accuracy ที่ 86.9% ซึ่งเป็นวิธีที่มีประสิทธิภาพมากที่สุด

**2.2.6 บทความวิจัยเรื่อง Churn Prediction: A Comparative Study Using KNN and Decision Trees**

โดย Mohammad A. Hassonah, Ali Rodan, Abdel-Karim Al-Tamimi, และ Jamal Alsakran (Hassonah et al., 2019)

งานวิจัยนี้เสนอการศึกษาเปรียบเทียบประสิทธิภาพของการเรียนรู้ของเครื่องระหว่าง 2 อัลกอริทึม ซึ่งได้แก่ Decision Tree และ K-Nearest Neighbor เพื่อทำนายการเลิกเป็นลูกค้าข้อมูลของบริษัทโทรคมนาคม โดยข้อมูลชุดข้อมูลประกอบด้วย 3,333 ตัวอย่างของลูกค้าและ 20 ตัวแปร โดยการตั้งค่าอัลกอริทึม KNN กำหนดให้ k=5 และการวัดแบบยุคลิดแบบผสม ส่วนอัลกอริทึม Decision Tree ใช้เกณฑ์ Gini Index ที่มีระดับความลึก 20 ระดับ จากผลการทดลองพบว่าอัลกอริทึม K-Nearest Neighbor มีค่า Accuracy ที่ 87% มีค่า Precision ที่ 61% มีค่า Recall ที่ 22% มีค่า F-Measure ที่ 33% และค่าพื้นที่ใต้เส้นโค้ง (AUC) มีประมาณ 0.82 ส่วนอัลกอริทึม Decision Tree มีค่า Accuracy ที่ 93% มีค่า Precision ที่ 77% มีค่า Recall ที่ 68% มีค่า F-Measure ที่ 73% และค่าพื้นที่ใต้เส้นโค้ง (AUC) มีประมาณ 0.86 ดังนั้นอัลกอริทึม Decision Tree จึงมีประสิทธิภาพที่ดีมากกว่าอัลกอริทึม K-Nearest Neighbor

**2.2.7 บทความวิจัยเรื่อง Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network**

โดย Xin Hu, Yanfei Yang, Lanhua Chen, และ Siru Zhu (Hu et al., 2020)

งานวิจัยนี้นำเสนอการออกแบบแบบจำลองการทำนายแบบผสมผสาน (Combined Prediction Model) โดยใช้แบบจำลองต้นไม้ตัดสินใจ (Decision Tree) และแบบจำลองโครงข่ายประสาทเทียม (Neural Network Model) ผสมรวมเป็นแบบจำลองเดียวกัน และนำไปเปรียบเทียบกับแบบจำลอง Decision Tree และ Neural Network Model ในการทำนายการเลิกเป็นลูกค้าของชุดข้อมูลลูกค้าซุปเปอร์มาร์เก็ต 2,681 คน จากการทดลองพบว่า แบบจำลอง Decision Tree มีค่า Accuracy ที่ 93.47% ,แบบจำลอง Neural Network มีค่า Accuracy ที่ 96.42% และการทำนายแบบผสมผสาน Combined Prediction Model มีค่า Accuracy ที่ 98.87% ซึ่งแสดงว่าแบบจำลองการทำนายแบบผสมผสานมีประสิทธิภาพการทำนายดีกว่าแบบจำลองอื่น ๆ

**2.2.8 บทความวิจัยเรื่อง Prediction of Customer Retention Rate Employing Machine Learning Techniques**

โดย Achintya Sharma, Deepak Gupta, Nikhil Nayak, Deepti Singh, และ Ankita Verma (Sharma et al., 2022)

งานวิจัยนี้นำเสนอการใช้เทคนิคการเรียนรู้ของเครื่อง ได้แก่ Logistic Regression, Support Vector Machine (SVM), Decision Tree, XGBoost, Random Forest, Light Gradient Boosting, Gradient Descent Boosting และ Cat Boost เพื่อวิเคราะห์ผลกระทบของการลดคุณลักษระในงานเพื่อการทำนายการเลิกเป็นลูกค้าในอุตสาหกรรมโทรคมนาคม ซึ่งมีชุดข้อมูลลูกค้า 7,043 คน และคุณลักษณะ 21 ชนิด ในการศึกษานี้เป็นการเปรียบเทียบการใช้เทคนิคการเลือกคุณลักษณะและการลดขนาดข้อมูล โดยการทดลองแบ่งออกเป็น 4 แบบ ได้แก่ ชุดข้อมูลเดิม (Original Dataset), PCA, Information Gain , และลดคุณลักษณะของชุดข้อมูล (Reduced Dataset) โดยวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy จากการทดลองพบว่าอัลกอริทึม XGBoost มีประสิทธิภาพดีที่โดยใช้วิธี Reduced Dataset มีค่า Accuracy ที่ 81.99%

**2.2.9 บทความวิจัยเรื่อง Customer churn model based on complementarity measure and random forest**

โดย Chen Zhang, Hong Li, Guangde Xu, และ Xuhui Zhu (Zhang et al., 2021)

งานวิจัยนี้นำเสนอการทำนายการเลิกเป็นลูกค้าของธนคาร โดยใช้ข้อมูลที่ถูกเก็บรวบรวมโดยธนาคารตั้งแต่เดือนกรกฎาคม 2018 ถึงมีนาคม 2019 ชุดข้อมูลประกอบด้วยข้อมูลลูกค้าหนึ่งล้านคนและตัวแปรทั้งหมด 181 แอตทริบิวต์ จากนั้นใช้วิธี Affinity Propagation (AP) Clustering เพื่อเลือกแอตทริบิวต์ที่เกี่ยวข้อง ทั้งนี้ได้ใช้อัลกอริทึมหลายแบบเพื่อเปรียบเทียบหาประสิทธิภาพได้แก่ SVM, Back Propagation (BP), CART, Deep Belief Network (DBN), Random Forest เป็นเทคนิคแบบ Bootstrap และ Random Forest ร่วมกับ Complementarity Measure (CM+RF) ซึ่งเป็นเทคนิคแบบ Pruning โดยแต่ละแบบจำลองได้จัดการกับปัญหาความไม่สมดุลของข้อมูลด้วยวิธี Random Over Sampling (ROS) และใช้เทคนิค Cross Validation โดยแบ่งข้อมูลออกเป็น 5 Fold ในงานวิจัยนี้ใช้การประเมินวัดประสิทธิภาพของแบบจำลองดังนี้ Accuracy, Precision, Recall, F-measure, AUC, AUPRC จากการทดลองพบว่าแบบจำลอง CM+RF เป็นวิธีที่มีประสิทธิภาพดีที่สุดมีค่า Accuracy ที่ 82.32%, ค่า Precision ที่ 86.25%, ค่า Recall ที่ 76.23%, ค่า F-measure ที่ 81.02%, ค่า AUC ที่ 0.821, ค่า AUPRC ที่ 0.857

**2.2.10 บทความวิจัยเรื่อง Machine Learning Based Telecom-Customer Churn Prediction**

โดย Pushkar Bhuse, Aayushi Gandhi, Parth Meswani, Riya Muni, และ Neha Katre (Bhuse et al., 2020)

งานวิจัยนี้นำเสนอการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) และการเรียนรู้เชิงลึก (Deep Learning) เพื่อทำนายการเลิกเป็นลูกค้าด้วยข้อมูลลูกค้าด้านโทรคมนาคม โดยเปรียบเทียบประสิทธิภาพของอัลกอริทึมแบบต่าง ๆ ได้แก่ Ridge Classifier, Random Forest, SVM, K-Nearest Neighbors (KNN), XGBoost, และ Deep Neural Networks จากการทดลองพบว่าแบบจำลอง Random Forest มีประสิทธิภาพดีที่สุดมีค่า Accuracy ที่ 90.96% แต่หลังจากใช้เทคนิคการค้นหาแบบกริด (Grid Search) พบว่ามีค่า Accuracy เพิ่มมากขึ้นเป็น 91.26%

ตาราง 3 แสดงผลสรุปของงานวิจัยที่เกี่ยวข้อง

| ลำดับ | ชื่องานวิจัย | วัตถุประสงค์ | แบบจำลองที่ใช้เปรียบเทียบ | แบบจำลองที่มีประสิทธิภาพดีที่สุด |
|---|---|---|---|---|
| 1 | Customer Churn Prediction using Machine Learning | การทำนายการเลิกสมัครใช้งานของลูกค้าธนาคาร | Logistic Regression, Decision Tree, Random Forest, Gradient Boosting | Random Forest มีค่า Accuracy ที่ 86% |
| 2 | Customer Churn Reasoning in Telecommunication Domain | การคาดการณ์การเลิกใช้งานของลูกค้าในอุตสาหกรรมโทรคมนาคม | Random Forest, Naïve Bayes, SVM, Ada Boost, Logistic Regression, K-Nearest Neighbor | Random Forest มีค่า Accuracy ที่ 89% |
| 3 | Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees | การคาดการณ์การเลิกเป็นลูกค้าในอุตสาหกรรมโทรคมนาคม | Decision Tree, Support Vector Machine, Multi-Layer Perceptron, Random Forest, Gradient Boosting | Random Forest ร่วมกับการใช้วิธี Over-Sampling มีค่า Accuracy ที่ 93.5% |

ตาราง 3 (ต่อ)

| ลำดับ | ชื่องานวิจัย | วัตถุประสงค์ | แบบจำลองที่ใช้เปรียบเทียบ | แบบจำลองที่มีประสิทธิภาพดีที่สุด |
|---|---|---|---|---|
| 4 | Research on Customer Churn Intelligent Prediction Model based on Borderline-SMOTE and Random Forest | การทำนายการสูญเสียลูกค้าของธนาคาร | Random Forest, KNN, Decision tree, Naïve Bayes | Random Forest ร่วมกับการใช้วิธี Borderline-SMOTE มีค่า OOB Error Rate ที่ 92.3%, AUC ที่ 92.1%, Precision ที่ 90.3%, Recall ที่ 94.4%, และ F-mean ที่ 92.3% |
| 5 | E-Commerce Customer Churn Prediction By Gradient Boosted Trees | การทำนายลูกค้าที่เลิกซื้อในบริการสั่งซื้ออาหารออนไลน์ในกรุงเตหะราน ประเทศอิหร่าน | Gradient Boosted Trees, KNN, Decision Trees, Naïve Bayes, Random Forest, Artificial Neural Network | Gradient Boosted Trees มีค่า Accuracy ที่ 86.9% |
| 6 | Churn Prediction: A Comparative Study Using KNN and Decision Trees | การทำนายการเลิกเป็นลูกค้าข้อมูลของบริษัทโทรคมนาคม | Decision Tree และ K-Nearest Neighbor | Decision Tree มีค่า Accuracy ที่ 93% มีค่า |
| 7 | Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network | การทำนายการเลิกเป็นลูกค้าของชุดข้อมูลลูกค้าซุปเปอร์มาร์เก็ต | Decision Tree, Neural Network, Combined Prediction Model (Decision Tree ร่วมกับ Neural Network) | Combined Prediction Model มีค่า Accuracy ที่ 98.87% |

ตาราง 3 (ต่อ)

| ลำดับ | ชื่องานวิจัย | วัตถุประสงค์ | แบบจำลองที่ใช้เปรียบเทียบ | แบบจำลองที่มีประสิทธิภาพดีที่สุด |
|---|---|---|---|---|
| 8 | Prediction of Customer Retention Rate Employing Machine Learning Techniques | การทำนายการเลิกเป็นลูกค้าในอุตสาหกรรมโทรคมนาคม | Logistic Regression, Support Vector Machine (SVM), Decision Tree, XGBoost, Random Forest, Light Gradient Boosting, Gradient Descent Boosting, Cat Boost | XGBoost โดยใช้วิธี Reduced Dataset มีค่า Accuracy ที่ 81.99% |
| 9 | Customer churn model based on complementarity measure and random forest | การทำนายการเลิกเป็นลูกค้าของธนคาร | SVM, Back Propagation (BP), CART, Deep Belief Network (DBN), Random Forest, Random Forest ร่วมกับ Complementarity Measure (CM+RF) | CM+RF มีค่า Accuracy ที่ 82.32% |
| 10 | Machine Learning Based Telecom-Customer Churn Prediction | การทำนายการเลิกเป็นลูกค้าด้วยข้อมูลลูกค้าด้านโทรคมนาคม | Ridge Classifier, Random Forest, Support Vector Classifier (SVC), K-nearest neighbors (KNN), XGBoost, Deep Neural Networks | Random Forest มีค่า Accuracy ที่ 91.26% |

# บทที่ 3

# การดำเนินการวิจัย

ในการวิจัยนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

3.1 กระบวนการทำงานของแบบจำลอง

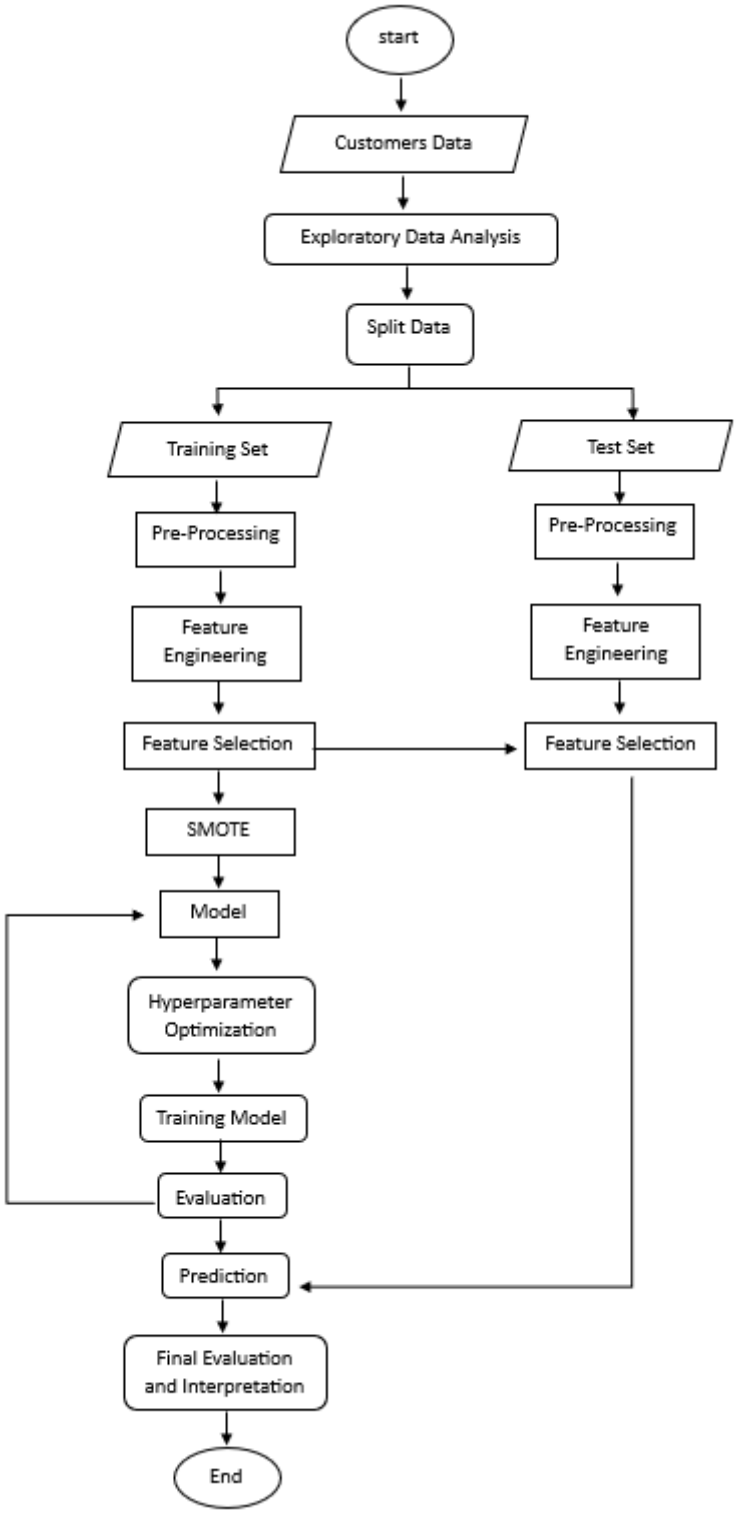
3.2 การเก็บรวบรวมข้อมูลและจัดการกับข้อมูล

3.3 กระบวนการสำรวจข้อมูล (Exploratory Data Analysis)

3.4 การเตรียมความพร้อมข้อมูล (Data Preprocessing)

3.5 การสร้างแบบจำลองเพื่อทำนายกลุ่ม

### 3.1 กระบวนการทำงานของแบบจำลอง




ภาพประกอบ 4 แสดงกระบวนการทำงานของแบบจำลองในงานวิจัยนี้

จากภาพประกอบ 4 แสดงขั้นตอนการสร้างแบบจำลองการทำนายแนวโน้มการเลิกเป็นลูกค้า โดยขั้นตอนแรกคือ การนำเข้าข้อมูลของลูกค้าบนเว็บไซต์ หลังจากนำเข้าแล้วจึงทำการสำรวจข้อมูล (Exploratory Data Analysis) เพื่อเข้าใจในข้อมูลเบื้องต้น จากนั้นการทำความสะอาดข้อมูล (Data Cleansing) ใช้วิธีเติมค่าว่างและลบค่าว่างบางส่วนทิ้ง จากนั้นทำการแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลสำหรับการเรียนรู้ของแบบจำลอง (Training Set) และข้อมูลสำหรับการทดสอบประสิทธิภาพของแบบจำลอง (Test Set) ในขั้นตอนการเตรียมพร้อมของข้อมูล (Data Preprocessing) และจัดการกับข้อมูลหมวดหมู่ (Categorical Features) แบบไม่มีลำดับให้อยู่ในรูปแบบตัวเลข โดยเลือกใช้วิธีการ One-Hot Encoding และปรับเปลี่ยนช่วงของข้อมูลตัวเลข โดยใช้วิธี Standard Scaler เพื่อให้ง่ายต่อการทำงาน จากนั้นทำการคัดเลือกคุณลักษณะหรือ Feature Selection ที่สำคัญ และมีการสังเคราะห์ข้อมูลของกลุ่มเป้าหมายที่มีน้อยให้เพิ่มข้อมูลมากขึ้นด้วย SMOTE โดยในงานวิจัยนี้เลือกใช้แบบจำลอง Logistic Regression, Support Vector Machines และ Random Forest พร้อมทั้งทำการปรับจูนพารามิเตอร์ด้วย เมื่อทำการประมวลผลการเรียนรู้เสร็จแล้ว ในขั้นตอนต่อไปจึงนำข้อมูลชุดทดสอบไปใช้ในการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix ทำการเปรียบเทียบค่าวัดประสิทธิภาพแต่ละแบบจำลอง เพื่อค้นหาแบบจำลองที่ดีที่สุดบนชุดข้อมูลทดสอบ จากนั้นทำการวิเคราะห์ตีความหมายของแบบจำลองหาตัวแปรผลที่สำคัญที่ส่งผลกระทบต่อการทำนายด้วยเทคนิคในการตีความคือ Local Interpretable Model-agnostic Explanations (LIME)

### 3.2 การเก็บรวบรวมข้อมูลและจัดการกับข้อมูล

ในงานวิจัยนี้ได้ใช้ข้อมูลประชากรของลูกค้าในเว็บไซต์แห่งหนึ่ง ซึ่งเป็นข้อมูลสาธารณะจาก Kaggle.com ซึ่งประกอบด้วย 23 แอททริบิวต์ มีจำนวนข้อมูลทั้งหมด 36,992 ตัวอย่าง และข้อมูลอยู่ในรูปแบบ Comma-Separated Values (CSV) โดยการทดลองได้ใช้ Google Colab และภาษา Python ตามภาพประกอบ 5, 6 และ 7 เป็นการแสดงตัวอย่างข้อมูลตั้งแต่แอททริบิวต์ที่ 1 ถึง 23 เพื่อดูภาพรวมของข้อมูลก่อนที่จะนำไปทำความสะอาดข้อมูลและคัดเลือกแอททริบิวต์ที่จำเป็นที่จะนำไปใช้ทดลองในแบบจำลอง

| | age | gender | security_no | region_category | membership_category | joining_date | joined_through_referral | referral_id |
|---|---|---|---|---|---|---|---|---|
| **0** | 18 | F | XW0DQ7H | Village | Platinum Membership | 2017-08-17 | No | xxxxxxxx |
| **1** | 32 | F | 5K0N3X1 | City | Premium Membership | 2017-08-28 | ? | CID21329 |
| **2** | 44 | F | 1F2TCL3 | Town | No Membership | 2016-11-11 | Yes | CID12313 |
| **3** | 37 | M | VJGJ33N | City | No Membership | 2016-10-29 | Yes | CID3793 |
| **4** | 31 | F | SVZXCWB | City | No Membership | 2017-09-12 | No | xxxxxxxx |
| **5** | 13 | M | PSG1LGF | City | Gold Membership | 2016-01-08 | No | xxxxxxxx |
| **6** | 21 | M | R3CX1EA | Town | Gold Membership | 2015-03-19 | Yes | CID24708 |
| **7** | 42 | M | 4UJ1551 | NaN | No Membership | 2016-07-12 | ? | CID56614 |
| **8** | 44 | M | 0481QNQ | Village | Silver Membership | 2016-12-14 | No | xxxxxxxx |
| **9** | 45 | F | ZHP4MCR | Town | No Membership | 2016-11-30 | No | xxxxxxxx |

ภาพประกอบ 5 แสดงตัวอย่างข้อมูลที่ใช้สำหรับแบบจำลองด้วยแอททริบิวต์ที่ 1 ถึง 8

| | preferred_offer_types | medium_of_operation | internet_option | last_visit_time | days_since_last_login | avg_time_spent | avg_transaction_value | avg_frequency_login_days |
|---|---|---|---|---|---|---|---|---|
| **0** | Gift Vouchers/Coupons | ? | Wi-Fi | 16:08:02 | 17 | 300.63 | 53005.25 | 17.0 |
| **1** | Gift Vouchers/Coupons | Desktop | Mobile_Data | 12:38:13 | 16 | 306.34 | 12838.38 | 10.0 |
| **2** | Gift Vouchers/Coupons | Desktop | Wi-Fi | 22:53:21 | 14 | 516.16 | 21027.00 | 22.0 |
| **3** | Gift Vouchers/Coupons | Desktop | Mobile_Data | 15:57:50 | 11 | 53.27 | 25239.56 | 6.0 |
| **4** | Credit/Debit Card Offers | Smartphone | Mobile_Data | 15:46:44 | 20 | 113.13 | 24483.66 | 16.0 |
| **5** | Gift Vouchers/Coupons | ? | Wi-Fi | 06:46:07 | 23 | 433.62 | 13884.77 | 24.0 |
| **6** | Gift Vouchers/Coupons | Desktop | Mobile_Data | 11:40:04 | 10 | 55.38 | 8982.50 | 28.0 |
| **7** | Credit/Debit Card Offers | Both | Fiber_Optic | 07:52:43 | 19 | 429.11 | 44554.82 | 24.0 |
| **8** | Without Offers | Smartphone | Fiber_Optic | 06:50:10 | 15 | 191.07 | 18362.31 | 20.0 |
| **9** | Gift Vouchers/Coupons | ? | Wi-Fi | 19:10:16 | 10 | 97.31 | 19244.16 | 28.0 |

ภาพประกอบ 6 แสดงตัวอย่างข้อมูลที่ใช้สำหรับแบบจำลองด้วยแอททริบิวต์ที่ 9 ถึง 16

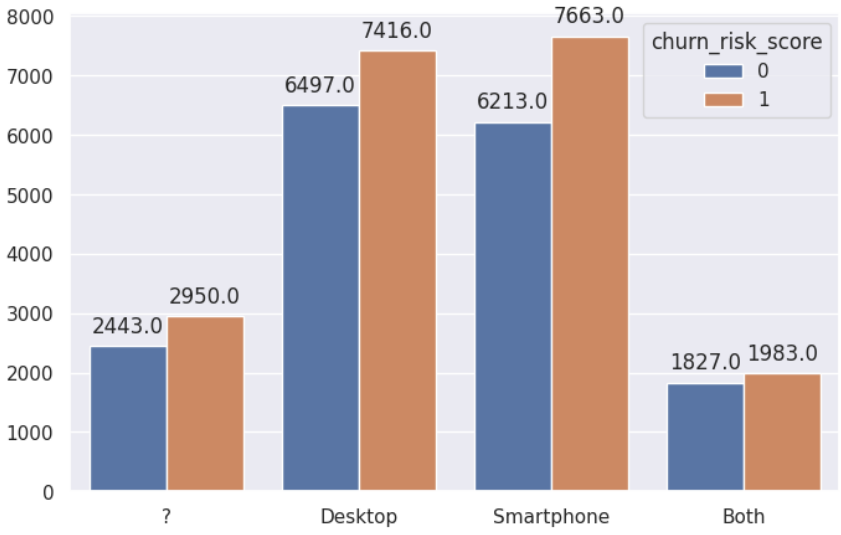
| | points_in_wallet | used_special_discount | offer_application_preference | past_complaint | complaint_status | feedback | churn_risk_score |
|---|---|---|---|---|---|---|---|
| **0** | 781.75 | Yes | Yes | No | Not Applicable | Products always in Stock | 0 |
| **1** | NaN | Yes | No | Yes | Solved | Quality Customer Care | 0 |
| **2** | 500.69 | No | Yes | Yes | Solved in Follow-up | Poor Website | 1 |
| **3** | 567.66 | No | Yes | Yes | Unsolved | Poor Website | 1 |
| **4** | 663.06 | No | Yes | Yes | Solved | Poor Website | 1 |
| **5** | 722.27 | Yes | No | Yes | Unsolved | No reason specified | 0 |
| **6** | 756.21 | Yes | No | Yes | Solved in Follow-up | No reason specified | 0 |
| **7** | 568.08 | No | Yes | Yes | Unsolved | Poor Product Quality | 1 |
| **8** | NaN | Yes | No | Yes | Solved in Follow-up | Poor Customer Service | 0 |
| **9** | 706.23 | No | Yes | Yes | No Information Available | Poor Customer Service | 1 |

ภาพประกอบ 7 แสดงตัวอย่างข้อมูลที่ใช้สำหรับแบบจำลองด้วยแอททริบิวต์ที่ 17 ถึง 23

```
[78]  data.dtypes

age                          int64
gender                      object
security_no                 object
region_category             object
membership_category         object
joining_date                object
joined_through_referral     object
referral_id                 object
preferred_offer_types       object
medium_of_operation         object
internet_option             object
last_visit_time             object
days_since_last_login        int64
avg_time_spent             float64
avg_transaction_value      float64
avg_frequency_login_days    object
points_in_wallet           float64
used_special_discount       object
offer_application_preference object
past_complaint              object
complaint_status            object
feedback                    object
churn_risk_score             int64
dtype: object
```

ภาพประกอบ 8 แสดง Data Type ของ ฟีเจอร์ที่ใช้ในงานวิจัย




ภาพประกอบ 9 แสดงตัวอย่างฟีเจอร์ที่มีความผิดปกติของข้อมูล

จากภาพประกอบ 8 และ 9 ทำการตรวจสอบหาค่าว่างและข้อมูลที่มีความผิดปกติของแต่ละฟีเจอร์ ซึ่งต้องทำการจัดการกับข้อมูลพวกนี้ก่อนนำไปใช้ในแบบจำลองเพื่อจะได้ไม่ส่งผลกระทบต่อแบบจำลองและประสิทธิภาพในการทำนายโดยฟีเจอร์ที่มีค่าว่างคือ region_category, preferred_offer_types, avg_frequency_login_days, points_in_wallet ตามภาพประกอบ 10 และฟีเจอร์ที่มีความผิดปกติของข้อมูลคือ medium_of_operation, joined_through_referral, days_since_last_login, avg_time_spent ดังภาพประกอบ 11

| index | count |
|---|---|
| age | 0 |
| gender | 0 |
| security_no | 0 |
| region_category | 5428 |
| membership_category | 0 |
| joining_date | 0 |
| joined_through_referral | 0 |
| referral_id | 0 |
| preferred_offer_types | 288 |
| medium_of_operation | 0 |
| internet_option | 0 |
| last_visit_time | 0 |
| days_since_last_login | 0 |
| avg_time_spent | 0 |
| avg_transaction_value | 0 |
| avg_frequency_login_days | 3522 |
| points_in_wallet | 3443 |
| used_special_discount | 0 |
| offer_application_preference | 0 |
| past_complaint | 0 |
| complaint_status | 0 |
| feedback | 0 |
| churn_risk_score | 0 |

ภาพประกอบ 10 แสดงจำนวนค่าว่างของแต่ละแอททริบิวต์

| | age | days_since_last_login | avg_time_spent | avg_transaction_value | avg_frequency_login_days | points_in_wallet | churn_risk_score |
|---|---|---|---|---|---|---|---|
| count | 36992.000000 | 36992.000000 | 36992.000000 | 36992.000000 | 33470.000000 | 33549.000000 | 36992.000000 |
| mean | 37.118161 | -41.915576 | 243.472334 | 29271.194003 | 15.976715 | 686.882199 | 0.540982 |
| std | 15.867412 | 228.819900 | 398.289149 | 19444.806226 | 9.215858 | 194.063624 | 0.498324 |
| min | 10.000000 | -999.000000 | -2814.109110 | 800.460000 | -43.652702 | -760.661236 | 0.000000 |
| 25% | 23.000000 | 8.000000 | 60.102500 | 14177.540000 | 9.000000 | 616.150000 | 0.000000 |
| 50% | 37.000000 | 12.000000 | 161.765000 | 27554.485000 | 16.000000 | 697.620000 | 1.000000 |
| 75% | 51.000000 | 16.000000 | 356.515000 | 40855.110000 | 23.000000 | 763.950000 | 1.000000 |
| max | 64.000000 | 26.000000 | 3235.578521 | 99914.050000 | 73.061995 | 2069.069761 | 1.000000 |

ภาพประกอบ 11 แสดงสถิติเชิงบรรยายของแต่ละแอททริบิวต์ที่ยังไม่ได้ทำความสะอาดข้อมูล

สำหรับฟีเจอร์ preferred_offer_types เก็บข้อมูลข้อเสนอของลูกค้ามีจำนวนตัวอย่างข้อมูลทั้งหมด 36,992 ตัวอย่าง พบว่ามี 3 ข้อเสนอคือ Gift Vouchers/Coupons หรือบัตรกำนัล/คูปองมีจำนวน 12,349 ตัวอย่าง และ Credit/Debit Card Offers หรือข้อเสนอบัตรเครดิต/เดบิตมีจำนวน 12,274 ตัวอย่าง และ Without Offers หรือไม่มีข้อเสนอที่ต้องการมีจำนวน 12,081 ตัวอย่าง และพบว่ามีจำนวนค่าว่างที่อยู่ในฟีเจอร์นี้อยู่ 288 ตัวอย่าง ซึ่งหาข้อมูลมาทดแทนไม่ได้จึงตัดสินใจลบค่าว่างในข้อมูลนี้ออก 288 ตัวอย่าง ซึ่งอยู่ในจำนวนที่ไม่มากเกินไป ซึ่งฟีเจอร์นี้เอง

เมื่อลบค่าว่างออกไปแล้วจะส่งผลทำให้ฟีเจอร์อื่น ๆ ที่อยู่ในแถวเดียวกันถูกลบออกไปด้วย ดังนั้น ฟีเจอร์ preferred_offer_types จึงเหลือตัวอย่างข้อมูลทั้งหมด 36,704 ตัวอย่าง

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

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

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

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

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

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

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

| | age | days_since_last_login | avg_time_spent | avg_transaction_value | avg_frequency_login_days | points_in_wallet | churn_risk_score |
|---|---|---|---|---|---|---|---|
| count | 34250.000000 | 34250.000000 | 34250.000000 | 34250.000000 | 34250.000000 | 34250.000000 | 34250.000000 |
| mean | 37.117635 | 66.369431 | 292.450552 | 29253.835764 | 14.932838 | 625.818783 | 0.539591 |
| std | 15.859211 | 223.623608 | 331.545641 | 19475.385199 | 9.338792 | 267.878174 | 0.498437 |
| min | 10.000000 | 1.000000 | 1.837399 | 800.460000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 23.000000 | 9.000000 | 71.470000 | 14163.282500 | 8.000000 | 576.902500 | 0.000000 |
| 50% | 37.000000 | 13.000000 | 174.005000 | 27469.545000 | 15.000000 | 681.430000 | 1.000000 |
| 75% | 51.000000 | 18.000000 | 370.717500 | 40784.997500 | 22.000000 | 757.137500 | 1.000000 |
| max | 64.000000 | 999.000000 | 3235.578521 | 99914.050000 | 73.061995 | 2069.069761 | 1.000000 |

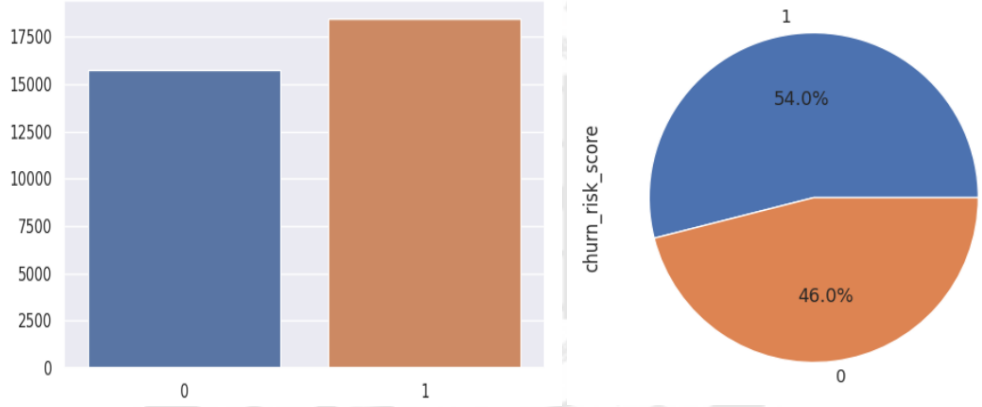
ภาพประกอบ 12 แสดงสถิติเชิงบรรยายของแต่ละแอททริบิวต์หลังจากทำความสะอาดข้อมูล

เมื่อจัดการกับค่าว่างโดยพิจารณาตามฟีเจอร์แล้ว ได้ทำการตรวจสอบว่ามีความซ้ำซ้อนของข้อมูลในแต่ละแถวด้วย ซึ่งพบว่าทั้งหมดของข้อมูลไม่พบแถวซ้ำกัน

เมื่อจัดการกับค่าว่างและตรวจสอบความซ้ำซ้อนของข้อมูลเรียบร้อยแล้ว เหลือข้อมูลทั้งหมด 34,250 ตัวอย่าง และ 19 ฟีเจอร์ ที่สามารถนำไปสำรวจข้อมูลและใช้ในการเรียนรู้ของแบบจำลองต่อไปได้ ดังภาพประกอบ 12

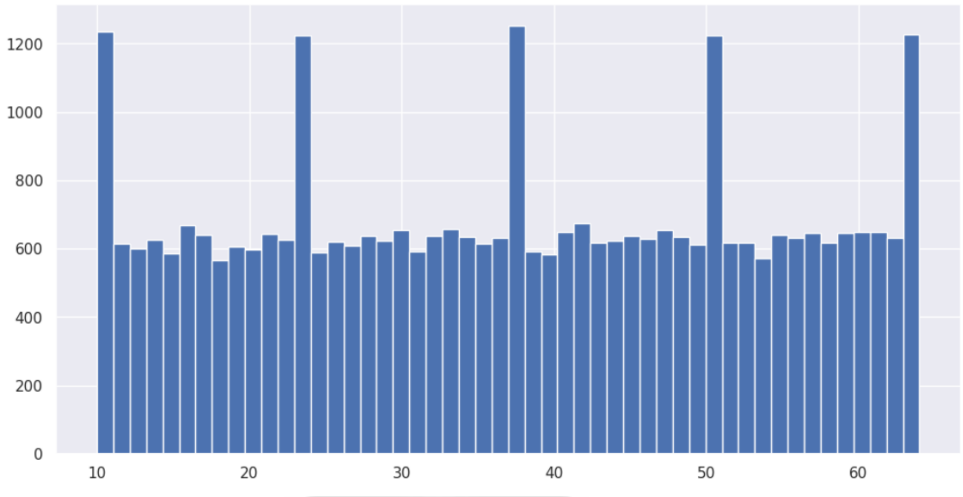
### 3.3 **กระบวนการสำรวจข้อมูล** (Exploratory Data Analysis)

การเข้าใจข้อมูลภายในจึงมีความสำคัญมาก การทำ Exploratory Data Analysis หรือ EDA จึงเป็นขั้นตอนที่ทำให้เราทราบถึงลักษณะของลูกค้าในแต่ละกลุ่ม โดยเริ่มต้นทำการสำรวจลูกค้าที่เป็นกลุ่มเป้าหมาย คือฟีเจอร์ churn_risk_score โดยพบว่ากลุ่ม 0 คือกลุ่มลูกค้ายังใช้บริการอยู่หรือ Exist มีจำนวน 15,769 ตัวอย่าง และกลุ่ม 1 คือกลุ่มลูกค้าเลิกใช้บริการแล้วหรือ Churn มีจำนวน 18,481 ตัวอย่าง จากการสำรวจกลุ่มที่ไม่เป็นลูกค้าแล้วมีจำนวนมากกว่ากลุ่มที่ยังเป็นลูกค้าอยู่ แสดงตามภาพประกอบ 13



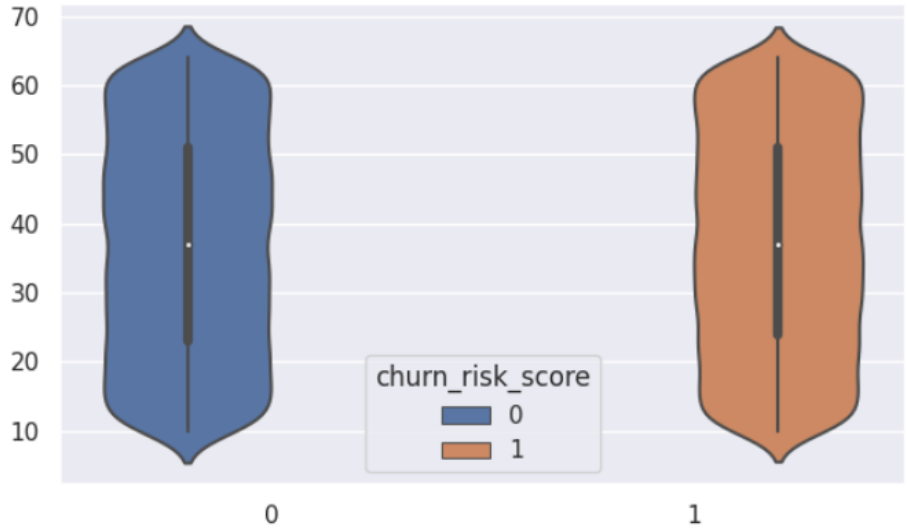

ภาพประกอบ 13 แสดงจำนวนกลุ่มเป้าหมายในรูปแบบกราฟแท่งและกราฟวงกลม

สำหรับฟีเจอร์ age เมื่อตรวจสอบอายุของลูกค้า พบว่ามีอายุตั้งแต่ 10 ถึง 64 ปี ซึ่งการกระจายของอายุไม่แตกต่างกันมาก จึงสรุปได้ว่าลูกค้าส่วนใหญ่อยู่ในช่วงอายุ 23 ถึง 51 ปี มีแนวโน้มที่จะมีค่าผิดปกติบางอย่างแต่ฐานลูกค้าจำนวนมากอยู่ในช่วงนี้ ตามภาพประกอบ 14




ภาพประกอบ 14 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามอายุในรูปแบบกราฟฮิสโตแกรม

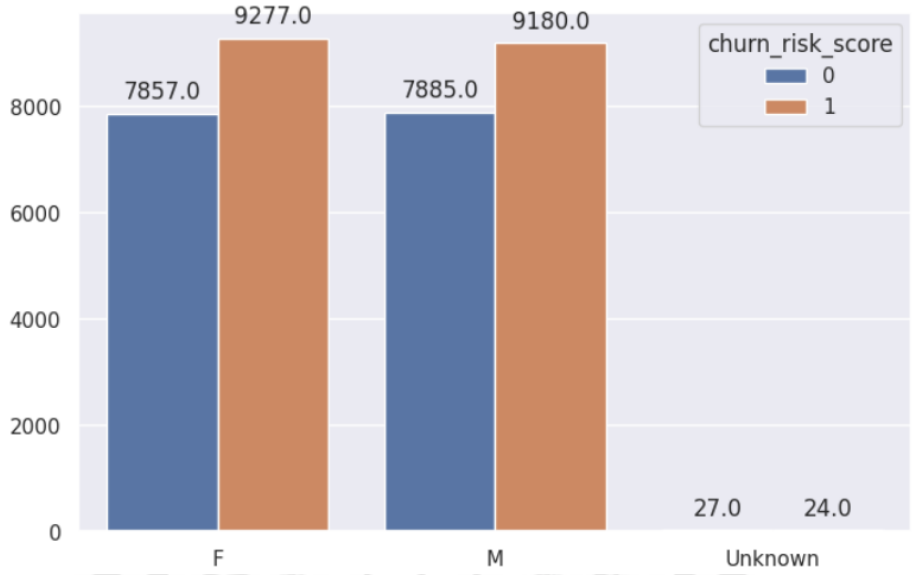
เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม 0 หรือ Exist มีจำนวนลูกค้า 15,769 คน มีลูกค้าที่อายุน้อยสุดคือ 10 ปี และอายุมากสุดคือ 64 ปี ส่วนลูกค้าในกลุ่ม 1 หรือ Churn มีจำนวนลูกค้าทั้งหมด 18,481 คน ซึ่งมีลูกค้าที่อายุน้อยสุดคือ 10 ปี และอายุมากสุดคือ 64 ปี โดยทุกช่วงอายุเป็นลูกค้า Churn มากกว่าลูกค้า Exist ตามภาพประกอบ 15




ภาพประกอบ 15 แสดงความหนาแน่นอายุของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟไวโอลิน

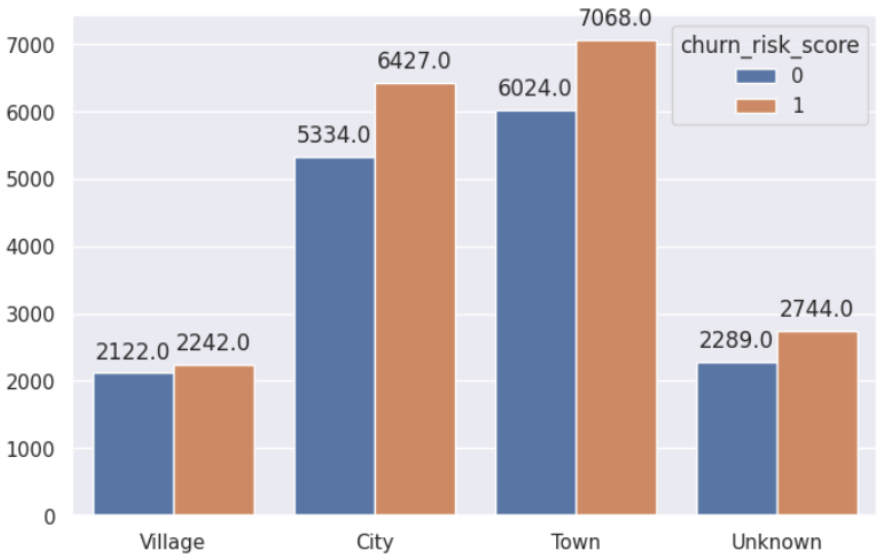
สำหรับฟีเจอร์ gender เมื่อตรวจสอบจำนวนจำนวนเพศลูกค้า พบว่าลูกค้า F หรือเพศหญิงมีจำนวน 17,134 คน ลูกค้า M หรือเพศชายมีจำนวน 17,065 คน และลูกค้า Unknown มีจำนวน 51 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าเพศหญิงเป็นลูกค้า Churn มากกว่า Exist และลูกค้า

เพศชายเป็นลูกค้า Churn มากกว่า Exist อีกเช่นกัน ส่วนลูกค้า Unknown เป็นลูกค้า Exist มากกว่า Churn ซึ่งไม่แตกต่างกันมาก ตามภาพประกอบ 16



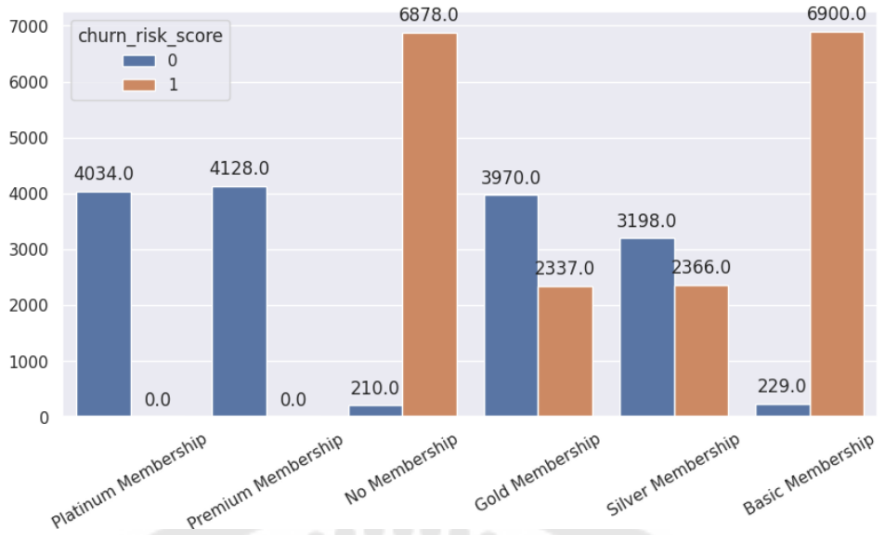

ภาพประกอบ 16 แสดงจำนวนเพศของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ region_category เมื่อตรวจสอบจำนวนพื้นที่อยู่อาศัยของลูกค้า พบว่าลูกค้าที่อยู่ใน Village หรือหมู่บ้านมีจำนวน 4,364 คน ลูกค้าที่อยู่ใน City หรือเมืองขนาดใหญ่มีจำนวน 11,761 คน ลูกค้าที่อยู่ใน Town หรือเมืองขนาดเล็กมีจำนวน 13,092 คน และลูกค้าที่อยู่ใน Unknown มีจำนวน 5,033 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าที่อยู่ใน Village มีลูกค้า Churn มากกว่า Exist อยู่เล็กน้อย ลูกค้าที่อยู่ใน City มีลูกค้า Churn มากกว่า Exist ลูกค้าที่อยู่ใน Town มีลูกค้า Churn มากกว่า Exist ลูกค้าที่อยู่ใน Unknown ลูกค้า Churn มากกว่า Exist ลูกค้าส่วนใหญ่อาศัยอยู่ที่ City และ Town ซึ่งเป็นแหล่งเมืองชุมชนเมือง และอาศัยอยู่ที่ Village น้อยที่สุด ตามภาพประกอบ 17



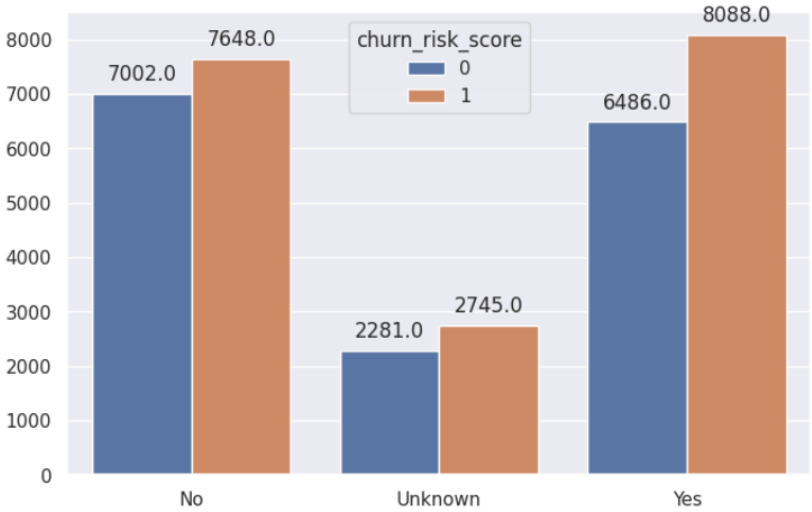

ภาพประกอบ 17 แสดงจำนวนพื้นที่อยู่อาศัยของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ membership_category เมื่อตรวจสอบจำนวนระดับสมาชิกลูกค้า พบว่า Platinum Membership หรือลูกค้าสมาชิกระดับแพลทินัมมีจำนวน 4,128 คน Premium Membership หรือลูกค้าสมาชิกระดับพรีเมียมมีจำนวน 4,034 คน No Membership หรือไม่ได้เป็นสมาชิกมีจำนวน 7,088 คน Gold Membership หรือลูกค้าสมาชิกระดับทองมีจำนวน 6,307 คน Silver Membership หรือลูกค้าสมาชิกระดับเงินมีจำนวน 5,564 คน Basic Membership หรือลูกค้าสมาชิกระดับเริ่มต้นมีจำนวน 7,129 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้า Platinum Membership และ Premium Membership เป็นลูกค้า Exist ทั้งหมด ซึ่งระดับลูกค้าที่สูงเป็นลูกค้าที่ภักดีซึ่งน่าจะเป็นเหตุผลที่เขายังใช้บริการต่อ แต่ลูกค้าส่วนใหญ่ที่ No Membership และ Basic Membership เป็นลูกค้า Churn จำนวนมากมีความแตกต่างกันอย่างชัดเจน ซึ่งน่าจะเป็นเหตุผลที่ไม่เป็นลูกค้าต่อเพราะสมาชิกไม่มีข้อผูกมัดหรือภักดีที่จะใช้ต่อ ส่วนลูกค้า Gold Membership และ Silver Membership เป็นลูกค้า Exist มากกว่า Churn แสดงตามภาพประกอบ 18



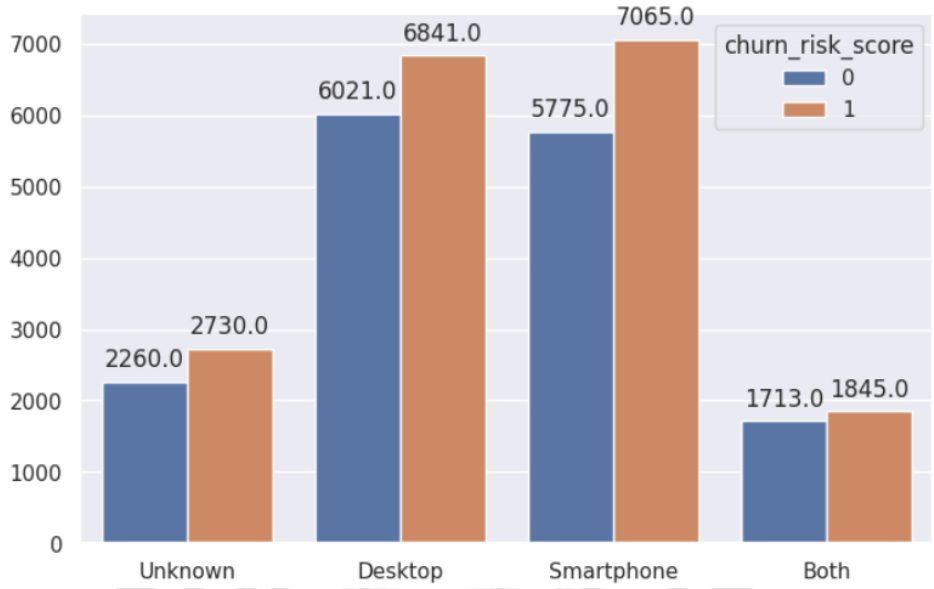

ภาพประกอบ 18 แสดงจำนวนระดับสมาชิกของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ joined_through_referral เมื่อตรวจสอบจำนวนลูกค้าที่เข้าร่วมเป็นสมาชิกโดย Code หรือ ID พบว่า Yes หรือลูกค้าที่เข้าร่วมเป็นสมาชิกโดย Code หรือ ID มีจำนวนทั้งหมด 14,574 คน No หรือลูกค้าที่ไม่ได้เข้าร่วมเป็นสมาชิกโดย Code หรือ ID มีจำนวน 14,650 คน ส่วนลูกค้า Unknown มีจำนวน 5,026 คน เมื่อแบ่งตามกลุ่มพบว่าทั้งลูกค้าที่เข้าร่วมและไม่ได้เข้าร่วมเป็นสมาชิกโดย Code หรือ ID เป็นลูกค้า Churn มากกว่า Exist ส่วนลูกค้า Unknown ก็เป็นเช่นเดียวกัน ตามภาพประกอบ 19



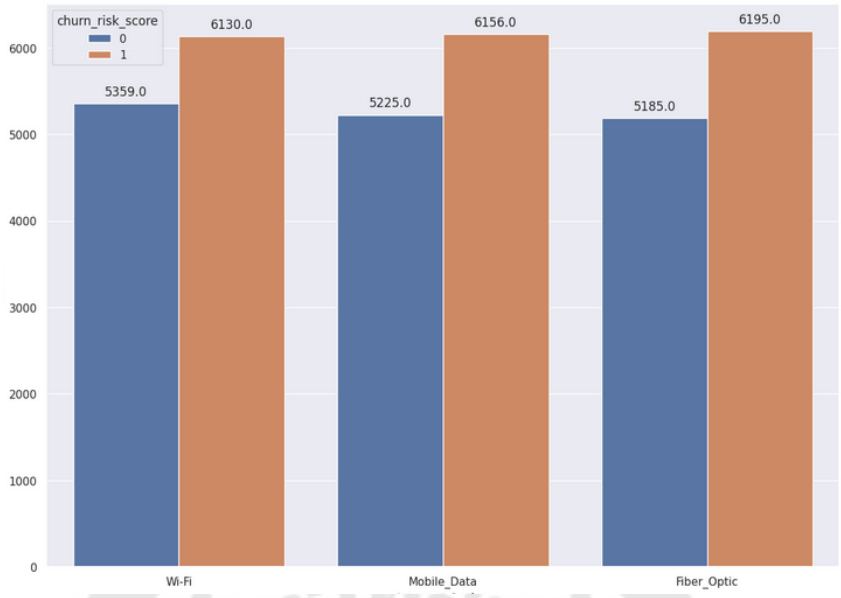

ภาพประกอบ 19 แสดงจำนวนลูกค้าที่เข้าร่วมเป็นสมาชิกโดย Code หรือ ID โดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ medium_of_operation เมื่อตรวจสอบจำนวนลูกค้าที่ใช้สื่อในการดำเนินการที่ใช้ทำธุรกรรม พบว่าลูกค้าใช้ Desktop หรือเดสก์ท็อปมีจำนวน 12,862 คน ลูกค้าใช้ Smartphone หรือสมาร์ตโฟนมีจำนวน 12,840 คน ส่วนลูกค้าที่ใช้ทั้งเดสก์ท็อปและสมาร์ตโฟนหรือ Both มีจำนวน 3,558 คน และลูกค้า Unknown มีจำนวน 4,990 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าที่ใช้ Desktop และลูกค้าที่ใช้ Smartphone ส่วนใหญ่เป็นลูกค้า Churn ส่วนที่ใช้ทั้งเดสก์ท็อปและสมาร์ตโฟนทั้งสองเป็นลูกค้า Churn ที่มากกว่าลูกค้า Exist ซึ่งต่างกันเล็กน้อย และลูกค้า Unknown เป็นลูกค้า Churn มากกว่าลูกค้า Exist เช่นกัน ตามภาพประกอบ 20



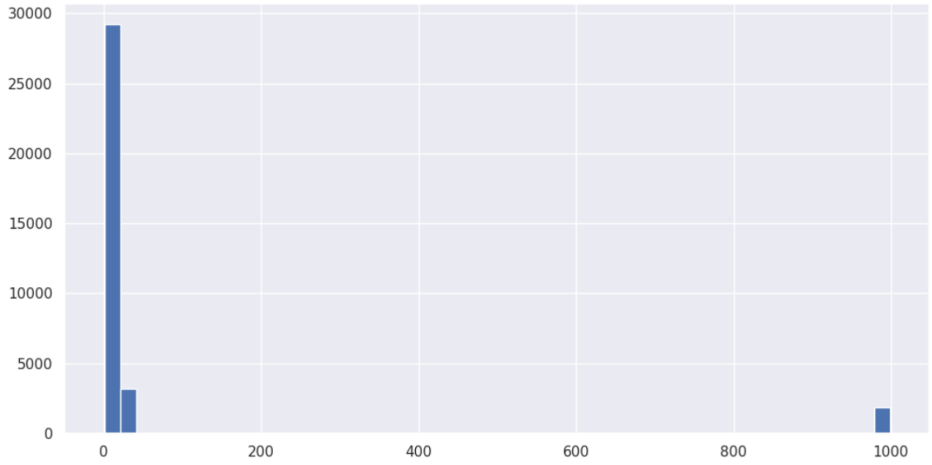

ภาพประกอบ 20 แสดงจำนวนลูกค้าที่ใช้สื่อในการดำเนินการที่ใช้ทำธุรกรรมโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ internet_option เมื่อตรวจสอบจำนวนลูกค้าที่ใช้บริการอินเทอร์เน็ต พบว่าลูกค้าใช้ Wi-Fi หรือการเข้าถึงอินเทอร์เน็ตผ่านเทคโนโลยีไร้สายหรือเราเตอร์เป็นจำนวนทั้งหมด 11,489 คน ลูกค้าใช้ Fiber_Optic หรือการเข้าถึงอินเทอร์เน็ตผ่านการใช้สายไฟเบอร์ออปติกเป็นจำนวน 11,380 คน และลูกค้าใช้ Mobile_Data หรือการเข้าถึงอินเทอร์เน็ตผ่านเครือข่ายเซลลูลาร์เป็นจำนวน 11,381 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าที่ใช้บริการอินเทอร์เน็ตทั้งสามประเภทเป็นลูกค้า Churn มากกว่าลูกค้า Exist ทั้งหมดเลย แสดงในภาพประกอบ 21



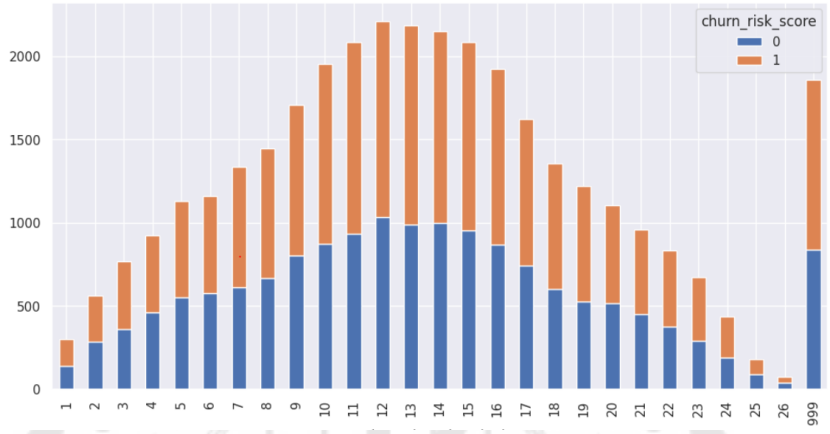

ภาพประกอบ 21 แสดงจำนวนลูกค้าที่ใช้บริการอินเทอร์เน็ตโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ days_since_last_login เมื่อตรวจสอบจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุด พบว่าจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุดคือ 1 วัน ส่วนจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุดเมื่อนานมาแล้วที่มากที่สุดคือ 999 วัน โดยจำนวน 12 วัน คือจำนวนวันที่ลูกค้าเข้าสู่ระบบกันมากที่สุด ดังภาพประกอบ 22



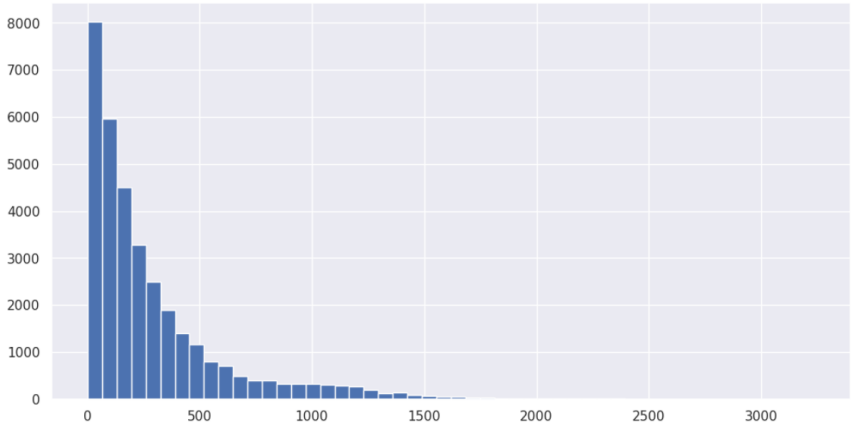

ภาพประกอบ 22 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุดในรูปแบบกราฟฮิสโตแกรม

ในภาพประกอบ 23 เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม 0 หรือ Exist มีจำนวนลูกค้าทั้งหมด 15,769 คน ลูกค้าในกลุ่ม 1 หรือ Churn มีจำนวนลูกค้า 18,481 คน โดยกลุ่มลูกค้าทั้ง 2 กลุ่มมีการกระจายตัวไม่แตกต่างกันคือจำนอนวันล่าสุดที่น้อยที่สุดคือ 1 วัน และจำนวนวันที่เข้าสู่ระบบนานสุดคือ 999 วัน และโดยส่วนมากจำนวนวันที่ลูกค้า Exist และลูกค้า Churn ที่เข้าสู่ระบบครั้งล่าสุดคือเมื่อ 12 วันมาแล้ว



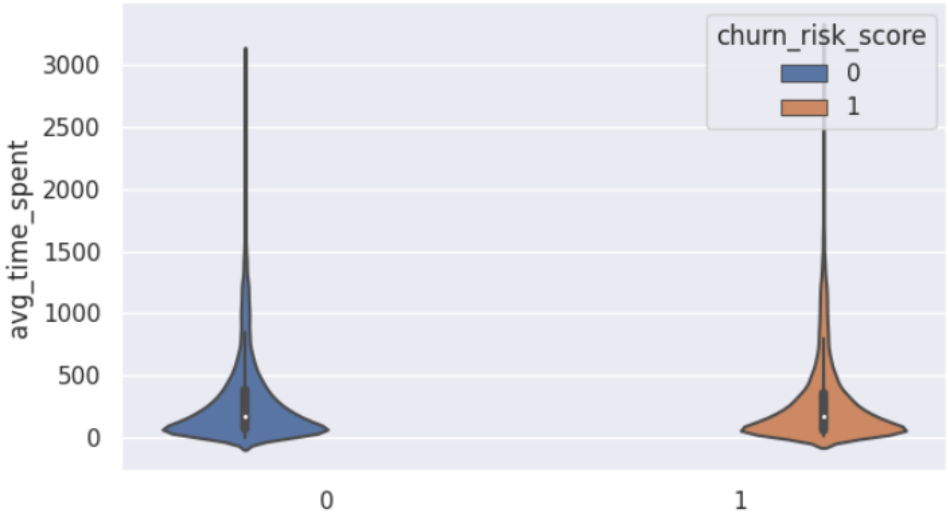

ภาพประกอบ 23 แสดงความหนาแน่นของจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุดโดยแบ่งตามกลุ่มในรูปแบบกราฟ Stacked

สำหรับฟีเจอร์ avg_time_spent เมื่อตรวจสอบเวลาที่ใช้บนเว็บไซต์โดยเฉลี่ยของลูกค้าพบว่าเวลาที่ใช้มากที่สุดคือ 3,235.57 นาที และเวลาที่ใช้น้อยที่สุดคือ 1.83 นาที ซึ่งโดยส่วนใหญ่ใช้เวลาอยู่ที่ 174.05 นาที ตามภาพประกอบ 24



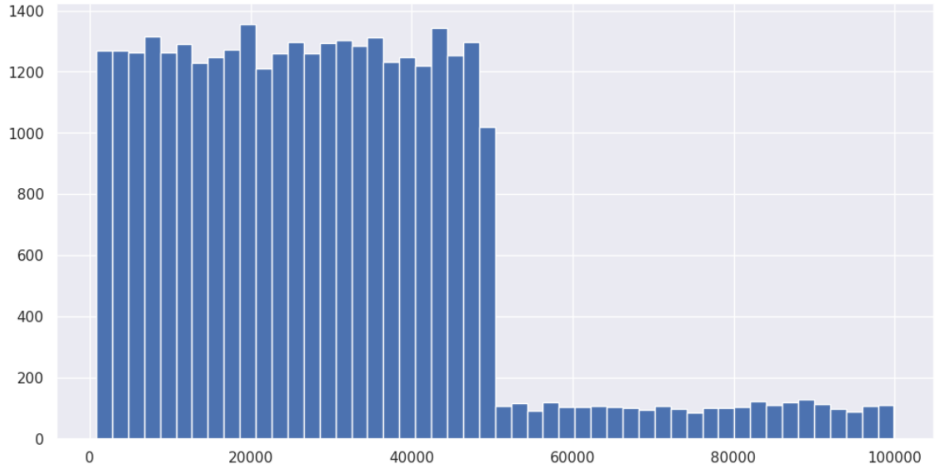

ภาพประกอบ 24 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุดในรูปแบบกราฟฮิสโตแกรม

ในภาพประกอบ 25 เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม 0 หรือ Exist มีจำนวนลูกค้าทั้งหมด 15,769 คน ซึ่งเวลาที่ใช้มากที่สุดคือ 3,040.41 นาที และเวลาที่ใช้น้อยที่สุดคือ 1.83 นาที ลูกค้าในกลุ่ม 1 หรือ Churn มีจำนวนลูกค้า 18,481 คน ซึ่งเวลาที่ใช้มากที่สุดคือ 3,235.57 นาที และเวลาที่ใช้น้อยที่สุดคือ 15.11 นาที โดยทั้งสองกลุ่มมีความหนาแน่นการใช้เวลาบนเว็บไซต์แทบไม่แตกต่างกัน โดยส่วนมากจำนวนเวลาที่ลูกค้า Exist ที่ใช้เวลาบนเว็บไซต์โดยเฉลี่ยคือ 180.50 นาที และส่วนมากจำนวนเวลาที่ลูกค้า Churn ที่ใช้เวลาบนเว็บไซต์โดยเฉลี่ยคือ 169.03 นาที



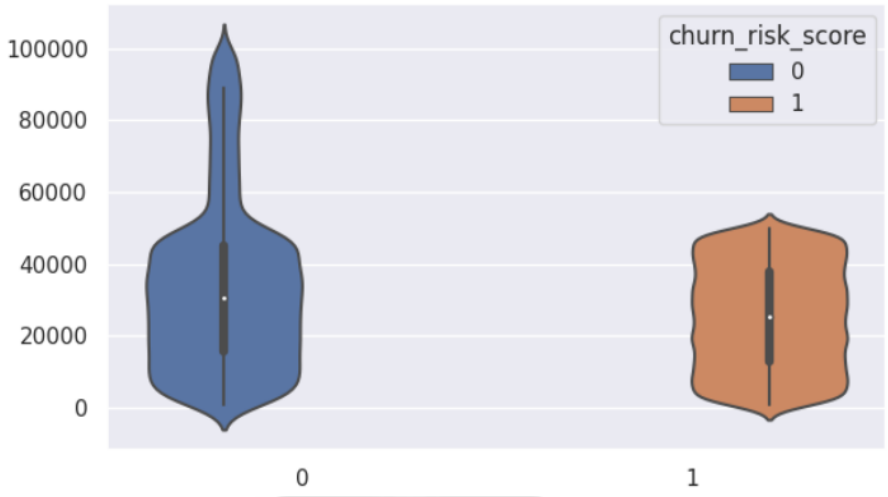

ภาพประกอบ 25 แสดงความหนาแน่นเวลาที่ใช้บนเว็บไซต์โดยเฉลี่ยของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟไวโอลิน

สำหรับฟีเจอร์ avg_transaction_value เมื่อตรวจสอบมูลค่าการซื้อโดยเฉลี่ยของลูกค้าพบว่าลูกค้าที่มีมูลค่าการซื้อโดยเฉลี่ยน้อยสุดคือ 800.46 ดอลลาร์ ส่วนลูกค้าที่มีมูลค่าการซื้อโดยเฉลี่ยมากสุดคือ 99,914.05 ดอลลาร์ และลูกค้าส่วนใหญ่มีมูลค่าการซื้อโดยเฉลี่ยอยู่ที่ 800 ดอลลาร์ ถึง 50,000 ดอลลาร์ ตามภาพประกอบ 26



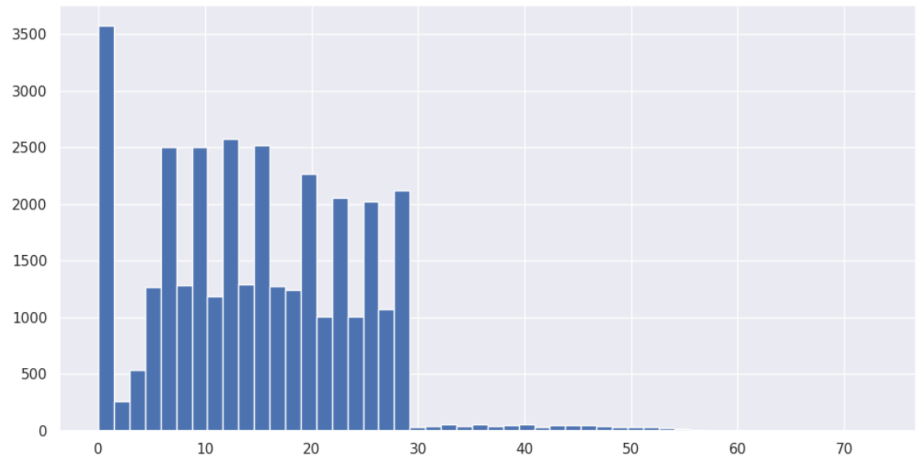

ภาพประกอบ 26 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามมูลค่าการซื้อโดยเฉลี่ยในรูปแบบกราฟฮิสโตแกรม

ในภาพประกอบ 27 เมื่อแบ่งตามกลุ่มลูกค้าพบว่าลูกค้าในกลุ่ม 0 หรือ Exist มีจำนวนลูกค้า 15,769 คน โดยลูกค้ามีมูลค่าการซื้อโดยเฉลี่ยมากที่สุดในคือ 99,914.05 ดอลลาร์ และมีมูลค่าการซื้อโดยเฉลี่ยที่น้อยที่สุดคือ 806.22 ดอลลาร์ ซึ่งจำนวนความหนาแน่นมูลค่าการซื้อโดยเฉลี่ยของลูกค้า Exist มีจำนวนที่มากกว่าลูกค้า Churn อีกทั้งยังมีช่วงที่ยาวกว่าอีกด้วย และลูกค้าในกลุ่ม 1 หรือ Churn มีจำนวนลูกค้า 18,481 คน โดยลูกค้ามีมูลค่าการซื้อโดยเฉลี่ยมากที่สุดในคือ 49,997.69 ดอลลาร์ และมีมูลค่าการซื้อโดยเฉลี่ยที่น้อยที่สุดคือ 800.46 ดอลลาร์ ซึ่งจำนวนความหนาแน่นมูลค่าการซื้อโดยเฉลี่ยของลูกค้า Churn มีจำนวนที่น้อยกว่า Exist และอยู่ในช่วงที่สั้นกว่า สามารถอนุมานได้ว่าลูกค้าที่มีมูลค่าการซื้อโดยเฉลี่ยกต่ำอาจมีแนวโน้มที่จะเป็นลูกค้า Churn




ภาพประกอบ 27 แสดงความหนาแน่นของมูลค่าการซื้อโดยเฉลี่ยของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟไวโอลิน

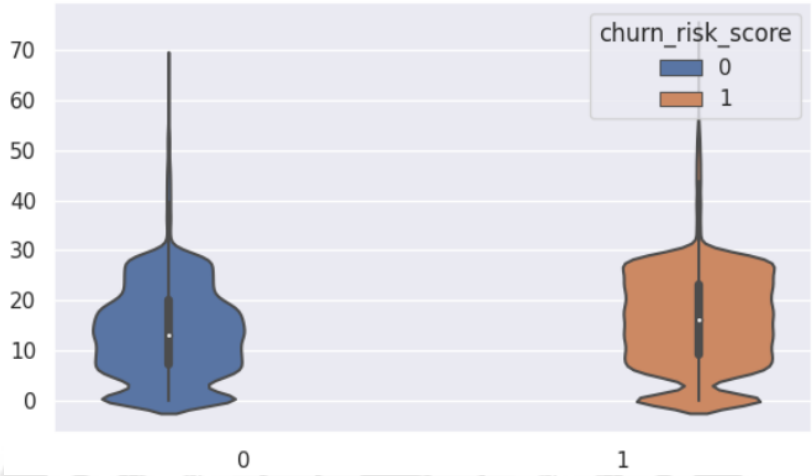
สำหรับฟีเจอร์ avg_frequency_login_days เมื่อตรวจสอบจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ย พบว่าจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยมีตั้งแต่ 0 ถึง 73.06 ครั้ง แสดงตามภาพประกอบ 28




ภาพประกอบ 28 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยในรูปแบบกราฟฮิสโตแกรม

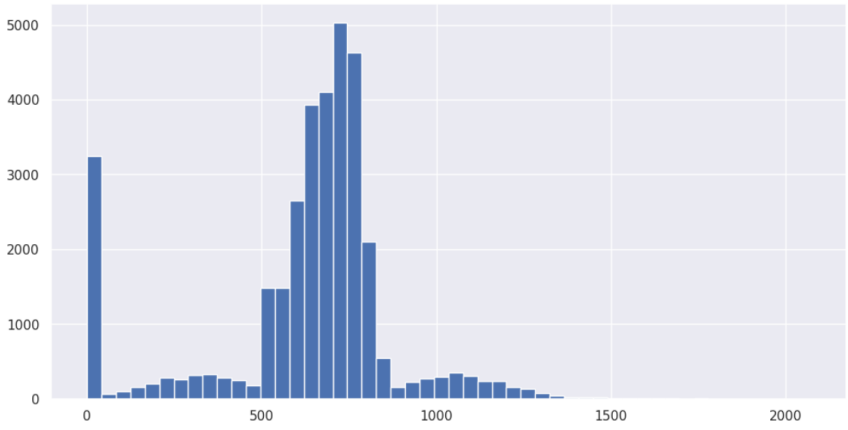
เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม 0 หรือ Exist มีจำนวนลูกค้า 15,769 คน โดยลูกค้าในกลุ่ม Exist ส่วนใหญ่มีจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ย 13 ครั้ง ซึ่งยังพบว่ามีจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยน้อยที่สุดคือ 0 ครั้ง และมีจำนวนครั้งที่ลูกค้าเข้าสู่

ระบบเว็บไซต์โดยเฉลี่ยมากที่สุดคือ 67.06 ครั้ง และลูกค้าในกลุ่ม 1 หรือ Churn มีจำนวนลูกค้าทั้งหมด 18,481 คน โดยลูกค้าในกลุ่ม Churn ส่วนใหญ่มีจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ย 16 ครั้ง ซึ่งยังพบว่ามีจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยน้อยที่สุดคือ 0 ครั้ง และมีจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยมากที่สุดคือ 73.06 ครั้ง สิ่งนี้บ่งชี้ว่าลูกค้า Churn มีความถี่ในการเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยสูงกว่า ดังภาพประกอบ 29



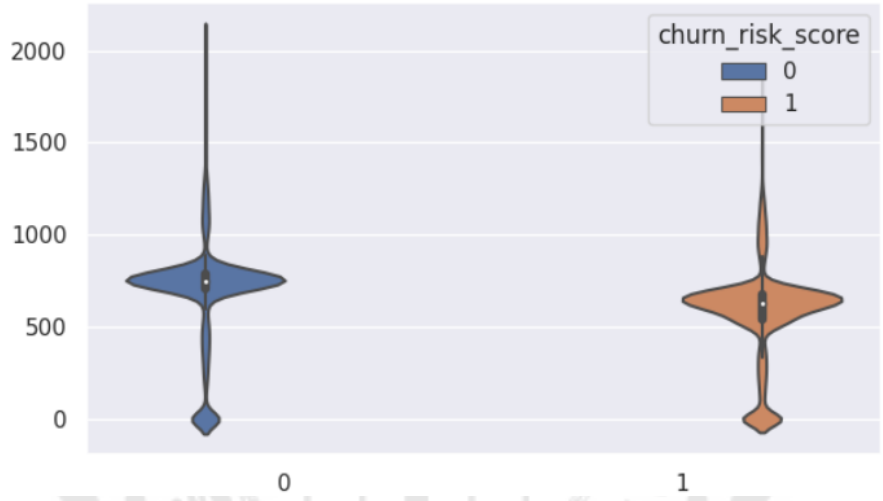

ภาพประกอบ 29 แสดงความหนาแน่นของจำนวนครั้งที่ลูกค้าเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยโดยแบ่งตามกลุ่มในรูปแบบกราฟไวโอลิน

สำหรับฟีเจอร์ points_in_wallet เมื่อตรวจสอบคะแนนสะสมของลูกค้า พบว่ามีคะแนนสะสมที่มากที่สุดคือ 2,069.06 คะแนน และมีคะแนนสะสมที่น้อยที่สุดคือ 0 คะแนน โดยลูกค้าส่วนใหญ่มีคะแนนสะสมประมาณ 681.43 คะแนน ตามภาพประกอบ 30



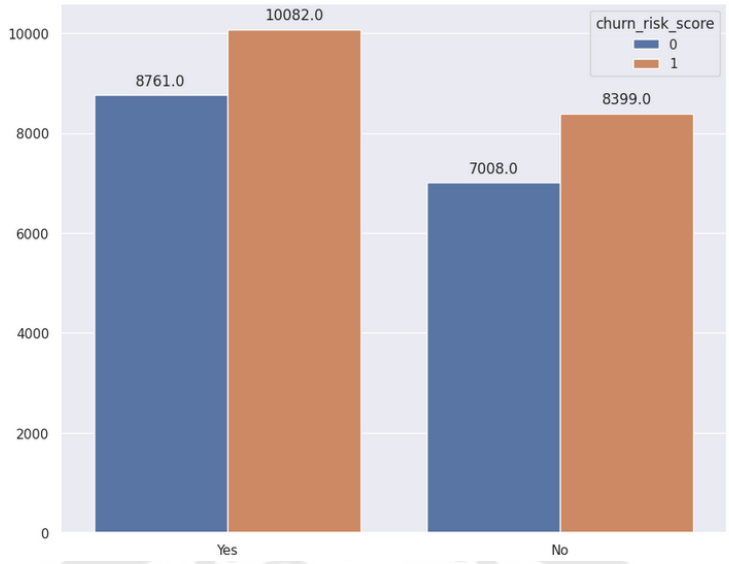

ภาพประกอบ 30 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามคะแนนสะสมที่ลูกค้าในรูปแบบกราฟฮิสโตแกรม

สำหรับภาพประกอบ 31 เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม 0 หรือ Exist มีจำนวนลูกค้า 15,769 คน โดยลูกค้าในกลุ่ม Exist ส่วนใหญ่มีคะแนนสะสมอยู่ที่ 749.33 คะแนน ซึ่งยังพบว่าคะแนนสะสมที่น้อยที่สุดคือ 0 คะแนน และคะแนนสะสมที่มากที่สุดคือ 2,069.06 คะแนน และลูกค้าในกลุ่ม 1 หรือ Churn มีจำนวนลูกค้า 18,481 คน โดยลูกค้าในกลุ่ม Churn ส่วนใหญ่มีคะแนนสะสมอยู่ที่ 629.10 คะแนน ซึ่งยังพบว่าคะแนนสะสมที่น้อยที่สุดคือ 0 คะแนนอีกเช่นกัน และคะแนนสะสมที่มากที่สุดคือ 1,816.93 คะแนน ตามภาพประกอบที่ 26 สามารถอนุมานได้ว่าลูกค้าที่มีคะแนนสะสมต่ำกว่าอาจมีแนวโน้มที่จะเป็นลูกค้า Churn



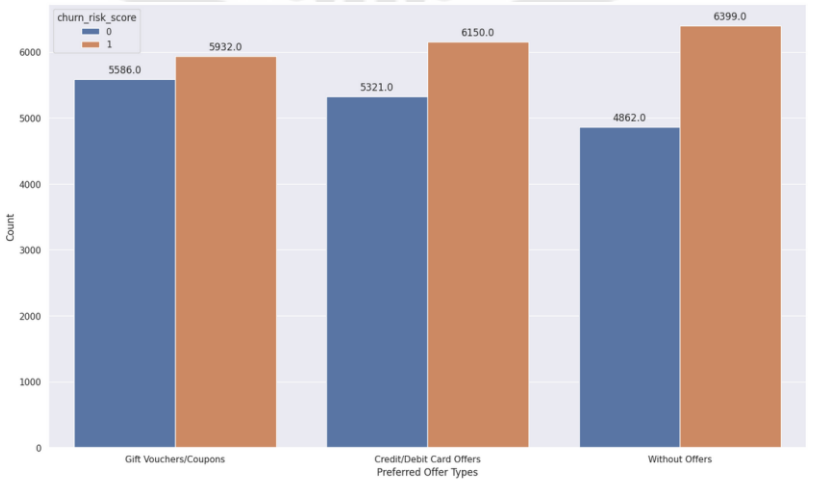

ภาพประกอบ 31 แสดงความหนาแน่นของคะแนนสะสมลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟไวโอลิน

สำหรับฟีเจอร์ used_special_discount เมื่อตรวจสอบจำนวนลูกค้าที่ต้องการใช้ส่วนลดพิเศษ พบว่า Yes หรือลูกค้าที่ต้องการใช้ส่วนลดพิเศษมีจำนวน 18,843 คน และ No หรือลูกค้าที่ไม่ต้องการใช้ส่วนลดพิเศษมีจำนวน 15,407 คน ซึ่งลูกค้าส่วนใหญ่ใช้ส่วนลดพิเศษ เมื่อแบ่งตามกลุ่มพบว่าลูกค้าต้องการใช้ส่วนลดเป็นลูก Churn มากกว่าลูกค้า Exist และลูกค้าที่ไม่ต้องการใช้ส่วนลดพิเศษก็ยังเป็นลูกค้า Churn มากกว่าลูกค้า Exist ตามภาพประกอบ 32



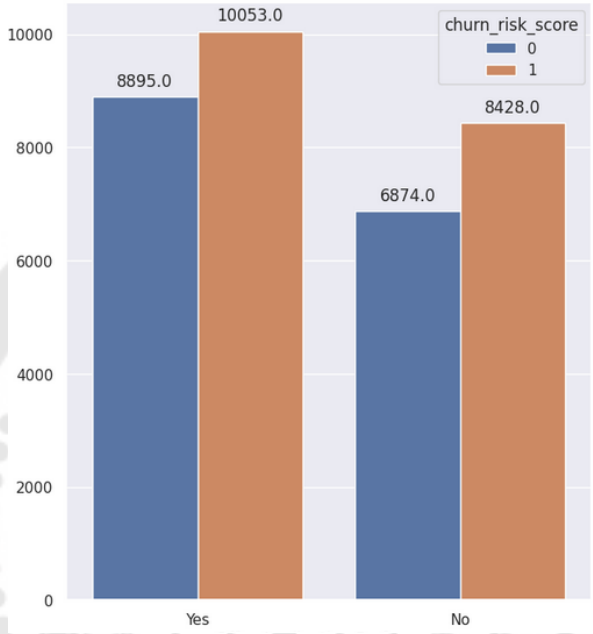

ภาพประกอบ 32 แสดงจำนวนลูกค้าที่ต้องการใช้ส่วนลดพิเศษโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ preferred_offer_types เมื่อทำการตรวจสอบจำนวนลูกค้าที่ต้องการข้อเสนอ พบว่าลูกค้าต้องการ Gift Vouchers/Coupons หรือบัตรกำนัล/คูปองเป็นจำนวน 11,518 คน ลูกค้าต้องการ Credit/Debit Card Offers หรือข้อเสนอบัตรเครดิต/เดบิตเป็นจำนวน 11,471 คน และลูกค้าที่ไม่มีข้อเสนอที่ต้องการ หรือ Without Offers เป็นจำนวน 11,261 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าทั้งสามข้อเสนอนี้เป็นลูกค้า Churn มากกว่าลูกค้า Exist แสดงตามภาพประกอบ 33



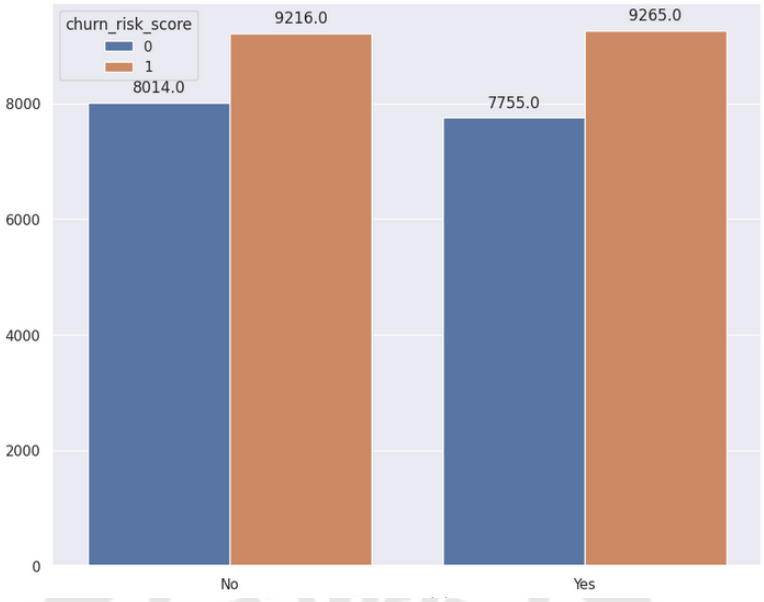

ภาพประกอบ 33 แสดงจำนวนลูกค้าที่ต้องการข้อเสนอโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ offer_application_preference ที่แสดงตามภาพประกอบ 34 เมื่อตรวจสอบจำนวนลูกค้าที่ต้องการใบเสร็จ พบว่า Yes หรือลูกค้าที่ต้องการใบเสร็จ ซึ่งมีจำนวนทั้งหมด 18,948 คน และ No หรือลูกค้าที่ไม่ต้องการใบเสร็จมีจำนวน 15,302 คน ซึ่งลูกค้าส่วนใหญ่ต้องการใบเสร็จ เมื่อแบ่งตามกลุ่มพบว่าลูกค้าต้องการใบเสร็จ Churn มากกว่าลูกค้า Exist และลูกค้าที่ไม่ต้องการใบเสร็จก็ยังเป็นลูกค้า Churn มากกว่าลูกค้า Exist



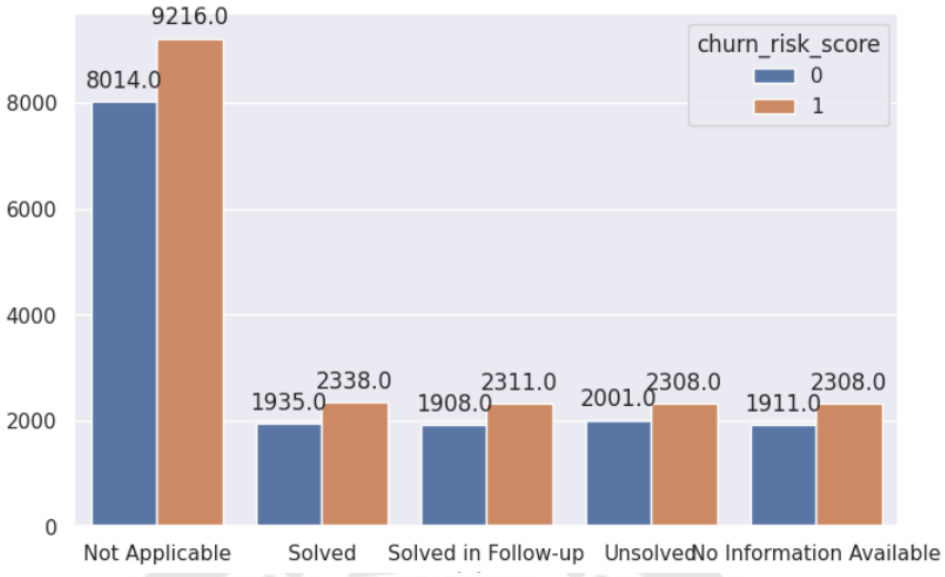

ภาพประกอบ 34 แสดงจำนวนลูกค้าที่ต้องการใบเสร็จโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ past_complaint ที่แสดงตามภาพประกอบ 35 เมื่อตรวจสอบจำนวนลูกค้าจากความต้องการการร้องเรียน พบว่า Yes หรือลูกค้ามีความต้องการร้องเรียนมีจำนวนทั้งหมด 17,230 คน และ No หรือลูกค้าไม่ต้องการร้องเรียนมีจำนวน 17,020 คน เมื่อแบ่งตามกลุ่มพบว่า No มีลูกค้า Churn มากกว่าลูกค้า Exist อีกทั้ง Yes ก็มีลูกค้า Churn มากกว่าลูกค้า Exist




ภาพประกอบ 35 แสดงจำนวนความต้องการร้องเรียนของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

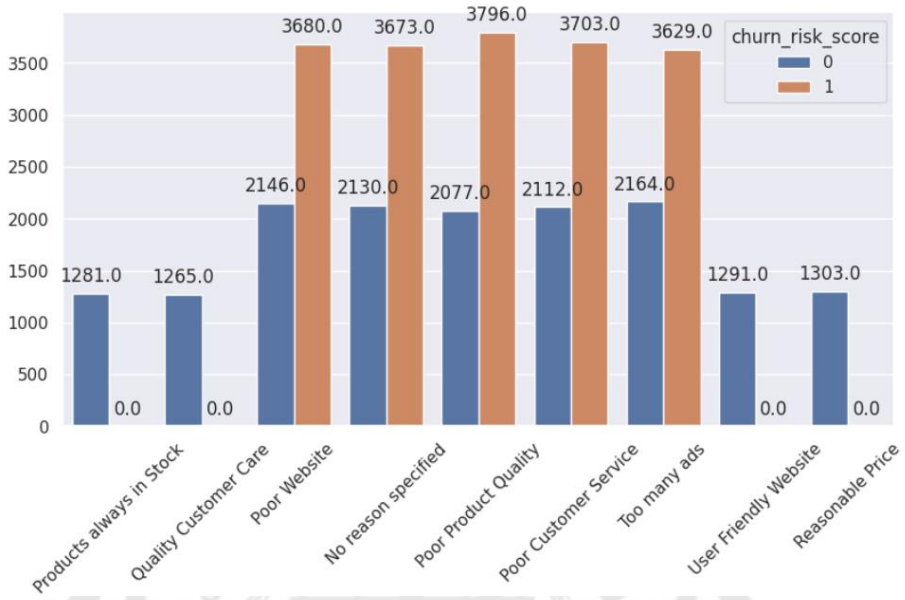
สำหรับฟีเจอร์ complaint_status เมื่อตรวจสอบจำนวนลูกค้าจากสถานะการร้องเรียน พบว่าลูกค้าที่ Not Applicable หรือไม่มีข้อร้องเรียนมีจำนวน 17,230 คน ลูกค้าที่ Solved หรือการร้องเรียนได้รับการแก้ไขแล้วมีจำนวน 4,273 คน ลูกค้าที่ Solved in Follow-up หรือการร้องเรียนไม่ได้รับการแก้ไขในตอนแรกแต่ได้รับการแก้ไขหลังมีจำนวน 4,219 คน ลูกค้าที่ Unsolved หรือการร้องเรียนยังไม่ได้รับการแก้ไขมีจำนวน 4,309 คน ลูกค้าที่ No Information Available หรือไม่มีข้อมูลเกี่ยวกับสถานการณ์ร้องเรียนมีจำนวน 4,219 คน เมื่อแบ่งตามกลุ่ม พบว่าทุกสถานะการร้องเรียนมีลูกค้า Churn มากกว่าลูกค้า Exist อยู่ทุกสถานะ ตามภาพประกอบ 36




ภาพประกอบ 36 แสดงจำนวนสถานะการร้องเรียนของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ feedback เมื่อตรวจสอบจำนวนลูกค้าที่แสดงความคิดเห็น พบว่าลูกค้าที่มีความคิดเห็นเกี่ยวกับ Products always in Stock หรือสินค้ามีในสต็อกเสมอมีจำนวน 1,281 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ Quality Customer Care หรือการบริการลูกค้าที่มีคุณภาพมีจำนวน 1,265 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ User Friendly Website หรือเว็บไซต์ใช้งานง่ายมีจำนวน 1,291 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ Reasonable Price หรือราคาสมเหตุสมผลมีจำนวน 1,303 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ Poor Website หรือเว็บไซต์ที่ไม่ดีมีจำนวนทั้งหมด 5,826 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ No reason specified หรือไม่ระบุสาเหตุมีจำนวน 5,803 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ Poor Product Quality หรือคุณภาพสินค้าไม่ดีมีจำนวน 5,873 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ Poor Customer Service หรือบริการลูกค้าไม่ดีมีจำนวน 5,815 คน ลูกค้าที่มีความคิดเห็นเกี่ยวกับ Too many ads หรือโฆษณามากเกินไปมีจำนวน 5,793 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้า Churn แสดงความคิดเห็นในทางลบที่ชัดเจน คือ เว็บไซต์ที่ไม่ดี คุณภาพสินค้าไม่ดี บริการลูกค้าไม่ดี และโฆษณามากเกินไป รวมทั้งไม่ระบุสาเหตุอีกด้วย ซึ่งการแสดงความคิดเห็นตรงนี้เองเป็นข้อชัดเจนแล้วว่า ลูกค้าถึงไม่กลับมาซื้อสินค้าจากเว็บไซต์อีก แต่กลับกันสินค้ามีในสต็อกเสมอ การบริการลูกค้าที่มีคุณภาพ เว็บไซต์ใช้งานง่ายและราคาสมเหตุสมผล การแสดงความคิดเห็นนี้เป็นการแสดงความคิดเห็นต่อเว็บไซต์ที่ดี อีกทั้งไม่

มีลูกค้า Churn คนไหนเลือกแสดงความคิดเห็นนี้ โดยการแสดงความคิดเห็นนี้ทั้งหมดเป็นลูกค้า Exist ดังภาพประกอบ 37




ภาพประกอบ 37 แสดงจำนวนแสดงความคิดเห็นของลูกค้าโดยแบ่งตามกลุ่มในรูปแบบกราฟแท่ง

### 3.4 การเตรียมความพร้อมข้อมูล (Data Preprocessing)

ก่อนการเตรียมข้อมูล ผู้วิจัยแบ่งข้อมูลโดยใช้คำสั่ง train_test_split ที่สัดส่วน 80% สำหรับข้อมูลในการเรียนรู้ (Training Set) ได้ข้อมูลทั้งหมด 27,400 ตัวอย่าง และสัดส่วน 20% สำหรับข้อมูลในการทดสอบ (Test Set) ได้ข้อมูลทั้งหมด 6,850 ตัวอย่าง โดยการเตรียมข้อมูลนี้เพื่อใช้กับชุดข้อมูลสำหรับการเรียนรู้เท่านั้น

#### 3.4.1 การเปลี่ยนรูปแบบข้อมูลแบบกลุ่มและตัวเลข

กระบวนการที่เลือกใช้ชื่อว่า make_column_transformer เป็นกระบวนการที่สามาถทำการเปลี่ยนแปลงข้อมูลกลุ่มแบบไม่มีลำดับเป็นข้อมูลตัวเลขจะใช้วิธี One-Hot Encoding และสามารถปรับเปลี่ยนช่วงของข้อมูลตัวเลขได้พร้อมกันจะใช้วิธี Standard Scaler เนื่องจากเหตุที่ต้องทำพร้อมกันเพราะมีความกังวลว่าจะเกิดปัญหาข้อมูลในชุดทดสอบรั่วไหล (Data Leakage) ดังนั้นจึงมีการวิธีใช้ make_column_transformer ร่วมกับวิธี make_pipeline เพื่อให้สามารถระบุฟีเจอร์ที่ต้องการทำกระบวนการเปลี่ยนแปลงข้อมูลกลุ่มแบบไม่มีลำดับเป็นตัวเลขและปรับช่วงข้อมูลได้อย่างปลอดภัย

สำหรับการเปลี่ยนแปลงข้อมูลกลุ่มแบบไม่มีลำดับเป็นข้อมูลตัวเลข (One-Hot Encoding) คือการเพิ่มฟีเจอร์ขึ้นมาจากค่าข้อมูล เช่น ฟีเจอร์ region_category มีข้อมูลอยู่ 4 ค่า

คือ City, Town, Village, และ Unknown เมื่อผ่านการเปลี่ยนแปลงข้อมูลแล้วจะได้ฟีเจอร์ใหม่ขึ้นมา 4 ฟีเจอร์คือ region_category_City, region_category_Town, region_category_Village, และ region_category_Unknown จะใช้ค่า 0 และ 1 ในการระบุว่ามีหรือไม่มีค่านั้น ๆ ตามภาพประกอบ 38

| | gender_F | gender_M | gender_Unknown | region_category_City | region_category_Town | region_category_Unknown | region_category_Village |
|---|---|---|---|---|---|---|---|
| **0** | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| **1** | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| **2** | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| **3** | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| **4** | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| **5** | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| **6** | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| **7** | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| **8** | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| **9** | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |

ภาพประกอบ 38 แสดงตัวอย่างการเปลี่ยนแปลงข้อมูลแบบกลุ่มไม่มีลำดับเป็นตัวเลขโดย One-Hot Encoding

สำหรับฟีเจอร์ที่เก็บข้อมูลเป็นตัวเลข ที่แสดงในภาพประกอบ 39 ซึ่งพบว่าค่าของข้อมูลอยู่ในช่วงที่ต่างกัน เช่น ฟีเจอร์ days_since_last_login หรือจำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุด อยู่ในช่วง 1 ถึง 999 วัน การที่ข้อมูลยังไม่ถูกทำให้อยู่ในช่วงเดียวกันนั้น สามารถส่งผลกระทบถึงประสิทธิภาพของอัลกอริทึมได้ ซึ่งวิธีที่ใช้คือ Standard Scaler ในงานวิจัยนี้

กระบวนการ Standard Scaler เป็นการปรับค่าข้อมูลตัวเลขโดยคำนวณจากค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐาน ดังสมการ (6)

$$\mathcal{X}\_scaled_{\mathrm{i}} = \frac{\mathcal{X}_{\mathrm{i}} - \mu}{\sigma} \tag{6}$$

โดยที่

$\mathcal{X}\_scaled_{\mathrm{i}}$ คือข้อมูล $\mathcal{X}$ ในชุดข้อมูลสำหรับเรียนรู้ตัวที่ $i$ ที่ผ่านการปรับค่าข้อมูล

$\mathcal{X}i$ คือข้อมูล $\mathcal{X}$ ชุดข้อมูลสำหรับเรียนรู้ตัวที่ $i$ ที่ยังไม่ผ่านการปรับข้อมูล

$\mu$ คือค่าเฉลี่ยของชุดข้อมูลสำหรับการเรียนรู้

$\sigma$ คือส่วนเบี่ยงเบนมาตรฐานของชุดข้อมูลสำหรับการเรียนรู้

| age | days_since_last_login | avg_time_spent | avg_transaction_value | avg_frequency_login_days | points_in_wallet |
|---|---|---|---|---|---|
| -0.445852 | -0.226125 | -0.701277 | 0.473173 | -0.530243 | 0.306509 |
| 1.129898 | -0.217210 | -0.635571 | -1.049709 | -0.209277 | 0.128616 |
| 1.445048 | -0.235040 | -0.591567 | -1.129004 | -0.102288 | 0.515283 |
| -1.013122 | -0.221667 | -0.327602 | 0.601155 | 0.218679 | 0.109552 |
| -0.256762 | -0.243955 | -0.726826 | 0.922259 | 0.753623 | 0.040618 |
| -1.454332 | -0.217210 | 0.523656 | -0.513159 | 0.111690 | 0.471618 |
| 0.499598 | -0.248412 | -0.773475 | 0.447931 | -1.600132 | 0.508036 |
| 0.184448 | -0.212753 | -0.507166 | 1.616168 | -1.386154 | 2.693017 |
| 1.571108 | -0.208295 | -0.530400 | -0.847744 | 0.432657 | 0.048867 |
| -1.202212 | -0.203838 | -0.520541 | 3.171923 | 0.432657 | 0.693993 |

ภาพประกอบ 39 แสดงข้อมูลตัวเลขหลังจากปรับค่าข้อมูลโดย Standard Scaler

### 3.4.2 การคัดเลือกคุณลักษณะ

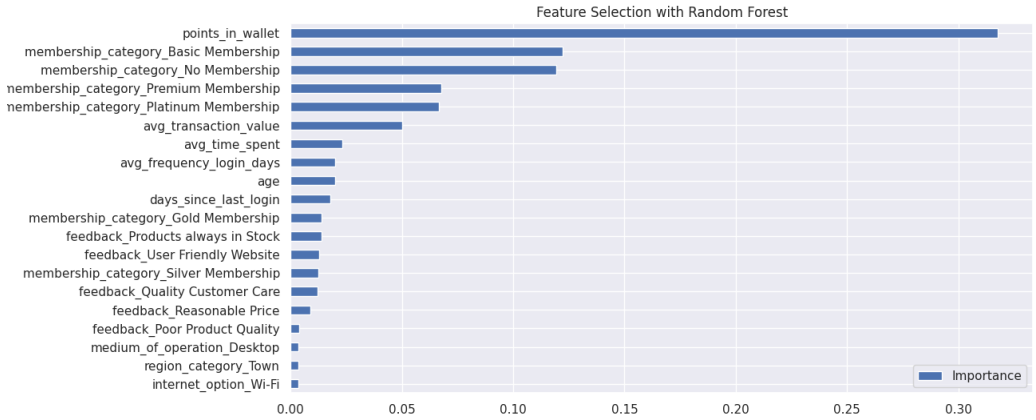
สำหรับการคัดเลือกคุณลักษณะหรือ Feature Selection โดยใช้เทคนิค Random Forest เลือกเฉพาะคุณลักษณะหรือ Feature ที่มีความสำคัญสุดจากชุดข้อมูลเดิม นำ Feature ที่มีความสำคัญต่ำออก เพื่อลดขนาดของข้อมูลเพื่อที่จะได้เพิ่มความเร็วในการประมวลผลของแบบจำลอง เนื่องจากมีจำนวนคุณลักษณะเป็นจำนวนมาก การวิจัยนี้จึงเลือกใช้ Feature เพียง 16 ตัว ที่ได้จากแบบจำลอง Random Forest โดยเรียงลำดับตามคะแนนความสำคัญ ดังตาราง 4 และภาพประกอบ 40

ตาราง 4 แสดงคุณลักษณะโดยเรียงลำดับคะแนนความสำคัญ ที่คำนวณจากแบบจำลอง Random Forest

| ลำดับ | คุณลักษณะ | ความหมายของคุณลักษณะ | คะแนนความสำคัญ |
|---|---|---|---|
| 1 | points_in_wallet | คะแนนสะสมที่ลูกค้าได้รับในแต่ละการทำธุรกรรม | 0.3176 |
| 2 | membership_category_ Basic Membership | ลูกค้าสมาชิกระดับเริ่มต้น | 0.1221 |
| 3 | membership_category_No Membership | ลูกค้าที่ไม่ได้เป็นสมาชิก | 0.1195 |

ตาราง 4 (ต่อ)

| ลำดับ | คุณลักษณะ | ความหมายของคุณลักษณะ | คะแนนความสำคัญ |
|---|---|---|---|
| 4 | membership_category_Premium Membership | ลูกค้าสมาชิกระดับพรีเมียม | 0.0676 |
| 5 | membership_category_Platinum Membership | ลูกค้าสมาชิกระดับแพลทินัม | 0.0666 |
| 6 | avg_transaction_value | มูลค่าการซื้อสะสมโดยเฉลี่ยของลูกค้า | 0.050 |
| 7 | avg_time_spent | เวลาที่ใช้เวลาบนเว็บไซต์โดยเฉลี่ยของลูกค้า (นาที) | 0.0232 |
| 8 | avg_frequency_login_days | จำนวนนับครั้งการเข้าสู่ระบบเว็บไซต์โดยเฉลี่ยของลูกค้า | 0.0199 |
| 9 | age | อายุของลูกค้า | 0.0198 |
| 10 | days_since_last_login | จำนวนวันที่ลูกค้าเข้าสู่ระบบครั้งล่าสุด | 0.0177 |
| 11 | membership_category_Gold Membership | ลูกค้าสมาชิกระดับทอง | 0.0141 |
| 12 | feedback_Products always in Stock | การแสดงความคิดเห็นของลูกค้าว่าสินค้ามีในสต็อกเสมอ | 0.0138 |
| 13 | feedback_User Friendly Website | การแสดงความคิดเห็นของลูกค้าว่าเว็บไซต์ใช้งานง่าย | 0.0129 |
| 14 | embership_category_Silver Membership | ลูกค้าสมาชิกระดับเงิน | 0.0126 |
| 15 | feedback_Quality Customer Care | การแสดงความคิดเห็นของลูกค้าว่าการบริการลูกค้าที่มีคุณภาพ | 0.0121 |
| 16 | feedback_Reasonable Price | การแสดงความคิดเห็นของลูกค้าว่าราคาสมเหตุสมผล | 0.0090 |

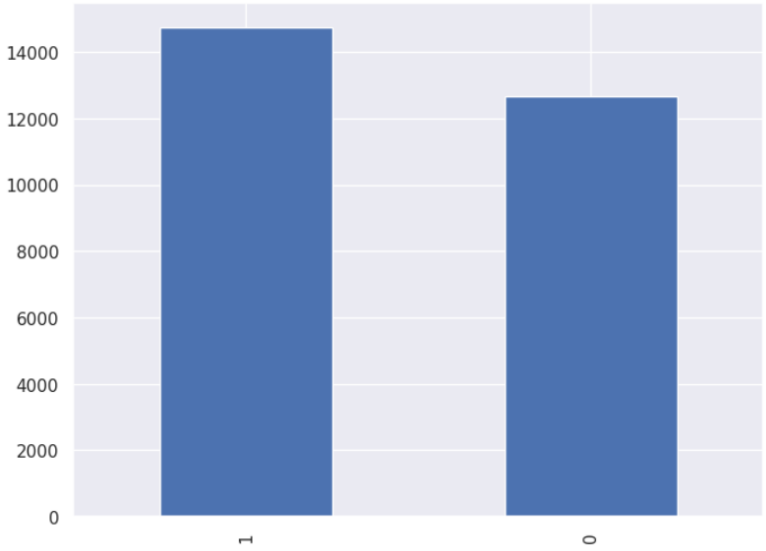



ภาพประกอบ 40 แสดงคะแนนของคุณลักษณะที่สำคัญ 20 อันดับแรกจากแบบจำลอง Random Forest

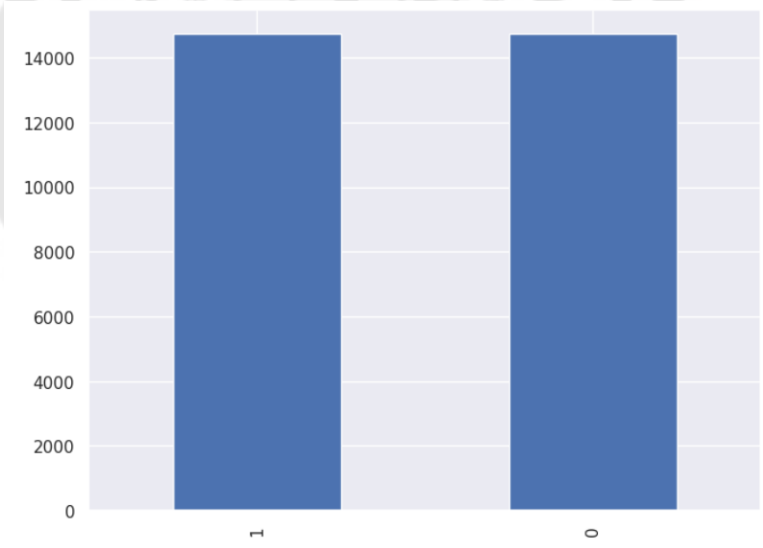
### 3.4.3 การแก้ไขปัญหาข้อมูลไม่สมดุล

จากการสำรวจข้อมูลข้างต้น พบว่ากลุ่มของลูกค้า หรือ churn_risk_score ที่เป็นเลเบลในงานวิจัยนี้ เกิดปัญหาความไม่สมดุลของข้อมูล หรือ Imbalanced Data คือการที่จำนวนกลุ่มของลูกค้ามีจำนวนไม่เท่ากัน ซึ่งอาจจะทำให้กระทบถึงประสิทธิภาพในการเรียนรู้ของแบบจำลองได้ ถ้ามีกลุ่มของลูกค้าใดมากกว่าอาจจะทำให้เกิดความลำเอียงของข้อมูลได้ เพื่อการแก้ไขปัญหานี้จึงเลือกใช้วิธีการทำให้กลุ่มของลูกค้ามีจำนวนเท่ากันก่อนนำไปใช้กับแบบจำลองการเรียนรู้ โดยเลือกใช้วิธี Synthetic Minority Oversampling Technique (SMOTE) โดยอาศัยหลักการ K-Nearest Neighbors ซึ่งได้กำหนดค่า k_neighbors=5 เพื่อสร้างขอบเขตสำหรับการสุ่มเพิ่มข้อมูลใหม่

ในสำหรับข้อมูลในการเรียนรู้ (Training Set) พบว่ากลุ่มลูกค้ามีจำนวนตัวอย่างทั้งหมด 27,400 ตัวอย่าง โดยกลุ่มที่ 1 คือ ลูกค้า Churn มีจำนวน 14,735 ตัวอย่าง และกลุ่มที่ 0 คือลูกค้า Exist มีจำนวน 12,665 ตัวอย่าง ซึ่งกลุ่มลูกค้ามีจำนวนไม่เท่ากัน แสดงดังภาพประกอบ 41 ซึ่งกลุ่มลูกค้า Exist มีจำนวนน้อยกว่า หลังจากการทำ SMOTE พบว่ากลุ่มลูกค้ามีจำนวนตัวอย่างทั้งหมด 29,470 ตัวอย่าง ซึ่งทำให้กลุ่มลูกค้าที่ 1 และกลุ่มลูกค้าที่ 0 มีจำนวนตัวอย่างของข้อมูลเท่ากันที่ 14,735 ตัวอย่าง เมื่อได้ข้อมูลที่สมดุลกันแล้ว จะช่วยลดปัญหาแบบจำลองที่ทำนายเป็นกลุ่มที่มีจำนวนข้อมูลมากกว่าได้ ดังภาพประกอบ 42




ภาพประกอบ 41 แสดงจำนวนข้อมูลกลุ่มลูกค้าของข้อมูลในการเรียนรู้ที่ยังไม่ผ่านการทำ SMOTE




ภาพประกอบ 42 แสดงจำนวนข้อมูลกลุ่มลูกค้าของข้อมูลในการเรียนรู้ที่ผ่านการทำ SMOTE

### 3.5 การสร้างแบบจำลองเพื่อทำนายกลุ่ม

การหาพารามิเตอร์ที่ดีที่สุดสำหรับแบบจำลองในการทำนาย โดยใช้ข้อมูลในการเรียนรู้ (Training Set) ทั้งหมดมีจำนวน 27,400 ตัวอย่าง และข้อมูลที่ใช้ในการทดสอบ (Test Set) ทั้งหมดมีจำนวน 6,850 ตัวอย่าง อีกทั้งชุดข้อมูลที่ได้ถูกจัดการกับการแก้ไขปัญหาความไม่สมดุลด้วย SMOTE มีข้อมูลทั้งหมด 29,470 ตัวอย่าง ซึ่งการเลือกหาชุดข้อมูลที่ดีที่สุดทำโดยใช้เทคนิค

Grid Search ซึ่งแบบจำลองจะถูกตรวจสอบด้วย Cross Validation ที่ทั้งหมด 5 Fold ในขั้นตอนการเรียนรู้ของแบบจำลอง เพื่อทำการปรับจูนพารามิเตอร์ โดยทำการทดลองกับแบบจำลองที่ใช้ คือ Logistic Regression, Support Vector Machines (SVM) และ Random Forest

สำหรับแบบจำลอง Logistic Regression ได้ทำการหา Hyperparameter โดยทำการ Tuning ปรับค่าพารามิเตอร์ให้เหมาะสมทั้งหมด 3 ตัวดังนี้ คือ

1. C คือ ค่าผกผันของการทำให้แบบจำลองเป็นมาตรฐาน ซึ่งในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น 0.001, 0.01, 0.1, 1 และ 10

2. Penality คือการระบุ norm ใน Penalization โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น l1 และ l2

3. Solver คือพารามิเตอร์ที่ใช้ในการระบุวิธีการคำนวณค่าพารามิเตอร์ โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น liblinear และ saga

เมื่อทำการปรับจูนพารามิเตอร์ ทั้ง 3 ตัว ครบแล้ว ซึ่งการจูนนี้ ปรับทั้งแบบจำลองที่ไม่ใช้ SMOTE ตลอดจนแบบจำลองที่ใช้ SMOTE โดยผลจากการปรับที่ได้ตามตาราง 5

ตาราง 5 แสดงผล HyperParameter ที่ได้จาก GridSearchCV ของแบบจำลอง Logistic Regression

| **พารามิเตอร์** | **แบบจำลองที่ไม่ใช้ SMOTE** | **แบบจำลองที่ใช้ SMOTE** |
|---|---|---|
| C | 0.01 | 0.001 |
| Penality | l2 | l1 |
| Solver | saga | saga |

สำหรับแบบจำลอง Support Vector Machines (SVM) ได้ทำการหา Hyperparameter โดยทำการ Tuning ปรับค่าพารามิเตอร์ให้เหมาะสมทั้งหมด 3 ตัวดังนี้ คือ

1. C คือการกำหนดขนาด ของ Regularization parameter โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น 0.1, 1 และ 10

2. Gamma คือค่าสัมประสิทธิ์ของ Kernel โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น 0.001, 0.01, 0.1

3. Kernel คือประเภทที่ใช้ในการแปลงมิติของข้อมูล โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น linear และ rbf

เมื่อทำการปรับจูนพารามิเตอร์ ทั้ง 3 ตัว ครบแล้ว ซึ่งการจูนนี้ ปรับทั้งแบบจำลองที่ไม่ใช้ SMOTE ตลอดจนแบบจำลองที่ใช้ SMOTE โดยผลจากการปรับที่ได้ตามตาราง 6

ตาราง 6 แสดงผล Hyperparameter ที่ได้จาก GridSearchCV ของแบบจำลอง Support Vector Machines (SVM)

| **พารามิเตอร์** | **แบบจำลองที่ไม่ใช้** SMOTE | **แบบจำลองที่ใช้** SMOTE |
|---|---|---|
| C | 10 | 10 |
| Gamma | 0.1 | 0.1 |
| Kernel | rbf | rbf |

สำหรับแบบจำลอง Random Forest ได้ทำการหา Hyperparameter โดยทำการ Tuning ปรับค่าพารามิเตอร์ให้เหมาะสมทั้งหมด 4 ตัวดังนี้ คือ

1. n_estimators คือการกำหนดจำนวนต้นไม้ที่ใช้ในการตัดสินใจ โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น 200 และ 500

2. max_features คือการกำหนดฟีเจอร์ที่ใช้ในการตัดสินใจ โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น auto, sqrt และ log2

3. max_depth คือจำนวนลำดับชั้นของต้นไม้แต่ละต้น โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น 4, 5, 6, 7 และ 8

4. criterion คือฟังก์ชั่นสำหรับการวัดคุณภาพของการแบ่งข้อมูล โดยในงานวิจัยนี้ได้กำหนดค่าพารามิเตอร์เริ่มต้นเป็น gini และ entropy

เมื่อทำการปรับจูนพารามิเตอร์ ทั้ง 4 ตัว ครบแล้ว ซึ่งการจูนนี้ ปรับทั้งแบบจำลองที่ไม่ใช้ SMOTE ตลอดจนแบบจำลองที่ใช้ SMOTE โดยผลจากการปรับที่ได้ตามตาราง 7

ตาราง 7 แสดงผล Hyperparameter ที่ได้จาก GridSearchCV ของแบบจำลอง Random Forest

| **พารามิเตอร์** | **แบบจำลองที่ไม่ใช้ SMOTE** | **แบบจำลองที่ใช้ SMOTE** |
|---|---|---|
| n_estimators | 500 | 200 |
| max_features | auto | auto |
| max_depth | 8 | 8 |
| criterion | entropy | gini |

ฉะนั้นหลังจากการทำ Tuning หา Hyperparameter ครบทั้งหมดทุกแบบจำลองที่จะใช้ในงานวิจัยนี้แล้ว เพื่อนำไปทำการเรียนรู้ในข้อมูลชุดเรียนรู้ หลังจากทำการเรียนรู้ทุกแบบจำลองได้สำเร็จเสร็จสิ้นเป็นอันเรียบร้อยแล้ว จึงได้ทำการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง

# บทที่ 4

# ผลการดำเนินการวิจัย

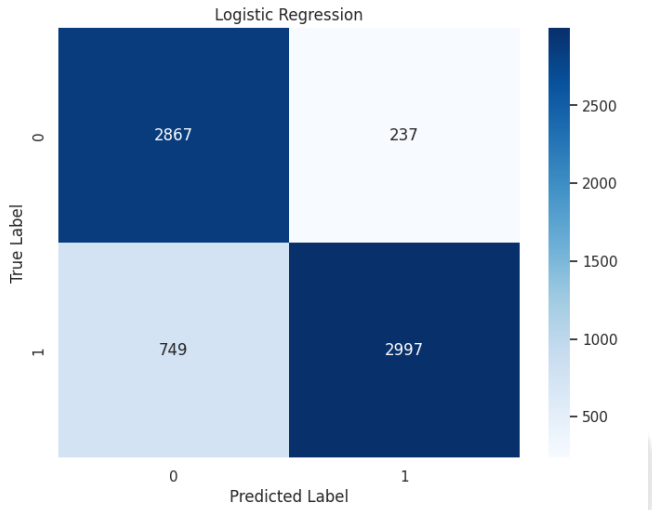
ในการวิจัยการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง ได้ใช้ข้อมูลทดสอบ (Test Set) มีจำนวนข้อมูลทั้งหมด 6,850 ตัวอย่าง ซึ่งประกอบด้วยกลุ่ม 0 คือลูกค้า Exist มีข้อมูลทั้งหมด 3,104 ตัวอย่างและกลุ่ม 1 คือลูกค้า Churn มีข้อมูลทั้งหมด 3,746 ตัวอย่าง ผู้วิจัยได้ดำเนินการวิจัยโดยศึกษาตามขั้นตอนต่าง ๆ ตลอดจนวัดประสิทธิภาพการทำงานของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix ได้ตามตาราง 8

ตาราง 8 แสดงผลลัพธ์ของแบบจำลองทั้งหมด

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

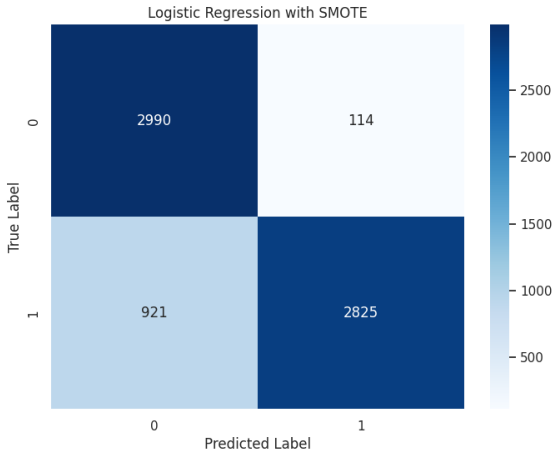
จากการศึกษาผลลัพธ์ของแบบจำลองในตาราง 8 จะเห็นได้ว่าผลลัพธ์ของแบบจำลอง Random Forest ทั้งแบบที่ใช้และไม่ใช้งาน SMOTE พบว่าแบบจำลองทั้งสองมีประสิทธิภาพ

โดยรวมดีที่สุด โดยที่ค่า Accuracy 92%, Precision 93%, Recall 92% และ F1-Score 93% ได้ค่าเท่ากันและระยะเวลาที่ใช้ในการเรียนรู้ไม่แตกต่างกันมาก ในขณะที่แบบจำลอง Logistic Regression ซึ่งใช้ระยะเวลาในการเรียนรู้ที่น้อยกว่า แต่มีประสิทธิภาพโดยรวมน้อยกว่าแบบจำลอง Random Forest ซึ่งแสดงให้เห็นว่าแบบจำลอง Random Forest เป็นทางเลือกที่เหมาะสมสำหรับข้อมูลตัวอย่าง นอกจากนี้ผู้วิจัยได้ทำ Confusion Matrix เพื่อตรวจสังเกตความถูกต้องและความผิดพลาดของตัวอย่างข้อมูลจากชุดทดสอบ แสดงตามผลลัพธ์ดังนี้



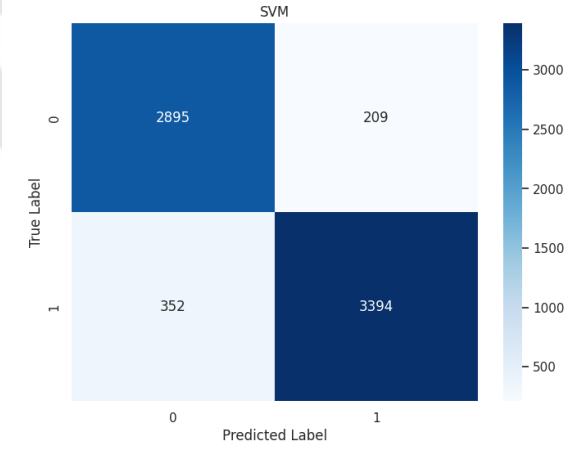

ภาพประกอบ 43 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง Logistic Regression

จากภาพประกอบ 43 แสดงผล Confusion Matrix ของแบบจำลอง Logistic Regression พบว่ามีประสิทธิภาพการทำนายผลในกลุ่มลูกค้า Exist (Label 0) เป็นจำนวน 2,867 ตัวอย่างข้อมูล และทำนายกลุ่มลูกค้า Churn (Label 1) เป็นจำนวน 2,997 ตัวอย่างข้อมูล แบบจำลองมีการทำนายที่ถูกต้อง 5,864 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 986 ตัวอย่างข้อมูล



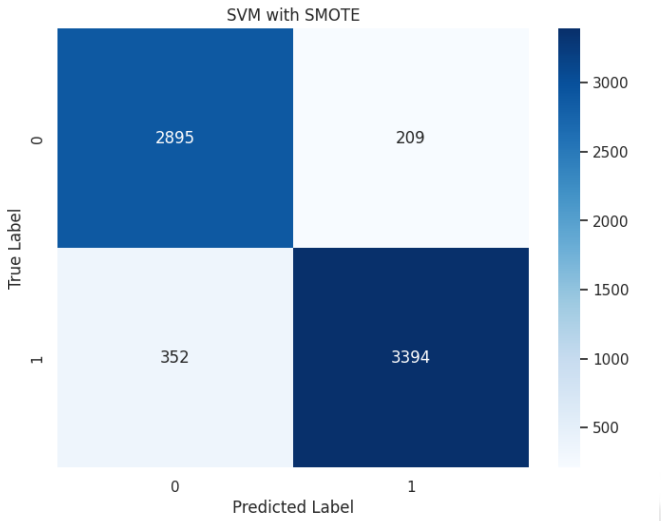

ภาพประกอบ 44 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง Logistic Regression ที่ใช้งาน SMOTE

จากภาพประกอบ 44 แสดงผล Confusion Matrix ของแบบจำลอง Logistic Regression ใช้งาน SMOTE พบว่ามีประสิทธิภาพในการทำนายผลได้ถูกต้องในกลุ่มลูกค้า Exist (Label 0) เป็นจำนวน 2,990 ตัวอย่างข้อมูลและทำนายได้ถูกต้องในกลุ่มลูกค้า Churn (Label 1) เป็นจำนวน 2,825 ตัวอย่างข้อมูล มีการทำนายที่ถูกต้อง 5,815 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 1,035 ตัวอย่างข้อมูล



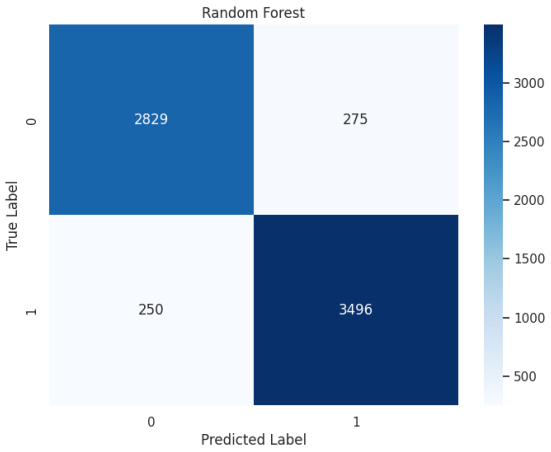

ภาพประกอบ 45 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง SVM

จากภาพประกอบ 45 แสดงผล Confusion Matrix ของแบบจำลอง SVM พบว่ามีประสิทธิภาพในการทำนายผลได้ถูกต้องในกลุ่มลูกค้า Exist (Label 0) เป็นจำนวน 2,895 ตัวอย่างข้อมูล และทำนายได้ถูกต้องในกลุ่มลูกค้า Churn (Label 1) เป็นจำนวน 3,394 ตัวอย่างข้อมูล ซึ่งแบบจำลองมีการทำนายที่ถูกต้อง 6,289 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 561 ตัวอย่างข้อมูล



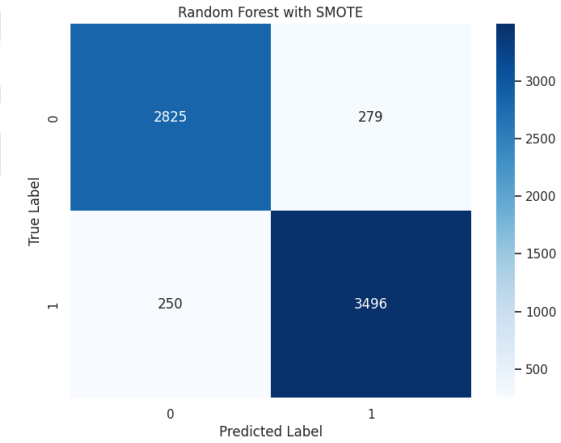

ภาพประกอบ 46 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง SVM ที่ใช้งาน SMOTE

จากภาพประกอบ 46 แสดงผล Confusion Matrix ของแบบจำลอง SVM ใช้งาน SMOTE พบว่ามีประสิทธิภาพในการทำนายผลได้ถูกต้องในกลุ่มลูกค้า Exist (Label 0) เป็นจำนวน 2,895 ตัวอย่างข้อมูลและทำนายได้ถูกต้องในกลุ่มลูกค้า Churn (Label 1) เป็นจำนวนทั้งหมด 3,394 ตัวอย่างข้อมูล มีการทำนายที่ถูกต้อง 6,289 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 561 ตัวอย่างข้อมูล




ภาพประกอบ 47 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง Random Forest

จากภาพประกอบ 47 แสดงผล Confusion Matrix ของแบบจำลอง Random Forest พบว่ามีประสิทธิภาพในการทำนายผลได้ถูกต้องในกลุ่มลูกค้า Exist (Label 0) เป็นจำนวน 2,829 ตัวอย่างข้อมูล และทำนายได้ถูกต้องในกลุ่มลูกค้า Churn (Label 1) เป็นจำนวน 3,496 ตัวอย่างข้อมูล ซึ่งแบบจำลองมีการทำนายที่ถูกต้อง 6,325 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 525 ตัวอย่างข้อมูล



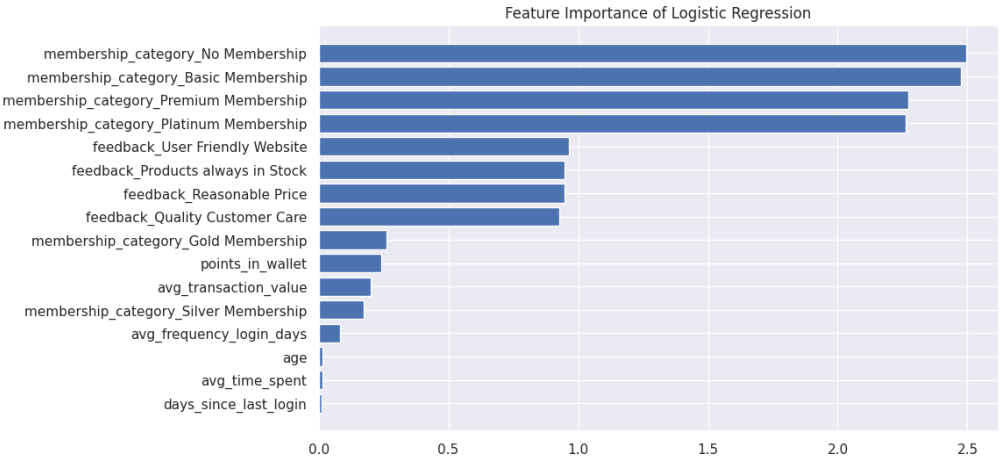

ภาพประกอบ 48 แสดงผลลัพธ์ Confusion Matrix จากการทำนายของแบบจำลอง Random Forest ที่ใช้งาน SMOTE

จากภาพประกอบ 48 แสดงผล Confusion Matrix ของแบบจำลอง Random Forest ใช้งาน SMOTE พบว่ามีประสิทธิภาพในการทำนายผลได้ถูกต้องในกลุ่มลูกค้า Exist (Label 0) เป็น

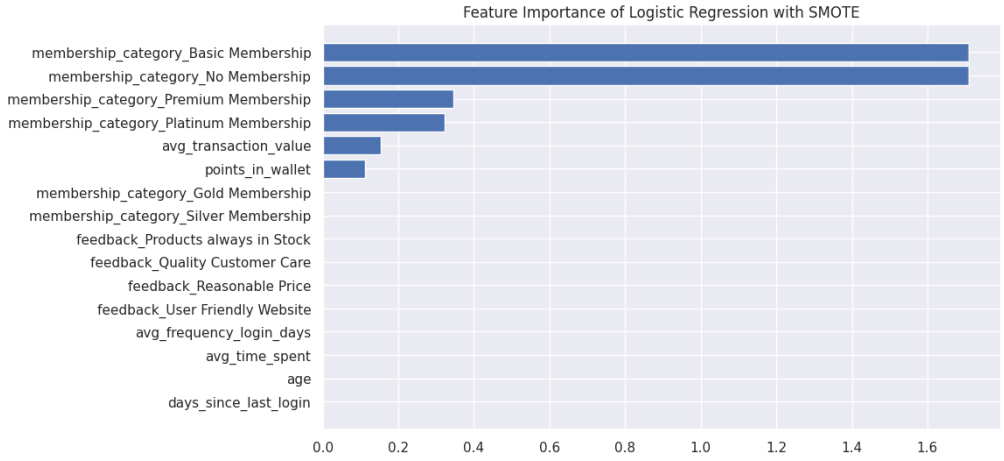
จำนวน 2,825 ตัวอย่างข้อมูลและทำนายได้ถูกต้องในกลุ่มลูกค้า Churn (Label 1) เป็นจำนวนทั้งหมด 3,496 ตัวอย่างข้อมูล ซึ่งแบบจำลองมีการทำนายที่ถูกต้อง 6,321 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 529 ตัวอย่างข้อมูล

ในขั้นตอนต่อไปนี้ ตามภาพประกอบ 49 และ 50 ผู้วิจัยได้ทำการแสดงผลความสำคัญของฟีเจอร์ในแต่ละแบบจำลองเพื่อเป็นการสำรวจให้เข้าใจว่า ฟีเจอร์ใดมีอิทธิพลในการกำหนดผลลัพธ์ให้กับกลุ่มลูกค้า Exist และ Churn โดยสำหรับแบบจำลอง Logistic Regression ฟีเจอร์ที่มีค่าสัมประสิทธิ์ (Coefficients) มากจะถือว่าเป็นคุณลักษณะที่สำคัญ ซึ่งแบบจำลอง Logistic Regression ไม่ใช้ SMOTE พบว่าฟีเจอร์ที่สำคัญต่อการจำแนกกลุ่มมีความเกี่ยวข้องกับระดับการเป็นสมาชิกของลูกค้าคือ membership_category_No Membership หรือลูกค้าที่ไม่ได้เป็นสมาชิก และ membership_category_Basic Membership หรือลูกค้าสมาชิกระดับเริ่มต้น

ส่วนแบบจำลอง Logistic Regression ใช้งาน SMOTE พบว่าฟีเจอร์ที่มีส่วนสำคัญต่อการจัดกลุ่มคือ membership_category_Basic Membership, membership_category_No Membership ซึ่งทั้ง 2 แบบจำลองนี้มุ่งเน้นให้ความสนใจไปที่ระดับของลูกค้าที่ส่งผลต่อการทำนายการเลิกเป็นลูกค้า



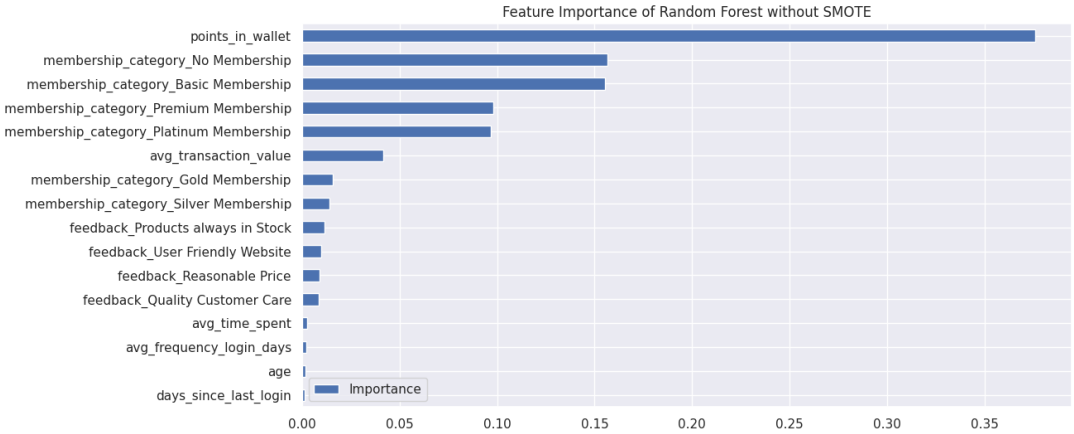

ภาพประกอบ 49 แสดงความสำคัญของคุณลักษณะจากแบบจำลอง Logistic Regression



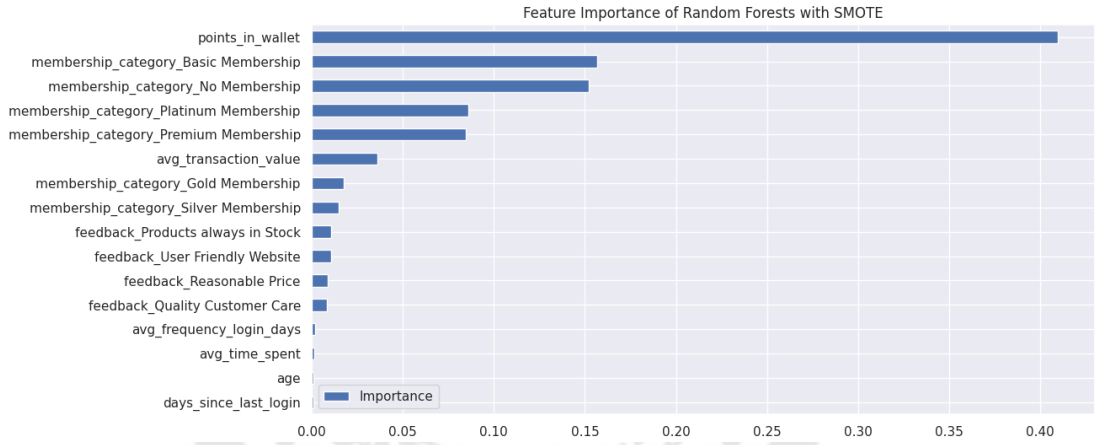

ภาพประกอบ 50 แสดงความสำคัญของคุณลักษณะจากแบบจำลอง Logistic Regression ที่ใช้งาน SMOTE

สำหรับแบบจำลอง SVM และ SVM ใช้งาน SMOTE เนื่องจากได้พารามิเตอร์ rbf (Radial Basis Function) ที่ได้จาก GridSearchCV ไม่สามารถแสดง Feature Importance ออกมาได้โดยตรงเหมือนกับแบบจำลองอื่น ๆ เหตุเพราะ rbf เป็นฟังก์ชั่นไม่ใช่เชิงเส้น คือมีการเปลี่ยนแปลงมิติของฟีเจอร์

สำหรับแบบจำลอง Random Forest แบบไม่ใช้ SMOTE พบว่าฟีเจอร์ที่มีความสำคัญต่อการจัดกลุ่มคือ points_in_wallet หรือคะแนนสะสมของลูกค้า และ membership_category_No Membership สำหรับแบบจำลอง Random Forest แบบใช้ SMOTE พบว่าฟีเจอร์ที่มีความสำคัญสุดคือ points_in_wallet และ membership_category_Basic Membership ซึ่งทั้ง 2 แบบจำลองนี้มีฟีเจอร์ที่ต่างจากแบบจำลองอื่นคือเน้นให้ความสำคัญกับคะแนนสะสมของลูกค้ามากที่สุด ตามภาพประกอบที่ 51 และ 52



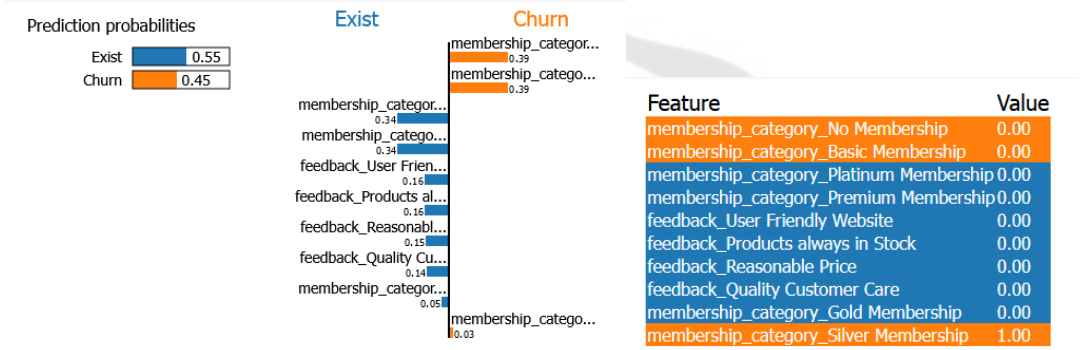

ภาพประกอบ 51 แสดงความสำคัญของคุณลักษณะจากแบบจำลอง Random Forest




ภาพประกอบ 52 แสดงความสำคัญของคุณลักษณะจากแบบจำลอง Random Forest ที่ใช้งาน SMOTE

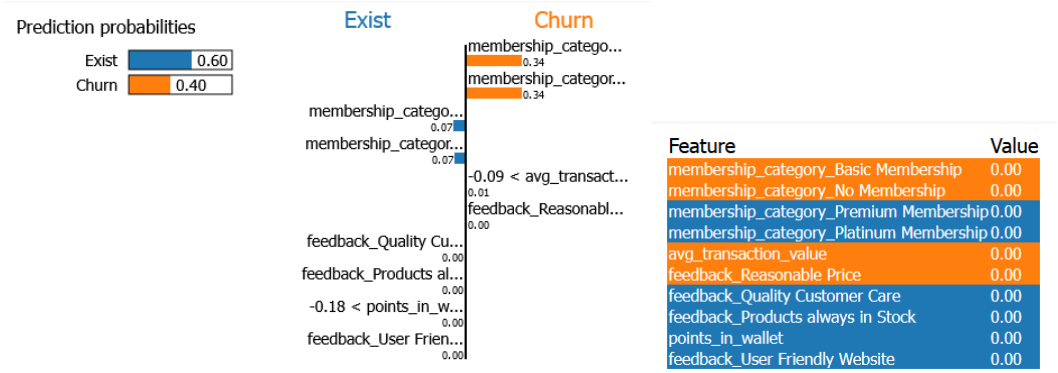
เมื่อทำการสำรวจฟีเจอร์ที่มีความสำคัญของแต่ละแบบจำลองเรียบร้อยแล้ว แสดงให้เห็นว่าฟีเจอร์ที่มีอิทธิพลในการจัดกลุ่มส่วนใหญ่คือ ระดับการเป็นสมาชิกของลูกค้าและคะแนนสะสมของลูกค้า หลังจากนั้นใช้เทคนิคอธิบายและตีความหมายของการทำงานแต่ละแบบจำลอง โดยผู้วิจัยเลือกใช้ไลบารี่ LIME (Local Interpretable Model-Agnostic Explanations) ในการหาการใช้งานบนฟีเจอร์บนข้อมูล 1 ตัว เพื่อเปรียบเทียบและสร้างความน่าเชื่อถือให้กับแบบจำลอง อีกทั้งยังวิเคราะห์หาความผิดพลาดที่เกิดขึ้นอีกด้วย เนื่องจากเหตุผลดังกล่าวนี้เป็นเหตุนำมาเพื่อเลือกหาแบบจำลองมีประสิทธิภาพมากที่สุด จากแนวคิดข้างต้น ผู้วิจัยต้องการพิจารณาว่าฟีเจอร์ใดบ้างที่ถูกนำมาใช้งาน จึงทำการเลือกข้อมูลแบบสุ่มมา 1 ข้อมูลจากข้อมูลในชุดทดสอบ (Test

Set) โดยเลือกข้อมูลในตำแหน่งของแถว (Index) ได้เป็นข้อมูลที่ 125 ซึ่งเลเบลของข้อมูลนี้เป็นกลุ่มลูกค้า Churn เมื่อทำการสำรวจแบบจำลอง Logistic Regression ที่ไม่ได้ใช้ SMOTE แสดงความน่าจะเป็นและการใช้ฟีเจอร์ดังภาพประกอบ 53 พบว่าแบบจำลองมีการทำนายผิดพลาดเกิดขึ้น คือทำนายเป็นกลุ่มลูกค้า Exist ด้วยความน่าจะเป็นถึง 0.55 ในขณะกลุ่มลูกค้า Churn มีความน่าจะเป็นเพียง 0.45 ซึ่งหมายความว่าลูกค้ารายนี้มีแนวโน้มที่จะถูกจำแนกเป็นลูกค้าที่ยังใช้บริการอยู่ เมื่อพิจารณาการใช้ฟีเจอร์ที่แบบจำลองนำมาอธิบายการจำแนกประเภทกลุ่มลูกค้า พบว่า membership_category_No Membership หรือลูกค้าสมาชิกระดับแพลทินัม ซึ่งให้ค่าน้ำหนักที่ 0.39 เป็นฟีเจอร์ที่มีอิทธิพลในการทำนายมากที่สุด



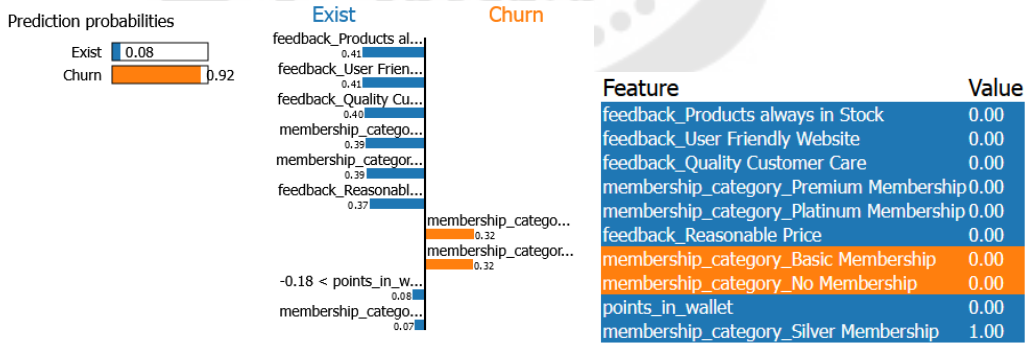

ภาพประกอบ 53 แสดงผลลัพธ์ด้วยวิธี LIME จากแบบจำลอง Logistic Regression

สำหรับแบบจำลอง Logistic Regression ที่ใช้ SMOTE แสดงความน่าจะเป็นและการใช้ฟีเจอร์ตามภาพประกอบ 54 พบว่าแบบจำลองมีการทำนายผิดพลาด โดยที่การทำนายเป็นกลุ่มลูกค้า Exist ด้วยความน่าจะเป็นถึง 0.60 ในขณะกลุ่มลูกค้า Churn ด้วยความน่าจะเป็นที่ 0.40 เมื่อพิจารณาฟีเจอร์ที่นำมาอธิบายการทำงานของแบบจำลองแสดงให้เห็นว่า membership_category_ Basic Membership หรือลูกค้าสมาชิกระดับเริ่มต้น ให้ค่าน้ำหนักที่ 0.34 ซึ่งมีอิทธิพลมากสุด อีกทั้งยังพบว่าแบบจำลองใช้ฟีเจอร์ไม่เหมาะสมในการจัดกลุ่ม พบว่า feedback_Reasonable Price หรือการแสดงความคิดเห็นของลูกค้าว่าราคาสมเหตุสมผล ถูกจัดให้เป็นฟีเจอร์ในกลุ่มลูกค้า Churn




ภาพประกอบ 54 แสดงผลลัพธ์ด้วยวิธี LIME จากแบบจำลอง Logistic Regression ที่ใช้งาน SMOTE

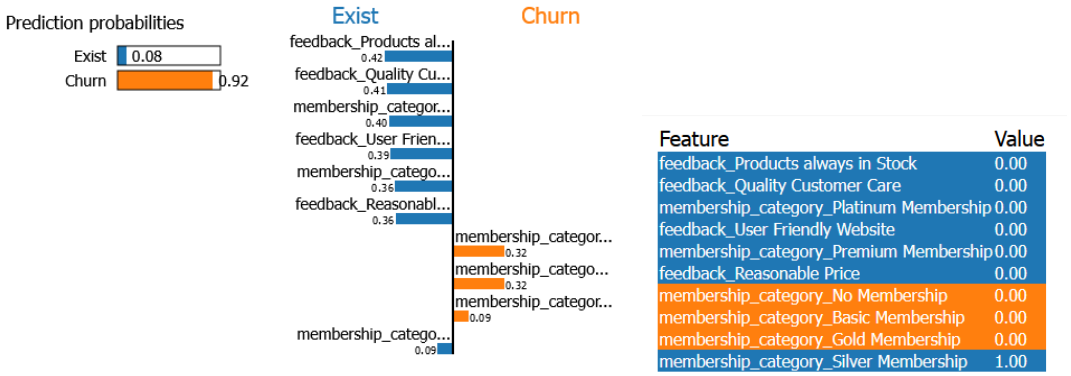
สำหรับแบบจำลอง SVM ที่ไม่ใช้ SMOTE แสดงความน่าจะเป็นและการใช้ฟีเจอร์ดังภาพประกอบ 55 พบว่าแบบจำลองมีการทำนายถูกต้อง โดยที่การทำนายเป็นกลุ่มลูกค้า Exist ด้วยความน่าจะเป็นถึง 0.08 ในขณะกลุ่มลูกค้า Churn ด้วยความน่าจะเป็นที่ 0.92 แสดงให้เห็นว่าลูกค้ารายนี้มีแนวโน้มที่จะเป็นลูกค้า Churn มากกว่า ฟีเจอร์ที่ถูกนำมาแบ่งกลุ่มพบว่า feedback_Products always in Stock หรือการแสดงความคิดเห็นของลูกค้าว่าสินค้ามีในสต็อกเสมอ มีค่าน้ำหนัก 0.41 ซึ่งเป็นฟีเจอร์ที่มีความสำคัญเป็นอันดับแรก ซึ่งแบบจำลองนี้ใช้ฟีเจอร์ที่แตกต่างไปจากแบบจำลองอื่น ๆ คือให้ความสำคัญไปที่ฟีเจอร์ feedback หรือการแสดงความคิดเห็นของลูกค้า




ภาพประกอบ 55 แสดงผลลัพธ์ด้วยวิธี LIME จากแบบจำลอง SVM

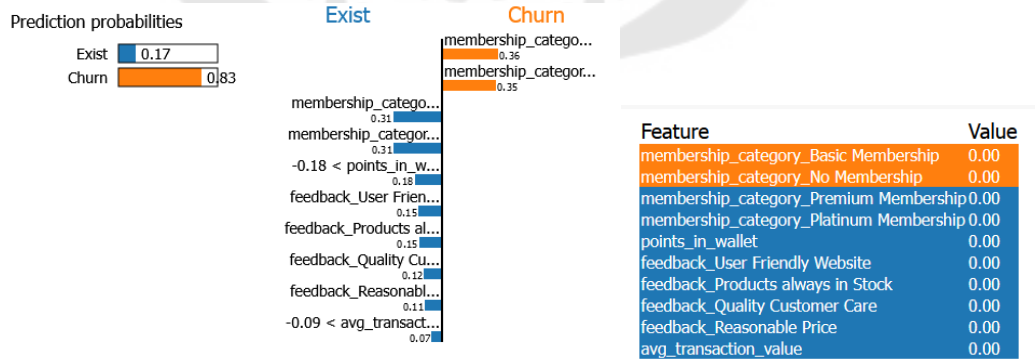
ส่วนแบบจำลอง SVM ที่ใช้ SMOTE แสดงความน่าจะเป็นและการใช้ฟีเจอร์ตามภาพประกอบ 56 พบว่าแบบจำลองมีการทำนาถูกต้อง โดยที่การทำนายเป็นกลุ่มลูกค้า Exist ด้วยความน่าจะเป็นถึง 0.08 ในขณะกลุ่มลูกค้า Churn ด้วยความน่าจะเป็นที่ 0.92 แสดงให้เห็นว่า

ลูกค้ารายนี้มีแนวโน้มที่จะเป็นลูกค้า Churn มากกว่า และฟีเจอร์ที่มีความสำคัญว่าในการจัดกลุ่มส่วนใหญ่เป็นการแสดงความคิดเห็นของลูกค้าและระดับสมาชิก ซึ่งฟีเจอร์ feedback_Products always in Stock มีอิทธิพลมากสุด




ภาพประกอบ 56 แสดงผลลัพธ์ด้วยวิธี LIME จากแบบจำลอง SVM ที่ใช้งาน SMOTE

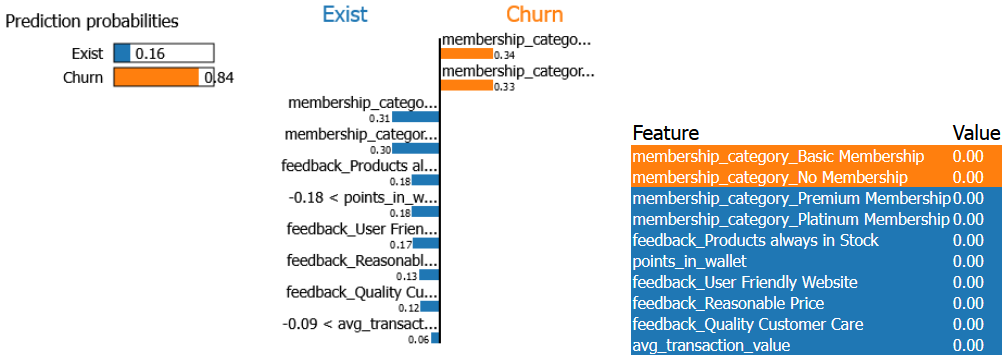
สำหรับแบบจำลอง Random Forest แบบไม่ใช้ SMOTE แสดงความน่าจะเป็นและการใช้ฟีเจอร์ตามภาพประกอบ 57 พบว่าแบบจำลองมีการทำนายถูกต้อง โดยที่การทำนายเป็นกลุ่มลูกค้า Exist ด้วยความน่าจะเป็นถึง 0.17 ในขณะกลุ่มลูกค้า Churn ด้วยความน่าจะเป็นที่ 0.83 แสดงให้เห็นว่าการทำนายนี้มีแนวโน้มที่จะเป็นลูกค้า Churn มากกว่า และฟีเจอร์ที่ถูกนำมาจัดว่าเป็นกลุ่มลูกค้าที่ไม่ใช้บริการแล้วได้อย่างถูกต้องที่มีความสำคัญเป็นอันดับแรกคือ membership_category_Basic Membership หรือลูกค้าสมาชิกระดับเริ่มต้น




ภาพประกอบ 57 แสดงผลลัพธ์ด้วยวิธี LIME จากแบบจำลอง Random Forest

แบบจำลอง Random Forest ที่ใช้ SMOTE แสดงความน่าจะเป็นและการใช้ฟีเจอร์ดังภาพประกอบ 58 พบว่าแบบจำลองมีการทำนายถูกต้อง โดยที่การทำนายเป็นกลุ่มลูกค้า Exist

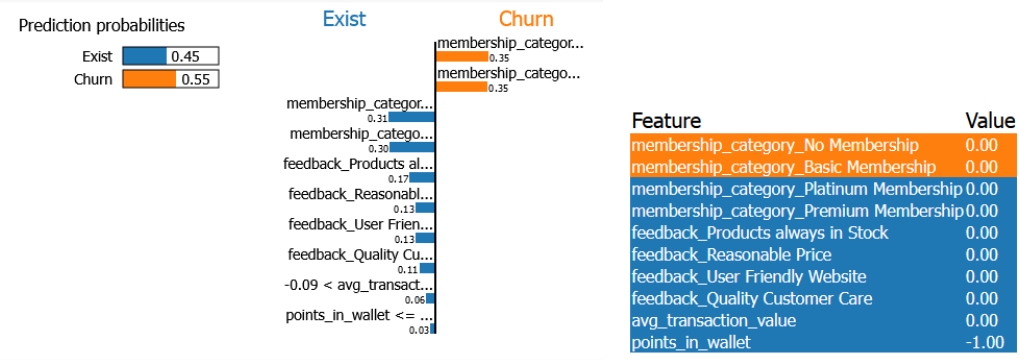
ด้วยความน่าจะเป็นถึง 0.16 ในขณะกลุ่มลูกค้า Churn ด้วยความน่าจะเป็นที่ 0.84 จากการพิจารณาด้วยฟีเจอร์ที่พบว่ามีแนวโน้มที่จะถูกจำแนกเป็นลูกค้ากลุ่ม Churn มากกว่าที่จะใช้บริการ ซึ่งฟีเจอร์ที่มีความสำคัญในการนำไปจัดกลุ่มคือ membership_category_Basic Membership หรือลูกค้าที่ไม่ได้เป็นสมาชิกเช่นเดียวกับที่ไม่ใช้ SMOTE




ภาพประกอบ 58 แสดงผลลัพธ์ด้วยวิธี LIME จากแบบจำลอง Random Forest ที่ใช้งาน SMOTE

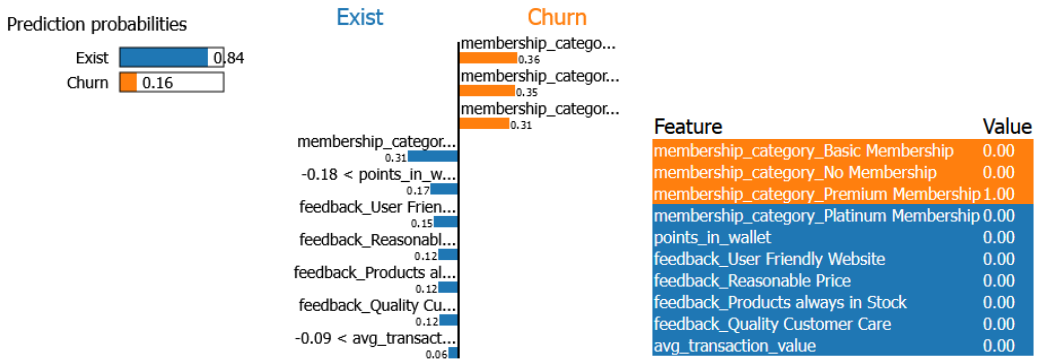
ในขั้นตอนต่อไปนี้ผู้วิจัยมีความต้องการเปรียบเทียบการอธิบายฟีเจอร์ด้วย LIME กับผลค่าความสำคัญของฟีเจอร์ (Feature Importance) ว่ามีความสอดคล้องกันหรือไม่ โดยวัดจากผลลัพธ์ของประสิทธิภาพของแบบจำลองที่ดีที่สุดในงานวิจัยนี้ จึงเลือกใช้ Random Forest แบบไม่ใช้ SMOTE นำมาพิจารณากับข้อมูลชุดทดสอบ ที่ยังไม่ได้เห็นมาก่อนหน้านี้ เพื่อตรวจสอบการทำงานของแบบจำลองนี้เพิ่มเติม ว่าฟีเจอร์ที่ใช้ในการอธิบายเหมือนกัน และมีความเหมาะสมที่ใช้ในการทำนายหรือไม่ พบว่า Feature Importance ของแบบจำลอง Random Forest แบบไม่ใช้ SMOTE ที่มีค่าความสำคัญที่ให้ค่าสูง 2 อันดับแรกคือ ฟีเจอร์ points_in_wallet หรือคะแนนสะสมของลูกค้า และ membership_category_No Membership หรือไม่ได้เป็นสมาชิก ข้อมูลแรกที่เลือกมาคือข้อมูลที่ 246 จากการสำรวจพบว่าการทำนายนี้ถูกต้องว่าเป็นลูกค้า Churn โดยที่ข้อมูลนี้มีแนวโน้มที่จะเป็นลูกค้ากลุ่ม Churn มากกว่าลูกค้า Exist ซึ่งมีความน่าจะเป็นถึง 0.55 โดยพบว่าฟีเจอร์ที่มีความสำคัญมากที่สุด 3 อันดับแรก ซึ่งลำดับที่ 1 คือ ฟีเจอร์ membership_category_No Membership และลำดับที่ 2 คือ membership_category_ Basic Membership ในส่วนลำดับที่ 3 คือ membership_category_Premium Membership หรือลูกค้าสมาชิกระดับแพลทินัม ซึ่งฟีเจอร์ทั้งหมดนี้มีอิทธิพลในการจัดกลุ่มลูกค้ามากที่สุด และเมื่อพิจารณาแล้วพบว่ามีความสอดคล้องกับ Feature Importance ของแบบจำลองนี้ แต่มีการเปลี่ยนแปลงลำดับของคุณลักษณะที่สำคัญเล็กน้อย ทำให้ membership_category_No Membership เป็นคุณลักษณะที่สำคัญที่สุดแทน มีน้ำหนักที่ 0.35 แต่กลับกัน ฟีเจอร์

points_in_wallet โดยแท้จริงแล้วมีค่าความสำคัญของฟีเจอร์สูงที่สุด ใน Feature Importance แต่กลับเป็นฟีเจอร์ที่มีความสำคัญเป็นอันดับท้ายใน LIME มีน้ำหนัก 0.03 ซึ่งเป็นที่ชัดเจนจากการสำรวจข้อมูลในบทที่ 3 ว่ากลุ่มลูกค้าระดับเริ่มต้นและที่ไม่ได้สมัครสมาชิกนั้นมีแต่กลุ่มลูกค้า Churn ซึ่งมีความแตกต่างจากระดับลูกค้าอื่น ๆ นี่อาจจะเป็นสาเหตุที่ทำให้แบบจำลองนี้ให้ความสำคัญกับการใช้คุณลักษณะระดับสมาชิกมากกว่าคะแนนสะสมของลูกค้าได้ ตามภาพประกอบ 59



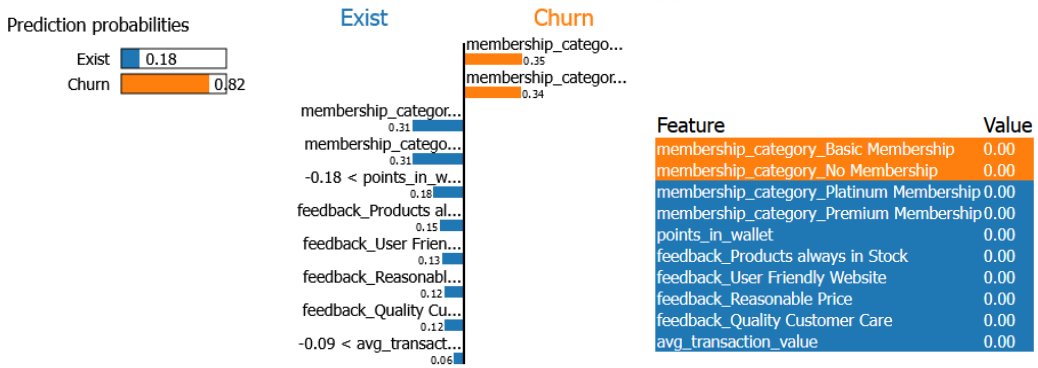

ภาพประกอบ 59 แสดงผลลัพธ์ด้วยวิธี LIME ของข้อมูลที่ 246 ในข้อมูลชุดทดสอบ จากแบบจำลอง Random Forest

สำหรับข้อมูลที่ 879 ตามภาพประกอบ 60 เลือกมาศึกษาเพิ่มเติมในการทำงานของแบบจำลองนี้ พบว่าข้อมูลนี้มีแนวโน้มที่จะเป็นลูกค้ากลุ่ม Exist มีความน่าจะเป็นถึง 0.84 เลยทีเดียว ซึ่งการทำนายนี้ถูกต้องว่าเป็นกลุ่ม Exist จากนั้นพิจารณาฟีเจอร์ที่ถูกนำมาใช้ในการจำแนกนี้ พบว่าส่วนใหญ่เป็นฟีเจอร์ที่เกี่ยวข้องกับระดับการเป็นสมาชิกเหมือนกับการใช้เครื่องมือ LIME ในข้อมูลที่ 246 โดยทั้ง 3 ฟีเจอร์เหล่านี้มีความสอดคล้องกับ Feature Importance แต่เมื่อสังเกตแล้วพบว่ามีความผิดพลาดที่ใช้ในการอธิบายเกิดขึ้น เนื่องในกลุ่มลูกค้า Churn นั้น ที่ได้จากการสำรวจข้อมูลไม่มีลูกค้าระดับพรีเมียมอยู่เลยแต่ข้อมูลชุดนี้จัดให้ลูกค้าระดับพรีเมียมเป็นกลุ่มลูกค้า Churn ซึ่งคุณลักษณะ membership_category_Premium Membership นี้ถูกจัดกลุ่มไม่สมเหตุสมผล



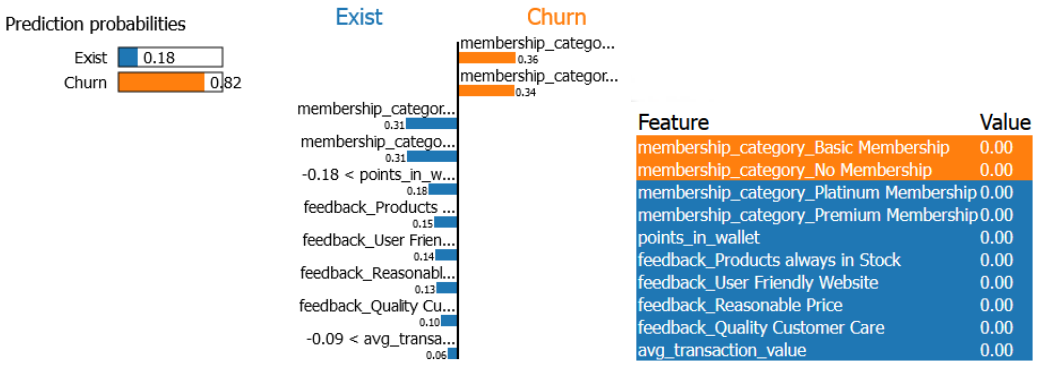

ภาพประกอบ 60 แสดงผลลัพธ์ด้วยวิธี LIME ของข้อมูลที่ 879 ในข้อมูลชุดทดสอบ จากแบบจำลอง Random Forest

สำหรับข้อมูลที่ 1784 ซึ่งมีการทำนายถูกต้องเป็นกลุ่มลูกค้า Churn พบว่าข้อมูลนี้มีแนวโน้มที่จะเป็นลูกค้ากลุ่ม Churn มีความน่าจะเป็น 0.82 ดังภาพประกอบ 61 แล้วฟีเจอร์ที่ถูกนำมาจำแนกว่าเป็นกลุ่มลูกค้านั้นให้ลำดับความสำคัญของคุณลักษณะเหมือนกับข้อมูลก่อนหน้านี้ ฟีเจอร์เหล่านี้มีความสอดคล้องกับค่าความสำคัญของฟีเจอร์แต่มีการเปลี่ยนแปลงลำดับเล็กน้อย โดยให้ความสำคัญกับ membership_category_ Basic Membership แทน แต่มีข้อสังเกตที่แตกต่างจากข้อมูลที่ 246 เพราะ points_in_wallet หรือคะแนนสะสมของลูกค้านั้น มีน้ำหนักที่ 0.18 ซึ่ง LIME ไม่ได้ให้ความสำคัญเป็นฟีเจอร์ลำดับสุดท้าย ทำให้ความน่าจะเป็นในการทำนายกลุ่ม Churn เพิ่มขึ้น จาก 0.55 เป็น 0.82 จากการสำรวจข้อมูลแล้วพบว่าคะแนนสะสมของลูกค้านั้นก็แสดงถึงความแตกต่างของกลุ่มลูกค้าได้ คือคะแนนสะสมของลูกค้าที่มีคะแนนสะสมมากส่วนใหญ่นั้นเป็นกลุ่มลูกค้าที่ยังใช้บริการอยู่



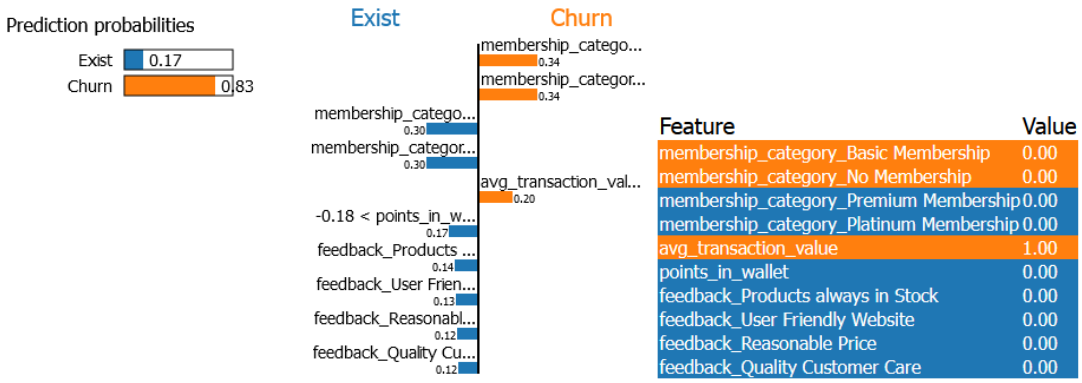

ภาพประกอบ 61 แสดงผลลัพธ์ด้วยวิธี LIME ของข้อมูลที่ 1784 ในข้อมูลชุดทดสอบ จากแบบจำลอง Random Forest

สำหรับข้อมูลที่ 3459 ซึ่งเลเบลของข้อมูลนี้คือกลุ่มลูกค้าที่ยังใช้บริการอยู่หรือ Exist โดยที่ข้อมูลนี้มีแนวโน้มที่จะเป็นกลุ่มลูกค้า Exist มีความน่าจะเป็น 0.18 ซึ่งแนวโน้มที่จะเป็นกลุ่มลูกค้า churn มีความน่าจะเป็น 0.82 แสดงตามภาพประกอบ 62 พบว่ามีการทำนายที่ผิดพลาดเกิดขึ้น ตามภาพประกอบที่ 62 แต่ฟีเจอร์นั้นมีความสอดคล้องกับ Feature Importance แต่มีการเปลี่ยนแปลงลำดับกันเล็กน้อย โดยฟีเจอร์ points_in_wallet มีน้ำหนัก 0.18 ซึ่งไม่ได้เป็นฟีเจอร์ที่มีความสำคัญอันดับแรก




ภาพประกอบ 62 แสดงผลลัพธ์ด้วยวิธี LIME ของข้อมูลที่ 3459 ในข้อมูลชุดทดสอบ จากแบบจำลอง Random Forest

สำหรับข้อมูลที่ 1255 พบว่าทำนายที่ผิดพลาดเกิดขึ้นอีกเช่นกัน ซึ่งเลเบลของข้อมูลนี้คือกลุ่มลูกค้าที่ยังใช้บริการอยู่ โดยที่ข้อมูลนี้มีแนวโน้มที่จะเป็นกลุ่มลูกค้า Exist มีความน่าจะเป็น 0.17 ซึ่งแนวโน้มที่จะเป็นกลุ่มลูกค้า churn มีความน่าจะเป็น 0.83 ตามภาพประกอบ 63 มีความแตกต่างจากข้อมูล 3459 โดยพบว่าฟีเจอร์ avg_transaction_value หรือมูลค่าการซื้อสะสมโดยเฉลี่ยของลูกค้า นั้นมีอิทธิพลต่อการจัดกลุ่มและถูกจัดให้อยู่ในกลุ่ม Churn จากการสำรวจข้อมูลเบื้องต้นในบทที่ 3 แล้วก็พบว่า มูลค่าการซื้อสะสมโดยเฉลี่ยของลูกค้านั้นก็แสดงความแตกต่างของกลุ่มลูกค้าได้ คือถ้ามีมูลค่าการซื้อสะสมมาก ก็จะยังคงเป็นลูกค้ากลุ่ม Exist อยู่ ถ้ามีมูลค่าที่สะสมน้อย ก็จะถูกจัดให้เป็นกลุ่ม Churn




ภาพประกอบ 63 แสดงผลลัพธ์ด้วยวิธี LIME ของข้อมูลที่ 1255 ในข้อมูลชุดทดสอบ จากแบบจำลอง Random Forest

อย่างไรก็ตาม บทที่ 4 ผลการดำเนินงานวิจัย มีการสรุปผลลัพธ์ของแบบจำลองทั้งหมดหมด 6 แบบจำลอง โดยมีการเปรียบเทียบประสิทธิภาพของแบบจำลอง และเปรียบเทียบในเรื่องของเวลาที่ใช้ในการเรียนรู้ในแต่ละแบบจำลองอีกด้วย รวมทั้งมีการแสดงการเปรียบเทียบของคุณลักษณะที่สำคัญ (Feature Importance) กับ Local Interpretable Model-agnostic Explanations (LIME) ซึ่งจากการเปรียบเทียบพบว่าแบบจำลอง Random Forest ที่ไม่ได้ใช้ข้อมูล SMOTE มีประสิทธิภาพดีที่สุด ในการทำนายการจำแนกประเภทของกลุ่มลูกค้าที่ยังใช้บริการอยู่กับลูกค้าที่เลิกใช้บริการไปแล้ว

# บทที่ 5

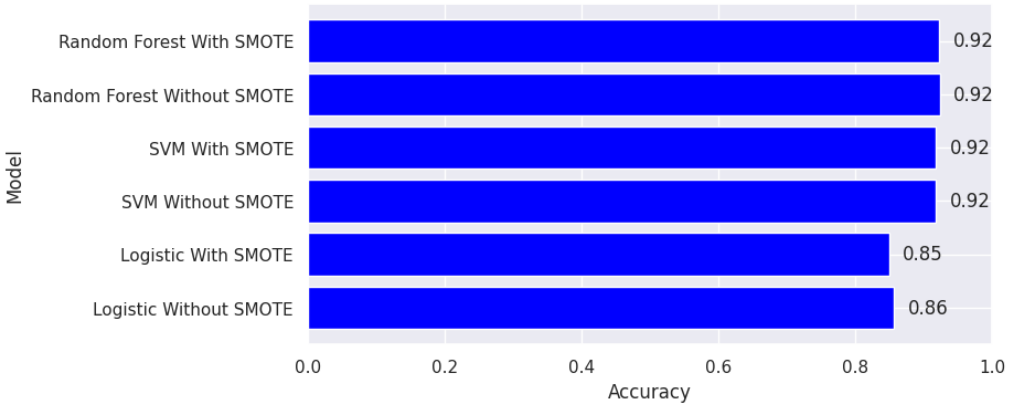
# สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในปัจจุบันมีการแข่งขันสำหรับธุรกิจ e-Commerce เป็นจำนวนมาก ประชากรโลกส่วนใหญ่เลือกที่จะบริโภคและอุปโภคในช่องทางออนไลน์กันมากยิ่งขึ้น ดังนั้นเพื่อให้ธุรกิจอยู่รอด จึงจำเป็นต้องค้นหาวิธีรักษาลูกค้าไว้ให้ได้ ในการวิจัยการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากร โดยใช้เทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้วัดประสิทธิภาพของแบบจำลองแต่ละอัลกอริทึมเพื่อเปรียบเทียบและทำการสรุปผลโดยแบ่งตามหัวข้อดังนี้

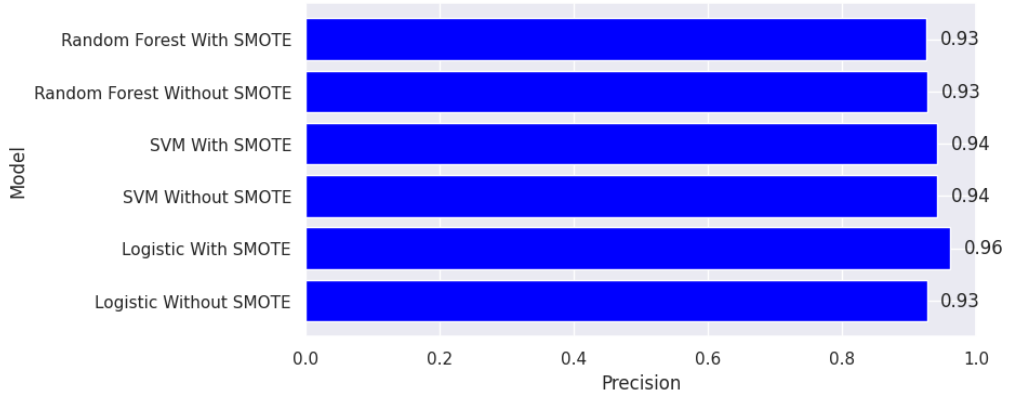
1. สรุปผลการวิจัย
2. อภิปรายผล
3. ข้อเสนอแนะ

## 5.1 สรุปผลการวิจัย

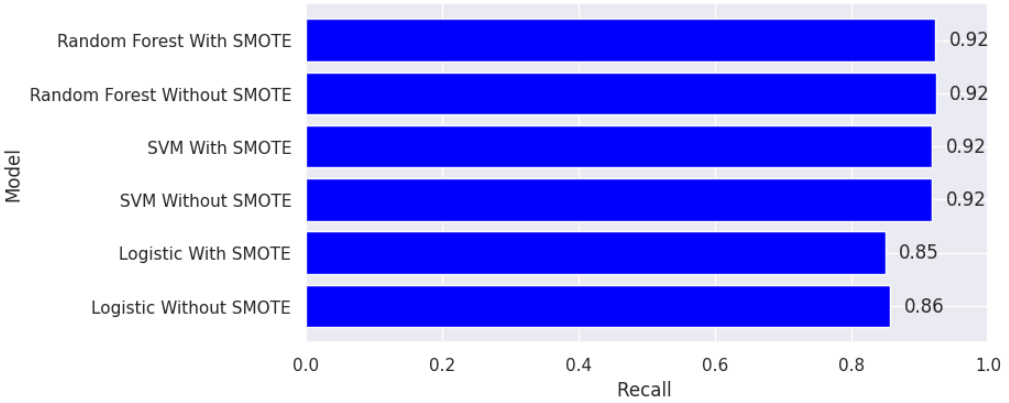
ในการวิจัยนี้ศึกษาการสร้างแบบจำลองการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรของลูกค้าบนเว็บไซต์แห่งหนึ่ง โดยใช้เทคนิคการเรียนรู้ของเครื่องเพื่อเปรียบเทียบประสิทธิภาพในการทำงานของแบบจำลอง ซึ่งมีทั้งหมด 3 แบบจำลองคือ Logistic Regression, Support Vector Machines (SVM) และ Random Forest แบบจำลองเหล่านี้ใช้การคัดเลือกคุณลักษณะและเทคนิค SMOTE ในการแก้ปัญหาความไม่สมดุลของข้อมูล จากผลการทดลองสรุปได้ว่าแบบจำลองที่ให้ประสิทธิภาพดีที่สุดคือ Random Forest ให้ผลลัพธ์ตามนี้ Accuracy 92%, Precision 93%, Recall 92% และ F1-Score 93% ในภาพประกอบ 64 ถึง 67 แสดงค่าวัดประสิทธิภาพของแบบจำลองทั้งหมด จากนั้นพิจารณา Confusion Matrix พบว่าแบบจำลองมีการทำนายที่ถูกต้อง 6,325 ตัวอย่างข้อมูล และการทำนายที่ไม่ถูกต้อง 525 ตัวอย่าง และยังค้นพบว่าแบบจำลอง Logistic Regression ที่ใช้ SMOTE ได้ใช้ระยะเวลาในการเรียนรู้ของแบบจำลองน้อยที่สุดคือ 0.09 วินาที



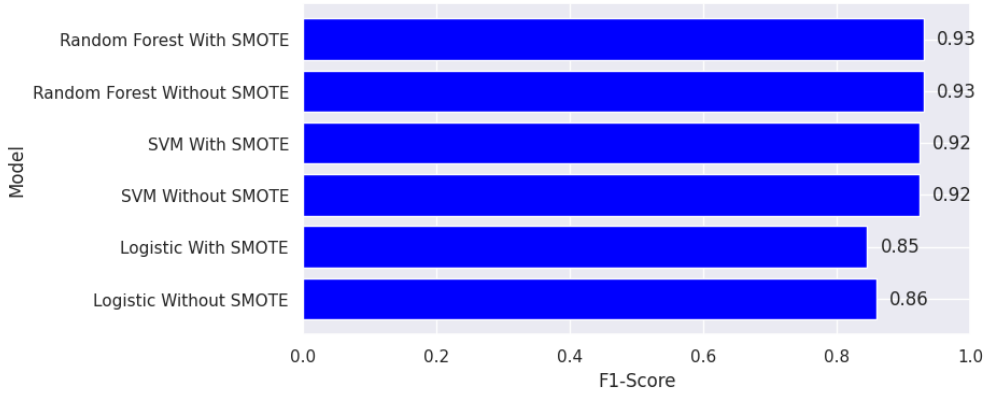

ภาพประกอบ 64 กราฟแท่งแสดงผลลัพธ์ค่า Accuracy ของแบบจำลองทั้งหมด




ภาพประกอบ 65 กราฟแท่งแสดงผลลัพธ์ค่า Precision ของแบบจำลองทั้งหมด




ภาพประกอบ 66 กราฟแท่งแสดงผลลัพธ์ค่า Recall ของแบบจำลองทั้งหมด




ภาพประกอบ 67 กราฟแท่งแสดงผลลัพธ์ค่า F1-Score ของแบบจำลองทั้งหมด

## 5.2 อภิปรายผลการวิจัย

งานวิจัยนี้ศึกษาการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง ในการวิจัยนี้ศึกษาการทำนายแนวโน้มการเลิกเป็นลูกค้าด้วยข้อมูลประชากรของลูกค้าบนเว็บไซต์แห่งหนึ่งที่นำมาจากแหล่งข้อมูลสาธารณะ Kaggle.com ซึ่งในงานวิจัยนี้เป็นปัญหาการจำแนกประเภทหรือ Classification ผู้วิจัยจึงเลือกแบบจำลองที่เป็นการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) โดยใช้แบบจำลองทั้งเชิงเส้นและไม่เชิงเส้นเพื่อนำมาเปรียบเทียบประสิทธิภาพทั้งหมด 3 แบบจำลองคือ Logistic Regression, Support Vector Machines (SVM) และ Random Forest โดยที่แบบจำลองเหล่านี้มีการคัดเลือกคุณลักษณะและมีการใช้วิธีการปรับความสมดุลของข้อมูลด้วยเทคนิค SMOTE อีกด้วย จากผลการทดลองพบว่าแบบจำลอง Random Forest เป็นแบบจำลองที่มีประสิทธิภาพที่ดีที่สุดร่วมกับจากการวิเคราะห์ Confusion Matrix ซึ่งพบว่าแบบจำลองทำนายกลุ่มถูกต้องมากที่สุดและทำนายผิดน้อยที่สุด อีกทั้งยังใช้เวลาในการทำงานไม่มาก แสดงให้เห็นว่าแบบจำลองนี้สามารถจำแนกกลุ่มได้อย่างแม่นยำและมีประสิทธิภาพสูงตอบสนองวัตถุประสงค์ของงาน

ในส่วนของงานวิจัยนี้ใช้ข้อมูลของลูกค้าบนเว็บไซต์แห่งหนึ่งจากแหล่งข้อมูลสาธารณะ โดยผู้วิจัยมีวิธีการทำความสะอาดข้อมูลในรูปแบบที่แตกต่างออกไป เช่นการแทนที่ค่าว่างด้วยค่า 0 และการแทนที่ข้อมูลแปลกด้วยคำว่า Unknown หลังจากนั้นมีการคัดเลือกคุณลักษณะด้วยวิธี Random Forest และสร้างชุดข้อมูล SMOTE เพื่อนำมาเปรียบเทียบกับแบบจำลองต่าง ๆ ให้มีความหลากหลายวิธี ต่อจากนั้นดูฟีเจอร์ที่ส่งผลต่อการทำนายจำแนกประเภทของกลุ่มลูกค้า ซึ่งดูฟีเจอร์ด้วย Feature Importance กับ LIME ว่ามีความสอดคล้องกันหรือไม่ เพื่อที่เป็นการให้แน่ใจ

ว่าเป็นฟีเจอร์ที่สำคัญหรือมีความจำเป็นต่อการนำไปวิเคราะห์หาข้อมูลเชิงลึกที่นำไปใช้ในการอธิบายที่จะสามารถช่วยให้เราตัดสินใจได้

ผู้วิจัยได้ทำแสดงคุณลักษณะที่สำคัญ (Feature Importance) ของแต่ละแบบจำลองเพื่อให้เข้าใจการตัดสินใจใช้คุณลักษณะในการทำนายของแบบจำลองในการเรียนรู้ พบว่าแบบจำลอง Logistic Regression ให้ค่าความสำคัญกับฟีเจอร์ membership_category_No Membership มากที่สุด และแบบจำลอง Logistic Regression แบบใช้ SMOTE พบว่าให้ค่าความสำคัญกับฟีเจอร์ membership_category_Basic Membership มากที่สุด ส่วนสำหรับแบบจำลอง SVM พบว่าหลังจากการปรับจูนพารามิเตอร์ ได้ Kernel Rbf ทำให้ไม่สามารถแสดง Feature Importance ออกมาได้ สำหรับ Random Forest ฟีเจอร์ที่มีอิทธิพลมากที่สุดในการแบ่งกลุ่ม คือ points_in_wallet ซึ่งให้ค่าสูงที่สุดทั้งแบบไม่ใช้และใช้ SMOTE

นอกจากนี้ผู้วิจัยยังใช้เครื่องมือ LIME ที่สามารถช่วยอธิบายแบบจำลองว่ามีการทำงานหรือตัดสินใจเลือกใช้ฟีเจอร์นำไปทำนาย และแสดงค่าความน่าจะเป็นของแต่ละกลุ่ม โดยผู้วิจัยทำการสุ่มข้อมูลมา 1 ข้อมูล นำมาจากชุดทดสอบ พบว่าเป็นเลเบลกลุ่ม 1 หรือลูกค้าเลิกใช้บริการแล้ว (Churn) สำหรับแบบจำลอง Logistic Regression ทำนายว่าข้อมูลนี้อยู่ในกลุ่มลูกค้ายังใช้บริการอยู่ (Exist) จากนั้นตรวจดูการตัดสินใจเลือกใช้ฟีเจอร์ของแบบจำลองในการทำนายพบว่าแบบจำลองให้ความสำคัญกับ membership_category_No Membership มากที่สุด สำหรับแบบจำลอง Logistic Regression ใช้งาน SMOTE ก็ทำนายว่าเป็นกลุ่มลูกค้ายังใช้บริการอยู่อีกเช่นกัน และให้ความสำคัญกับฟีเจอร์ membership_category_Basic Membership มากที่สุด สำหรับแบบจำลอง SVM ทั้งแบบไม่ใช้และใช้งาน SMOTE นั้นทำนายว่าข้อมูลนี้เป็นกลุ่มลูกค้าเลิกใช้บริการแล้ว ในส่วนฟีเจอร์ที่แบบจำลองให้ความสำคัญมากที่สุดคือ feedback_Products always in Stock สำหรับแบบจำลอง Random Forest ทั้งไม่ใช้และใช้งาน SMOTE ซึ่งทำนายว่าข้อมูลนี้เป็นกลุ่มลูกค้าเลิกใช้บริการแล้ว ในส่วนฟีเจอร์ที่แบบจำลองให้ความสำคัญมากที่สุดคือ membership_category_Basic Membership

ต่อไปพิจารณาสังเกตดูว่าผลลัพธ์ของคุณลักษณะที่สำคัญ (Feature Importance) ของแบบจำลองและผลลัพธ์จากการใช้เครื่องมือ LIME มีความสอดคล้องกันหรือไม่ โดยผู้วิจัยได้เลือกแบบจำลองที่มีประสิทธิภาพดีที่สุดคือ แบบจำลอง Random Forest พบว่ากลุ่มลูกค้าเลิกใช้บริการแล้วหรือ Churn มีความสอดคล้องกันของการใช้วิธี LIME ซึ่งพบว่าฟีเจอร์ที่ใช้ในการแบ่งกลุ่มเป็นระดับสมาชิก ในส่วนกลุ่มลูกค้าที่ยังใช้บริการอยู่หรือ Exist ซึ่งก็พบว่ามีความสอดคล้องกันจากการใช้วิธี LIME และพบว่าฟีเจอร์ที่ใช้ในการแบ่งกลุ่มเป็นระดับสมาชิกอีกเช่นกัน

แต่มีการเปลี่ยนแปลงลำดับความสำคัญของคุณลักษณะกันเล็กน้อย ผู้วิจัยมองว่าการใช้ฟีเจอร์เหล่านี้ในการจัดกลุ่มนั้นสมเหตุสมผลซึ่งระดับของลูกค้านั้นสามารถแบ่งกลุ่มลูกค้าได้ชัดเจนจากการสำรวจข้อมูลในบทที่ 3

การวิจัยในครั้งนี้แสดงให้เห็นว่าการสำรวจข้อมูลแต่ละลักษณะของแต่ละกลุ่มนั้น ช่วยให้เข้าใจกลุ่มลูกค้าได้ดีขึ้น โดยพบว่ากลุ่มลูกค้าเลิกใช้บริการแล้วมีความชัดเจนในระดับการเป็นสมาชิกของลูกค้าหรือ membership_category ที่ได้สำรวจ พบว่าลูกค้าที่เลิกใช้บริการนั้นเป็นลูกค้าที่ระดับไม่สูงและไม่ได้เป็นสมาชิกด้วย อีกทั้ง points_in_wallet หรือคะแนนสะสมส่วนใหญ่ของลูกค้ากลุ่มนี้ก็ยังมีคะแนนสะสมน้อยกว่ากลุ่มลูกค้าที่ยังใช้บริการอยู่ และการพิจารณาคุณลักษณะอื่นร่วมด้วยเช่น feedback หรือการแสดงความคิดเห็นของลูกค้า พบว่าลูกค้าเลิกใช้บริการนั้นยังแสดงความคิดเห็นของสินค้าและเว็บไซต์ไปในแง่ลบอีกด้วย

ดังนั้นในการใช้เครื่องมือ LIME มาช่วยในการอธิบายจึงช่วยทำให้นักการตลาดสามารถนำไปตัดสินใจและหาแนวทางวิธีต่าง ๆ เพื่อรักษาฐานลูกค้าไว้ให้ได้ ดังนั้นนักการตลาดสามารถใช้แนวทางนี้เพื่อวางแผนล่วงหน้าและป้องกันไม่ให้ลูกค้าเลิกใช้บริการได้ ตัวอย่างเช่น สามารถใช้ข้อมูลนี้เพื่อติดต่อลูกค้าแต่ละระดับสมาชิกหรือที่ไม่ได้สมัครสมาชิกแต่ลงทะเบียนเข้าสู่เว็บไซต์ที่ใช้บริการ เพื่อเสนอโปรโมชั่นหรือส่วนลดพิเศษ เพื่อให้ลูกค้ามียอดการซื้อที่มากขึ้นเพื่อเป็นการกระตุ้นให้สะสมคะแนนเป็นทางอ้อม หรือแก้ไขปัญหาที่ลูกค้ากำลังประสบอยู่ และยังสามารถใช้ข้อมูลนี้เพื่อปรับปรุงผลิตภัณฑ์และบริการให้ตอบสนองความต้องการของลูกค้าได้มากขึ้น เพื่อกระตุ้นลูกค้าให้ใช้บริการตลอดไม่หายไปไหน

### 5.3 ข้อเสนอแนะ

1. ผู้ที่สนใจศึกษาต่อผลการวิจัยนี้ อาจลองปรับเปลี่ยนวิธีการเตรียมความพร้อมของข้อมูลให้เหมาะสมกับงานวิจัยของตนเอง โดยพิจารณาจากปัจจัยต่าง ๆ เช่น ลักษณะของข้อมูล วัตถุประสงค์ของการสร้างแบบจำลอง เป็นต้น เช่น เปลี่ยนวิธีลบค่าว่าง การแทนที่ข้อมูล หรือปรับเปลี่ยนข้อมูลที่เป็นตัวเลข อาจเพิ่มประสิทธิภาพของแบบจำลองได้

2. ในการวิจัยนี้ ได้มีการใช้แบบจำลองการเรียนรู้ของเครื่องแบบมีผู้สอน 3 แบบ ได้แก่ Logistic Regression, SVM และ Random Forest ผู้ที่อยากศึกษาเพิ่มเติม ตัวอย่างเช่น อาจลองใช้แบบจำลอง XGboost และ Catboost หรือใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนเผื่อได้แบบจำลองที่มีประสิทธิภาพในการทำนายที่แม่นยำมากยิ่งขึ้น

3. ในการวิจัยนี้ ได้มีการใช้ค่าประสิทธิภาพของแบบจำลอง ได้แก่ Accuracy, Precision, Recall, F1-Score และ Confusion Matrix ผู้สนใจที่อยากนำไปศึกษาต่อ ตัวอย่างเช่น

อาจลองใช้วิธีอื่น ๆ ในการประเมินประสิทธิภาพของแบบจำลอง เช่น ROC Curve และ Precision-Recall Curve เพื่อเพิ่มแนวทางในการตัดสินใจเลือกแบบจำลองที่ดี

# บรรณานุกรม


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

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

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

Hu, X., Yang, Y., Chen, L., & Zhu, S. (2020). Research on a Customer Churn Combination

Prediction Model Based on Decision Tree and Neural Network. 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 129–132. Retrieved from https://doi.org/10.1109/ICCCBDA49378.2020.9095611

Mahesh, B. (2018). Machine Learning Algorithms—A Review. 9(1). Retrieved from https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

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

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

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

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

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

Stehani, S., Karunya, N., Ranjan, D. R. J. B., Sumathipala, S., & Sandanayake, T. C. (2020). Customer Churn Reasoning in Telecommunication Domain. 2020 International

Conference on Image Processing and Robotics (ICIP), 1–5. Retrieved from https://doi.org/10.1109/ICIP48927.2020.9367342

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

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

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

สิรภัทร เกาฏีระ. (ม.ป.ป.). ถึงเวลาแล้วหรือไม่? ที่คำว่า Shopping จะเป็นเพียงอดีต...เมื่อพฤติกรรม

ผู้บริโภคไม่เหมือนเดิม. สืบค้นจาก https://www.krungsri.com/th/wealth/krungsri-prime/privileges/articles/shopping-will-be-past-consumer-behavior-not-same

# ประวัติผู้เขียน