การจัดกลุ่มลูกค้าบริษัทยานยนต์ด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

AUTOMOBILE CUSTOMER SEGMENTATION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES

กาญจนมาส เปลี่ยนสกุล

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2564

การจัดกลุ่มลูกค้าบริษัทยานยนต์ด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

กาญจนมาส เปลี่ยนสกุล

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2564

AUTOMOBILE CUSTOMER SEGMENTATION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES

KANJANAMAS PLIENSAKUL

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

2021

สารนิพนธ์

เรื่อง

การจัดกลุ่มลูกค้าบริษัทยานยนต์ด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง

ของ

กาญจนมาส เปลี่ยนสกุล

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา)

..... ประธาน

(อาจารย์ ดร.ดวงดาว วิชาดากุล)

..... กรรมการ

(อาจารย์ ดร.ศุภร คนธภักดี)

ชื่อเรื่อง	การจัดกลุ่มลูกค้าบริษัทยานยนต์ด้วยข้อมูลประชากรโดยใช้เทคนิคการเรียนรู้ของเครื่อง
ผู้วิจัย	กาญจนมาส เปลี่ยนสกุล
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2564
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. นุวีย์ วิวัฒนวัฒนา

การจัดกลุ่มลูกค้าหรือ Customer Segmentation ถือเป็นกลยุทธ์สำคัญสำหรับการขับเคลื่อนธุรกิจ โดยเฉพาะกับธุรกิจที่มีการแข่งขันสูง เนื่องจากลูกค้าแต่ละคนมีความแตกต่างกัน นักการตลาดสามารถเข้าถึงและเข้าใจลูกค้าได้มากขึ้นผ่านการจัดกลุ่มลูกค้าที่มีพฤติกรรมหรือลักษณะบางอย่างคล้ายกัน ทำให้สามารถตอบสนองความต้องการของลูกค้ากลุ่มนั้น ๆ ได้ ในงานวิจัยนี้ได้เลือกใช้ข้อมูลประชากรสาธารณะของลูกค้าภายในบริษัทยานยนต์แห่งหนึ่ง เพื่อนำมาจัดกลุ่มทั้งหมด 4 กลุ่ม โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) ซึ่งแบบจำลองที่เลือกใช้คือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest และ Extreme Gradient Boosting (XGBoost) ร่วมกับการแก้ปัญหาข้อมูลไม่สมดุลด้วยวิธีการ SMOTE จากนั้นวัดประสิทธิภาพการทำงานด้วยค่า Accuracy, Precision, Recall, F1-Score และสังเกตความถูกต้องและความผิดพลาดที่เกิดขึ้นด้วย Confusion Matrix จากผลการทดลองพบว่าแบบจำลอง Random Forest ร่วมกับการใช้ SMOTE ให้ประสิทธิภาพที่ดีที่สุดที่ค่า Accuracy 48.75% Precision 48.10% Recall 48.75% และ F1-Score ที่ 48.31% สำหรับแบบจำลองที่ใช้เวลาในการเรียนรู้น้อยที่สุดคือ Naïve Bayes นอกจากนี้ยังมีการตรวจสอบความสำคัญของคุณลักษณะและตีความแบบจำลองด้วยเทคนิค LIME และ SHAP เพื่อเพิ่มความน่าเชื่อถือให้กับแบบจำลอง

คำสำคัญ : การจัดกลุ่มลูกค้า, เทคนิคการเรียนรู้ของเครื่อง, ข้อมูลประชากร

Title	AUTOMOBILE CUSTOMER SEGMENTATION USING DEMOGRAPHIC DATA BASED ON MACHINE LEARNING TECHNIQUES
Author	KANJANAMAS PLIENSAKUL
Degree	MASTER OF SCIENCE
Academic Year	2021
Thesis Advisor	Assistant Professor Nuwee Wiwatwattana , Ph.D.

Due to the fact that each consumer is unique, customer segmentation is an important strategy for organizations, especially those with a high level of competition. The marketing team can reach customers with similar behavior or characteristics through customer segmentation, allowing teams to address the customer demands. In this study, supervised machine learning techniques were used to divide a publicly available dataset from an automobile manufacturer into four categories. Based on the demographic data, the classification techniques consisted of Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost), along with the improvement of imbalanced data using SMOTE. The model with the greatest score, according to the test data, was Random Forest utilizing SMOTE, with 48.75% accuracy, 48.10% precision, 48.75% recall and an F1-Score of 48.31%. Naive Bayes required the least amount of time to learn the data. In addition, the features of highlighted importance and interpreted models used LIME and SHAP to improve model reliability.

Keyword : Customer segmentation, Machine learning techniques, Demographic data

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.นุวีย์ วิวัฒนวัฒนา อาจารย์ที่ปรึกษา ที่ให้คำปรึกษาตั้งแต่เริ่มต้นจนเสร็จสมบูรณ์และช่วยตรวจสอบความถูกต้องในด้านข้อมูลทางวิชาการ รวมถึงตรวจสอบความเรียบร้อย ความสวยงามของการใช้คำในสารนิพนธ์ในทุกขั้นตอน

ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์และอาจารย์ทุกท่าน ที่ให้ความรู้และคำแนะนำที่เป็นประโยชน์ในการปรับปรุงสารนิพนธ์ให้ดียิ่งขึ้น

ขอกราบขอบพระคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ สำหรับทุนสนับสนุนการนำเสนอผลงานวิจัยของนิสิตบัณฑิตศึกษาในงานประชุมวิชาการ ทำให้ได้รับประสบการณ์ที่ดีในการเผยแพร่และแลกเปลี่ยนความรู้กับผู้นำเสนอท่านอื่น ๆ

สุดท้ายนี้ขอขอบพระคุณครอบครัวของผู้วิจัยที่ให้โอกาสในการศึกษาและเป็นกำลังใจให้จนสำเร็จการศึกษา รวมถึงขอบคุณพี่ ๆ ในสาขาวิชาที่คอยให้ความช่วยเหลือและคอยให้คำแนะนำทั้งในช่วงเวลาเรียนและช่วงเวลาในการทำรูปเล่มสารนิพนธ์

กาญจนมาส เปลี่ยนสกุล

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและความเป็นมาของวิจัย

ในทุก ๆ ธุรกิจ สิ่งที่สำคัญต่อการขับเคลื่อนธุรกิจคือลูกค้า แต่นอกจากการหาลูกค้าใหม่แล้ว วิธีที่เป็นที่นิยมและให้ประสิทธิภาพที่ดีคือการรักษาลูกค้าเก่าที่ทำรายได้ให้กับธุรกิจ การมองหาลูกค้าใหม่ค่อนข้างทำได้ยาก เนื่องจากมีธุรกิจหลากหลายบนตลาด การแข่งขันสูง การโน้มน้าวลูกค้าใหม่จำเป็นต้องใช้ค่าใช้จ่ายจำนวนมาก และจากการสำรวจของ Globalwebindex (Globalwebindex, 2020) พบว่าลูกค้าในปัจจุบันเข้าถึงได้ยากขึ้น 48% โดยการใช้ Ad-blocker และลูกค้าค่อนข้างกังวลเกี่ยวกับความปลอดภัยและความเป็นส่วนตัว ทำให้ธุรกิจไม่สามารถเข้าถึงลูกค้าได้ทุกคน การจัดกลุ่มลูกค้าจึงเป็นวิธีการที่เหมาะสมในการรักษาลูกค้าเก่า ทำให้เราสามารถรับรู้ถึงลักษณะและพฤติกรรมต่าง ๆ ของลูกค้า ทราบถึงความแตกต่างของลูกค้าแต่ละกลุ่ม เพื่อที่จะสามารถใช้แผนการดำเนินงานทางธุรกิจต่าง ๆ ที่เหมาะสมกับลูกค้าได้อย่างคุ้มค่าใช้จ่ายมากที่สุด

การจัดกลุ่มลูกค้าหรือ Customer Segmentation ถือเป็นกลยุทธ์สำคัญสำหรับธุรกิจในสมัยนี้ โดยเฉพาะกับธุรกิจที่มีการแข่งขันสูง เนื่องจากลูกค้าแต่ละคนมีความแตกต่างกัน การจัดกลุ่มลูกค้าเป็นวิธีที่กำหนดกลุ่มเป้าหมายที่ชัดเจน ทำให้นักการตลาดสามารถเข้าถึงและเข้าใจลูกค้าได้มากขึ้นผ่านการจัดกลุ่มลูกค้าที่มีพฤติกรรมหรือลักษณะบางอย่างคล้ายกัน ทำให้สามารถใช้กลยุทธ์ต่าง ๆ ที่เหมาะสมกับแต่ละกลุ่มและสามารถตอบสนองความต้องการของลูกค้าและเพิ่มประสิทธิภาพในการสื่อสารกับลูกค้าได้ สามารถประเมินจำนวนลูกค้าที่มีความสนใจต่อสินค้าและบริการนั้น ๆ ทำให้ธุรกิจสามารถจัดหาสินค้าและบริการให้เพียงพอต่อความต้องการของลูกค้าและยังส่งผลให้ธุรกิจสามารถประเมินยอดขายและกำไรที่อาจเกิดขึ้นได้อีกด้วย แต่หากไม่ทำการจัดกลุ่มและศึกษาลูกค้า การโฆษณาหรือนำเสนอสินค้าและบริการที่ได้รับความนิยมสูงสุดให้แก่ลูกค้าทุกคนเหมือนว่าเป็นวิธีที่ไม่มีประสิทธิภาพ เนื่องจากลูกค้าแต่ละคนมีความชอบความต้องการที่แตกต่างกัน การนำเสนอแต่เพียงสินค้าและบริการที่มีความนิยมสูงสุดโดยรวมนั้นไม่สามารถตอบสนองความต้องการของลูกค้าที่แตกต่างกันได้ อาจยิ่งทำให้เกิดผลเสียต่อธุรกิจ

การจัดกลุ่มลูกค้ามีหลายประเภทขึ้นอยู่กับข้อมูลที่ใช้ อาจใช้ข้อมูลประชากรหรือว่า Demographic Data ในการจัดกลุ่ม เช่น อายุ เพศ อาชีพ รายได้ หรือใช้ข้อมูลทางภูมิศาสตร์ (Geographical Data) ในการจัดกลุ่ม เช่น พิกัดทางภูมิศาสตร์ละติจูด ลองจิจูด พื้นที่ เขต เมือง ประเทศ หรือใช้ข้อมูลด้านจิตวิทยาของลูกค้า (Psychographic Data) ในการจัดกลุ่ม เช่น ความ

สนใจของลูกค้า ไลฟ์สไตล์ บุคลิกภาพ แต่นับว่าเป็นข้อมูลที่หาได้ค่อนข้างยาก หรือใช้ข้อมูลพฤติกรรมของลูกค้า (Behavioral Data) ในการจัดกลุ่ม เช่น การรับบริการเสริมต่าง ๆ จากเครือข่ายโทรศัพท์มือถือ วิธีการซื้อขายบนช่องทางออนไลน์ การใช้ข้อมูลแต่ละประเภทมีข้อดีข้อเสียที่แตกต่างกันไป เช่น หากใช้ข้อมูลด้านภูมิศาสตร์ของลูกค้าทำให้สามารถแบ่งกลุ่มลูกค้าตามพื้นที่อยู่อาศัยและเสนอสินค้าหรือบริการที่เหมาะสมกับที่อยู่อาศัยของลูกค้าได้ แต่ข้อมูลทางภูมิศาสตร์เพียงอย่างเดียวไม่สามารถแสดงถึงความสนใจของลูกค้าได้ หรือหากใช้ข้อมูลด้านพฤติกรรมของลูกค้าสามารถวิเคราะห์ได้ง่าย เนื่องจากเป็นข้อมูลเกี่ยวกับพฤติกรรมการใช้จ่ายของลูกค้า ทำให้ธุรกิจสามารถนำข้อมูลมาวิเคราะห์และเพิ่มยอดขายและกำไรได้ง่ายขึ้นแต่พฤติกรรมลูกค้าสามารถเปลี่ยนแปลงได้ตลอดเวลา ซึ่งข้อมูลทั้ง 4 รูปแบบไม่ได้มีรูปแบบใดที่ดีที่สุด ถึงแม้ดูเหมือนว่าการใช้ข้อมูลพฤติกรรมของลูกค้าจะเข้าใจลูกค้าได้มากกว่า แต่ในความเป็นจริงแต่ละประเภทมีความสำคัญที่แตกต่างกันออกไป เช่น ตัวอย่างจาก ThinkwithGoogle (Gevelber, 2015) เขียนไว้ว่าคนที่ค้นหาอุปกรณ์กีฬา 56% เป็นผู้หญิง คนที่ค้นหาสินค้าประเภทผลิตภัณฑ์บำรุงผิวส่วนใหญ่เป็นผู้ชาย ซึ่งการที่เรารู้สิ่งเหล่านี้ได้ เราจำเป็นต้องมีข้อมูลทั้งประเภทประชากรและข้อมูลพฤติกรรมของลูกค้า อนึ่งชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้คือข้อมูลลูกค้าของบริษัทยานยนต์แห่งหนึ่ง เป็นรูปแบบข้อมูลประเภทประชากร (Demographic Data) ที่ได้รับความนิยมมากที่สุด ซึ่งข้อดีของการใช้ข้อมูลประชากรคือสามารถกำหนดกลุ่มของลูกค้าได้ง่าย เป็นข้อมูลที่เกี่ยวกับลูกค้าโดยตรงและเปลี่ยนแปลงได้ยาก แสดงให้เห็นความแตกต่างของลูกค้าได้ชัดเจน

โดยปกติแล้ว การจัดกลุ่มลูกค้าใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนเนื่องจากข้อมูลที่มีในความเป็นจริงนั้นไม่มีเลเบล ทำให้เราไม่สามารถรู้ได้ว่าลูกค้าคนใดอยู่กลุ่มใด แต่ในงานวิจัยนี้เราใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอนในการจัดกลุ่มและทำนายกลุ่มของลูกค้าใหม่ของบริษัทยานยนต์ เพื่อเพิ่มประสิทธิภาพในการทำงานให้แก่ทีมการตลาด ในการเสนอสินค้าที่คาดว่าเหมาะสมและสร้างความพึงพอใจแก่ลูกค้า โดยใช้ข้อมูลประชากรของลูกค้าในการจัดกลุ่ม เช่น หากลูกค้าจัดอยู่ในกลุ่มประชากรชนชั้นกลาง อาจนำเสนอเป็นรถยนต์ขนาดเล็ก ในขณะที่หากลูกค้าที่มีรายได้ค่อนข้างสูง อาจทำการแนะนำรถยนต์ที่มีราคาแพงขึ้น หรือหากพิจารณาจากอายุ สำหรับสินค้าที่มีสีสันอาจถูกแนะนำแก่ลูกค้าในกลุ่มที่มีอายุน้อยกว่า และเนื่องจากข้อมูลที่ใช้เป็นข้อมูลที่เก็บในรูปแบบตารางที่มีเลเบล หากใช้เทคนิคแบบมีผู้สอนทำให้สามารถวัดผลประสิทธิภาพการทำงานของแบบจำลองได้แม่นยำมากกว่าเทคนิคการเรียนรู้แบบไม่มีผู้สอน และเพื่อเพิ่มความน่าเชื่อถือให้กับแบบจำลอง ผู้วิจัยได้เลือกใช้การอธิบายแบบจำลองด้วยค่า Feature Importance หรือค่าความสำคัญของฟีเจอร์ที่แบบจำลองใช้ในการเรียนรู้ รวมถึงมีการตีความการ

ทำงานของแบบจำลองด้วยเครื่องมือที่ชื่อว่า Local Interpretable Model-agnostic Explanations หรือ LIME และ Shapley Additive Explanations หรือ SHAP เพื่อตรวจสอบการใช้งานฟีเจอร์บนข้อมูล 1 ข้อมูล โดยการแสดงค่าความสำคัญของฟีเจอร์และการอธิบายแบบจำลองช่วยในการแสดงพฤติกรรมการทำงานของแบบจำลองที่มองไม่เห็นให้ผู้อ่านหรือแม้แต่ตัวผู้วิจัยเองสามารถอ่านและเข้าใจได้ง่ายขึ้น

1.2 จุดประสงค์ของงานวิจัย

1. เพื่อจัดกลุ่มลูกค้าโดยใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน สังเกตความเหมือนกันของลูกค้าแต่ละกลุ่ม เพื่อที่สามารถใช้กลยุทธ์ทางการตลาดต่างๆให้เหมาะสมกับแต่ละกลุ่ม
2. เพื่อศึกษาและเปรียบเทียบการทำงานและประสิทธิภาพของแต่ละอัลกอริทึมบนข้อมูลชุดเดียวกัน โดยแบบจำลองที่เลือกใช้ในการเปรียบเทียบคือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, และ Extreme Gradient Boosting (XGBoost)
3. เพื่อวิเคราะห์ความสำคัญของคุณลักษณะและการตีความแบบจำลองที่ได้ด้วยการอธิบายแบบจำลอง

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้ใช้ข้อมูลประชากรของลูกค้าภายในบริษัทยานยนต์แห่งหนึ่งจากแหล่งข้อมูลสาธารณะ Kaggle.com ซึ่งประกอบด้วย 11 ตัวแปรรวมเลเบลกลุ่มลูกค้าและข้อมูลลูกค้าทั้งหมด 10,695 คน โดยเป้าหมายคือต้องการจัดกลุ่มลูกค้าออกเป็น 4 กลุ่มคือ A, B, C และ D เพื่อใช้กลยุทธ์ทางการตลาดให้เหมาะสมและตอบสนองความต้องการของลูกค้าแต่ละกลุ่มได้สูงสุด โดยแบบจำลองที่เลือกใช้คือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, และ Extreme Gradient Boosting (XGBoost) เพื่อนำมาเปรียบเทียบประสิทธิภาพการทำงาน ซึ่งทั้ง 5 ตัวเป็นแบบจำลองสำหรับงานจัดหมวดหมู่ของข้อมูล จากนั้นทำการสำรวจความเหมือนและแตกต่างของลูกค้าแต่ละกลุ่มเพื่อหาข้อมูลเชิงลึกที่เป็นประโยชน์ และเนื่องจากงานวิจัยนี้เป็นการศึกษาปัญหาด้านการจัดหมวดหมู่ของข้อมูลหรือ Classification จึงใช้การวัดประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score

1. ตัวแปรที่ศึกษา

- ID หรือรหัสลูกค้า
- Gender หรือเพศของลูกค้า (Male/Female)
- Ever_Married หรือสถานภาพสมรสของลูกค้า (Yes/No)
- Age หรืออายุของลูกค้า (ปี)
- Graduated หรือการจบการศึกษาของลูกค้า (Yes/No)
- Profession หรืออาชีพของลูกค้า
- Work_Experience หรือประสบการณ์ทำงานของลูกค้า (ปี)
- Spending_Score หรือคะแนนการใช้จ่ายของลูกค้า (High/Average/Low)
- Family_Size หรือจำนวนสมาชิกในครอบครัวโดยนับรวมลูกค้า
- Var_1 หรือประเภทของลูกค้าแบบไม่ระบุชื่อ
- Segmentation หรือกลุ่มของลูกค้า (เลเบลหรือคลาส)

1.4 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถจัดกลุ่มและทำนายกลุ่มของลูกค้าใหม่ได้อย่างมีประสิทธิภาพ เพื่อที่สามารถนำเสนอสินค้าและบริการได้ตรงต่อความสนใจของลูกค้า

2. สามารถระบุกลุ่มลูกค้าได้ชัดเจนเพื่อลดต้นทุนในการทำการตลาดและจัดสรรสินค้าและบริการให้เพียงพอต่อความต้องการของลูกค้า

3. สามารถนำแบบจำลองไปใช้จัดกลุ่มลูกค้าในความดูแลขององค์กรอื่น ๆ ที่มีการเก็บข้อมูลประชากรได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

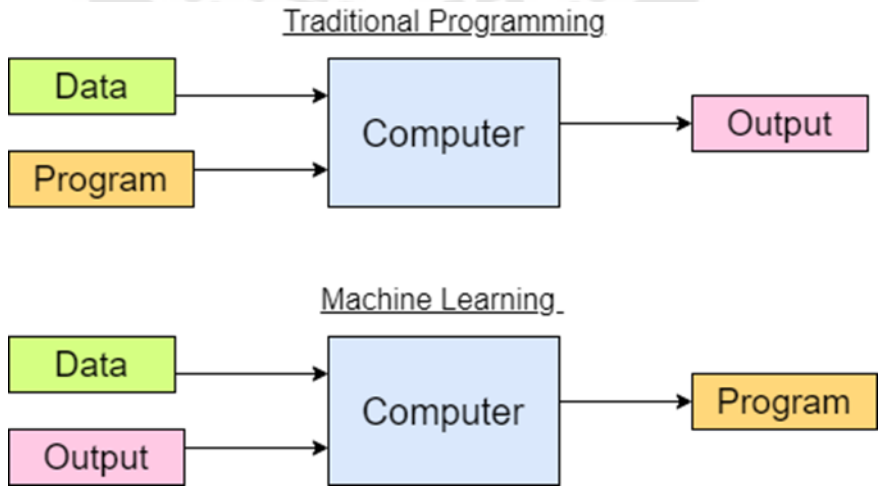
ในบทนี้ ผู้วิจัยได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มลูกค้าโดยใช้เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน เพื่อพัฒนาแบบจำลองให้มีประสิทธิภาพสูงสุด โดยมีหัวข้อดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

เทคนิคการเรียนรู้ของเครื่องหรือ Machine Learning คือการให้คอมพิวเตอร์เรียนรู้ด้วยข้อมูลและตัวอย่างต่างๆเพื่อสร้างแบบจำลองออกมาใช้ในการทำนายข้อมูล โดยปกติของการเขียนโปรแกรมทั่วไปจะใส่ข้อมูลและโปรแกรมให้คอมพิวเตอร์เรียนรู้ สิ่งที่ได้ออกมาจากการเรียนรู้คือผลลัพธ์หรือ Output แต่ใน Machine Learning จะใส่ข้อมูลและผลลัพธ์เข้าไปในคอมพิวเตอร์ เมื่อคอมพิวเตอร์เรียนรู้จะสามารถสร้างโมเดลออกมาได้และนำไปใช้ต่อในการทำนายข้อมูลที่คอมพิวเตอร์ไม่เคยเรียนรู้ โดยการทำงานของ Machine Learning สามารถอธิบายได้ดังภาพประกอบ 1



ภาพประกอบ 1 แสดงความแตกต่างระหว่างกระบวนการเขียนโปรแกรมทั่วไปและการเรียนรู้แบบ Machine Learning

ซึ่งข้อมูลที่ Machine Learning สามารถเรียนรู้ได้สามารถแบ่งเป็น 2 รูปแบบ คือ Labelled Data หรือข้อมูลที่มีคำตอบสมบูรณ์ เช่น มีภาพเอ็กซเรย์ของคนไข้และมีคำตอบว่าภาพนี้คือคนไข้ที่ติดเชื้อโควิด หรือมีอีเมล 1 ฉบับและมีคำตอบไว้ว่าอีเมลฉบับนี้เป็นสแปม ซึ่งคำตอบที่ข้อมูลมีมาจากผู้เชี่ยวชาญเกี่ยวกับชุดข้อมูลนั้น ๆ ที่ได้ทำการเขียนกำกับไว้ อีกรูปแบบหนึ่งคือ Unlabelled Data หรือข้อมูลที่ไม่มีคำตอบสมบูรณ์ เช่น มีภาพเอ็กซเรย์ของคนไข้แต่ไม่มีคำตอบไว้ว่าภาพเอ็กซเรย์ภาพใดคือคนไข้ที่ติดเชื้อหรือไม่ติดเชื้อ ซึ่งข้อมูลในความเป็นจริงมักเป็นแบบ Unlabelled Data เนื่องจากการให้ผู้เชี่ยวชาญกำกับคำตอบของข้อมูลนั้นใช้ค่าใช้จ่ายที่ค่อนข้างสูงและเวลาที่ค่อนข้างมาก

Machine Learning สามารถแบ่งออกเป็น 4 กลุ่ม โดยใช้วิธีการเรียนรู้ข้อมูลเป็นตัวแบ่งประเภทและตัวอย่างงานของแต่ละกลุ่มแสดงดังภาพประกอบ 2 ดังนี้

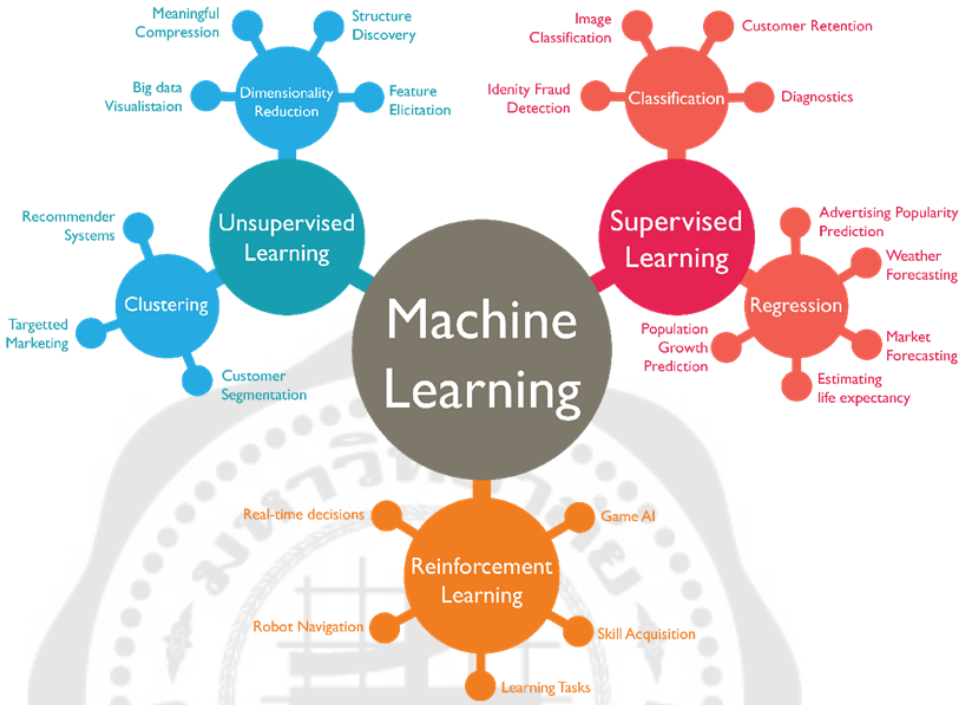
1. การเรียนรู้ของเครื่องแบบมีผู้สอนหรือ Supervised Machine Learning คือการที่คอมพิวเตอร์เรียนรู้โดยใช้ข้อมูลแบบ Labelled Data โดยการเรียนรู้แบบมีผู้สอนสามารถแบ่งเป็น 2 ปัญหาหลักคือปัญหา Regression คือผลลัพธ์ที่ได้จะเป็นตัวเลขที่มีความต่อเนื่องหรือจำนวนจริง เช่น การทำนายราคาบ้าน ผลลัพธ์ที่ได้คือราคาของบ้านซึ่งเป็นจำนวนจริง อีกหนึ่งปัญหาคือปัญหา Classification คือผลลัพธ์ที่ได้จะเป็นตัวเลขที่ไม่ต่อเนื่องกัน เช่น การทำนายลูกค้าที่มีแนวโน้มย้ายค่าย ผลลัพธ์ที่ได้คือย้ายค่ายหรือไม่ย้าย ซึ่งในงานวิจัยนี้ได้เลือกใช้การเรียนรู้แบบมีผู้สอนในการจัดการกับปัญหา Classification คือการทำนายและจัดกลุ่มลูกค้า

2. การเรียนรู้ของเครื่องแบบไม่มีผู้สอนหรือ Unsupervised Machine Learning คือการที่คอมพิวเตอร์เรียนรู้โดยใช้ข้อมูลแบบ Unlabelled Data โดยการเรียนรู้แบบไม่มีผู้สอนสามารถแบ่งเป็น 2 ปัญหาหลักคือปัญหา Clustering หรือการจัดกลุ่มโดยใช้ความเหมือนของข้อมูล เช่น การจัดกลุ่มลูกค้าที่มีลักษณะการใช้จ่ายคล้ายกัน อีกปัญหาคือปัญหา Association คือการหากฎความเชื่อมโยงของพฤติกรรมหรือสิ่งของ เช่น หากมีพฤติกรรม A มักจะมีพฤติกรรม B ด้วย

3. การเรียนรู้ของเครื่องแบบกึ่งมีผู้สอนหรือ Semi-Supervised Machine Learning คือการที่คอมพิวเตอร์เรียนรู้โดยใช้ข้อมูลทั้งแบบ Labelled และ Unlabelled Data ร่วมกัน สาเหตุที่ใช้ร่วมกันอาจเป็นเพราะมี Labelled Data ในจำนวนน้อยมาก

4. การเรียนรู้ของเครื่องแบบ Reinforcement Learning คือการที่คอมพิวเตอร์เรียนรู้จากสภาพแวดล้อมไปเรื่อยๆเพื่อให้ได้ผลลัพธ์ที่ดีสุดโดยไม่ใช้ทั้ง Labelled และ Unlabelled Data

ตัวอย่างที่มีชื่อเสียงคือหุ่นยนต์แข่งหมากล้อม Alpha Go ที่เกิดจากการเรียนรู้ด้วยตัวเองผ่านการจำลองการแข่งขันเป็นจำนวนแสนถึงล้านรอบ



ภาพประกอบ 2 แสดงการแบ่งประเภทภายใน Machine Learning และตัวอย่างงาน

ที่มา (Minaphinant, 2018)

ใน Machine Learning มีอัลกอริทึมมากมายที่เหมาะสมกับแต่ละปัญหา แต่เนื่องจากในงานวิจัยนี้สนใจปัญหาด้าน Classification จึงขออธิบายเฉพาะอัลกอริทึมที่ใช้ในงานนี้ นั่นคือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, และ Extreme Gradient Boosting (XGBoost)

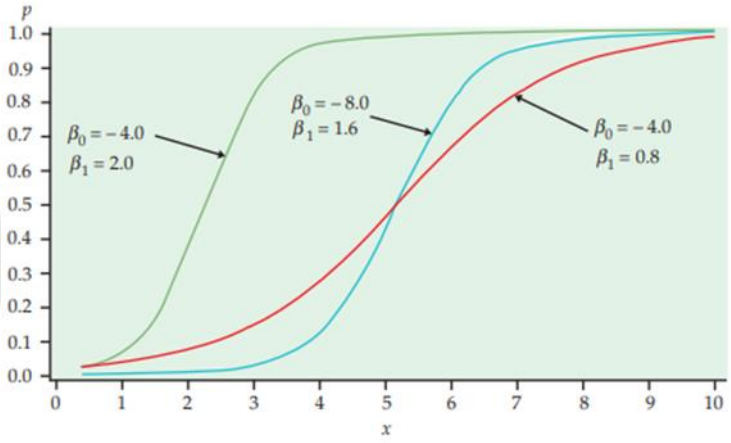
2.1.1 Logistic Regression

การวิเคราะห์การถดถอยแบบโลจิสติกหรือ Logistic Regression คืออัลกอริทึมที่ใช้จัดการกับปัญหา Classification โดยเป็นอัลกอริทึมเชิงเส้น (Linear Classification) ที่ปรับเปลี่ยนมาจากสมการเส้นตรงโดยใช้ Sigmoid Function หรือ Logistic Function มาแทนดังภาพประกอบ 3 โดยค่าที่ได้ออกมาคือค่าความน่าจะเป็น 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



ภาพประกอบ 3 แสดงกราฟ Logistic Regression

ที่มา (Moore, MacCabe, & Craig, 2017)

2.1.2 Naïve Bayes

Naïve Bayes คืออัลกอริทึมที่ใช้หลักการความน่าจะเป็นซึ่งตั้งอยู่บนทฤษฎีของ Bayes ถ้าหากเทียบกับการจัดการปัญหา Classification ด้วยอัลกอริทึมเชิงเส้น อัลกอริทึมเชิงเส้นจะมีการแปลงข้อมูลไปเป็นจุดไว้บน Feature Space จากนั้นทำการสร้าง Decision Boundary ขึ้นมาเพื่อแบ่งข้อมูลเป็นกลุ่ม ๆ ถือเป็นอัลกอริทึมทางเรขาคณิต แต่ Naïve Bayes ไม่มีการแปลงข้อมูลไปเป็นจุด แต่ใช้หลักการความน่าจะเป็น

ทฤษฎีของเบย์หรือ Bayes' Theorem คือการคำนวณความน่าจะเป็นแบบมีเงื่อนไข โดยใช้ค่าธรรมชาติของปราฏการณ์พิจารณาร่วมกับค่าธรรมชาติของแบบจำลอง เช่น ถ้าไฟดับแสดงว่าฝนอาจตกหรือเกิดปัญหาที่หม้อแปลง ความน่าจะเป็นที่ไฟดับเนื่องจากฝนตกคือ 0.2 และความน่าจะเป็นที่ไฟดับเนื่องจากเกิดปัญหาที่หม้อแปลงคือ 0.6 วันนี้ไฟดับ หากให้เราคาดเดาสาเหตุที่ทำให้ไฟดับ เราอาจตอบว่าเพราะเกิดปัญหาที่หม้อแปลงเนื่องจากความน่าจะเป็น

มากกว่า นั่นคือค่าธรรมชาติของแบบจำลอง แต่เราต้องพิจารณาร่วมกับค่าธรรมชาติของปรากฏการณ์ด้วย นั่นคือสมมติว่าโดยปกติแล้วความน่าจะเป็นที่ฝนตกมีค่า 0.8 แต่ความน่าจะเป็นที่เกิดปัญหากับหม้อแปลงมีเพียง 0.2 ดังนั้นเมื่อพิจารณาร่วมกันจะพบว่าสาเหตุที่ไฟดับอาจเป็นเพราะฝนตกมากกว่า โดยความน่าจะเป็นจากทฤษฎีของเบย์สามารถคำนวณได้ดังสมการ (2)

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (2)$$

โดยที่

x คือเหตุการณ์ที่เกิดขึ้นจากสาเหตุ y

y คือสาเหตุที่ทำให้เกิดเหตุการณ์ x

$P(y|x)$ คือความน่าจะเป็นของสาเหตุ y ในเหตุการณ์ x เรียกว่า Posterior Probability

$P(x|y)$ คือความน่าจะเป็นของเหตุการณ์ x จากสาเหตุ y เรียกว่า Likelihood Probability

$P(y)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ y เรียกว่า Priori Probability

$P(x)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ x เรียกว่า Marginal Probability

การใช้ทฤษฎีของเบย์ในปัญหา Classification จะมองว่า x คือฟีเจอร์ ซึ่งแต่ละฟีเจอร์สามารถมีความสัมพันธ์กันได้ เช่น ถ้าหากอายุมาก ประสบการณ์ทำงานอาจมากตามด้วย แต่การที่แต่ละฟีเจอร์มีความสัมพันธ์กันทำให้การคำนวณมีความยากมากขึ้น ดังนั้นจึงมีสมมติฐานว่าให้แต่ละฟีเจอร์เป็นอิสระต่อกัน ดังนั้นจะได้ดังสมการ (3)

$$\text{Class of Sample } x = \underset{i}{\operatorname{argmax}} \prod_k P(x_k|y_i)P(y_i) \quad (3)$$

โดยที่

x_k คือฟีเจอร์ k ฟีเจอร์

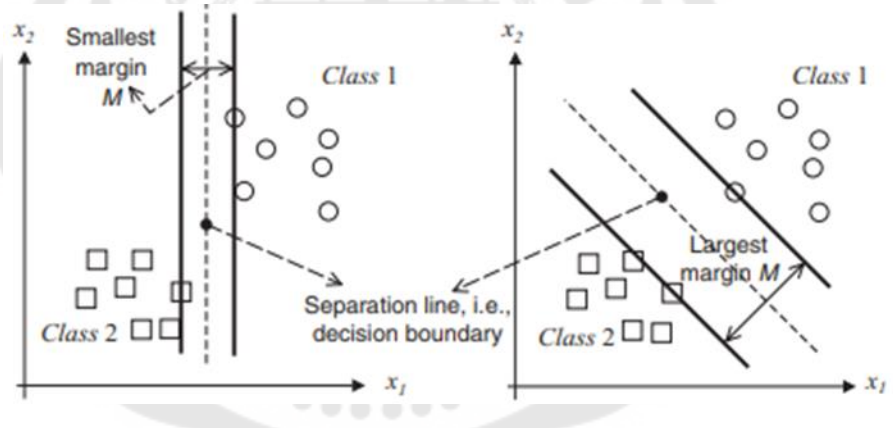
y_i คือคลาสหรือเลเบลที่ i

2.1.3 Support Vector Machine (SVM)

Support Vector Machine หรือ SVM ถือเป็นอัลกอริทึมเชิงเส้นที่มีประสิทธิภาพสูงในการจัดการกับปัญหา Classification เหมาะกับข้อมูลขนาดไม่ใหญ่มากประมาณหลักหมื่น โดย SVM จะแตกต่างจากอัลกอริทึมเชิงเส้นทั่วไปในเรื่องของการเลือกและสร้าง Decision Boundary

คือ SVM เลือก Decision Boundary ที่ให้ค่า Error น้อยที่สุดและ Maximal Margin ต่างกับอัลกอริทึมเชิงเส้นอื่นที่เลือก Decision Boundary โดยพิจารณาจากเส้นที่ให้ค่า Error น้อยที่สุดเท่านั้น นอกจากนี้การสร้าง Decision Boundary ของอัลกอริทึมเชิงเส้นอื่น ๆ ใช้ทุกจุดในการสร้าง ทำให้หากชุดข้อมูลนั้นมี Outlier จะถูกนำมาสร้าง Decision Boundary ด้วย ทำให้อาจเกิดความผิดพลาดในการแยกคลาสสูงกว่า SVM เนื่องจาก SVM ใช้เพียงบางจุดในการสร้าง Decision Boundary นั่นคือจุดที่เป็น Maximal Margin จุดที่ถูกนำมาสร้าง Decision Boundary เรียกว่า Support Vector ดังนั้น การนำจุดที่ Maximal Margin มาใช้ทำให้ Outlier ไม่มีบทบาทในการสร้าง Decision Boundary

Margin คือระยะห่างที่มากที่สุดของจุดที่ใกล้ Decision Boundary มากที่สุด การที่มี Maximal Margin จะช่วยลดความแปรปรวนหรือ Variance ของแบบจำลอง เพราะจุดอยู่ห่างจาก Decision Boundary ทำให้ไม่เกิดความสับสนในการตัดสินใจคลาสของจุดนั้น แสดงดังภาพประกอบ 4



ภาพประกอบ 4 แสดงให้เห็นถึง Maximal margin, Support Vector และ Decision Boundary

ที่มา (Kecman, 2005)

การพิจารณา Margin สามารถเขียนได้ดังสมการ (4) ดังนี้

$$\min_{d} d(x_n, w^T x + d) \tag{4}$$

โดยที่

x_n คือจุดข้อมูลใดๆ

$w^T x + d$ คือสมการเส้นตรงของ Decision Boundary

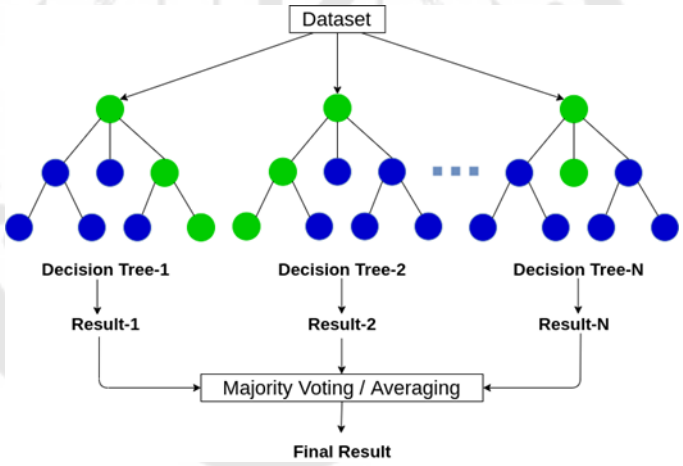
$d(x_n, w^T x + d)$ คือระยะทางระหว่างจุดใดๆถึง Decision Boundary

ดังนั้นจะได้ Maximal Margin ที่สามารถเขียนในรูปสมการดังนี้ (5)

$$\underset{w,d}{\operatorname{argmax}} \min_{n} d(x_n, w^T x + d) \quad (5)$$

2.1.4 Random Forest

Random Forest คืออัลกอริทึมที่ใช้ในการแก้ปัญหา Classification ที่สามารถอ่านเข้าใจได้ง่ายและมีประสิทธิภาพสูง โดย Random Forest สร้างมาจากต้นไม้ในการตัดสินใจ (Decision Tree) หลายต้น โดยแต่ละต้นได้รับข้อมูลที่แตกต่างกันในการเรียนรู้ ซึ่งข้อมูลที่แตกต่างกันนั้นได้มาจากการสุ่มจากชุดข้อมูลเดิม เพื่อเพิ่มจำนวนให้แต่ละต้นสามารถเรียนรู้ได้อย่างมีประสิทธิภาพ เมื่อแต่ละต้นได้ผลการทำนายออกมา จากนั้นใช้วิธีการ Majority Vote คือตัดสินใจจากเสียงส่วนมากเพื่อหาผลลัพธ์ที่เป็นไปได้มากที่สุด โดยการใช้ต้นไม้หลายต้นในการช่วยตัดสินใจสามารถช่วยในการลด Error และความแปรปรวนหรือ Variance ได้



ภาพประกอบ 5 แสดงการทำงานของ Random Forest

ที่มา (Ampadu, 2021)

การใช้ Random Forest อยู่ภายใต้สมมติฐานคือทุกต้นไม้หรือแบบจำลองต้องเป็นอิสระต่อกันจึงจะทำงานได้ดี ดังนั้นการทำให้สมมติฐานนั้นเป็นจริงได้ต้องอาศัยวิธีการสุ่มข้อมูลแบบ Booststrapping นั่นคือการนำชุดข้อมูลมาสุ่มในแนวนอนแบบใส่คืน (Random With Replacement) ให้ต้นไม้แต่ละต้นเรียนรู้ เช่น มีข้อมูล 10,000 เรคคอร์ด มี 5 แบบจำลอง สัดส่วนในการเลือกข้อมูลคือสุ่ม 2 ใน 3 ของข้อมูล นั่นคือแต่ละแบบจำลองจะได้ข้อมูล 6,666 เรคคอร์ด

ในการเรียนรู้ ทั้ง 5 แบบจำลองจะได้ข้อมูลที่มีความแตกต่างกันพอสมควรอยู่ในเกณฑ์ที่ยอมรับได้ จากนั้นทำการสุ่มในแนวตั้งแบบไม่ใส่คืน (Random Without Replacement) คือสุ่มฟีเจอร์ เช่น มี 10 ฟีเจอร์ ต้องการสุ่มมา 5 ฟีเจอร์ในการใช้ จากนั้นไปใช้รวมกับข้อมูลที่เราสุ่มไว้ด้านบน ได้ว่า ต้นไม้ที่ 1 ใช้ข้อมูล 6,666 เรคคอร์ดและ 5 ฟีเจอร์ ในการเรียนรู้ โดยการทำงานของ Random Forest แสดงดังภาพประกอบ 5

2.1.5 Extreme Gradient Boosting (XGBoost)

จากข้างต้น Random Forest ถือเป็นวิธีการ Bagging ซึ่งใช้เทคนิค Booststrapping ในการทำให้แต่ละแบบจำลองอิสระต่อกัน แต่ XGBoost ใช้วิธีการ Boosting คือการให้น้ำหนักแต่ละแบบจำลองเพื่อให้แต่ละแบบจำลองอิสระต่อกันและการให้น้ำหนักที่ไม่เท่ากันยังทำให้ได้ข้อมูลที่แตกต่างกันได้ และอีกสิ่งที่ Boosting แตกต่างจาก Bagging คือ Boosting เป็น Sequential Training คือต้องให้แบบจำลอง 1 เรียนรู้ทั้งชุดข้อมูลเสร็จก่อนจึงให้แบบจำลองตัวอื่นเรียนรู้ต่อไปได้ แต่ Bagging สามารถทำทุกแบบจำลองพร้อมกันได้

XGBoost เป็นที่นิยมสำหรับการแข่งขันบน Kaggle ใช้วิธีการ Boosting นั่นคือการทำงานคือการนำต้นไม้หลาย ๆ ต้นมาทำงานต่อเนื่องกันโดยมีการเพิ่มน้ำหนักให้ข้อมูลที่ถูกทำนายผิด และลดน้ำหนักให้ข้อมูลที่มีการทำนายถูก เพื่อให้การเรียนรู้ครั้งต่อไปมีการเรียนรู้ความผิดพลาดจากการเรียนรู้รอบก่อนหน้า และมีความสามารถในการลดปัญหา Overfitting สามารถจัดการกับค่าว่างได้อัตโนมัติ

2.1.6 การวัดประสิทธิภาพการทำงานของอัลกอริทึม

ในงานวิจัยนี้เป็นการจัดการปัญหา Classification ซึ่งสิ่งที่ต้องการทำนายอยู่ในรูปคลาส (Categorical Target Variable) จึงใช้การวัดประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score ซึ่งคำนวณได้จาก Confusion Matrix ดังตาราง 1

ตาราง 1 แสดง Confusion Matrix

		Actual	
		Positive	Negative
Predict	Positive	TP	FP
	Negative	FN	TN

โดยที่

TP คือสิ่งที่แบบจำลองทำนายว่าเป็นกลุ่มนั้นจริง

FP คือสิ่งที่แบบจำลองทำนายว่าเป็นกลุ่มนั้นแต่ความจริงไม่ใช่

FN คือสิ่งที่แบบจำลองทำนายว่าไม่ใช่กลุ่มนั้นแต่ความจริงใช่

TN คือสิ่งที่แบบจำลองทำนายว่าไม่ใช่กลุ่มนั้นและความจริงไม่ใช่

2.1.6.1 Accuracy

คืออัตราส่วนของการทำนายถูกกับจำนวนข้อมูลทั้งหมด แสดงได้ดังสมการ (6)

$$Accuracy = \frac{\text{ข้อมูลที่ทำนายถูก}}{\text{ข้อมูลทั้งหมด}} \quad (6)$$

2.1.6.2 Precision

คืออัตราส่วนของการทำนายว่าเป็นกลุ่มนั้นแล้วถูกกับจำนวนข้อมูลที่ถูกทายว่าเป็นกลุ่มนั้นทั้งหมด แสดงได้ดังสมการ (7)

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

2.1.6.3 Recall

คืออัตราส่วนของการทำนายว่าเป็นกลุ่มนั้นแล้วถูกกับข้อมูลทั้งหมดของกลุ่มนั้น แสดงได้ดังสมการ (8)

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

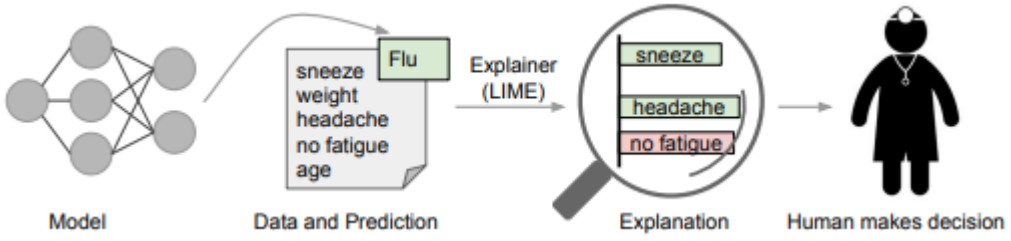
2.1.6.4 F1-Score

คือค่าเฉลี่ยถ่วงน้ำหนักระหว่าง Precision และ Recall แสดงได้ดังสมการ (9)

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

2.1.7 Local Interpretable Model-agnostic Explanations

Local Interpretable Model-agnostic Explanations หรือ LIME คืออัลกอริทึมที่สามารถอธิบายการทำนายของแบบจำลองในรูปแบบที่สามารถเข้าใจและเชื่อถือได้ โดยแสดงออกมาในรูปแบบข้อความหรือรูปภาพที่แสดงความสัมพันธ์ระหว่างส่วนประกอบของข้อมูล (เช่น คำในข้อความหรือฟีเจอร์) และการทำนายของแบบจำลอง ซึ่งทำการอธิบายทีละ 1 ข้อมูล



ภาพประกอบ 6 แสดงการทำงานของ LIME

ที่มา (Ribeiro, Singh, & Guestrin, 2016)

จากภาพประกอบ 6 แสดงขั้นตอนการอธิบายการทำนายข้อมูล 1 ข้อมูลด้วย LIME โดยแบบจำลองนี้ต้องการทำนายโรคของคนไข้ ซึ่งทำนายได้ว่าคนไข้เป็นไข้ (Flu) จากนั้น LIME ได้ทำการแสดงน้ำหนักของแต่ละอาการออกมา เห็นได้ว่าอาการจาม (Sneeze) และอาการปวดหัว (Headache) ใช้สีเขียวในการแสดง หมายความว่า 2 อาการนี้มีผลต่อการทำนาย ในขณะที่การไม่มีความเหนื่อยล้า (No Fatigue) ใช้สีแดงซึ่งหมายความว่าให้ผลลัพธ์ที่ตรงข้ามกัน ด้วยการใช้ LIME ทำให้แพทย์สามารถมองเห็นและเข้าใจการทำงานของแบบจำลองและทำให้แพทย์สามารถตัดสินใจได้ว่าสามารถเชื่อแบบจำลองได้หรือไม่

2.1.8 Shapley Additive Explanations

Shapley Additive Explanations หรือ SHAP คืออัลกอริทึมที่ใช้ในการอธิบายการใช้ฟีเจอร์ในการทำนายบน 1 ข้อมูลจากการใช้ค่าความสำคัญของคุณลักษณะหรือ Feature Importance ว่ามีฟีเจอร์ใดบ้างที่ส่งผลต่อผลลัพธ์ ซึ่งผลลัพธ์ในงานนี้คือกลุ่มหรือ Segmentation ซึ่ง SHAP ช่วยเพิ่มความน่าเชื่อถือให้กับแบบจำลองเพราะสามารถทำให้ผู้นำแบบจำลองนี้ไปใช้สามารถอ่านและทำความเข้าใจแบบจำลองได้ โดยผู้วิจัยเลือกใช้การแสดงผลแบบ TreeExplainer() ซึ่งสามารถใช้ได้กับแบบจำลองที่ใช้ต้นไม้ในการตัดสินใจเพราะต้องการใช้ในการ

แสดงผลของ XGBoost โดยสามารถแสดงผลแยกกลุ่มได้ว่าฟีเจอร์นั้น ๆ มีอิทธิพลต่อกลุ่มใดมากที่สุด

2.2 งานวิจัยที่เกี่ยวข้อง

การศึกษางานวิจัยที่เกี่ยวข้อง เราสนใจงานวิจัยที่ใช้การเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) กับข้อมูลลูกค้า บางงานวิจัยอาจใช้ข้อมูลด้านพฤติกรรมซึ่งแตกต่างกับงานวิจัยนี้ที่ใช้ข้อมูลด้านประชากร บางงานวิจัยอาจใช้การเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Learning) แต่เนื่องจากใช้ข้อมูลเกี่ยวกับลูกค้าเหมือนกัน เราจึงศึกษาและดูวิธีการบางอย่างจากงานวิจัยเหล่านี้

1. บทความวิจัยเรื่อง Integrated Churn Prediction and Customer Segmentation Framework for Telco Business (Wu, Yau, Ong, & Chong, 2021)

ในงานวิจัยนี้ได้ทำการทำนายลูกค้าว่าลูกค้าคนใดมีแนวโน้มย้ายค่ายโดยทำการทดลองบนข้อมูล 3 ชุดที่แตกต่างกัน ภายในข้อมูลประกอบด้วยข้อมูลประเภทประชากรและข้อมูลด้านพฤติกรรมการใช้บริการของลูกค้า ในการทดลองนี้ใช้ แบบจำลองทั้งหมด 6 ตัวคือ Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Adaboost, และ Multi-layer Perceptron เนื่องจากแบบจำลอง 6 ตัวนี้ทำงานได้ดีกับงานวิจัยเกี่ยวกับการทำนายการย้ายค่ายและจัดการปัญหาข้อมูลไม่สมดุลด้วยวิธีการ SMOTE เพื่อช่วยเพิ่มประสิทธิภาพของแบบจำลอง ในบทความนี้ผู้เขียนได้วัดประสิทธิภาพการทำงานของแบบจำลองทั้งก่อนและหลังใช้ SMOTE โดยใช้ค่า Accuracy, Precision, Recall, F1-score และ AUC ซึ่งผลจากการใช้ SMOTE ให้ประสิทธิภาพที่ดีกว่า

จากนั้นก่อนการจัดกลุ่ม มีการใช้ Bayesian จากการคำนวณค่า Odds Ratio ในการกำจัดบางฟีเจอร์ที่ไม่ส่งผลต่อการย้ายค่ายออกไปไม่นำไปใช้ในการจัดกลุ่ม โดยฟีเจอร์ที่ให้ค่าบวกมากแสดงว่าฟีเจอร์มีความสัมพันธ์กับการย้ายค่ายมาก ถ้าฟีเจอร์ให้ค่าลบมากแสดงว่าฟีเจอร์นั้นมีความน่าจะเป็นที่ทำให้ลูกค้าไม่ย้ายค่าย ซึ่งฟีเจอร์ที่ถูกกำจัดเป็นฟีเจอร์ที่ให้ค่าใกล้ศูนย์ทั้งในทางบวกและลบ เพราะแสดงว่าไม่มีความสัมพันธ์กับการย้ายค่าย

จากนั้นนำลูกค้าที่ถูกทำนายมาแล้วว่ามีแนวโน้มย้ายค่าย มาจัดกลุ่มโดยใช้ K-means Clustering โดยใช้ Elbow Method และเปรียบเทียบค่า Silhouette เพื่อพิจารณาค่า K ที่เหมาะสม ได้ว่า K เท่ากับ 3 คือแบ่งลูกค้าออกเป็น 3 กลุ่ม จะสามารถอธิบายลักษณะของแต่ละกลุ่มจากการดูฟีเจอร์ จากนั้นใช้มาตรการต่าง ๆ ที่เหมาะสมกับแต่ละกลุ่ม เพื่อตอบสนองความต้องการของลูกค้าและรักษาลูกค้าไว้

2. บทความวิจัยเรื่อง Random Forest-based Approach for Classifying Customers in Social CRM (Lamrhari, Elghazi, & El Faker, 2020)

ในงานวิจัยนี้ผู้เขียนใช้แบบจำลอง Random Forest ในการจัดการกับปัญหา Classification ที่มีหลายคลาสในบริบทของ CRM ในการทดลองนี้ได้จัดกลุ่มลูกค้าออกเป็น 3 คลาสหรือ 3 กลุ่มคือกลุ่ม Prospects, Satisfied และ Unsatisfied Customers โดยใช้ข้อมูลจากบนโซเชียล

การระบุว่าลูกค้าแต่ละคนมีความพึงพอใจหรือไม่กับบริษัท ทำให้ทางบริษัทสามารถใช้กลยุทธ์ต่าง ๆ เพื่อเปลี่ยนลูกค้าที่ไม่พึงพอใจเป็นลูกค้าที่มีความพึงพอใจได้ สาเหตุที่ผู้เขียนได้เลือกใช้แบบจำลอง Random Forest เพราะว่ามีการใช้เทคนิค Bagging และ Bootstrapping ที่สามารถจัดการกับค่าว่างได้ดี

ผู้เขียนยังได้เปรียบเทียบประสิทธิภาพการทำงานของแบบจำลอง Random Forest กับอีก 3 แบบจำลองคือ Artificial Neuron Network (ANN), Support Vector Machine (SVM) และ K-Nearest Neighbors (KNN) ด้วยค่า Accuracy, Sensitivity, Specificity, False Positive Rate และ False Negative Rate ผลการทดลองพบว่า Random Forest ให้ประสิทธิภาพที่ดีกว่าด้วยค่า Accuracy ที่ 98.46%, Sensitivity ที่ 97.69%, Specificity ที่ 98.84%, False Positive Rate ที่ 1.15% และ False Negative Rate ที่ 2.30%

3. บทความวิจัยเรื่อง An Analysis of Blessed Friday Sale at a Retail Store Using Classification Models (Shaukat et al., 2021)

ในงานวิจัยนี้ผู้เขียนได้นำเสนอการเปรียบเทียบเทคนิคที่ใช้ในการทำนายช่วงอายุของลูกค้าในวัน Blessed Friday เนื่องจากข้อมูลที่ใช้เป็นข้อมูลที่มีเลเบลจึงใช้หลักการการเรียนรู้ของเครื่องแบบมีผู้สอน เพื่อทำนายช่วงอายุของลูกค้าที่มีความสนใจในการซื้อสินค้าในวัน Blessed Friday โดยต้องการทำนายลูกค้าเป็นช่วงอายุมี 3 คลาสคือ Young, Middle และ Old อัลกอริทึมต่าง ๆ ที่นำมาใช้ คือ Decision Tree, K-Nearest Neighbor, Naïve Bayes และ Neural Network เพื่อเปรียบเทียบและเลือกใช้อัลกอริทึมที่เหมาะสม

ในการทดลองนี้ผู้เขียนใช้ RapidMiner ในการวัดประสิทธิภาพของแบบจำลอง พบว่า Decision Tree ให้ค่า Accuracy ที่สูงที่สุดเมื่อแบ่ง Training Set 70% และ Test Set 30% คือ 88.22% และทำนายผิดอยู่ที่ 11.78% ในขณะที่ Neural Network ให้ค่า Accuracy ที่ 81.69% KNN ให้ค่า Accuracy 65.82% และ Naïve Bayes ให้ค่า Accuracy ที่ 64.84% ข้อมูลตรงนี้จะช่วยให้นักการตลาดสามารถเลือกใช้แบบจำลองได้ถูกต้องเพื่อเพิ่มยอดขายในกับบริษัทของตนเอง

4. บทความวิจัยเรื่อง Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa (Umuhoza, Ntirushwamaboko, Awuah, & Birir, 2020)

ในงานวิจัยนี้ต้องการจัดกลุ่มลูกค้าผู้ถือบัตรเครดิตโดยใช้เทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนเพราะข้อมูลที่ใช้ในการทดลองเป็นข้อมูลพฤติกรรมการใช้บัตรเครดิตของลูกค้าซึ่งไม่มีเลเบล ผู้เขียนเชื่อว่าหากจัดกลุ่มตามพฤติกรรมการใช้บัตรเครดิตของลูกค้าจะสามารถกระตุ้นการใช้งานบัตรเครดิตภายในแอฟริกาได้

ผู้เขียนเลือกใช้ K-means Algorithm ในการจัดกลุ่ม เริ่มต้นด้วยการคำนวณค่า Elbow, Silhouette และ Calinski-Harbaz เพื่อพิจารณาจำนวนคลัสเตอร์ พบว่าจำนวนคลัสเตอร์ที่เหมาะสมคือ K เท่ากับ 4 คือแบ่งลูกค้าออกเป็น 4 กลุ่ม จากนั้นใช้ PCA ในการลดมิติของข้อมูล ผลการคำนวณได้ว่ามี 9 Principle Component ที่นำไปใช้ในกระบวนการ K-means ต่อไป

เมื่อจัดกลุ่มแล้วได้กลุ่มทั้ง 4 ดังนี้ กลุ่มที่ 1 มีลักษณะการใช้จ่ายวันต่อวัน ใช้เงินไปกับเรื่องทั่วไป เช่น การดูแลสุขภาพ การสื่อสาร กลุ่มที่ 2 ส่วนใหญ่มักใช้จ่ายไปกับสินค้าแฟชั่น มีบ้างบางครั้งที่ใช้จ่ายกับการอุปโภคบริโภค กลุ่มที่ 3 มีแนวโน้มจะเป็นผู้ใช้จ่ายสูงสุด ใช้จ่ายตั้งแต่เรื่องการศึกษาไปจนถึงเรื่องเครื่องประดับเพชรพลอย และกลุ่มที่ 4 มักใช้บัตรไปในทางจำกัด คือใช้กับการสร้างบ้านหรือเฟอร์นิเจอร์

5. บทความวิจัยเรื่อง Customer Segmentation With Machine Learning : New Strategy For Targeted Actions (Abidar, Zaidouni, & Ennouaary, 2020)

ในงานวิจัยนี้ผู้เขียนต้องการจัดกลุ่มลูกค้าเพื่อเข้าใจความต้องการและพฤติกรรมของลูกค้าด้วยข้อมูลการซื้อขายของลูกค้าภายในร้านค้าปลีกออนไลน์แห่งหนึ่งในช่วงเวลา 8 เดือน โดยผู้เขียนได้นำเสนอวิธีการใหม่โดยตั้งอยู่บนหลัก RFM โมเดลโดยใช้ค่า Recency, Frequency และ Monetary ในการให้คะแนนและดูการกระจายตัวของลูกค้า จากนั้นใช้ K-means Clustering Algorithm ในการจัดกลุ่ม จากการคำนวณ Elbow Method ทำให้ได้ค่า K คือ 3 เป็นค่าที่เหมาะสมคือแบ่งลูกค้าออกเป็น 3 กลุ่ม

กลุ่มที่ 1 ชื่อว่า High Segment เป็นกลุ่มที่ให้ค่า Frequency และ Monetary สูง หมายความว่ากลุ่มนี้ทำได้รายได้ค่อนข้างดีให้กับบริษัท กลุ่มที่ 2 ชื่อว่า Mid Segment เป็นกลุ่มที่ให้ค่า Frequency ค่อนข้างน้อย หมายความว่าทำรายได้ให้บริษัทได้ไม่มาก และกลุ่มที่ 3 ชื่อว่า Low Segment เป็นกลุ่มลูกค้าที่ไม่เป็นทางการคือมีการซื้อสินค้าน้อยครั้ง

6. บทความวิจัยเรื่อง Customer Segmentation Strategy for Rail Freight Market : The Case of Turkish State Railways (Zeybek, 2018)

ในงานวิจัยนี้ผู้เขียนต้องการจัดกลุ่มลูกค้าของบริการขนส่งทางรถไฟโดยใช้ข้อมูลด้านความพึงพอใจและแนวคิดต่อบริการขนส่งทางรถไฟในตุรกี โดยผู้เขียนทำการสร้างแบบสอบถามขึ้นมาเพื่อสัมภาษณ์ลูกค้าทั้งหมด 258 คน ส่วนแรกของแบบสอบถามจะถามเกี่ยวกับข้อมูลภูมิหลัง และส่วนที่สองถามเกี่ยวกับความพึงพอใจของลูกค้าต่อบริการขนส่งทางรถไฟ

ผู้เขียนเลือกใช้ K-means Clustering ใน Statistical Package for the Social Sciences หรือ SPSS ในการจัดกลุ่มลูกค้าได้เป็น 6 กลุ่มคือ Extremely Loyals, Loyals, Better Price and Quality Seekers, Bounds, Potential to Lose และ Quitter

กลุ่ม Extremely Loyals เมื่อเทียบกับทุกกลุ่ม เป็นกลุ่มที่มีความพึงพอใจกับบริการมากที่สุดและมีแนวโน้มแนะนำบริการนี้ให้กับผู้อื่น กลยุทธ์ที่นำมาใช้ในการเพิ่มความพึงพอใจให้กลุ่มนี้อาจเป็นการปรับปรุงอุปกรณ์ทางเทคนิคให้ดีขึ้นและพิจารณาระยะทางระหว่างศูนย์ให้บริการและที่อยู่ของลูกค้า

กลุ่ม Loyals เป็นกลุ่มที่มีลูกค้ามากที่สุด มีความพึงพอใจกับบริการแต่ยังไม่มากเท่ากับกลุ่มแรก มีแนวโน้มที่ยังคงใช้บริการต่อไป กลยุทธ์ที่ใช้กับกลุ่มนี้คือการปรับปรุงราคาให้เหมาะสมและลงทุนกับอุปกรณ์ทางเทคนิคเพิ่มขึ้น

กลุ่ม Better Price and Quality Seekers เป็นกลุ่มที่มีแนวโน้มที่ยังคงใช้บริการต่อไปและแนะนำให้ผู้อื่นใช้บริการ แต่ลูกค้าในกลุ่มนี้มีการพูดถึงต้องการราคาและคุณภาพที่ดีขึ้น กลยุทธ์ที่ใช้กับกลุ่มนี้คือปรับราคาให้เหมาะสมและเพิ่มการเข้าถึงอุปกรณ์ทางเทคนิค

กลุ่ม Bounds เป็นกลุ่มที่ไม่พึงพอใจต่อราคาและการให้บริการแต่ยังคงเลือกใช้บริการนี้เนื่องจากไม่มีบริการอื่นให้เลือกใช้ มีความไม่พึงพอใจกับอุปกรณ์ทางเทคนิค ลูกค้าในกลุ่มนี้พูดถึงบริการว่าใช้เทคโนโลยีไม่คุ้มค่าและควรลงทุนอุปกรณ์ทางเทคนิคเพิ่มขึ้น กลยุทธ์ที่ใช้กับกลุ่มนี้คือปรับปรุงคุณภาพให้สอดคล้องกับราคา

กลุ่ม Potential to Lose เป็นกลุ่มที่ไม่พึงพอใจต่อคุณภาพของบริการ มีแนวโน้มเลิกใช้บริการสูงเนื่องจากไม่พอใจกับราคาและคุณภาพ กลยุทธ์ที่ใช้กับกลุ่มนี้คือมุ่งความสนใจไปที่ความต้องการของลูกค้าเพื่อโน้มน้าวให้ลูกค้าใช้บริการต่อ

กลุ่ม Quitter เป็นกลุ่มที่ไม่พึงพอใจกับบริการและได้เลิกใช้บริการแล้วเนื่องจากไม่พึงพอใจในเรื่องราคาและคุณภาพ เวลาขนส่งและเรื่องความปลอดภัย ซึ่งค่อนข้างยากที่จะโน้มน้าวให้ลูกค้ากลุ่มนี้กลับมาใช้บริการอีก

7. บทความวิจัยเรื่อง Behavior Segmentation based Micro-Segmentation Approach for Health Insurance Industry (Nandapala, Jayasena, & Rathnayaka, 2020)

ในงานวิจัยนี้ผู้เขียนได้ใช้การจัดกลุ่มแบบ Micro-Segmentation เพื่อจัดการกับข้อมูลด้านประกันภัยสุขภาพโดยข้อมูลที่ใช้เป็นประเภทข้อมูลประชากรและข้อมูลด้านพฤติกรรมของผู้ถือประกันภายใต้กระบวนการ RFM analysis

Micro-Segmentation คือการจัดกลุ่มในขั้นที่สูงกว่าการจัดกลุ่มทั่วไป ซึ่งช่วยทำให้ธุรกิจสามารถเข้าใจความต้องการของลูกค้าและสามารถนำเสนอสินค้าหรือบริการได้ตรงตามความต้องการ

เริ่มต้นผู้เขียนใช้ข้อมูลด้านประชากรในการนับจำนวนผู้ถือประกันโดยแบ่งตามเพศ พบว่าส่วนใหญ่เป็นเพศชาย จากนั้นนำข้อมูลผู้ถือประกันเพศชายมาแสดงโดยแบ่งตามช่วงอายุ พบว่าส่วนใหญ่อยู่ในช่วงอายุ 36-45 ปี จากนั้นนำข้อมูลเกี่ยวกับโรคที่อยู่ภายในช่วงอายุ 36-45 ปี มาแสดง พบว่าผู้ที่เป็นโรคเกี่ยวกับระบบทางเดินอาหารมีจำนวนมากที่สุด จากนั้นใช้ RFM Analysis กับผู้ถือประกันที่เป็นเพศชายอายุ 36-45 ปีที่มีโรคเกี่ยวกับระบบทางเดินอาหาร จะได้เป็น RFM Score ออกมา เช่น RFM Score 444 หมายความว่าผู้ถือประกันคนนั้นมีการเคลมล่าสุดไม่นานมานี้ มีการเคลมบ่อยครั้งและจ่ายค่าเคลมค่อนข้างมาก เป็นกลุ่มที่ทำให้กำไรของบริษัทลดลงเนื่องจากมีการจ่ายค่าเคลมค่อนข้างมาก กลยุทธ์ที่ใช้กับกลุ่มนี้คือทางผู้ให้ประกันควรหาวิธีที่สามารถลดค่าเคลมลง เพื่อเพิ่มความพึงพอใจของลูกค้า จากนั้นทำการค้นหาและแสดงว่าผู้ถือประกันคนไหนมีลักษณะแย่ที่สุด ผู้ถือประกันคนไหนมีโอกาสเลิกใช้บริการหรือเลิกใช้แล้ว ผู้ถือประกันคนไหนทำรายได้ให้บริษัทมากที่สุด ผู้ถือประกันคนไหนมีความภักดีต่อบริษัทและผู้ถือประกันคนไหนเป็นผู้ถือประกันที่ดีที่สุดในกลุ่ม ซึ่งจะทำให้บริษัทสามารถใช้กลยุทธ์ต่าง ๆ เพื่อจัดการกับการเคลมได้ดียิ่งขึ้น

ตาราง 2 แสดงผลการทดลองของงานวิจัยที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
1.	Integrated Churn Prediction and Customer Segmentation Framework for Telco Business (Shuli Wu, Wei-Chuen Yau, Thian-Song Ong and Siew-Chin Chong, 2021)	ผู้เขียนต้องการทำนายการย้ายค่ายของลูกค้า จากนั้นนำลูกค้าที่ถูกทำนายว่ามีแนวโน้มในการย้ายค่ายมาจัดกลุ่มเพื่อใช้กลยุทธ์ทางการตลาดในการรักษาลูกค้า โดยผู้เขียนได้ทำการทดลองบนชุดข้อมูล 3 ชุดที่แตกต่างกัน	Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, AdaBoost, Multiple Layer Perceptron พร้อมกับการใช้ SMOTE สำหรับการทำนายการย้ายค่าย Clustering ที่ K=3 สำหรับการจัดกลุ่ม	ข้อมูลชุดที่ 1 Adaboost ให้ผลการคำนวณค่า Accuracy, Precision, F1-Score และ AUC ที่ดีที่สุด คือ 77.19%, 55.44%, 63.11% และ 84.52% ตามลำดับ Logistic Regression ให้ค่า Recall ที่ดีที่สุดคือ 78.76%

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
				ข้อมูลชุดที่ 2 Random Forest ให้ผลการคำนวณค่า Accuracy, Precision, Recall, F1-Score และ AUC ที่ดีที่สุดคือ 93.60%, 74.63%, 80.71%, 77.20% และ 91.40% ตามลำดับข้อมูล ชุดที่ 3 Random Forest ให้ผลการคำนวณค่า Accuracy ที่ดีที่สุดคือ 63.09%

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
				Adaboost ให้ผลการคำนวณค่า Precision ที่ดีที่สุดคือ 36.53% Logistic Regression ให้ค่า Recall และ AUC ที่ดีที่สุดคือ 53.67% และ 58.66% ตามลำดับ Multi Layer Perceptron ให้ค่า F1-Score ที่ดีที่สุดคือ 42.84%

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
2.	Random Forest-based Approach for Classifying Customers in Social CRM (Soumaya Lamrhari, Hamid Elghazi and Abdellatif El Faker, 2020)	ผู้เขียนต้องการจัดกลุ่มลูกค้าโดยใช้วิธีการจัดการกับปัญหา Classification แบบมากกว่า 2 คลาส ในบริบทของ CRM โดยใช้ข้อมูลบนโซเชียล	Random Forest, ANN, SVM, KNN ในการทำนายกลุ่มลูกค้าซึ่งมี 3 กลุ่มคือ Prospect, Satisfied และ Unsatisfied Customer	Random Forest ให้ผลการคำนวณค่า Accuracy, Sensitivity, Specificity, False Positive Rate และ False Negative Rate ที่ดีที่สุดคือ 98.46%, 97.69%, 98.84%, 1.15% และ 2.3% ตามลำดับ
3.	An Analysis of Blessed Friday Sale at a Retail Store Using Classification Models	ผู้เขียนต้องการทำนายช่วงอายุของลูกค้าที่มีการใช้จ่ายในวัน Blessed-Friday	Decision Tree, KNN, Naïve Bayes และ Neural Network	Decision Tree ให้ค่า Accuracy ที่ดีที่สุดคือ 88.22% Neural Network ให้ค่า Accuracy ที่ 81.69%

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
	(Kamran Shaukat, Suhuai Luo, Nadir Abbas, Talha Mahboob Alam, Muhammad Ehtesham Tahir, and Ibrahim, 2021)			KNN ให้ค่า Accuracy ที่ 65.82% Naïve Bayes ให้ค่า Accuracy ที่ 64.84%
4.	Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa	ผู้เขียนต้องการจัดกลุ่มลูกค้าผู้ถือบัตรเครดิตในแอฟริกา โดยใช้ข้อมูลพฤติกรรมการใช้บัตรเครดิตของลูกค้า	ใช้ PCA ในการลดมิติของข้อมูล K-means Clustering ที่ K=4 สำหรับการจัดกลุ่ม	ได้ลูกค้าเป็น 4 กลุ่มตามพฤติกรรมการใช้บัตรเครดิต โดยกลุ่มที่ 1 ใช้จ่ายกับเรื่องทั่วไป กลุ่มที่ 2 ใช้จ่ายกับสินค้าแฟชั่น

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
	(Eric Umuhoza, Dominique Ntirushwamaboko, Jane Awuah and Beatrice Birir, 2020)			กลุ่มที่ 3 มีแนวโน้มจะเป็นกลุ่มที่ใช้จ่ายสูงสุด และกลุ่มที่ 4 ใช้จ่ายกับการสร้างหรือตกแต่งบ้าน
5.	Customer Segmentation With Machine Learning : New Strategy For Targeted Actions (Lahcen Abidar, Dounia Zaidouni and Abdeslam Ennouaary, 2020)	ผู้เขียนต้องการจัดกลุ่มลูกค้าของร้านค้าปลีกออนไลน์แห่งหนึ่ง เพื่อเข้าใจความต้องการของลูกค้าโดยการใช้ข้อมูลการซื้อขายในการจัดกลุ่ม	RFM Analysis K-Means Clustering ที่ K=3 สำหรับการจัดกลุ่ม	ได้ลูกค้าเป็น 3 กลุ่มตามพฤติกรรมการซื้อสินค้า โดยกลุ่มที่ 1 เป็นกลุ่มที่ทำรายได้ให้กับร้านค้าสูงที่สุด กลุ่มที่ 2 มีการใช้จ่ายน้อยครั้งและกลุ่มที่ 3 คือลูกค้าที่ไม่ใช้ลูกค้าประจำ มีการใช้จ่ายน้อยที่สุด

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
6.	Customer segmentation strategy for rail freight market : The case of Turkish State Railways (Hulya Zeybek, 2018)	ผู้เขียนต้องการจัดกลุ่มลูกค้าผู้ใช้บริการขนส่งทางรถไฟโดยการใช้ข้อมูลความพึงพอใจและแนวคิดต่อบริการของขนส่งทางรถไฟ	K-means Clustering ที่ K=6	ได้ลูกค้าเป็น 6 กลุ่มตามความพึงพอใจ โดยกลุ่มที่ 1 มีความพึงพอใจกับบริการสูงที่สุดและมีแนวโน้มที่จะแนะนำบริการให้กับผู้อื่นต่อ กลุ่มที่ 2 เป็นกลุ่มที่มีลูกค้ามากที่สุดและมีแนวโน้มจะใช้บริการต่อไป กลุ่มที่ 3 ต้องการคุณภาพและราคาที่ดีขึ้น กลุ่มที่ 4 ไม่พึงพอใจต่อราคาและการให้บริการแต่ยังคงใช้บริการต่อไป

ตาราง 2 (ต่อ)

ลำดับ	ชื่องานวิจัย	วัตถุประสงค์	แบบจำลอง	ผลการทดลอง
				กลุ่มที่ 5 มีแนวโน้มจะเลิกใช้บริการสูง และกลุ่มที่ 6 เลิกใช้บริการแล้ว
7.	Behavior Segmentation based Micro-Segmentation Approach for Health Insurance Industry (E.Y.L Nandapala, K.P.N Jayasena, R.M.K.T Rathnayaka, 2020)	ผู้เขียนต้องการจัดกลุ่มผู้ถือประกันภัยด้านสุขภาพโดยใช้ข้อมูลประชากรและข้อมูลด้านพฤติกรรมของลูกค้า	Micro-Segmentation RFM Analysis	ทราบถึงผู้ถือประกันที่มีโอกาสจะเลิกใช้บริการ ผู้ถือประกันที่ทำรายได้ให้บริษัทมากที่สุด ผู้ถือประกันที่มีความภักดีต่อบริษัท และผู้ถือประกันที่เป็นผู้ถือประกันที่ดีที่สุดในกลุ่ม

บทที่ 3
การดำเนินการวิจัย

ในงานวิจัยนี้ ผู้วิจัยได้ดำเนินตามขั้นตอนดังนี้

3.1 กระบวนการทำงานของแบบจำลอง

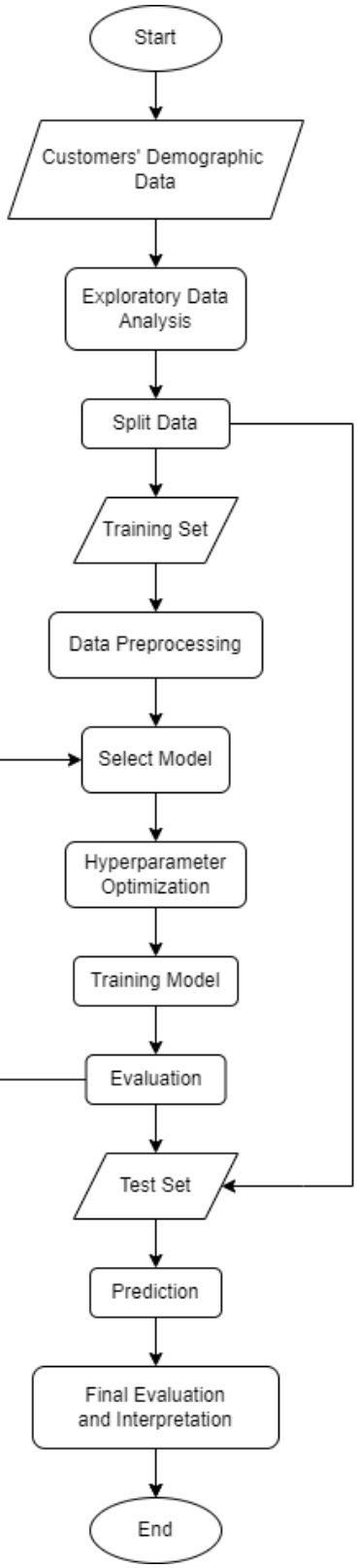
3.2 การเก็บรวบรวมข้อมูลและการตรวจสอบข้อมูล

3.3 การสำรวจข้อมูล (Exploratory Data Analysis)

3.4 การเตรียมข้อมูล (Data Preprocessing)

3.5 การสร้างแบบจำลองเพื่อทำการจัดกลุ่มลูกค้า

3.1 กระบวนการทำงานของแบบจำลอง



ภาพประกอบ 7 แสดงกระบวนการทำงานของแบบจำลองในงานวิจัยนี้

จากภาพประกอบ 7 ได้แสดงถึงกระบวนการทำงานของแบบจำลอง เริ่มต้นจากการนำเข้าข้อมูลประชากรของลูกค้าภายในบริษัทยานยนต์ จากนั้นทำการสำรวจข้อมูลหรือ Exploratory Data Analysis เพื่อทำความเข้าใจข้อมูล จากนั้นทำการแบ่งข้อมูลเป็นสองชุดคือ ข้อมูลสำหรับการเรียนรู้ของแบบจำลองหรือ Training Set และข้อมูลสำหรับการทดสอบประสิทธิภาพของแบบจำลองหรือ Test Set จากนั้นทำการเตรียมข้อมูลหรือ Data Preprocessing ที่ข้อมูลสำหรับการเรียนรู้ของแบบจำลองเท่านั้น เพื่อทำให้ข้อมูลอยู่ในรูปแบบที่ง่ายต่อการทำงานและประสิทธิภาพของแบบจำลอง กระบวนการที่ใช้คือการเปลี่ยนข้อมูลกลุ่มให้อยู่ในรูปแบบตัวเลข การสังเคราะห์ข้อมูลในกลุ่มที่มีน้อยให้เพิ่มขึ้น และการปรับให้สเกลของข้อมูลอยู่ในช่วงใกล้เคียงกัน

ขั้นตอนต่อไปคือการเลือกใช้แบบจำลอง โดยในงานวิจัยนี้เลือกใช้แบบจำลองทั้งในรูปแบบเชิงเส้นและไม่เชิงเส้นคือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest และ Extreme Gradient Boosting (XGBoost) เพื่อประสิทธิภาพของแบบจำลอง จึงใช้การปรับจูนพารามิเตอร์ร่วมด้วย เมื่อทำการเรียนรู้เสร็จเรียบร้อย จึงนำข้อมูลชุดทดสอบมาใช้ในการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix จากนั้นเปรียบเทียบค่าวัดประสิทธิภาพของแต่ละแบบจำลอง เพื่อหาแบบจำลองที่ทำงานได้ดีที่สุดบนชุดข้อมูลนี้ พร้อมทั้งวิเคราะห์ความผิดพลาดที่เกิดขึ้นจากการทำนายโดยใช้ Confusion Matrix นอกจากนี้ยังมีการวิเคราะห์และตีความหมายของแบบจำลองด้วย Feature Importance และการใช้เทคนิคในการตีความ คือ Local Interpretable Model-agnostic Explanations หรือ LIME และ Shapley Additive Explanations หรือ SHAP

3.2 การเก็บรวบรวมข้อมูลและจัดการกับข้อมูล

ในงานวิจัยนี้ได้ใช้ข้อมูลประชากรของลูกค้าภายในบริษัทยานยนต์แห่งหนึ่งซึ่งเป็นชุดข้อมูลสาธารณะจาก Kaggle.com (Kash, 2020) ประกอบด้วย 11 แอททริบิวต์ ถูกเก็บแยกไว้ 2 ไฟล์คือไฟล์ Train.csv ใช้สำหรับการเรียนรู้ของแบบจำลอง มีข้อมูลทั้งหมด 8,068 แถว และไฟล์ Test.csv ใช้สำหรับการทดสอบประสิทธิภาพของแบบจำลอง มีข้อมูลทั้งหมด 2,627 แถว

ตาราง 3 แสดงแอททริบิวต์ของข้อมูล

ลำดับ	ชื่อแอททริบิวต์	ข้อมูลภายในแอททริบิวต์	คำอธิบาย
1.	ID	เลขรหัสที่ไม่ซ้ำกัน	รหัสลูกค้า
2.	Gender	Male / Female	เพศของลูกค้า
3.	Ever_Married	Yes / No	สถานภาพสมรสของลูกค้า
4.	Age	ปี	อายุของลูกค้า
5.	Graduated	Yes / No	การจบการศึกษาของลูกค้า
6.	Profession	Artist / Doctor / Engineer / Entertainment / Executive / Healthcare / Homemaker / Lawyer / Marketing	อาชีพของลูกค้า

ตาราง 3 (ต่อ)

ลำดับ	ชื่อแอททริบิวต์	ข้อมูลภายในแอททริบิวต์	คำอธิบาย
7.	Work_Experience	ปี	ประสบการณ์ทำงานของลูกค้า
8.	Spending_Score	High / Average / Low	ระดับการใช้จ่ายของลูกค้า
9.	Family_Size	จำนวนคน	จำนวนสมาชิกในครอบครัวของลูกค้าโดยนับรวมตัวลูกค้า
10.	Var_1	Cat_1 / Cat_2 / Cat_3 / Cat_4 / Cat_5 / Cat_6 / Cat_7	กลุ่มของลูกค้าที่ทางบริษัทระบุไว้แบบนิรนาม
11.	Segmentation (Label)	A / B / C / D	กลุ่มของลูกค้า

ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
462809	Male	No	22	No	Healthcare	1	Low	4	Cat_4	D
462643	Female	Yes	38	Yes	Engineer		Average	3	Cat_4	A
466315	Female	Yes	67	Yes	Engineer	1	Low	1	Cat_6	B
461735	Male	Yes	67	Yes	Lawyer	0	High	2	Cat_6	B
462669	Female	Yes	40	Yes	Entertainment		High	6	Cat_6	A
461319	Male	Yes	56	No	Artist	0	Average	2	Cat_6	C
460156	Male	No	32	Yes	Healthcare	1	Low	3	Cat_6	C
464347	Female	No	33	Yes	Healthcare	1	Low	3	Cat_6	D
465015	Female	Yes	61	Yes	Engineer	0	Low	3	Cat_7	D
465176	Female	Yes	55	Yes	Artist	1	Average	4	Cat_6	C

ภาพประกอบ 8 แสดงตัวอย่างข้อมูลในไฟล์ Train.csv 10 แถวแรก

ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
458989	Female	Yes	36	Yes	Engineer	0	Low	1	Cat_6	B
458994	Male	Yes	37	Yes	Healthcare	8	Average	4	Cat_6	A
458996	Female	Yes	69	No		0	Low	1	Cat_6	A
459000	Male	Yes	59	No	Executive	11	High	2	Cat_6	B
459001	Female	No	19	No	Marketing		Low	4	Cat_6	A
459003	Male	Yes	47	Yes	Doctor	0	High	5	Cat_4	C
459005	Male	Yes	61	Yes	Doctor	5	Low	3	Cat_6	D
459008	Female	Yes	47	Yes	Artist	1	Average	3	Cat_6	D
459013	Male	Yes	50	Yes	Artist	2	Average	4	Cat_6	B
459014	Male	No	19	No	Healthcare	0	Low	4	Cat_6	B

ภาพประกอบ 9 แสดงตัวอย่างข้อมูลในไฟล์ Test.csv 10 แถวแรก

เนื่องจากข้อมูลที่ได้มามี 2 ไฟล์ นั่นคือ Train.csv และ Test.csv ที่แสดงดังภาพประกอบ 8 และ 9 เริ่มต้นนำข้อมูล 2 ไฟล์นี้รวมเป็นตารางเดียวกัน เพื่อง่ายต่อการตรวจสอบและจัดการกับค่าว่างต่างๆ พบว่าเมื่อรวมแล้วได้ทั้งหมด 10,695 แถว

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A

ภาพประกอบ 10 แสดงตัวอย่างข้อมูล 5 แถวแรกเมื่อรวม 2 ไฟล์เข้าด้วยกัน

จากภาพประกอบ 10 พบว่ามีค่า Nan หรือค่าว่าง ที่อาจทำให้เกิดปัญหาในการให้แบบจำลองเรียนรู้ภายหลังและอาจเกิดผลกระทบต่อประสิทธิภาพในการจัดกลุ่มให้ลูกค้า จึงทำการจัดการกับค่าว่างของแต่ละฟีเจอร์ด้วยวิธีที่แตกต่างกัน ขั้นแรกทำการตรวจสอบว่ามีค่าว่างที่ฟีเจอร์ใดบ้าง พบว่าฟีเจอร์ที่มีค่าว่างคือ Ever_Married, Graduated, Profession, Work_Experience, Family_Size และ Var_1 ตามภาพประกอบ 11

	Null Value
ID	0
Gender	0
Ever_Married	190
Age	0
Graduated	102
Profession	162
Work_Experience	1098
Spending_Score	0
Family_Size	448
Var_1	108
Segmentation	0

ภาพประกอบ 11 แสดงจำนวนค่าว่างของแต่ละแอททริบิวต์

สำหรับฟีเจอร์ Ever_Married เก็บข้อมูลเป็น Yes สำหรับลูกค้าที่แต่งงานแล้ว และ No สำหรับลูกค้าที่ยังไม่ได้แต่งงาน ดังนั้นเมื่อลูกค้าไม่ได้กรอกข้อมูลตรงนี้มาให้ จึงเลือกแทนด้วยค่าที่ไม่ส่งผลกระทบต่อความรู้สึกลูกค้าและการทำงานของแบบจำลอง นั่นคือค่า No

สำหรับฟีเจอร์ Graduated เก็บข้อมูลเป็น Yes สำหรับลูกค้าที่จบการศึกษาแล้ว และ No สำหรับลูกค้าที่ยังไม่จบการศึกษา ดังนั้นเมื่อพบว่าลูกค้าไม่ได้กรอกข้อมูลตรงนี้มาให้ อาจเข้าใจได้ว่าลูกค้ายังไม่จบการศึกษา จึงเลือกแทนด้วยค่า No เช่นเดียวกับฟีเจอร์ Ever_Married

สำหรับฟีเจอร์ Profession เก็บข้อมูลอาชีพของลูกค้า ซึ่งจากข้อมูลทั้งหมดพบว่าลูกค้าของบริษัทยานยนต์แห่งนี้มีลูกค้าจาก 9 สายอาชีพ คือ Artist หรือด้านศิลปิน Doctor หรือแพทย์ Engineer หรือวิศวกร Entertainment หรือด้านบันเทิง Executive หรือด้านบริหาร Healthcare หรือด้านสายสุขภาพ Homemaker หรือพ่อบ้านแม่บ้าน Lawyer หรือทนายความและ Marketing หรือด้านการตลาด สำหรับฟีเจอร์นี้เราพิจารณาลบแถวที่พบค่าว่าง เนื่องจากไม่สามารถหาค่าแทนข้อมูลมาทดแทนได้ จำนวนแถวที่ต้องลบคือ 162 แถว ซึ่งอยู่ในจำนวนที่ไม่มากเกินไป

สำหรับฟีเจอร์ Work_Experience เก็บข้อมูลจำนวนปีประสบการณ์การทำงานของลูกค้า นอกจากพบว่ามีจำนวนค่าว่างมากถึง 1,098 แถวแล้ว ยังพบว่าลูกค้ากรอกข้อมูลในช่องนี้ว่า 0 ปี เป็นจำนวนมากถึง 3,032 แถว ซึ่งหากใช้ค่าเฉลี่ยในการแทนค่าว่าง ค่า 0 ที่มีมากเกินไปอาจทำให้ค่าเฉลี่ยเป็นค่าที่ไม่เหมาะสมสำหรับแทนข้อมูล อีกปัญหาที่พบคือลูกค้าส่วนใหญ่ที่กรอกค่าเป็น 0 คือลูกค้าในกลุ่มที่อายุสูง ซึ่งเมื่อคิดค่าเฉลี่ยแยกตามอายุทำให้ขัดแย้งกับความเป็นจริงตรงที่พบว่าลูกค้าที่อายุน้อยกว่ามีประสบการณ์ทำงานมากกว่าลูกค้าที่อายุมาก ดังนั้นจึงพิจารณาและตัดสินใจลบฟีเจอร์นี้ ถึงแม้ประสบการณ์ทำงานอาจเป็นข้อมูลที่สามารถช่วยในการจัดกลุ่มลูกค้าได้ดี แต่หากข้อมูลมีความไม่สมบูรณ์และไม่ตรงกับความเป็นจริง อาจทำให้เกิดผลเสียต่อการทำงานของแบบจำลองได้

สำหรับฟีเจอร์ Family_Size เก็บข้อมูลเป็นจำนวนสมาชิกในครอบครัวของลูกค้าโดยนับรวมตัวลูกค้าเองด้วย ค่าที่เหมาะสมสำหรับแทนข้อมูลที่เป็นตัวเลขคือค่าเฉลี่ย จึงเลือกใช้ค่าเฉลี่ยของจำนวนสมาชิกภายในครอบครัวของลูกค้าทั้งหมดคือ 2.845259 คน หรือประมาณ 3 คน

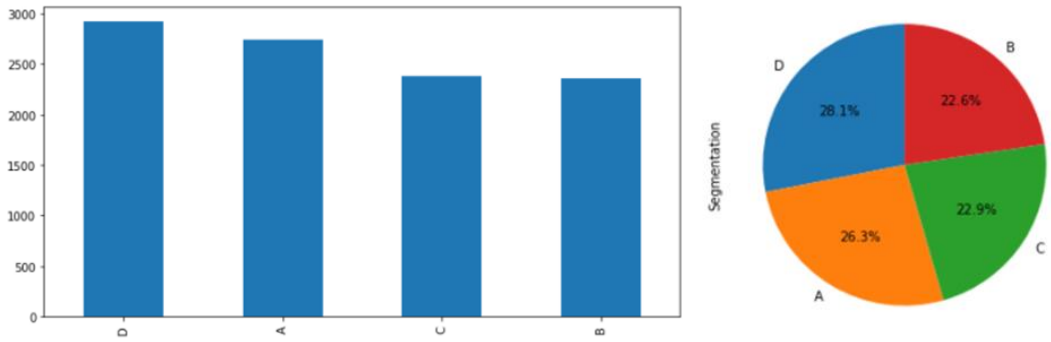
สำหรับฟีเจอร์ Var_1 เก็บข้อมูลเป็นกลุ่มของลูกค้าที่ทางบริษัทระบุไว้แบบนิรนาม ซึ่งมีทั้งหมด 7 กลุ่มคือ Cat_1, Cat_2, Cat_3, Cat_4, Cat_5, Cat_6 และ Cat_7 เหมือนกับกรณีของฟีเจอร์ Profession ที่ไม่สามารถหาค่าแทนข้อมูลได้ จึงพิจารณาการลบแถวที่พบค่าว่างออก จำนวนแถวที่หายไปคือ 108 แถวซึ่งอยู่ในเกณฑ์ที่รับได้

เมื่อจัดการกับค่าว่างโดยพิจารณาตามฟีเจอร์แล้ว พบแถวซ้ำทั้งหมด 38 แถว อาจเกิดจากความผิดพลาดในขั้นตอนการกรอกข้อมูล เนื่องจากเป็นการซ้ำกันทั้งแถว จึงพิจารณาลบออกเพื่อป้องกันความผิดพลาดในอนาคต

เมื่อจัดการกับค่าว่างเรียบร้อยแล้ว เหลือข้อมูลทั้งหมด 10,390 แถว 10 ฟีเจอร์ ที่สามารถนำไปสำรวจข้อมูลและใช้ในการเรียนรู้ของแบบจำลองต่อไปได้

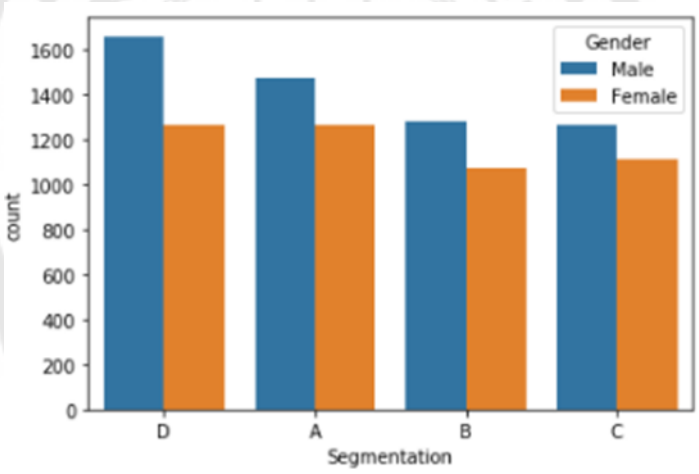
3.3 **การสำรวจข้อมูล** (Exploratory Data Analysis)

เพื่อเป็นการเข้าใจข้อมูลและลูกค้าภายในบริษัทมากขึ้น การทำ Exploratory Data Analysis หรือ EDA จึงเป็นขั้นตอนที่สำคัญมาก ทำให้เราสามารถทราบถึงลักษณะของลูกค้าในแต่ละกลุ่มได้ โดยเริ่มต้นทำการสำรวจจำนวนลูกค้าในแต่ละกลุ่ม พบว่าลูกค้าในกลุ่ม D มีจำนวนมากที่สุดคือ 2,920 คน ลำดับต่อมาคือกลุ่ม A มีลูกค้าจำนวน 2,737 คน กลุ่ม C มีลูกค้าจำนวน 2,380 คน และกลุ่ม B มีจำนวนลูกค้าน้อยที่สุดอยู่ที่ 2,353 คน ซึ่งแสดงในภาพประกอบ 12



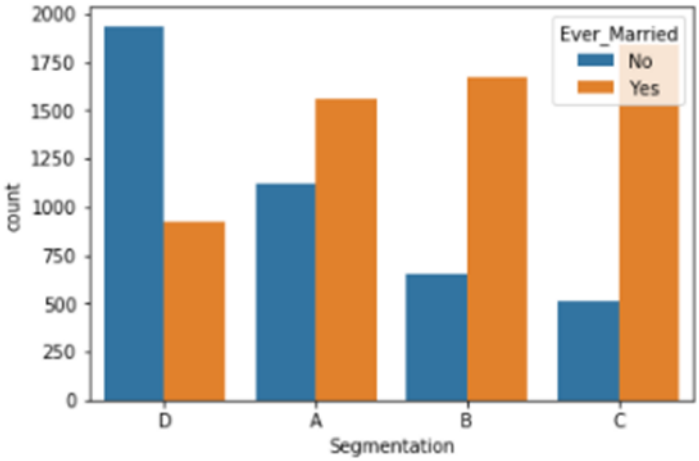
ภาพประกอบ 12 แสดงจำนวนลูกค้าในแต่ละกลุ่มในรูปแบบกราฟแท่งและกราฟวงกลม

สำหรับฟีเจอร์ Gender เมื่อตรวจสอบจำนวนลูกค้าโดยแบ่งตามเพศ พบว่ามีลูกค้าเพศชาย 5,681 คน และเพศหญิง 4,709 คน เมื่อแบ่งตามกลุ่มพบว่าทุกกลุ่มมีจำนวนเพศชายมากกว่าเพศหญิงในสัดส่วนที่คล้ายกันดังภาพประกอบ 13



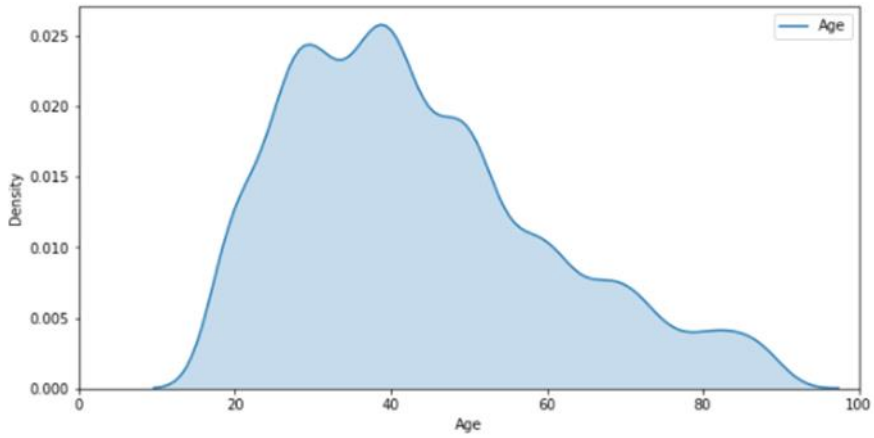
ภาพประกอบ 13 แสดงจำนวนลูกค้าในแต่ละกลุ่มโดยแบ่งตามเพศในรูปแบบกราฟแท่ง

สำหรับภาพประกอบ 14 แสดงฟีเจอร์ Ever_Married ที่ตรวจสอบจำนวนลูกค้าโดยแบ่งตามสถานภาพสมรส พบว่ามีลูกค้าที่แต่งงานแล้ว 5,991 คน และลูกค้าที่ยังไม่ได้แต่งงาน 4,218 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม D ส่วนใหญ่ยังไม่ได้แต่งงาน กลุ่ม A ส่วนใหญ่แต่งงานแล้วแต่สัดส่วนไม่ได้ต่างกับลูกค้าที่ยังไม่ได้แต่งงานมากนัก ในขณะที่ลูกค้าในกลุ่ม B และ C ส่วนใหญ่แต่งงานแล้ว เมื่อเทียบกับลูกค้าที่ยังไม่ได้แต่งงานจำนวนค่อนข้างต่างกันมาก



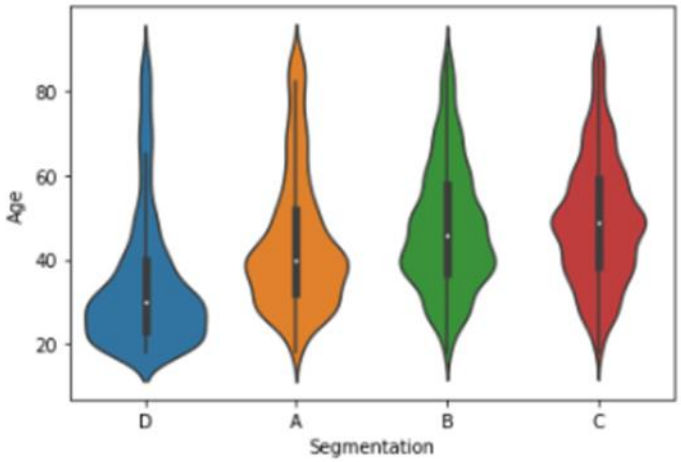
ภาพประกอบ 14 แสดงจำนวนลูกค้าในแต่ละกลุ่มโดยแบ่งตามสถานภาพสมรสในรูปแบบกราฟแท่ง

ในภาพประกอบ 15 แสดงฟีเจอร์ Age ที่ตรวจสอบจำนวนลูกค้าภายในบริษัทพบว่าลูกค้าภายในบริษัทมีตั้งแต่อายุ 18-89 ปี แต่โดยส่วนใหญ่ลูกค้ามีอายุในช่วง 30-40 ปี



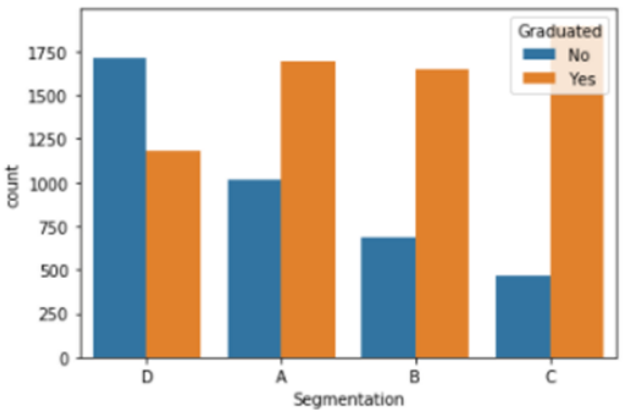
ภาพประกอบ 15 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามอายุ

เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม D ส่วนใหญ่อยู่ในช่วงอายุ 20-30 ปี ซึ่งสอดคล้องกับด้านบนที่สำรวจตามฟีเจอร์ Ever_Married ที่พบว่าลูกค้าในกลุ่ม D ยังไม่ได้แต่งงาน ลูกค้าส่วนใหญ่ในกลุ่ม A อยู่ในช่วงอายุ 30-40 ปี ลูกค้าส่วนใหญ่ในกลุ่ม B มีอายุ 40-50 ปี และลูกค้าในกลุ่ม C ส่วนใหญ่อยู่ในช่วงอายุ 50-60 ปีแสดงดังภาพประกอบ 16



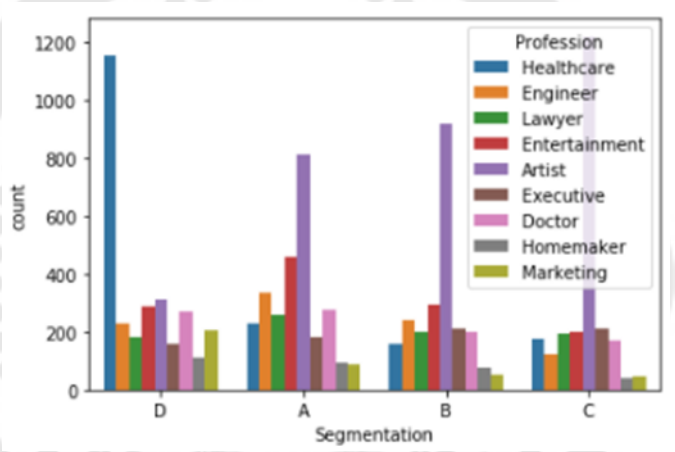
ภาพประกอบ 16 แสดงความหนาแน่นของลูกค้าในแต่ละกลุ่มโดยแบ่งตามอายุในรูปแบบกราฟไวโอลิน

สำหรับฟีเจอร์ Graduated ที่แสดงดังภาพประกอบ 17 เมื่อตรวจสอบจำนวนลูกค้าโดยแบ่งตามการจบการศึกษาของลูกค้าพบว่ามีลูกค้าที่จบการศึกษาแล้วจำนวน 6,424 คน และลูกค้าที่ยังไม่จบการศึกษาจำนวน 3,874 คน เมื่อแบ่งตามกลุ่มพบว่าลูกค้าในกลุ่ม D ส่วนใหญ่ยังไม่จบการศึกษา ซึ่งสอดคล้องกับการสำรวจแบ่งตามสถานภาพสมรสและอายุที่ว่าลูกค้าในกลุ่ม D ยังไม่ได้แต่งงานและมีอายุน้อยที่สุดเมื่อเทียบกับอีก 3 กลุ่ม ในขณะที่ลูกค้าในกลุ่ม A, B และ C ส่วนใหญ่จบการศึกษาแล้ว



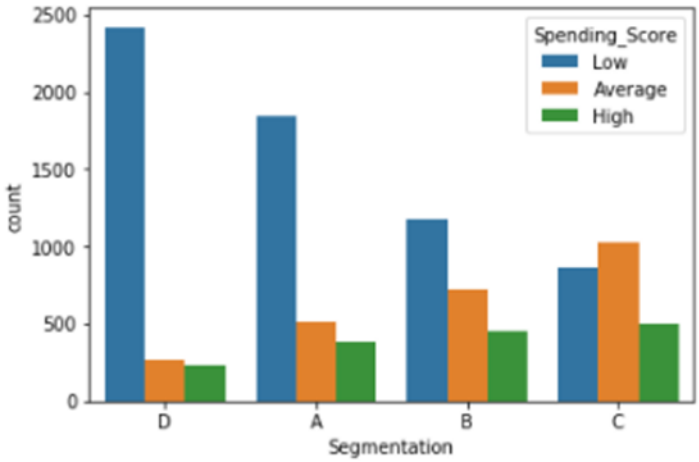
ภาพประกอบ 17 แสดงจำนวนลูกค้าในแต่ละกลุ่มโดยแบ่งตามการจบการศึกษาในรูปแบบกราฟแท่ง

สำหรับภาพประกอบ 18 แสดงฟีเจอร์ Profession ที่ตรวจสอบจำนวนลูกค้าโดยแบ่งตามอาชีพพบว่าลูกค้าส่วนใหญ่ประกอบอาชีพด้านศิลปิน มีจำนวนถึง 3,273 คน ประกอบอาชีพด้านสุขภาพจำนวน 1,717 คน ประกอบอาชีพด้านบันเทิง 1,241 คน ประกอบอาชีพด้านวิศวกรจำนวน 925 คน ประกอบอาชีพแพทย์ 919 คน ประกอบอาชีพทนายความ 834 คน ประกอบอาชีพด้านบริหาร 764 คน ประกอบอาชีพด้านการตลาด 398 คน และประกอบอาชีพพ่อบ้านหรือแม่บ้านจำนวน 319 คน เมื่อแบ่งลูกค้าตามกลุ่มพบว่าลูกค้าส่วนใหญ่ในกลุ่ม D ประกอบอาชีพด้านสุขภาพ กลุ่ม A และ B มีลักษณะคล้ายกันคือลูกค้าส่วนใหญ่ประกอบอาชีพด้านศิลปินและบันเทิง ในขณะที่กลุ่ม C ลูกค้าส่วนใหญ่ประกอบอาชีพด้านศิลปิน สำหรับอาชีพอื่น ๆ มีจำนวนที่ไม่แตกต่างกันมากนักและมีจำนวนลูกค้าที่ประกอบอาชีพพ่อบ้านแม่บ้านและการตลาดที่น้อยมาก



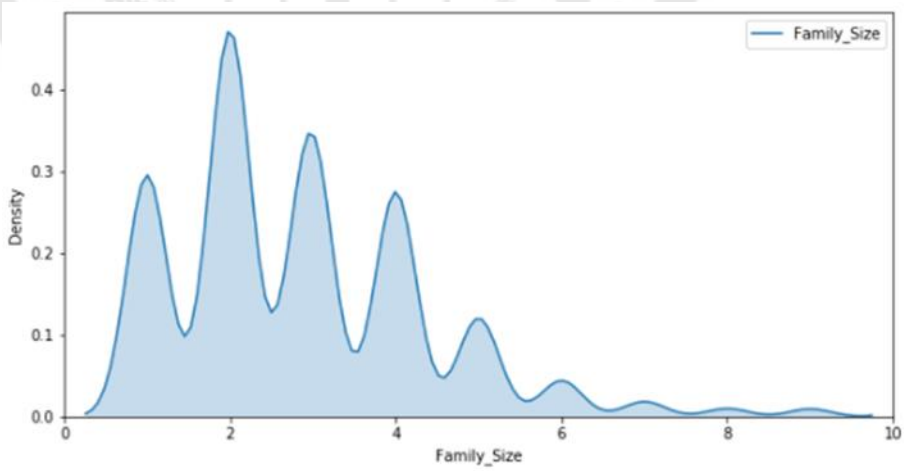
ภาพประกอบ 18 แสดงจำนวนลูกค้าในแต่ละกลุ่มโดยแบ่งตามอาชีพในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ Spending_Score เมื่อตรวจสอบจำนวนลูกค้าโดยแบ่งตามระดับการใช้จ่ายพบว่าลูกค้าส่วนใหญ่ใช้จ่ายในระดับต่ำหรือ Low มีจำนวนถึง 6,297 คน รองลงมาอยู่ในระดับปานกลางหรือ Average มีจำนวน 2,525 คน และลำดับสุดท้ายคือลูกค้าอยู่ในระดับการใช้จ่ายสูงจำนวน 1,568 คน เมื่อแบ่งลูกค้าตามกลุ่มที่แสดงดังภาพประกอบ 19 พบว่าลูกค้าในกลุ่ม A และ D ส่วนใหญ่อยู่ในระดับการใช้จ่ายต่ำ ลูกค้าในกลุ่ม B ส่วนใหญ่อยู่ในระดับการใช้จ่ายต่ำเช่นกันแต่จำนวนไม่ต่างจากลูกค้าในระดับการใช้จ่ายปานกลางและสูงมากเท่ากลุ่ม A และ D ในขณะที่กลุ่ม C ลูกค้าส่วนใหญ่มีระดับการใช้จ่ายปานกลาง



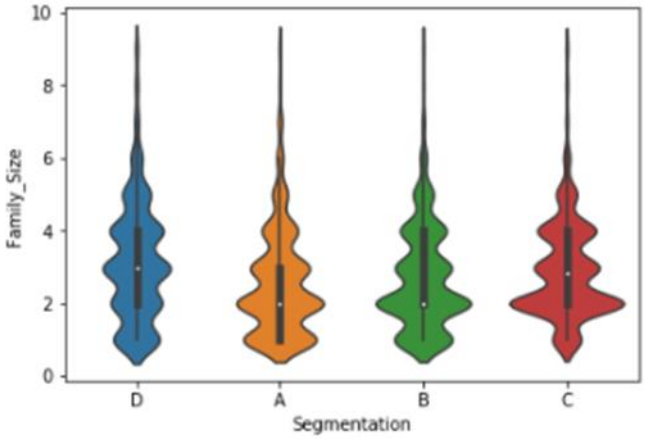
ภาพประกอบ 19 แสดงจำนวนลูกค้าในแต่ละกลุ่มโดยแบ่งตามระดับการใช้จ่ายในรูปแบบกราฟแท่ง

สำหรับฟีเจอร์ Family_Size ที่แสดงดังภาพประกอบ 20 เมื่อตรวจสอบจำนวนลูกค้าภายในบริษัทพบว่าสมาชิกในครอบครัวของลูกค้ามีตั้งแต่ 1 คนคือตัวลูกค้าเองไปจนถึง 9 คน โดยลูกค้าส่วนใหญ่มีสมาชิกในครอบครัว 2 คนนับรวมตัวเองแล้ว



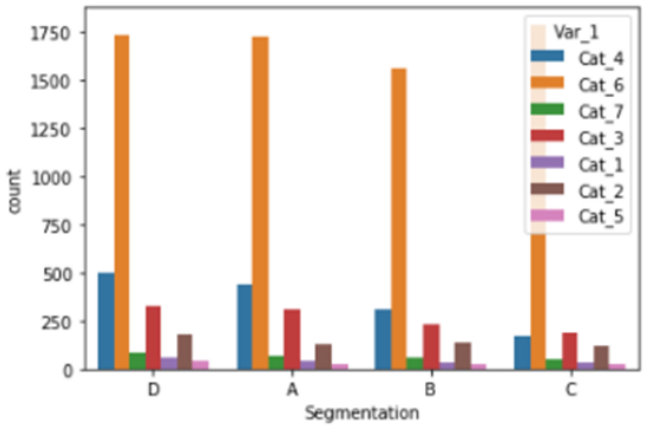
ภาพประกอบ 20 แสดงความหนาแน่นของลูกค้าโดยแบ่งตามจำนวนสมาชิกในครอบครัว

เมื่อสำรวจตามกลุ่มตามภาพประกอบ 21 พบว่าลูกค้าในกลุ่ม D ส่วนใหญ่มีสมาชิกในครอบครัวจำนวน 3-4 คน กลุ่ม A และ B ส่วนใหญ่มีจำนวนสมาชิกในครอบครัวประมาณ 1-2 คน และกลุ่ม C ส่วนใหญ่มีสมาชิกในครอบครัวประมาณ 2 คน



ภาพประกอบ 21 แสดงความหนาแน่นของลูกค้าในแต่ละกลุ่มโดยแบ่งตามจำนวนสมาชิกในครอบครัวในรูปแบบกราฟไวโอลิน

สำหรับฟีเจอร์ Var_1 เมื่อตรวจสอบจำนวนลูกค้าโดยแบ่งตามกลุ่มที่ทางบริษัทระบุไว้แบบนิรนามพบว่าลูกค้าส่วนใหญ่ถูกจัดอยู่ในกลุ่ม Cat_6 จำนวนถึง 6,813 คน ถูกจัดให้อยู่ในกลุ่ม Cat_4 จำนวน 1,427 คน ถูกจัดให้อยู่ในกลุ่ม Cat_3 จำนวน 1,056 คน ถูกจัดให้อยู่ในกลุ่ม Cat_2 จำนวน 559 คน ถูกจัดให้อยู่ในกลุ่ม Cat_7 จำนวน 257 คน ถูกจัดให้อยู่ในกลุ่ม Cat_1 จำนวน 165 คน และถูกจัดให้อยู่ในกลุ่ม Cat_5 จำนวน 113 คน เมื่อสำรวจลูกค้าแบ่งตามกลุ่มแสดงดังภาพประกอบ 22 พบว่าทุกกลุ่มมีลักษณะคล้ายกันคือส่วนใหญ่อยู่ในกลุ่ม Cat_6 ตามด้วย Cat_4 และ Cat_3 ตามลำดับ ดังนั้นฟีเจอร์นี้อาจไม่สามารถบอกถึงความแตกต่างระหว่างกลุ่ม A, B, C และ D ได้



ภาพประกอบ 22 แสดงจำนวนลูกค้าในแต่ละกลุ่มโดยแบ่งตามฟีเจอร์ Var_1 ในรูปแบบกราฟแท่ง

3.4 **การเตรียมข้อมูล** (Data Preprocessing)

ก่อนการเตรียมข้อมูล ผู้วิจัยแบ่งข้อมูลด้วย Train_Test_Split ที่สัดส่วน 80% สำหรับข้อมูลในการเรียนรู้ได้ข้อมูลทั้งหมด 8,312 ข้อมูลและ 20 % สำหรับข้อมูลในการทดสอบได้ข้อมูลทั้งหมด 2,078 ข้อมูล โดยการเตรียมข้อมูลดังต่อไปนี้ใช้กับชุดข้อมูลสำหรับการเรียนรู้เท่านั้น

3.4.1 **การเปลี่ยนรูปแบบข้อมูลแบบกลุ่มและตัวเลข**

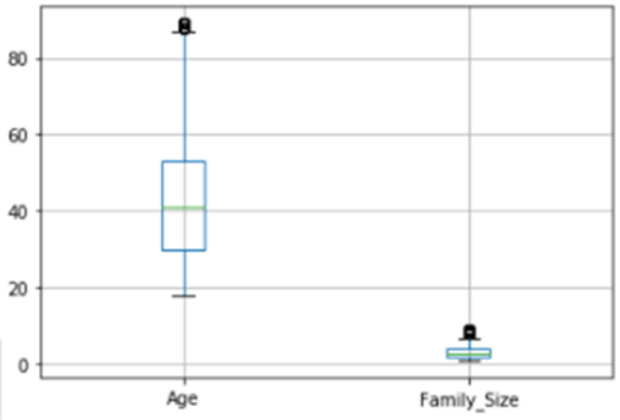
เพื่อเพิ่มประสิทธิภาพในการเรียนรู้ของแบบจำลอง การทำให้ข้อมูลแบบกลุ่มหรือ Categorical Data เปลี่ยนเป็นอยู่ในรูปข้อมูลตัวเลขหรือ Numerical Data เป็นขั้นตอนที่สำคัญช่วยให้แบบจำลองสามารถให้ค่าความน่าจะเป็นได้ง่ายขึ้น ดังนั้นเราจึงพิจารณาทำการปรับเปลี่ยนข้อมูลในฟีเจอร์ที่เก็บแบบกลุ่ม นั่นคือฟีเจอร์ Gender, Ever_Married, Graduated, Profession, Spending_Score และ Var_1 ให้เป็นตัวเลข และสำหรับฟีเจอร์ที่เก็บข้อมูลเป็นตัวเลขนั่นคือฟีเจอร์ Age และ Family_Size พบว่าข้อมูลอยู่ในช่วงที่ต่างกัน ดังนั้นก่อนนำข้อมูลตัวเลขให้แบบจำลองเรียนรู้ต้องมีการปรับเปลี่ยนช่วงของข้อมูลเพื่อประสิทธิภาพในการทำงานของแบบจำลองเช่นเดียวกัน

กระบวนการที่เลือกใช้ชื่อว่า ColumnTransformer เป็นกระบวนการที่สามารถทำการเปลี่ยนแปลงข้อมูลกลุ่มเป็นข้อมูลตัวเลขหรือ One Hot Encoding และสามารถปรับเปลี่ยนช่วงของข้อมูลตัวเลขได้พร้อมกัน สาเหตุที่ต้องทำพร้อมกันเนื่องจากไม่สามารถทำการปรับช่วงข้อมูลตัวเลขทั้งชุดข้อมูลได้ เนื่องจากกังวลว่าจะเกิดปัญหาข้อมูลในชุดทดสอบรั่วไหลหรือว่า Data Leakage ดังนั้นจึงมีการใช้ ColumnTransformer ร่วมกับ Pipeline เพื่อให้สามารถระบุฟีเจอร์ที่ต้องการทำกระบวนการเปลี่ยนแปลงข้อมูลกลุ่มเป็นข้อมูลตัวเลขและปรับช่วงข้อมูลได้อย่างปลอดภัย

สำหรับการเปลี่ยนแปลงข้อมูลกลุ่มเป็นข้อมูลตัวเลขหรือ One Hot Encoding คือการเพิ่มฟีเจอร์ขึ้นมาจากค่าข้อมูล เช่น ฟีเจอร์ Gender เก็บข้อมูล 2 ค่าคือ Female และ Male เมื่อผ่านการเปลี่ยนแปลงข้อมูลจะได้ฟีเจอร์ใหม่ขึ้นมา 2 ฟีเจอร์คือ Gender_Female และ Gender_Male หากลูกค้าคนใดเป็นเพศหญิงจะเก็บค่า 1 ที่ฟีเจอร์ Gender_Female และเก็บค่า 0 ที่ Gender_Male โดยทำการเปลี่ยนแปลงข้อมูลกลุ่มทุกฟีเจอร์

สำหรับการพิจารณาฟีเจอร์ที่เก็บข้อมูลเป็นตัวเลข เราพบว่าค่าของข้อมูลอยู่ในช่วงที่ต่างกัน คือฟีเจอร์อายุที่อยู่ในช่วง 18-89 ปี และจำนวนสมาชิกในครอบครัวหรือฟีเจอร์ Family_Size อยู่ในช่วง 1-9 คนแสดงดังภาพประกอบ 23 การที่ข้อมูลยังไม่ถูกทำให้อยู่ในช่วงเดียวกันนั้น สามารถส่งผลกระทบถึงประสิทธิภาพของอัลกอริทึมที่ใช้ระยะทางในการคำนวณ เช่น K-Nearest Neighbor (KNN) หรือ Support Vector Machine (SVM) ซึ่งในงานวิจัยนี้มีการ

เลือกใช้ SVM ดังนั้นจึงทำการปรับช่วงข้อมูลให้อยู่ในช่วงเดียวกันแสดงดังภาพประกอบ 24 แต่หากเป็นอัลกอริทึมที่ไม่ได้ใช้ระยะทางในการคำนวณ เช่น Random Forest อาจไม่จำเป็นสำหรับการทำขั้นตอนนี้ สามารถทำได้แต่อาจไม่ช่วยในการเพิ่มประสิทธิภาพการทำงาน



ภาพประกอบ 23 แสดงช่วงของข้อมูลในฟีเจอร์ Age และ Family_Size ก่อนการปรับช่วงข้อมูลในรูปแบบแผนภาพกล่อง

กระบวนการที่เลือกใช้คือ StandardScaler เป็นการปรับค่าข้อมูลตัวเลขโดยคำนวณจากค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐาน ดังสมการ (10)

$$x\_scaled_i = \frac{x_i - \mu}{\sigma} \quad (10)$$

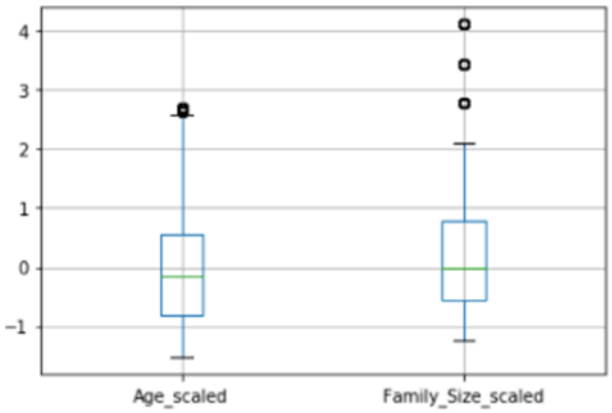
โดยที่

x_scaled_i คือข้อมูล x ในชุดข้อมูลสำหรับเรียนรู้ตัวที่ i ที่ผ่านการปรับค่าข้อมูล

x_i คือข้อมูล x ชุดข้อมูลสำหรับเรียนรู้ตัวที่ i ที่ยังไม่ผ่านการปรับข้อมูล

μ คือค่าเฉลี่ยของชุดข้อมูลสำหรับการเรียนรู้

σ คือส่วนเบี่ยงเบนมาตรฐานของชุดข้อมูลสำหรับการเรียนรู้

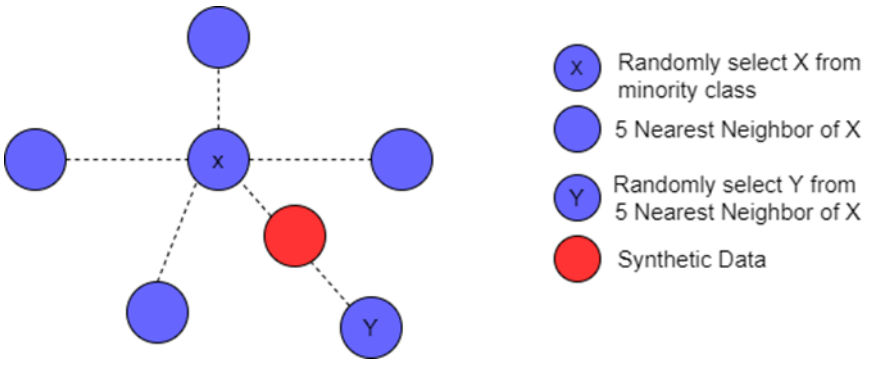


ภาพประกอบ 24 แสดงช่วงของข้อมูลในฟีเจอร์ Age และ Family_Size หลังปรับช่วงข้อมูลในรูปแบบแผนภาพกล่อง

3.4.2 การแก้ไขปัญหาข้อมูลไม่สมดุล

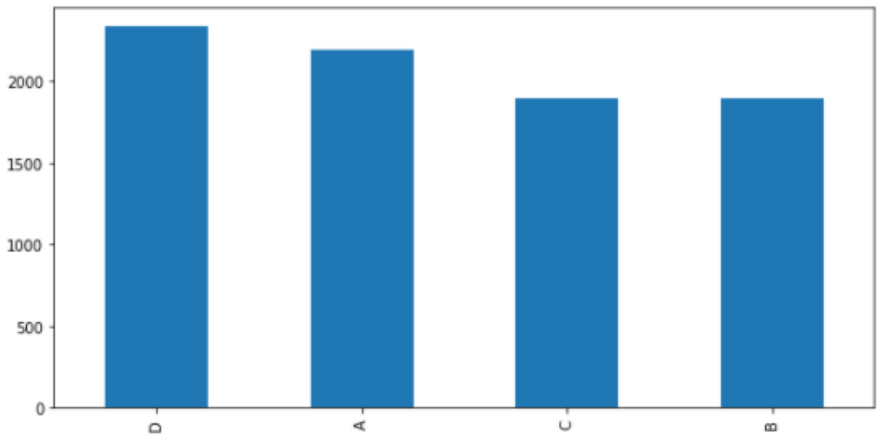
จากการสำรวจข้อมูลข้างต้นเราพบว่าข้อมูลกลุ่ม Segmentation ที่เป็นเลเบลในงานนี้ เกิดปัญหาข้อมูลไม่สมดุลหรือ Imbalanced Data นั่นคือจำนวนลูกค้าในแต่ละกลุ่มมีจำนวนไม่เท่ากัน ซึ่งอาจกระทบถึงประสิทธิภาพในการเรียนรู้ของแบบจำลองได้ นั่นคือแบบจำลองอาจทำนายกลุ่มให้ลูกค้าใหม่ด้วยกลุ่มที่มีจำนวนมากกว่า เพื่อแก้ไขปัญหานี้จึงใช้วิธีการทำให้ข้อมูลลูกค้าในแต่ละกลุ่มมีจำนวนเท่ากันก่อนนำไปให้แบบจำลองเรียนรู้ ซึ่งวิธีการที่เลือกใช้ชื่อว่า Synthetic Minority Oversampling Technique (SMOTE)

การทำงานของ SMOTE ที่แสดงดังภาพประกอบ 25 คือการสังเคราะห์ข้อมูลใหม่ขึ้นมาจากข้อมูลกลุ่มที่มีจำนวนน้อย ซึ่งอาศัยการทำงานของอัลกอริทึมแบบเพื่อนบ้านที่ใกล้ที่สุดหรือ K-Nearest Neighbor วิธีการคือเลือกข้อมูล X แบบสุ่มขึ้นมาจากข้อมูลในกลุ่มที่มีจำนวนน้อยหรือ Minority Class จากนั้นหาเพื่อนบ้านที่ใกล้ที่สุด ในงานนี้เลือกใช้ 5 เพื่อนบ้านที่ใกล้ที่สุด จากนั้นสุ่มข้อมูล Y ที่อยู่ในเพื่อนบ้าน 5 ข้อมูล ข้อมูลใหม่จะถูกสังเคราะห์ระหว่างข้อมูล X และ Y ทำให้ข้อมูลใหม่จะอยู่ระหว่างข้อมูลในกลุ่มที่มีจำนวนน้อยด้วยกันและมีคุณสมบัติคล้ายข้อมูลกลุ่มเดิม

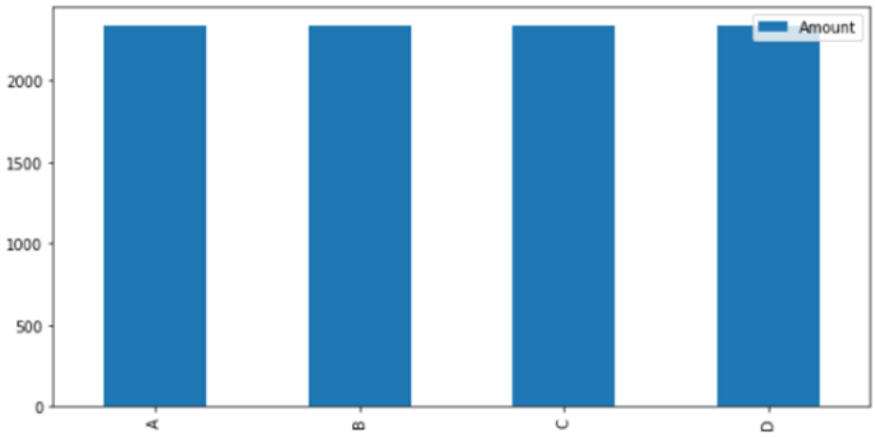


ภาพประกอบ 25 แสดงการสังเคราะห์ข้อมูลใหม่ด้วยวิธีการ SMOTE

ในชุดข้อมูลสำหรับทดสอบแบบจำลองพบว่าแต่ละกลุ่มมีจำนวนลูกค้าไม่เท่ากัน แสดงดังภาพประกอบ 26 โดยที่กลุ่มที่เยอะที่สุดมีจำนวน 2,335 คน และกลุ่มที่น้อยที่สุดมีลูกค้า 1,893 คน หลังจากการทำ SMOTE ทำให้เราได้ข้อมูลเพิ่มเป็นทุกกลุ่มมีจำนวนเท่ากันคือ 2,335 คนดังภาพประกอบ 27 ดังนั้นจำนวนข้อมูลทั้งหมดที่แบบจำลองได้รับเพื่อการเรียนรู้คือ 9,340 ข้อมูล เมื่อข้อมูลในแต่ละกลุ่มมีความสมดุลกันแล้ว จะช่วยลดปัญหาที่แบบจำลองทำนายเป็นกลุ่มที่มีจำนวนลูกค้ามากกว่าได้



ภาพประกอบ 26 แสดงจำนวนข้อมูลของลูกค้าแต่ละกลุ่มชุดที่ให้แบบจำลองเรียนรู้ที่ยังไม่ได้ผ่านการทำ SMOTE



ภาพประกอบ 27 แสดงจำนวนข้อมูลของลูกค้าแต่ละกลุ่มชุดที่ให้แบบจำลองเรียนรู้ที่ผ่านการทำ SMOTE เรียบร้อยแล้ว

3.5 สร้างแบบจำลองเพื่อทำการจัดกลุ่มลูกค้า

เมื่อเตรียมข้อมูลสำหรับการเรียนรู้เรียบร้อยแล้ว ขั้นตอนต่อไปจึงนำข้อมูลเข้าสู่แบบจำลอง โดยใช้ข้อมูลในการเรียนรู้ทั้งหมด 8,312 ข้อมูลและข้อมูลสำหรับการทดสอบทั้งหมด 2,078 ข้อมูล ร่วมกับการทำ Cross Validation ที่ 10 fold เพื่อเลือกชุดข้อมูลที่ดีที่สุดในการเรียนรู้ สำหรับชุดข้อมูลที่ผ่านการแก้ไขปัญหาข้อมูลไม่สมดุลด้วย SMOTE ได้ข้อมูลสำหรับการเรียนรู้ทั้งหมด 9,340 ข้อมูล โดยทำการทดลองตามแบบจำลองที่เลือกใช้คือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest และ Extreme Gradient Boosting (XGBoost) ตามลำดับร่วมกับการปรับจูนพารามิเตอร์

สำหรับแบบจำลอง Logistic Regression ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 3 ตัวดังนี้ คือ

1. C คือส่วนกลับของค่าคงที่ที่กำหนดขนาดของพจน์ Penalty เช่น จากพจน์ของ Penalty ดังสมการที่ 11

$$\alpha\|w\|^2 \tag{11}$$

โดยที่

α	คือค่าคงที่ที่กำหนดขนาดของ $\|w\|^2$
w	คือความชันของแบบจำลอง
$\|w\|^2$	คือพจน์ L2 Regularization

หากค่า α มากหมายความว่ามีการให้ความสำคัญที่พจน์ Penalty มาก ทำให้แบบจำลองลดความยึดติดกับชุดข้อมูลในการเรียนรู้ลง ในทางตรงกันข้ามหากค่า α

น้อยหมายความว่าให้ความสำคัญกับพจน์ Penalty น้อย ทำให้แบบจำลองค่อนข้างยึดติดกับชุดข้อมูลในการเรียนรู้ ในขณะที่ค่า C แสดงดังสมการที่ คือ 12

$$C = \frac{1}{\alpha} \tag{12}$$

คือส่วนกลับของค่าคงที่ α ทำให้ค่า C ที่ได้มีค่าน้อยลง ทำให้ความสำคัญกับพจน์ Penalty ลดลงไปด้วย

2. Penalty คือพจน์ในการช่วยทำให้แบบจำลองลดการเกิดเหตุการณ์ Overfitting กับชุดข้อมูลสำหรับการเรียนรู้ โดยทำการปรับจูนระหว่าง L1 Regularization ดังสมการ 13 หรือ L2 Regularization ดังสมการ 12

$$\alpha\|w\| \tag{13}$$

โดยที่

α	คือค่าคงที่ที่กำหนดขนาดของ $\|w\|$
w	คือความชันของแบบจำลอง
$\|w\|$	คือพจน์ L1 Regularization

3. Solver คือชื่อเรียกอัลกอริทึมที่ต้องการเรียกใช้ในการทำงานของแบบจำลอง โดยในการวิจัยนี้เลือกปรับจูนระหว่าง Newton-cg, Liblinear และ Lbfgs ซึ่งในการใช้อัลกอริทึมแต่ละอันนั้นมีข้อจำกัดเกี่ยวกับการเลือกใช้พจน์ Penalty คือสามารถเลือกใช้ได้บางตัวเท่านั้นดังตารางที่ 4 อ้างอิงจาก (Bishop, 2006)

ตาราง 4 แสดงการเลือกใช้อัลกอริทึมและพจน์ Penalty ที่สามารถใช้ร่วมกันได้

อัลกอริทึม	**พจน์** Penalty
Newton-cg	L2 Regularization หรือไม่ใช้พจน์ Penalty
Liblinear	L1 หรือ L2 Regularization

ตาราง 4 (ต่อ)

อัลกอริทึม	พจน์ Penalty
Lbfgs	L2 Regularization หรือไม่ใช้พจน์ Penalty

และเมื่อทำการปรับจูนพารามิเตอร์ครบทั้ง 3 ตัวแล้ว ได้ผลการปรับจูนออกมาดังตารางที่ 5 โดยปรับจูนทั้งหมด 2 แบบจำลองคือแบบจำลองที่ใช้และไม่ใช้ SMOTE

ตาราง 5 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง Logistic Regression

พารามิเตอร์	แบบจำลองที่ไม่ใช้ SMOTE	แบบจำลองที่ใช้ SMOTE
C	1.0	0.1
Penalty	L1 Regularization	L2 Regularization
Solver	Liblinear	Newton-cg

สำหรับแบบจำลอง Naïve Bayes ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 1 ตัวดังนี้คือ

1. Var_smoothing คือการใส่ค่าเพื่อป้องกันเหตุการณ์ที่ความน่าจะเป็นของบางคลาสเป็น 0 เนื่องจากอาจไม่มีข้อมูลของคลาสนั้นๆในชุดข้อมูลสำหรับการเรียนรู้ เมื่อทำการปรับจูนพารามิเตอร์ทำให้ได้ค่าออกมาดังตารางที่ 6

ตาราง 6 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง Naïve Bayes

พารามิเตอร์	**แบบจำลองที่ไม่ใช้ SMOTE**	**แบบจำลองที่ใช้ SMOTE**
Var_smoothing	0.023	0.035

สำหรับแบบจำลอง Support Vector Machine (SVM) ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 3 ตัวดังนี้

1. C คือการกำหนดขนาดของการทำ Regularization โดยใช้คู่กับ L2 Regularization
2. Kernel คือฟังก์ชันที่ใช้ในการแปลงมิติของข้อมูล โดยเลือกระหว่าง Poly หรือ Rbf
3. Gamma คือสัมประสิทธิ์ของ Kernel

เมื่อทำการปรับจูนพารามิเตอร์ทั้งหมดสำหรับแบบจำลอง SVM ทำให้ได้ผลออกมาดังตารางที่ 7

ตาราง 7 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง SVM

พารามิเตอร์	**แบบจำลองที่ไม่ใช้ SMOTE**	**แบบจำลองที่ใช้ SMOTE**
C	100	100

ตาราง 7 (ต่อ)

พารามิเตอร์	แบบจำลองที่ไม่ใช้ SMOTE	แบบจำลองที่ใช้ SMOTE
Kernel	Rbf	Rbf
Gamma	0.01	0.01

สำหรับแบบจำลอง Random Forest ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 4 ตัวดังนี้คือ

1. N_estimators คือการกำหนดจำนวนต้นไม้ที่ใช้ในการตัดสินใจ
2. Max_features คือการกำหนดจำนวนฟีเจอร์ที่ใช้ในการตัดสินใจ โดยมีทั้งหมด 3 แบบคือ
   - Auto คือไม่มีข้อกำหนด ใช้ทุกฟีเจอร์ที่แบบจำลองมองว่าเหมาะสม
   - Sqrt คือใช้ฟีเจอร์จำนวนรากที่สองของจำนวนฟีเจอร์ทั้งหมด
   - Log2 คือใช้ฟีเจอร์จำนวนลอการิทึมฐานสองของจำนวนฟีเจอร์ทั้งหมด
3. Max_depth คือการกำหนดจำนวนชั้นของต้นไม้ในการตัดสินใจ
4. Criterion คือสูตรที่ใช้ในการวัดคุณภาพของการแบ่งพาร์ทิชัน (Partition) ของต้นไม้ที่ใช้ในการตัดสินใจ โดยมีทั้งหมด 2 ค่าคือ
   - Gini คือค่าที่ใช้ในการวัดความสะอาดของพาร์ทิชันที่ถูกแบ่งโดยใช้ฟีเจอร์หนึ่ง โดยฟีเจอร์ที่ให้ค่า Gini ต่ำหมายความว่ามีความสะอาดมาก
   - Entropy คือค่าวัดความไม่แน่นอนของข้อมูล ซึ่งความไม่แน่นอนหมายถึงจำนวนข้อมูลที่ทำนายผิด ดังนั้นเราต้องการฟีเจอร์ที่ให้ค่า Entropy ต่ำ

เมื่อทำการปรับจูนพารามิเตอร์ทั้งหมดสำหรับแบบจำลอง Random Forest ทำให้ได้ผลออกมาดังตารางที่ 8

ตาราง 8 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง Random Forest

พารามิเตอร์	**แบบจำลองที่ไม่ใช้ SMOTE**	**แบบจำลองที่ใช้ SMOTE**
N_estimators	200	200
Max_features	Auto	Sqrt
Max_depth	8	8
Criterion	Gini	Gini

สำหรับแบบจำลอง Extreme Gradient Boosting ได้ทำการปรับจูนพารามิเตอร์ทั้งหมด 4 ตัวดังนี้ คือ

1. N_estimators คือการกำหนดจำนวนต้นไม้ที่ใช้ในการตัดสินใจ
2. Max_depth คือการกำหนดจำนวนชั้นของต้นไม้ในการตัดสินใจ
3. Learning_rate คือค่าในการกำหนดน้ำหนักของการเปลี่ยนแปลงแบบจำลองใน 1 รอบ

เนื่องจาก XGBoost และ Random Forest เป็นแบบจำลองที่ใช้ต้นไม้ในการตัดสินใจเหมือนกัน ดังนั้นพารามิเตอร์จึงมีความคล้ายกัน เมื่อทำการปรับพารามิเตอร์ทำให้ได้ผลออกมาดังตารางที่ 9

ตาราง 9 แสดงพารามิเตอร์ที่ได้จากการใช้ GridSearchCV ของแบบจำลอง XGBoost

พารามิเตอร์	**แบบจำลองที่ไม่ใช้** SMOTE	**แบบจำลองที่ใช้** SMOTE
N_estimators	50	100
Max_depth	5	7
Learning_rate	0.1	0.1

จากนั้นเมื่อทำการปรับจูนพารามิเตอร์ครบทุกแบบจำลองแล้ว ให้แบบจำลองได้เรียนรู้กับข้อมูลชุดเรียนรู้ต่อไป หลังจากทำการเรียนรู้สำเร็จได้ทำการวัดประสิทธิภาพของแบบจำลองด้วยค่า Accuracy, Precision, Recall, F1-Score และแสดงผลการทำนายทั้งถูกและผิดด้วย Confusion Matrix

บทที่ 4
ผลการดำเนินการวิจัย

ในการวัดประสิทธิภาพของแบบจำลอง ผู้วิจัยใช้ข้อมูลชุดทดสอบทั้งหมด 2,078 แถว ประกอบด้วยลูกค้าจากกลุ่ม D 585 คน ลูกค้าจากกลุ่ม A 546 คน ลูกค้าจากกลุ่ม C 487 คนและลูกค้าจากกลุ่ม B 460 คน จากนั้นวัดประสิทธิภาพการทำงานด้วยค่า Accuracy, Precision, Recall, F1-Score และ Confusion Matrix ได้ตามตารางที่ 10 ดังนี้

ตาราง 10 แสดงผลการวัดประสิทธิภาพจากชุดข้อมูลทดสอบ

ชื่อแบบจำลอง	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	**เวลาที่ใช้ในการเรียนรู้ (วินาที)**
Logistic Regression Without SMOTE	47.11	46.33	47.11	45.02	67.03
Logistic Regression With SMOTE	47.21	46.66	47.21	46.54	45.87
Naïve Bayes Without SMOTE	46.01	44.92	46.01	43.73	**18.51**

ตาราง 10 (ต่อ)

ชื่อแบบจำลอง	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	เวลาที่ใช้ในการเรียนรู้ (วินาที)
Naïve Bayes With SMOTE	46.10	45.11	46.10	44.18	20.07
SVM Without SMOTE	47.88	47.20	47.88	46.91	5722.58
SVM With SMOTE	47.88	47.20	47.88	46.91	5810.13
Random Forest Without SMOTE	47.70	46.51	47.70	46.56	345.89
Random Forest With SMOTE	**48.75**	**48.10**	**48.75**	**48.31**	392.62

ตาราง 10 (ต่อ)

ชื่อแบบจำลอง	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	เวลาที่ใช้ในการเรียนรู้ (วินาที)
XGBoost Without SMOTE	48.60	47.82	48.60	47.85	1716.96
XGBoost With SMOTE	46.68	46.01	46.68	46.20	1880.13

จากตารางที่ 10 ได้ทำการเปรียบเทียบผลการวัดประสิทธิภาพของทุกแบบจำลองพบว่า Random Forest ร่วมกับการใช้ SMOTE ให้ประสิทธิภาพที่ดีที่สุดเมื่อวัดด้วยค่า Accuracy, Precision, Recall และ F1-Score ในขณะที่แบบจำลอง Naïve Bayes ใช้เวลาในการเรียนรู้ที่สั้นที่สุด นอกจากนี้ผู้วิจัยได้ทำการแสดงผล Confusion Matrix เพื่อสังเกตจำนวนความถูกต้องและความผิดพลาดในการทำนายที่เกิดขึ้นดังนี้

ตาราง 11 แสดง Confusion Matrix ของแบบจำลอง Logistic Regression ทั้งแบบใช้และไม่ใช้ SMOTE

ไม่ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**369**	124	65	73
	A	150	286	**188**	107
	B	19	34	61	44
	C	47	102	146	263

ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**355**	105	57	69
	A	142	250	139	76
	B	37	79	112	78
	C	51	112	**152**	264

จากตารางที่ 11 พบว่ากลุ่มที่มีการทำนายถูกมากที่สุดคือกลุ่ม D ของทั้งสองแบบจำลอง ในขณะที่การทำนายผิดมากที่สุดของแบบจำลองที่ไม่ใช้ SMOTE เกิดที่การทำนายจากกลุ่ม B เป็นกลุ่ม A ที่มีจำนวนมากถึง 188 ข้อมูลและสำหรับแบบจำลองที่ใช้ SMOTE เกิดการทำนายผิดมากที่สุดที่กลุ่ม B เป็นกลุ่ม C ที่จำนวน 152 ข้อมูล จากข้อมูลตรงนี้อาจมองได้ว่ากลุ่ม B เกิดความผิดพลาดในการทำนายมากที่สุด

ตาราง 12 แสดง Confusion Matrix ของแบบจำลอง Naïve Bayes ทั้งแบบใช้และไม่ใช้ SMOTE

ไม่ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**375**	147	66	81
	A	102	176	100	46
	B	28	68	80	35
	C	80	155	**214**	325

ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**379**	146	66	80
	A	90	166	91	45
	B	38	78	97	46
	C	78	156	**206**	316

จากตารางที่ 12 พบว่ากลุ่มที่เกิดการทำนายถูกมากที่สุดคือกลุ่ม D ของทั้งสองแบบจำลอง ในขณะที่การทำนายผิดมากที่สุดของแบบจำลองที่ใช้และไม่ใช้ SMOTE เกิดที่การทำนายผิดจากกลุ่ม B เป็นกลุ่ม C มากถึง 214 และ 206 ข้อมูล ซึ่งกลุ่ม B ยังคงเป็นกลุ่มที่เกิดการทำนายผิดมากที่สุด

ตาราง 13 แสดง Confusion Matrix ของแบบจำลอง SVM ทั้งแบบใช้และไม่ใช้ SMOTE

ไม่ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**381**	137	66	73
	A	145	255	**157**	91
	B	30	82	111	79
	C	29	72	126	244

ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**372**	119	59	71
	A	155	281	**172**	98
	B	30	71	97	73
	C	28	75	132	245

จากตารางที่ 13 พบว่ากลุ่มที่เกิดการทำนายถูกมากที่สุดของทั้งสองแบบจำลองยังคงเป็นกลุ่ม D และการทำนายผิดมากที่สุดของทั้งสองแบบจำลองคือการทำนายผิดจากกลุ่ม B เป็นกลุ่ม A ที่จำนวน 157 และ 172 ข้อมูล ซึ่งกลุ่ม B ยังคงพบการทำนายผิดพลาดมากที่สุด

ตาราง 14 แสดง Confusion Matrix ของแบบจำลอง Random Forest ทั้งแบบใช้และไม่ใช้ SMOTE

ไม่ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**382**	136	64	73
	A	133	262	**168**	88
	B	39	68	90	71
	C	31	80	138	255

ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**381**	133	64	75
	A	138	262	**163**	84
	B	33	70	95	75
	C	33	81	138	253

จากตารางที่ 14 พบว่ากลุ่มที่เกิดการทำนายถูกมากที่สุดของทั้งสองแบบจำลองยังคงเป็นกลุ่ม D และการทำนายผิดมากที่สุดของทั้งสองแบบจำลองคือการทำนายผิดจากกลุ่ม B เป็นกลุ่ม A ที่จำนวน 168 และ 163 ข้อมูล

ตาราง 15 แสดง Confusion Matrix ของแบบจำลอง XGBoost ทั้งแบบใช้และไม่ใช้ SMOTE

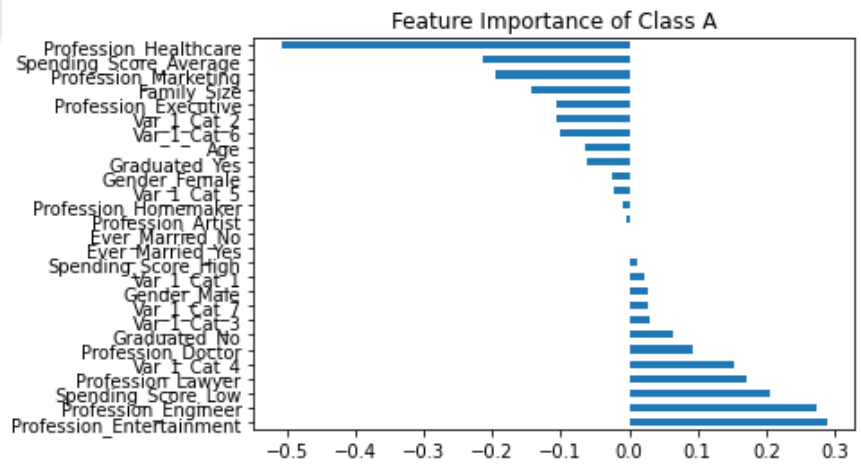
ไม่ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**387**	142	75	73
	A	132	253	**145**	86
	B	41	78	122	80
	C	25	73	118	248

ใช้ SMOTE		Actual			
		D	A	B	C
Predicted	D	**366**	137	72	76
	A	**138**	233	135	78
	B	46	98	130	92
	C	35	78	123	241

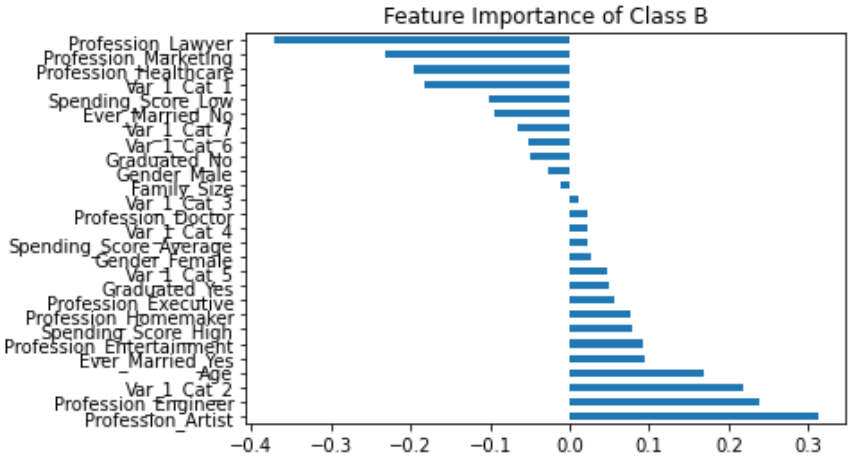
จากตารางที่ 15 พบว่ากลุ่มที่เกิดการทำนายถูกมากที่สุดของทั้งสองแบบจำลองยังคงเป็นกลุ่ม D สำหรับการทำนายผิดมากที่สุดของแบบจำลองไม่ใช้ SMOTE เกิดที่การทำนายผิดจากกลุ่ม B เป็นกลุ่ม A จำนวน 145 ข้อมูลและแบบจำลองที่ใช้ SMOTE เกิดการทำนายผิดจากกลุ่ม D เป็นกลุ่ม A จำนวน 138 ข้อมูล

จากการสำรวจผลจาก Confusion Matrix เห็นได้ว่ากลุ่ม D เกิดการทำนายถูกมากที่สุด แต่ในข้อมูลชุดทดสอบมีข้อมูลของลูกค้าในกลุ่ม D มากที่สุดเช่นกัน จึงอาจเป็นอีกหนึ่งเหตุผลที่ทำให้กลุ่ม D มีจำนวนการทำนายถูกมากที่สุด และยังพบว่ากลุ่มที่เกิดการทำนายผิดมากที่สุดคือกลุ่ม B จากทุก ๆ แบบจำลอง

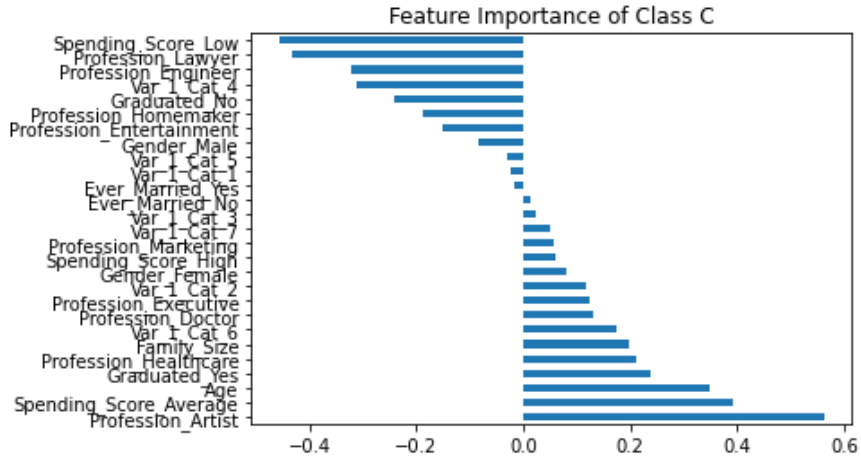
ต่อไปผู้วิจัยได้ทำการแสดงผลความสำคัญของฟีเจอร์ของแต่ละแบบจำลองเพื่อสำรวจว่าแต่ละกลุ่มถูกจัดด้วยฟีเจอร์ใดบ้าง สำหรับแบบจำลอง Logistic Regression ดูความสำคัญของแต่ละฟีเจอร์จากค่าสัมประสิทธิ์พบว่ากลุ่ม A ใช้ฟีเจอร์ Profession_Entertainment มากที่สุด กลุ่ม B และ C ใช้ฟีเจอร์ Profession_Artist มากที่สุดและกลุ่ม D ใช้ฟีเจอร์ Profession_Lawyer มากที่สุดในการพิจารณาจัดกลุ่มโดยเหมือนกันทั้งแบบใช้และไม่ใช้ SMOTE ซึ่งลำดับความสำคัญของฟีเจอร์ทั้งหมดแสดงดังภาพประกอบที่ 28 ถึง 31 ดังนี้



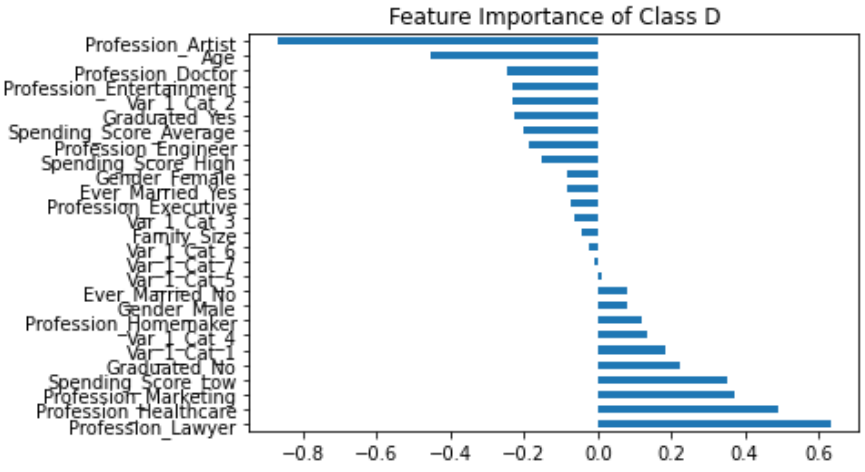
ภาพประกอบ 28 แสดงความสำคัญของฟีเจอร์ในการจัดกลุ่ม A ของแบบจำลอง Logistic Regression



ภาพประกอบ 29 แสดงความสำคัญของฟีเจอร์ในการจัดกลุ่ม B ของแบบจำลอง Logistic Regression

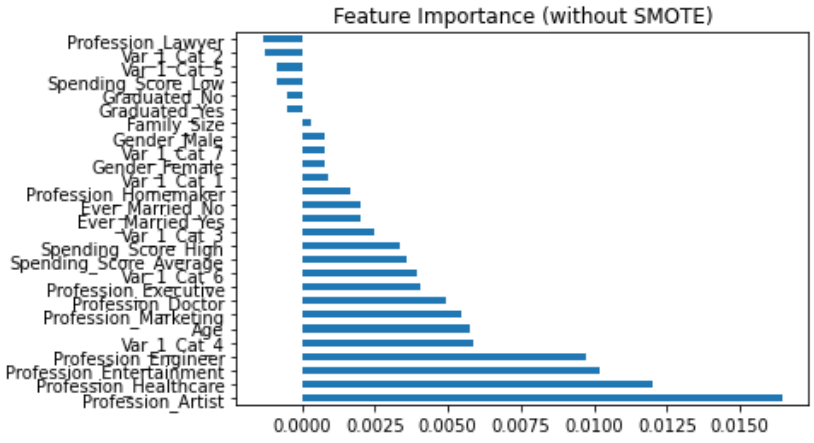


ภาพประกอบ 30 แสดงความสำคัญของฟีเจอร์ในการจัดกลุ่ม C ของแบบจำลอง Logistic Regression

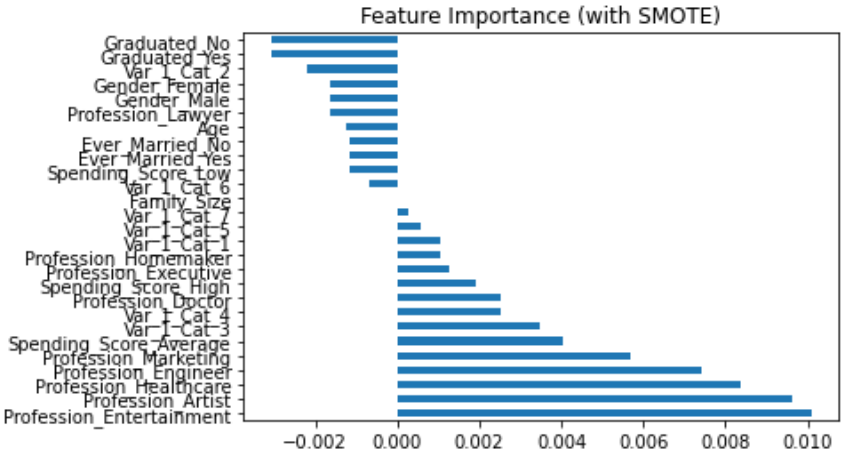


ภาพประกอบ 31 แสดงความสำคัญของฟีเจอร์ในการจัดกลุ่ม D ของแบบจำลอง Logistic Regression

สำหรับแบบจำลอง Naïve Bayes แบบไม่ใช้ SMOTE พบว่าฟีเจอร์ที่สำคัญในการจัดกลุ่มคือฟีเจอร์ Profession_Artist และฟีเจอร์ที่สำคัญสำหรับแบบจำลองที่ใช้ SMOTE คือ Profession_Entertainment โดยลำดับความสำคัญของฟีเจอร์ทั้งหมดแสดงดังภาพประกอบ 32 และ 33 ซึ่งเห็นได้ว่าค่าความสำคัญของแต่ละฟีเจอร์ค่อนข้างน้อยมาก ฟีเจอร์ที่ให้ค่าสูงสุดมีค่าเพียง 0.01 โดยประมาณ นี่อาจเป็นสาเหตุที่ทำให้แบบจำลอง Naïve Bayes ให้ประสิทธิภาพที่ต่ำที่สุดเมื่อเทียบกับแบบจำลองอื่น



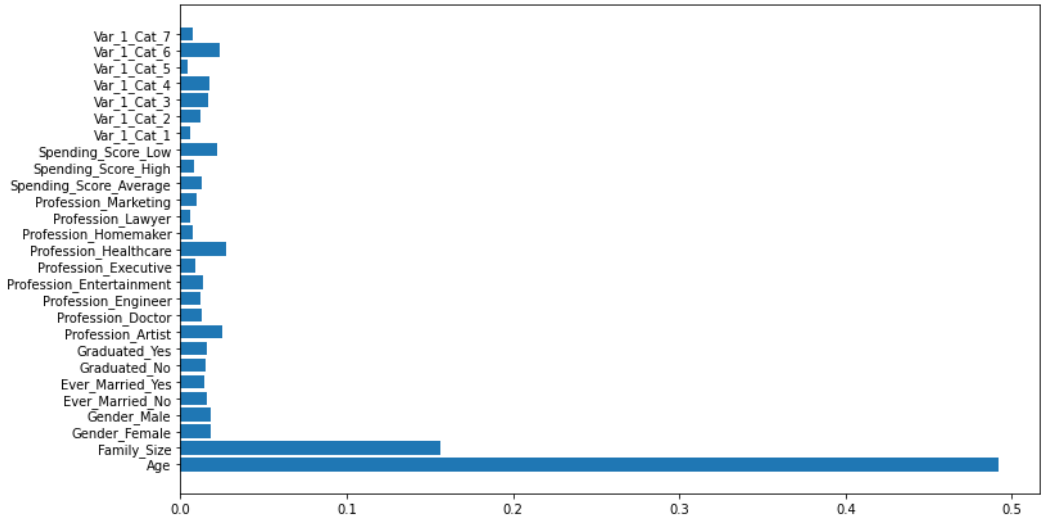
ภาพประกอบ 32 แสดงความสำคัญของฟีเจอร์ของแบบจำลอง Naïve Bayes ที่ไม่ใช้ SMOTE



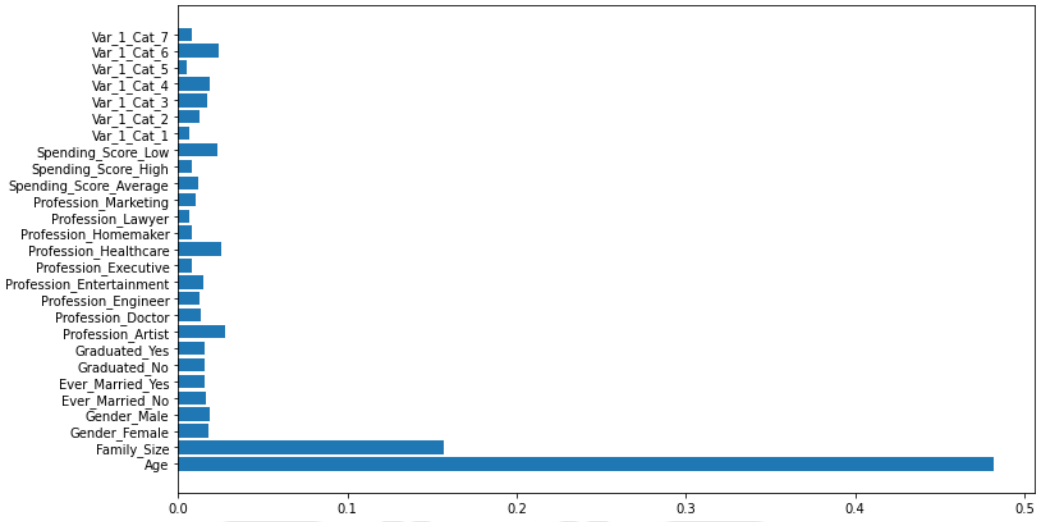
ภาพประกอบ 33 แสดงความสำคัญของฟีเจอร์ของแบบจำลอง Naïve Bayes ที่ใช้ SMOTE

สำหรับแบบจำลอง SVM เนื่องจากการทำ GridSearchCV ทำให้ได้พารามิเตอร์ Kernel คือ Rbf ซึ่งเป็นฟังก์ชั่นที่ไม่ใช่เชิงเส้น (Non-Linear) คือมีการเปลี่ยนแปลงมิติของฟีเจอร์ไป ทำให้ไม่สามารถแสดงความสำคัญของแต่ละฟีเจอร์ได้

สำหรับแบบจำลอง Random Forest ทั้งแบบใช้และไม่ใช้ SMOTE พบว่าฟีเจอร์ที่สำคัญต่อการจัดกลุ่มคือ Age ตามด้วย Family_Size ซึ่งค่อนข้างแตกต่างจากแบบจำลองข้างต้นที่เน้นการใช้ฟีเจอร์เกี่ยวกับอาชีพ เมื่อพิจารณาประกอบกับการสำรวจข้อมูลจากบทที่ 2 เห็นได้ว่าแต่ละกลุ่มอยู่ในช่วงอายุที่ค่อนข้างแตกต่างกันและมีจำนวนสมาชิกในครอบครัวค่อนข้างแตกต่างกันเช่นกัน จึงอาจเป็นสาเหตุให้สองฟีเจอร์นี้สามารถจำแนกกลุ่มของลูกค้าได้แม่นยำเมื่อเทียบกับแบบจำลองอื่นๆ โดยลำดับความสำคัญของแต่ละฟีเจอร์แสดงดังภาพประกอบ 34 และ 35 ดังนี้

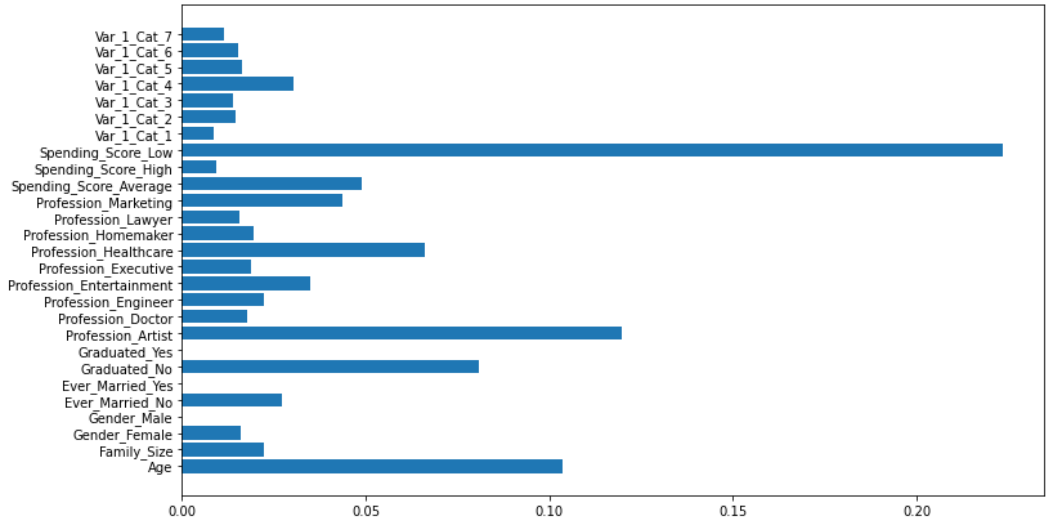


ภาพประกอบ 34 แสดงความสำคัญของฟีเจอร์ของแบบจำลอง Random Forest ที่ไม่ใช้ SMOTE

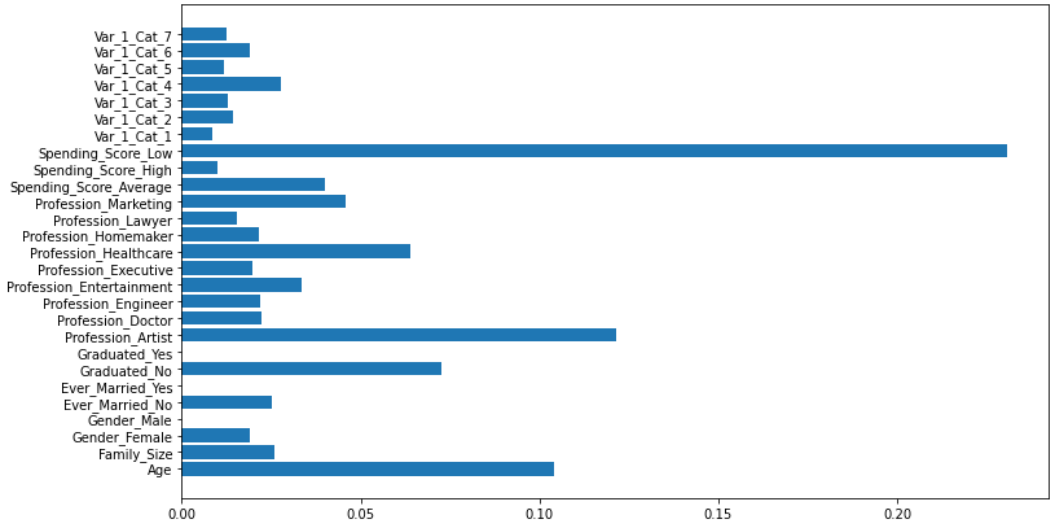


ภาพประกอบ 35 แสดงความสำคัญของฟีเจอร์ของแบบจำลอง Random Forest ที่ใช้ SMOTE

สำหรับแบบจำลอง XGBoost ทั้งแบบใช้และไม่ใช้ SMOTE พบว่าฟีเจอร์ที่สำคัญต่อการจัดกลุ่มมากที่สุดเกี่ยวกับการใช้จ่ายของลูกค้าคือ Spending_Score_Low ตามมาด้วยฟีเจอร์เกี่ยวกับอาชีพคือ Profession_Artist โดยลำดับความสำคัญของฟีเจอร์ทั้งหมดแสดงดังภาพประกอบ 36 และ 37 ดังนี้



ภาพประกอบ 36 แสดงความสำคัญของฟีเจอร์ของแบบจำลอง XGBoost ที่ไม่ใช้ SMOTE

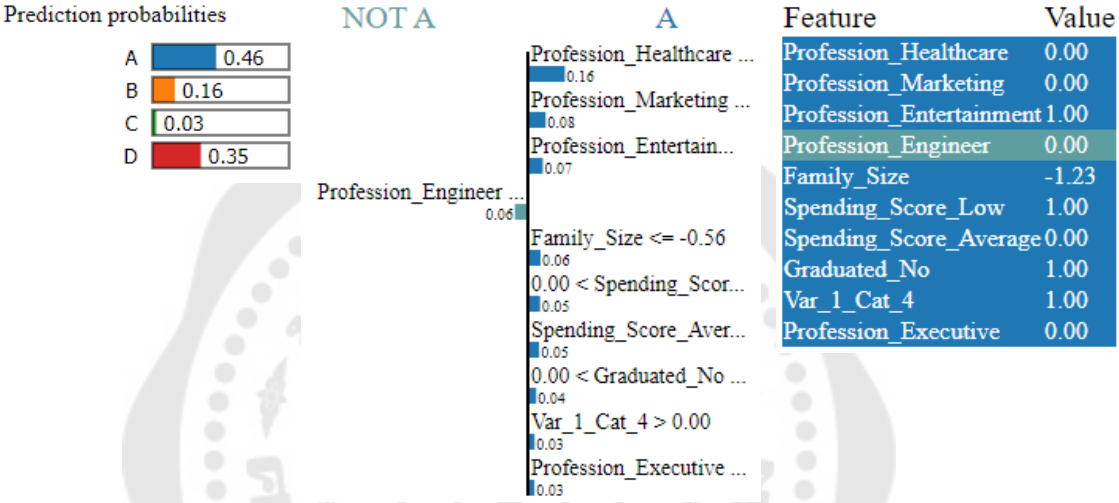


ภาพประกอบ 37 แสดงความสำคัญของฟีเจอร์ของแบบจำลอง XGBoost ที่ใช้ SMOTE

เมื่อสำรวจความสำคัญของฟีเจอร์ทุกแบบจำลองแล้วพบว่าส่วนใหญ่ฟีเจอร์ที่สำคัญต่อการจัดกลุ่มคือฟีเจอร์เกี่ยวกับอาชีพ และ Random Forest ที่ใช้ฟีเจอร์เกี่ยวกับอายุและจำนวนสมาชิกในครอบครัว

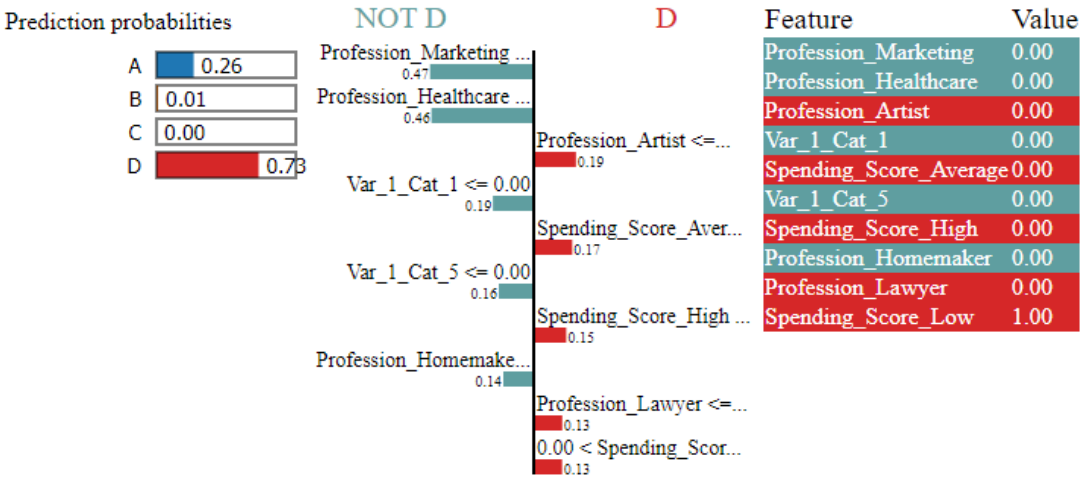
จากนั้นเพื่อทำการวิเคราะห์ความผิดพลาดที่เกิดขึ้น ผู้วิจัยเลือกใช้ไลบารี่ LIME (Local Interpretable Model-agnostic Explanations) ในการหาการใช้งานฟีเจอร์บนข้อมูล 1 ตัวและเปรียบเทียบแต่ละแบบจำลองเพื่อหาว่าในการทำนายนั้นมีฟีเจอร์ใดบ้างที่ถูกนำมาใช้พิจารณา โดยผู้วิจัยทำการเลือกข้อมูลแบบสุ่มมา 1 ข้อมูลจากข้อมูลชุดทดสอบ ได้เป็นข้อมูลที่ 163 ที่พบ

การทำนายผิดเกิดขึ้น ซึ่งเลเบลของข้อมูลนี้คือกลุ่ม B แต่เมื่อทำการสำรวจที่การทำงานของแบบจำลองพบว่า Logistic Regression ทำนายเป็นกลุ่ม A ด้วยความน่าจะเป็นถึง 0.46 ในขณะที่ความน่าจะเป็นในการทำนายได้กลุ่ม B มีเพียง 0.16 เมื่อพิจารณาการใช้ฟีเจอร์ในฝั่ง NOT A พบว่ามีเพียงฟีเจอร์ Profession_Engineer ซึ่งให้ค่าน้ำหนักที่ 0.06 ทำให้อาจไม่เพียงพอให้แบบจำลองพิจารณาว่าเป็นกลุ่ม B ได้ โดยความน่าจะเป็นและการใช้ฟีเจอร์ของแบบจำลอง Logistic Regression แสดงทั้งหมดดังภาพประกอบ 38 ดังนี้



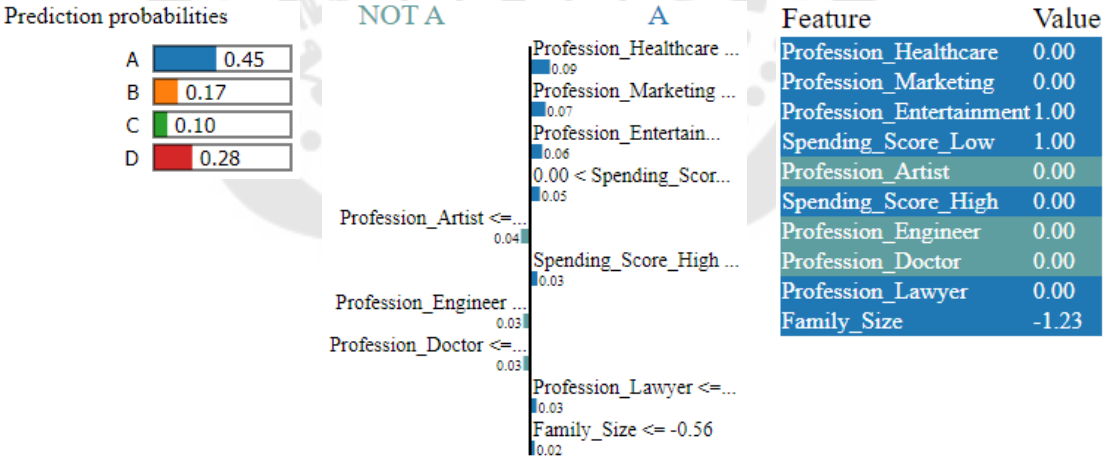
ภาพประกอบ 38 แสดงผลจากการใช้ LIME สำหรับแบบจำลอง Logistic Regression

สำหรับแบบจำลอง Naïve Bayes เมื่อแสดงผลการทำงานของแบบจำลองพบว่าแบบจำลองทำนายข้อมูลนี้ว่าเป็นกลุ่ม B ซึ่งเป็นเลเบลจริงด้วยความน่าจะเป็นเพียง 0.01 ในขณะที่ทำนายว่าเป็นกลุ่ม D ด้วยความน่าจะเป็นถึง 0.73 เมื่อพิจารณาฟีเจอร์ในฝั่ง NOT D พบว่าฟีเจอร์ส่วนใหญ่เกี่ยวกับอาชีพและกลุ่มที่ทางบริษัทให้ไว้แบบนิรนาม โดยความน่าจะเป็นและการใช้ฟีเจอร์ของแบบจำลอง Naïve Bayes แสดงทั้งหมดดังภาพประกอบ 39



ภาพประกอบ 39 แสดงผลจากการใช้ LIME สำหรับแบบจำลอง Naïve Bayes

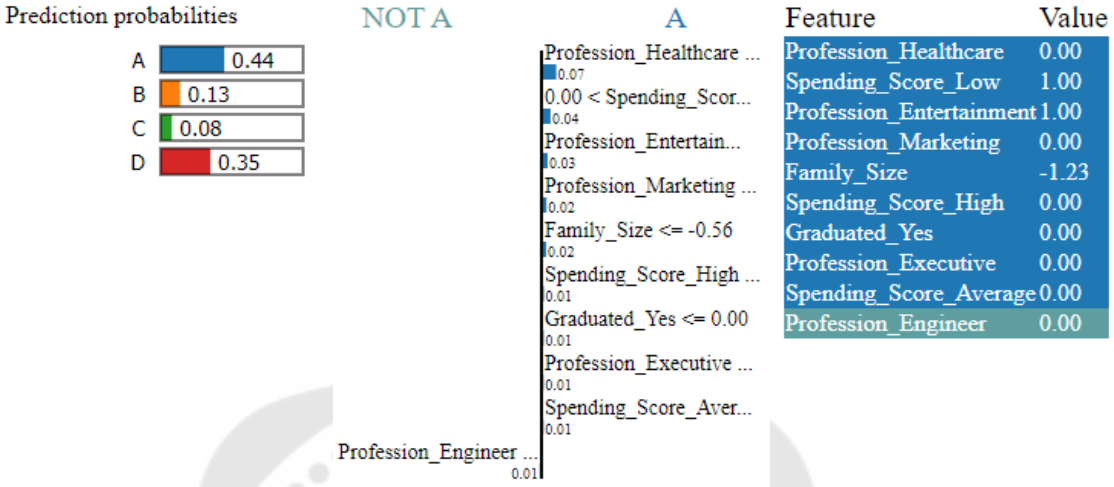
สำหรับแบบจำลอง SVM เมื่อแสดงผลการทำงานของแบบจำลองพบว่าแบบจำลองทำนายข้อมูลนี้ว่าเป็นกลุ่ม A ด้วยความน่าจะเป็น 0.45 ในขณะที่ทำนายเป็นกลุ่ม B ซึ่งเป็นเลเบลจริงด้วยความน่าจะเป็น 0.17 จากนั้นทำการสำรวจฟีเจอร์โดยฟีเจอร์และความน่าจะเป็นทั้งหมดแสดงดังภาพประกอบ 40 สามารถเห็นได้ว่าฟีเจอร์ส่วนใหญ่ที่นำมาใช้มีผลต่อการทำนายเป็นกลุ่ม A



ภาพประกอบ 40 แสดงผลจากการใช้ LIME สำหรับแบบจำลอง SVM

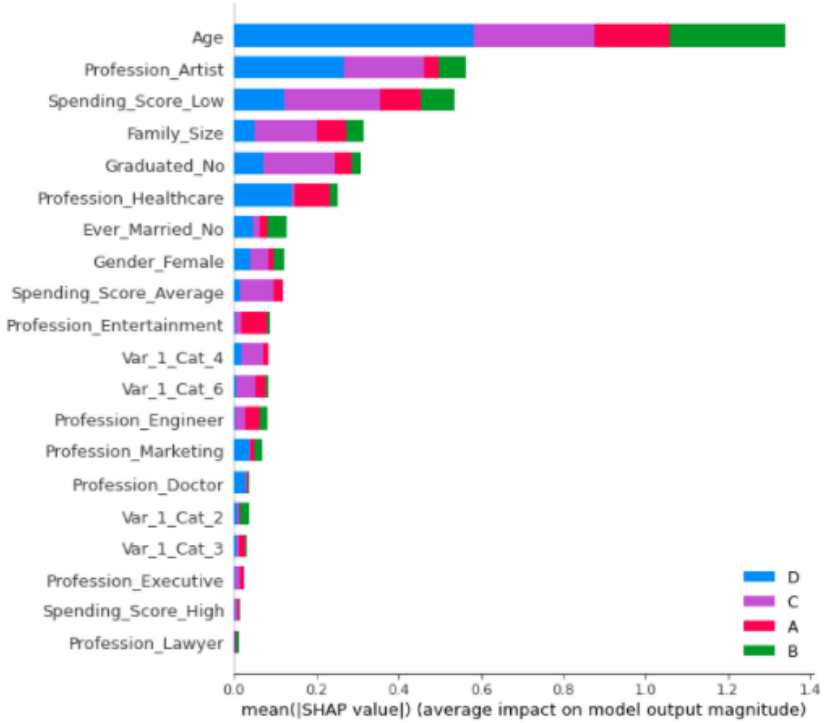
สำหรับแบบจำลอง Random Forest เมื่อแสดงผลการทำงานของแบบจำลองพบว่าแบบจำลองทำนายข้อมูลนี้ว่าเป็นกลุ่ม A ด้วยความน่าจะเป็น 0.44 ในขณะที่ทำนายเป็นกลุ่ม B ซึ่งเป็นเลเบลจริงด้วยความน่าจะเป็นเพียง 0.13 เมื่อสำรวจการใช้ฟีเจอร์แสดงดังภาพประกอบที่

41 พบว่ามีเพียงฟีเจอร์ Profession_Engineer ด้วยน้ำหนัก 0.01 ในฝั่ง NOT B ซึ่งอาจทำให้ไม่เพียงพอต่อแบบจำลองในการทำนายเป็นกลุ่ม B ได้



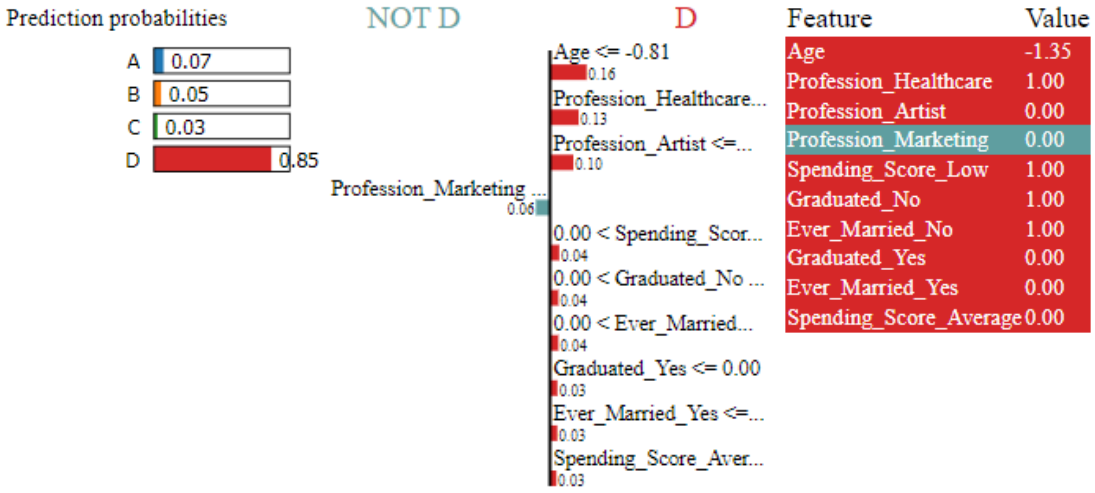
ภาพประกอบ 41 แสดงผลจากการใช้ LIME สำหรับแบบจำลอง Random Forest

สำหรับแบบจำลอง XGBoost เกิดปัญหาบางอย่างเนื่องจากเมื่อทำการเรียนรู้ของแบบจำลองทำให้จำนวนฟีเจอร์เปลี่ยนไปไม่สามารถใช้ LIME ได้ ผู้วิจัยจึงเปลี่ยนมาใช้ SHAP หรือ Shapley Additive Explanations ซึ่งเป็นอีกเครื่องมือหนึ่งในการอธิบายการทำนายของแบบจำลองได้ และเนื่องจาก XGBoost เป็นแบบจำลองที่ใช้ต้นไม้ในการตัดสินใจ ทำให้สามารถใช้งานกับคำสั่ง SHAP ดังต่อไปนี้ได้ โดยจากภาพประกอบ 42 แสดงถึงฟีเจอร์ที่มีอิทธิพลต่อแต่ละกลุ่ม จากภาพประกอบเห็นได้ว่าฟีเจอร์ Age ให้ค่าสูงที่สุดหมายความว่าเป็นฟีเจอร์ที่มีความสำคัญในการจัดกลุ่มมากที่สุด โดยที่มีอิทธิพลต่อการจัดกลุ่ม D ที่มากที่สุด ตามด้วยกลุ่ม C และ B ในค่าที่ใกล้เคียงกันและมีอิทธิพลต่อกลุ่ม A น้อยที่สุด หรืออีกฟีเจอร์ที่ให้ค่าสูงรองจาก Age คือ Profession_Artist ที่มีอิทธิพลต่อกลุ่ม D มากที่สุดตามด้วยกลุ่ม C, B และ A



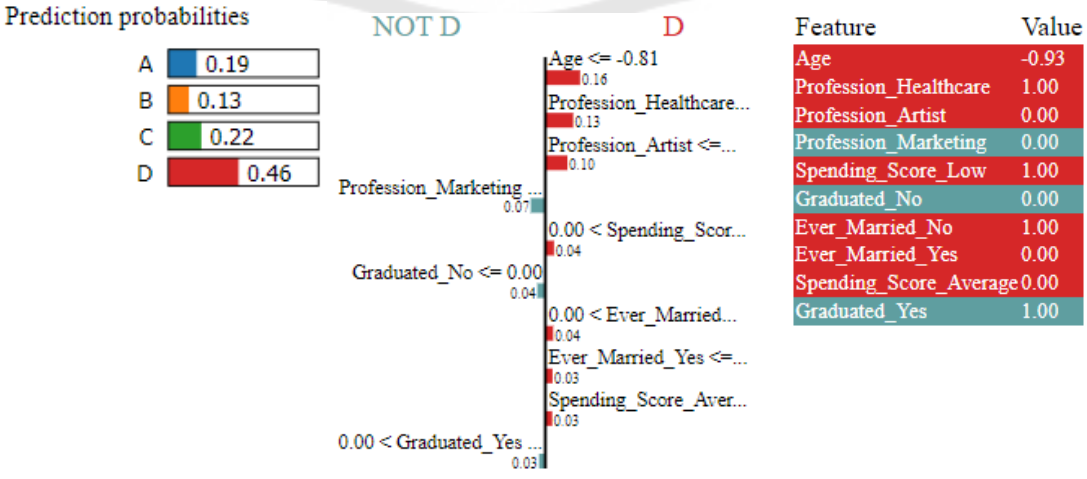
ภาพประกอบ 42 แสดงผลจากการใช้ SHAP สำหรับแบบจำลอง XGBoost

ขั้นตอนต่อไปผู้วิจัยต้องการพิจารณาผลจากการใช้ LIME ว่ามีความสอดคล้องกับผลจากค่าความสำคัญของฟีเจอร์หรือ Feature Importance หรือไม่ โดยจากผลการวัดประสิทธิภาพพบว่าแบบจำลองที่ดีที่สุดคือ Random Forest ร่วมกับการใช้ SMOTE ดังนั้นผู้วิจัยจึงเลือกแบบจำลองนี้มาใช้ในการพิจารณาด้วยการเลือกข้อมูลที่มีการทำนายถูกจากข้อมูลชุดทดสอบมาแสดงเพิ่มเติม ซึ่งจากผลค่าความสำคัญของฟีเจอร์สำหรับแบบจำลอง Random Forest พบว่าฟีเจอร์ที่ให้ค่าสูงที่สุดคือฟีเจอร์ Age หรืออายุของลูกค้าตามด้วยฟีเจอร์ Family_Size หรือจำนวนสมาชิกภายในครอบครัว ข้อมูลแรกที่เลือกมาคือข้อมูลที่ 2,003 ที่พบการทำนายถูกที่กลุ่ม D จากภาพประกอบที่ 43 พบว่าฟีเจอร์ที่ถูกนำมาใช้ในการทำนายเป็นกลุ่ม B มากที่สุด 3 อันดับแรกคือ Age, Profession_Healthcare และ Profession_Artist ซึ่งมีความสอดคล้องกับค่าความสำคัญของฟีเจอร์ รวมทั้งยังสอดคล้องกับการสำรวจข้อมูลที่ผ่านมาที่เราพบว่าฟีเจอร์เกี่ยวกับอายุและอาชีพค่อนข้างแสดงความแตกต่างของกลุ่ม D ออกจากกลุ่มอื่นได้ดีอีกด้วย นี่อาจเป็นสาเหตุทำให้ค่าความน่าจะเป็นในการทำนายเป็นกลุ่ม D สูงถึง 0.85



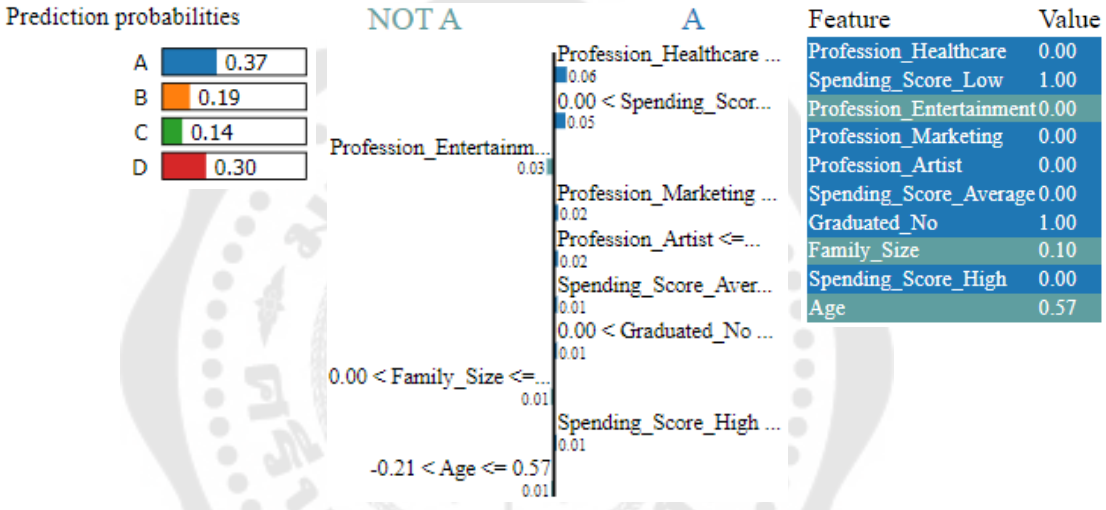
ภาพประกอบ 43 แสดงผลจากการใช้ LIME ของข้อมูลที่ 2,003 ในข้อมูลชุดทดสอบ

ต่อไปทำการเลือกข้อมูลที่ 1,114 ที่พบการทำนายถูกที่กลุ่ม D เช่นกัน จากภาพประกอบที่ 44 แสดงฟีเจอร์ที่ใช้ในการทำนายและค่าความน่าจะเป็นในการทำนายพบว่าฟีเจอร์ที่ถูกนำมาใช้มากที่สุดยังคงเป็น Age, Profession_Healthcare และ Profession_Artist แต่ในฝั่ง NOT D พบฟีเจอร์ Graduated_No และ Graduated_Yes ที่แตกต่างจากข้อมูลที่ 2,003 ซึ่ง 2 ฟีเจอร์ที่เพิ่มขึ้นมาในฝั่ง NOT D นี้ทำให้ความน่าจะเป็นในการทำนายเป็นกลุ่ม D ลดลงจาก 0.85 เหลือ 0.46 ซึ่งจากการสำรวจข้อมูลเราพบว่าฟีเจอร์การจบการศึกษาของลูกค้าเป็นฟีเจอร์ที่แสดงความแตกต่างระหว่างกลุ่ม D และกลุ่มอื่น ๆ ด้วยเช่นกัน การที่ฟีเจอร์การจบการศึกษาถูกนำไปใช้ในการทำนายฝั่ง NOT D อาจทำให้ขาดฟีเจอร์ที่ช่วยในการจำแนกกลุ่ม D ไป



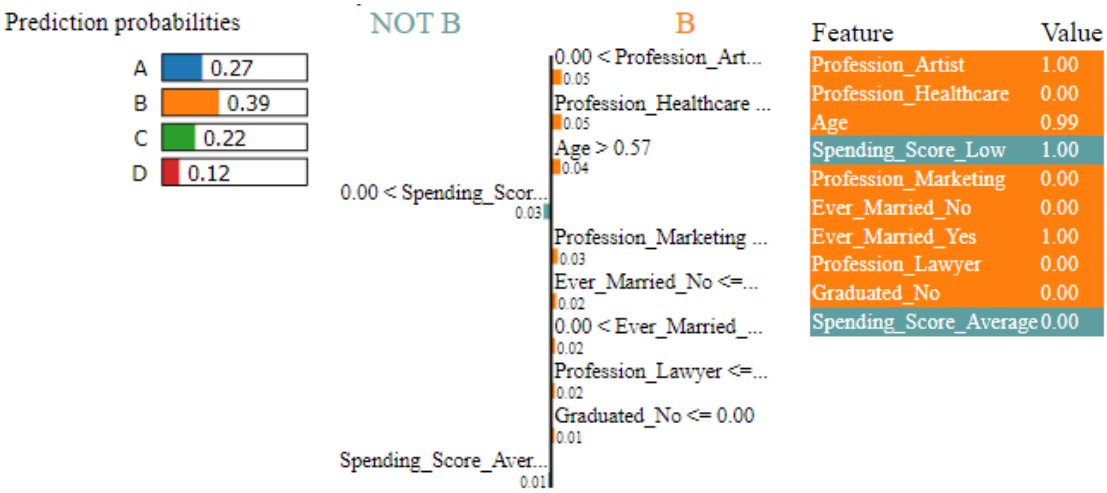
ภาพประกอบ 44 แสดงผลจากการใช้ LIME ของข้อมูลที่ 1,114 ในข้อมูลชุดทดสอบ

ต่อไปทำการเลือกข้อมูลที่มีการทำนายถูกที่กลุ่มอื่นพิจารณาร่วมด้วยคือข้อมูลที่ 550 จากการพิจารณาฟีเจอร์และความน่าจะเป็นตามภาพประกอบ 45 พบว่าฟีเจอร์ที่มีค่าสูงที่สุดคือ Profession_Healthcare, Spending_Score_Low และ Profession_Entertainment ซึ่งทั้ง 3 ฟีเจอร์นี้ไม่สอดคล้องกับค่าความสำคัญของฟีเจอร์เท่าที่ควร รวมถึงจากการสำรวจพบว่าลูกค้ากลุ่ม A ส่วนใหญ่ประกอบอาชีพด้านศิลปิน ในขณะที่แบบจำลองให้ความสำคัญกับอาชีพด้านสายสุขภาพมากกว่า ซึ่งเมื่อพิจารณาความน่าจะเป็นพบว่าความน่าจะเป็นในการทำนายเป็นกลุ่ม A และ D ค่อนข้างใกล้เคียงกัน อาจเกิดจากที่ความเป็นจริงกลุ่มที่สอดคล้องกับอาชีพด้ายสายสุขภาพคือกลุ่ม D มากกว่ากลุ่ม A



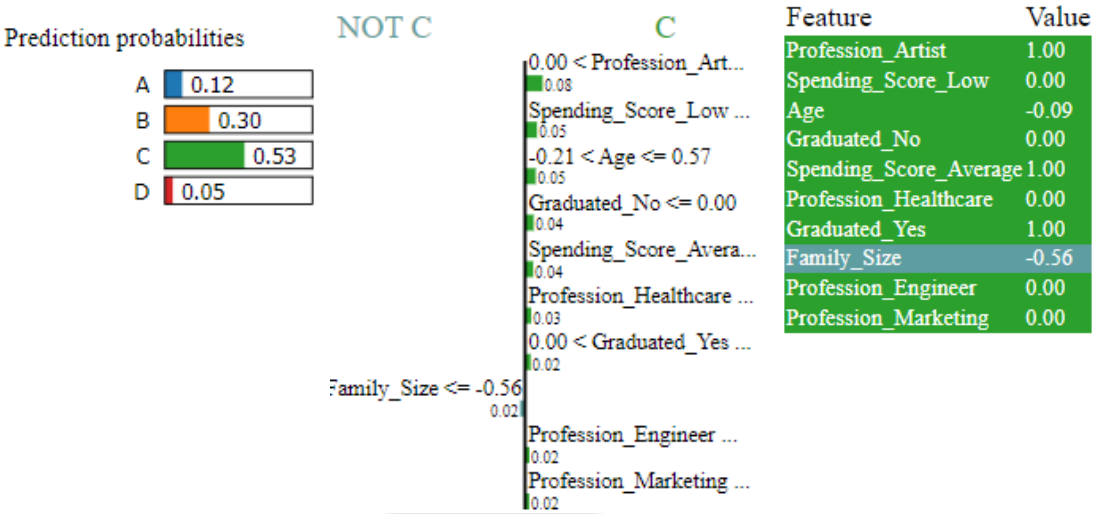
ภาพประกอบ 45 แสดงผลจากการใช้ LIME ของข้อมูลที่ 550 ในข้อมูลชุดทดสอบ

ข้อมูลถัดไปที่นำมาแสดงคือข้อมูลที่ 8 ซึ่งมีการทำนายถูกที่กลุ่ม B จากการพิจารณาฟีเจอร์และความน่าจะเป็นตามภาพประกอบ 46 พบว่าฟีเจอร์ที่ถูกนำมาใช้ในการทำนายว่าเป็นกลุ่ม B ด้วยค่าที่สูงที่สุดคือ Profession_Artist, Profession_Healthcare และ Age ซึ่งเมื่อเปรียบเทียบกับค่าความสำคัญของฟีเจอร์พบว่า Profession_Artist ไม่ใช่ฟีเจอร์ที่ได้ค่าสูง อาจเป็นเพราะค่าความสำคัญของฟีเจอร์นั้นมาจากข้อมูลชุดสำหรับการเรียนรู้ทั้งชุด แต่หากพิจารณาจากการสำรวจข้อมูลพบว่าลูกค้าส่วนใหญ่ในกลุ่ม B ประกอบอาชีพด้านศิลปิน ดังนั้นจึงมองว่าการที่แบบจำลองเลือกใช้ฟีเจอร์ Profession_Artist เป็นหลักจึงค่อนข้างมีความสมเหตุสมผล



ภาพประกอบ 46 แสดงผลจากการใช้ LIME ของข้อมูลที่ 8 ในข้อมูลชุดทดสอบ

ข้อมูลถัดไปที่เลือกมาแสดงคือข้อมูลที่ 210 ซึ่งเกิดการทำนายถูกที่กลุ่ม C จากฟีเจอร์และความน่าจะเป็นที่แสดงดังภาพประกอบ 47 พบว่าฟีเจอร์ที่ถูกนำมาใช้ในการทำนายเป็นกลุ่ม C มากที่สุดได้แก่ฟีเจอร์ Profession_Artist, Spending_Score_Low, Age, Graduated_No, Spending_Score_Average, Profession_Healthcare และ Graduated_Yes เช่นเดียวกับข้อมูลที่ 5 ที่พบว่า Profession_Artist อาจไม่ใช่ฟีเจอร์ที่มีค่าสูงที่สุดจากการพิจารณาค่าความสำคัญของฟีเจอร์ แต่เนื่องจากการสำรวจพบว่าลูกค้าส่วนใหญ่ในกลุ่ม C ประกอบอาชีพด้านศิลปิน จึงมองว่ามีความสมเหตุสมผล ในขณะที่จากการสำรวจพบว่าลูกค้าส่วนใหญ่ในกลุ่ม C อยู่ในระดับการใช้จ่ายปานกลางแต่ผลของ LIME แสดงว่าแบบจำลองให้ความสำคัญกับระดับการใช้จ่ายต่ำมากกว่า ซึ่งค่อนข้างขัดแย้งกับสิ่งที่พบจากการสำรวจ เมื่อตรวจสอบร่วมกับอีกหลายๆข้อมูลที่มีการทำนายถูกที่กลุ่ม C พบว่าใช้ฟีเจอร์ที่คล้ายกันจึงมองว่าหากแบบจำลองให้ความสำคัญกับระดับการใช้จ่ายปานกลางมากกว่าอาจทำให้สามารถทำนายได้ดียิ่งขึ้น



ภาพประกอบ 47 แสดงผลจากการใช้ LIME ของข้อมูลที่ 210 ในข้อมูลชุดทดสอบ

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

เนื่องจากการแข่งขันทางธุรกิจที่มากขึ้นทำให้ลูกค้ามีตัวเลือกที่มากขึ้นเช่นกัน ธุรกิจจึงจำเป็นต้องพัฒนาการตลาดเพื่อรักษาและค้นหาลูกค้า ผู้วิจัยมองว่าการจัดกลุ่มลูกค้าเป็นอีกวิธีหนึ่งที่ช่วยให้ธุรกิจสามารถเข้าใจและเข้าถึงลูกค้าได้จึงจัดทำวิจัยนี้ขึ้นมาเพื่อศึกษาและเปรียบเทียบการทำงานของแบบจำลองเพื่อทำนายกลุ่มของลูกค้าบริษัทยานยนต์แห่งหนึ่งซึ่งใช้ข้อมูลประเภทประชากร โดยคาดหวังว่าแบบจำลองสามารถนำไปใช้ในการจัดกลุ่มลูกค้าใหม่ของบริษัทในอนาคตได้ ผู้วิจัยได้ทำการสอนแบบจำลองและวัดประสิทธิภาพการทำงานของแบบจำลองและทำการสรุปผลแบ่งตามหัวข้อดังนี้

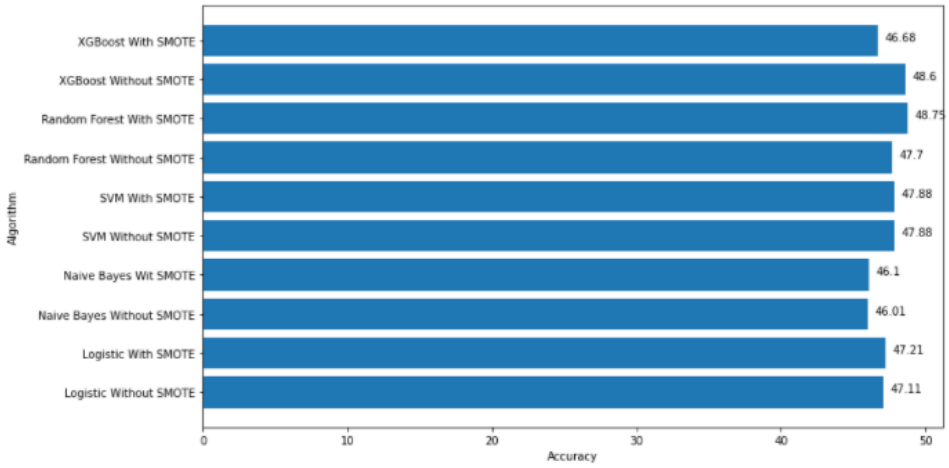
5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการวิจัย

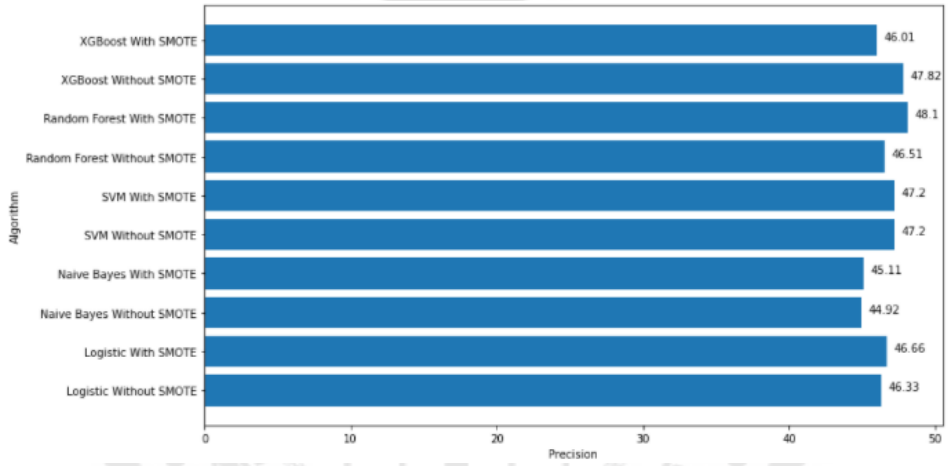
5.3 ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

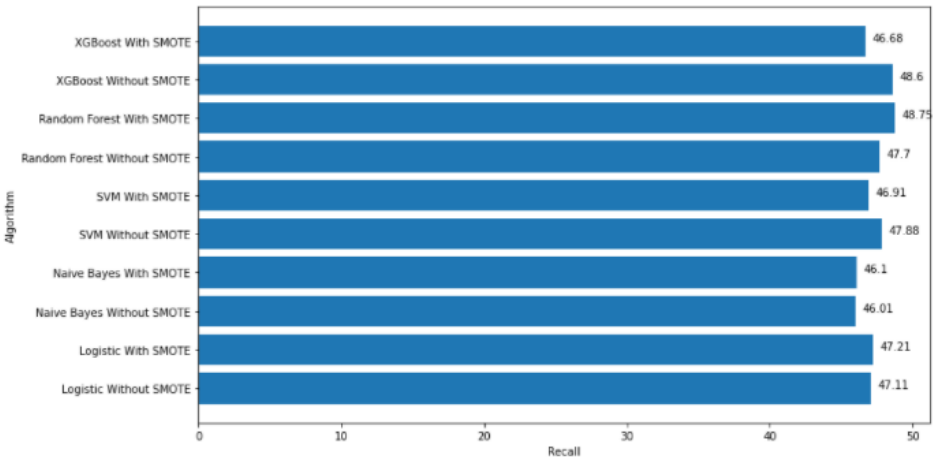
งานวิจัยนี้ศึกษาการจัดกลุ่มลูกค้าบริษัทยานยนต์แห่งหนึ่งซึ่งข้อมูลที่ใช้คือข้อมูลประเภทประชากรของลูกค้า โดยใช้เทคนิคการเรียนรู้แบบมีผู้สอนหรือ Supervised Learning พร้อมทั้งทำการเปรียบเทียบประสิทธิภาพการทำงานของแบบจำลองทั้งหมด 5 แบบจำลองคือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest และ Extreme Gradient Boosting ซึ่งเป็นแบบจำลองที่เหมาะสมกับงานแยกประเภทของข้อมูลร่วมกับการปรับจูนพารามิเตอร์ด้วย GridSearchCV และทำการจัดการกับปัญหาข้อมูลไม่สมดุลหรือ Imbalanced Data ด้วย SMOTE เมื่อวัดประสิทธิภาพด้วยค่า Accuracy, Precision, Recall และ F1-Score พบว่าแบบจำลอง Random Forest ร่วมกับการใช้ SMOTE ให้ผลลัพธ์ที่ดีที่สุดที่ Accuracy 48.75%, Precision 48.10%, Recall 48.75% และ F1-Score 48.31% ซึ่งค่าวัดประสิทธิภาพทั้งหมดแสดงดังภาพประกอบ 48 ถึง 51 นอกจากนี้ได้แสดงผลเวลาที่แต่ละแบบจำลองใช้ในการเรียนรู้พบว่า Naïve Bayes ใช้เวลาน้อยที่สุดคือ 18.51 วินาที และจากการพิจารณา Confusion Matrix พบว่ากลุ่มที่เกิดการทำนายถูกมากที่สุดคือกลุ่ม D ในขณะที่การทำนายผิดมากที่สุดเกิดที่กลุ่ม B ที่มักถูกทำนายผิดเป็นกลุ่ม A หรือกลุ่ม C



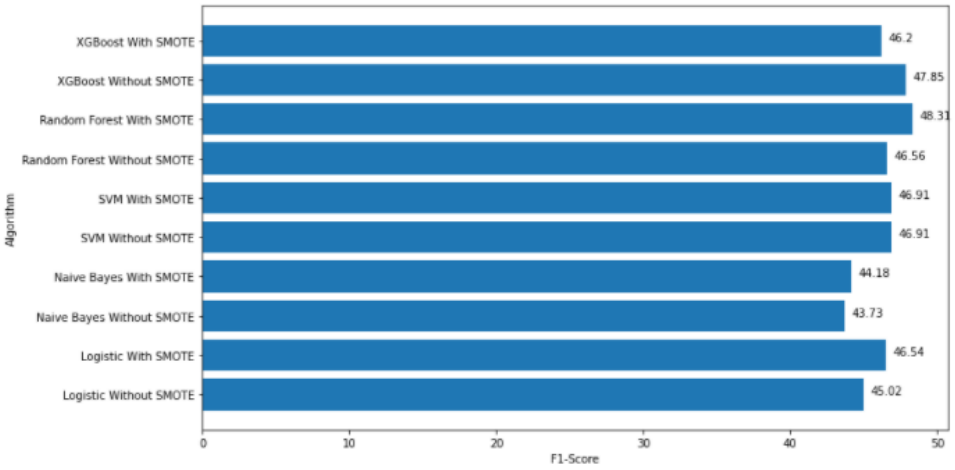
ภาพประกอบ 48 แสดงค่า Accuracy ของทุกแบบจำลอง



ภาพประกอบ 49 แสดงค่า Precision ของทุกแบบจำลอง



ภาพประกอบ 50 แสดงค่า Recall ของทุกแบบจำลอง



ภาพประกอบ 51 แสดงค่า F1-Score ของทุกแบบจำลอง

5.2 อภิปรายผลการวิจัย

งานวิจัยนี้ศึกษาการจัดกลุ่มลูกค้าบริษัทยานยต์แห่งหนึ่งซึ่งข้อมูลที่ใช้คือข้อมูลประเภทประชากรของลูกค้าโดยทำการเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด 5 แบบจำลองคือ Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest และ Extreme Gradient Boosting ซึ่งเป็นแบบจำลองที่ใช้ในงานจำแนกประเภทข้อมูลหรือ Classification โดยผู้วิจัยเลือกทั้งแบบเชิงเส้นและไม่เชิงเส้นเพื่อนำมาเปรียบเทียบประสิทธิภาพกัน จากผลการทดลองพบว่าแบบจำลอง Random Forest ให้ประสิทธิภาพที่ดีที่สุด จากนั้นทำการพิจารณา Confusion Matrix เพื่อสังเกตจำนวนการทำนายถูกและผิดของแต่ละแบบจำลอง ซึ่งพบว่ากลุ่ม D เกิดการทำนายถูกมากที่สุด แต่เมื่อตรวจสอบจำนวนข้อมูลแต่ละกลุ่มในข้อมูลชุดทดสอบพบว่ามีข้อมูลกลุ่ม D มากที่สุดเช่นกัน อาจเป็นอีกหนึ่งสาเหตุที่ทำให้กลุ่ม D มีจำนวนการทำนายถูกสูงที่สุด ในขณะที่กลุ่มที่เกิดการทำนายผิดมากที่สุดคือกลุ่ม B ที่มักถูกทำนายเป็นกลุ่ม A หรือ C ซึ่งเมื่อสำรวจข้อมูลพบว่ากลุ่ม A, B และ C มีลักษณะบางอย่างคล้ายกัน เช่น จากการสำรวจลูกค้าตามสถานภาพสมรสพบว่าลูกค้าส่วนใหญ่ในกลุ่ม A, B และ C ยังไม่ได้แต่งงาน ในขณะที่ลูกค้าส่วนใหญ่ในกลุ่ม D แต่งงานแล้ว หรือในการสำรวจลูกค้าตามอาชีพพบว่าลูกค้าส่วนใหญ่ในกลุ่ม A, B และ C ประกอบอาชีพด้านศิลปิน ในขณะที่ลูกค้าส่วนใหญ่ในกลุ่ม D ประกอบอาชีพด้านสายสุขภาพ นี่อาจเป็นสาเหตุให้แบบจำลองทำนายกลุ่ม B เป็นกลุ่ม A หรือ C ได้

ผู้วิจัยได้ทำการแสดงค่าความสำคัญของแต่ละฟีเจอร์ที่แบบจำลองใช้ในการเรียนรู้หรือ Feature Importance เพื่อให้เข้าใจการทำงานของแบบจำลองมากขึ้น โดยที่แบบจำลอง Logistic Regression สามารถดูความสำคัญของฟีเจอร์ได้จากค่าสัมประสิทธิ์ของฟีเจอร์แต่ละกลุ่ม ซึ่ง

ได้ผลว่าฟีเจอร์ที่มีอิทธิพลต่อการทำนายเป็นกลุ่ม A มากที่สุดคือฟีเจอร์ Profession_Entertainment สำหรับกลุ่ม B และ C คือฟีเจอร์ Profession_Artist และกลุ่ม D คือ Profession_Lawyer ซึ่งได้ฟีเจอร์ที่สำคัญของแต่ละกลุ่มเหมือนกันทั้งแบบใช้และไม่ใช้ SMOTE สำหรับแบบจำลอง Naïve Bayes แบบไม่ใช้ SMOTE ได้ว่าฟีเจอร์ Profession_Artist ให้ค่าสูงที่สุดและแบบใช้ SMOTE ได้ว่าฟีเจอร์ Profession_Entertainment ให้ค่าที่สูงที่สุด สำหรับแบบจำลอง SVM ไม่สามารถแสดงค่าความสำคัญของฟีเจอร์ได้เนื่องจากการปรับจูนพารามิเตอร์ทำให้ได้ Kernel Rbf ซึ่งทำให้ข้อมูลถูกแปลงมิติไปไม่ใช่เชิงเส้นจึงไม่สามารถแสดงค่าน้ำหนักของฟีเจอร์ออกมาได้ สำหรับแบบจำลอง Random Forest ทั้งแบบใช้และไม่ใช้ SMOTE ได้ผลว่าฟีเจอร์ Age มีผลต่อการจัดกลุ่มมากที่สุดและแบบจำลองสุดท้าย XGBoost ได้ผลว่าฟีเจอร์ Spending_Score_Low มีผลต่อการจัดกลุ่มมากที่สุดทั้งแบบใช้และไม่ใช้ SMOTE

นอกจากนี้ผู้วิจัยได้ใช้เครื่องมือ LIME และ SHAP ที่ช่วยในการอธิบายการเลือกใช้ฟีเจอร์ในการทำนายหนึ่งครั้งและแสดงค่าความน่าจะเป็นในการทำนายแต่ละกลุ่ม โดยทำการสุ่มข้อมูลจากชุดทดสอบขึ้นมา 1 ข้อมูลซึ่งมีเลเบลว่าอยู่กลุ่ม B สำหรับแบบจำลอง Logistic Regression, SVM และ Random Forest ทำนายว่าข้อมูลนี้อยู่กลุ่ม A และเมื่อสำรวจการเลือกใช้ฟีเจอร์พบว่าแบบจำลองให้ความสำคัญกับฟีเจอร์ Profession_Healthcare มากที่สุด สำหรับแบบจำลอง Naïve Bayes ทำนายว่าข้อมูลนี้อยู่กลุ่ม D และแบบจำลองให้ความสำคัญกับฟีเจอร์ Profession_Marketing มากที่สุด สำหรับแบบจำลอง XGBoost พบปัญหาบางอย่างจากการเรียนรู้ของแบบจำลองที่ทำให้จำนวนฟีเจอร์ไม่สัมพันธ์กับข้อมูลจึงไม่สามารถใช้ LIME ได้ ผู้วิจัยจึงเลือกใช้ SHAP ในการอธิบายภาพรวมพบว่าฟีเจอร์ที่สำคัญในการทำนายคือฟีเจอร์ Age และส่งผลต่อกลุ่ม D มากที่สุด

จากนั้นเกิดข้อสงสัยว่าระหว่างค่าความสำคัญของฟีเจอร์หรือ Feature Importance และผลจากการใช้ LIME หรือ SHAP มีความสอดคล้องกันหรือไม่ ผู้วิจัยจึงทำการสุ่มข้อมูลที่ทำนายถูกจากแบบจำลอง Random Forest ร่วมกับการใช้ SMOTE เพิ่มเติมเนื่องจากเป็นแบบจำลองที่ให้ประสิทธิภาพที่ดีที่สุดเพื่อสำรวจฟีเจอร์ พบว่ากลุ่ม D พบความสอดคล้องกันคือจากการใช้ LIME พบฟีเจอร์เกี่ยวกับอายุ อาชีพด้านสายสุขภาพและศิลปินซึ่งทั้ง 3 ฟีเจอร์นี้ให้ค่าที่สูงเมื่อตรวจสอบที่ค่าความสำคัญของฟีเจอร์ ในขณะที่กลุ่ม A พบความไม่สอดคล้องกันเท่าที่ควรคือจากการใช้ LIME พบฟีเจอร์ระดับการใช้จ่ายต่ำ อาชีพด้านสายสุขภาพและบันเทิง สำหรับกลุ่ม B พบฟีเจอร์อาชีพด้านศิลปินที่อาจไม่ใช่ฟีเจอร์ที่ได้ค่าสูงจากค่าความสำคัญของฟีเจอร์ แต่จากการสำรวจข้อมูลพบว่าลูกค้าในกลุ่ม B ส่วนใหญ่ประกอบอาชีพด้านศิลปิน จึงมองว่าการที่ใช้ฟีเจอร์อาชีพ

ศิลปินเป็นหลักค่อนข้างสมเหตุสมผล สำหรับกลุ่ม C คล้ายกับกรณีของกลุ่ม B คือพบฟีเจอร์อาชีพด้านศิลปินและจากการสำรวจพบว่ากลุ่ม C ลูกค้าส่วนใหญ่ประกอบอาชีพศิลปิน จึงมองว่ามีความสมเหตุสมผลที่เลือกใช้ฟีเจอร์นี้เช่นกัน

ลักษณะของแต่ละกลุ่มที่ได้จากการสำรวจคือกลุ่ม A และ B มีลักษณะที่คล้ายกันคือ ลูกค้าส่วนใหญ่อายุ 30-40 ปี ทำงานด้านศิลปินและมีการใช้จ่ายต่ำ สำหรับกลุ่ม C ลูกค้ามีอายุมากที่สุดคือส่วนใหญ่มีอายุ 50 ปี ทำงานด้านศิลปินและมีการใช้จ่ายปานกลาง สุดท้ายกลุ่ม D ลูกค้ามีอายุน้อยที่สุดคือส่วนใหญ่มีอายุ 20-30 ปี ยังไม่ได้แต่งงาน ทำงานด้านสายสุขภาพและมีการใช้จ่ายต่ำ

สำหรับการใช้ LIME และ SHAP นอกจากช่วยให้ผู้อ่านผลเข้าใจการทำงานของแบบจำลองแล้ว ยังช่วยสร้างความน่าเชื่อถือให้กับแบบจำลอง หากต้องนำแบบจำลองไปใช้ต่อในทางด้านธุรกิจ LIME และ SHAP สามารถทำให้นักการตลาดเชื่อมั่นในแบบจำลองและตัดสินใจใช้กลยุทธ์ต่างๆที่เหมาะสมกับแต่ละกลุ่มได้ เช่น จากแบบจำลอง XGBoost ใช้ SHAP ในการอธิบายผล พบว่าฟีเจอร์ Age มีความสำคัญต่อการจัดกลุ่มมากที่สุด จากการสำรวจข้อมูลในบทที่ 3 พบว่าลูกค้ากลุ่ม D อายุน้อยที่สุด อาจนำเสนอสินค้าที่มีสีสันให้เลือกซื้อ ในขณะที่ลูกค้ากลุ่ม C อายุมากที่สุด อาจนำเสนอสินค้าที่มีสีเรียบ เน้นการใช้งานระยะยาว

5.3 ข้อเสนอแนะ

1. จากผลการวิจัยพบว่าค่าที่ได้จากการวัดประสิทธิภาพต่าง ๆ ค่อนข้างน้อย จากการสำรวจข้อมูลพบว่าข้อมูลแต่ละกลุ่มมีความคล้ายกัน นี่อาจเป็นสาเหตุที่ทำให้แบบจำลองเกิดความผิดพลาดในการทำนายข้อมูลได้ หรือถ้าหากได้มีโอกาสเป็นผู้ออกแบบการเก็บข้อมูล อาจเก็บข้อมูลประเภทอื่นร่วมด้วย เช่น ข้อมูลความสนใจหรือพฤติกรรมของลูกค้าที่อาจช่วยอธิบายลักษณะของลูกค้าแต่ละคนได้ดียิ่งขึ้นเพื่อนำไปสู่การทำนายที่แม่นยำมากขึ้น

2. ในงานวิจัยนี้ใช้แบบจำลองสำหรับงานที่มีเลเบลหรือเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน แต่ในงานด้านจัดกลุ่มลูกค้ายังนิยมใช้แบบจำลองสำหรับเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนด้วย หากลองใช้เทคนิคการเรียนรู้แบบไม่มีผู้สอนแล้วนำผลลัพธ์มาเปรียบเทียบกัน อาจทำให้เห็นข้อมูลเชิงลึกบางอย่างเพิ่มขึ้น

3. ในอนาคตอาจมีการทดลองใช้แบบจำลองร่วมกับแอพพลิเคชันในการรับข้อมูลจากลูกค้าเพื่อความสะดวกและรวดเร็วในการวิเคราะห์และจัดทำการนำเสนอสินค้าและบริการต่อไป

บรรณานุกรม

Abidar, L., Zaidouni, D., & Ennouaary, A. (2020). *Customer Segmentation With Machine Learning*. Paper presented at the Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications.

Ampadu, H. (2021). Random Forest Understanding. https://ai-pool.com/a/s/random-forests-understanding

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1). New York: Springer New York, NY.

Gevelber, L. (2015). Why consumer intent is more powerful than demographics. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/why-consumer-intent-more-powerful-than-demographics/

Globalwebindex. (2020). Customer Segmentation 101 A Guide for Brands https://www.gwi.com/hubfs/Downloads/Customer%20Segmentation%20101_b.pdf

Kash. (2020). Customer Segmentation Classification. https://www.kaggle.com/kaushiksuresh147/customer-segmentation

Kecman, V. (2005). *Support Vector Machines: Theory and Applications*. Berlin: Springer.

Lamrhari, S., Elghazi, H., & El Faker, A. (2020). *Random Forest-based Approach for Classifying Customers in Social CRM*. Paper presented at the 2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD).

Minaphinant, V. (2018). Machine Learning. Retrieved from https://medium.com/investic/machine-learning-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-fa8bf6663c07

Moore, D. S., MacCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). New York: W.H. Freeman, Macmillan Learning.

Nandapala, E. Y. L., Jayasena, K. P. N., & Rathnayaka, R. M. K. T. (2020). *Behavior Segmentation based Micro-Segmentation Approach for Health Insurance Industry*.

Paper presented at the 2020 2nd International Conference on Advancements in Computing (ICAC).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "*Why Should I Trust You?*". Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Shaukat, K., Luo, S., Abbas, N., Mahboob Alam, T., Ehtesham Tahir, M., & Hameed, I. A. (2021). *An Analysis of Blessed Friday Sale at a Retail Store Using Classification Models*. Paper presented at the 2021 The 4th International Conference on Software Engineering and Information Management.

Umuhoza, E., Ntirushwamaboko, D., Awuah, J., & Birir, B. (2020). Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa. *SAIEE Africa Research Journal,* 111(3), 95-101.

Wu, S., Yau, W.-C., Ong, T.-S., & Chong, S.-C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. *IEEE Access,* 9, 62118-62136.

Zeybek, H. (2018). Customer segmentation strategy for rail freight market: The case of Turkish State Railways. *Research in Transportation Business & Management,* 28, 45-53.

ประวัติผู้เขียน

ชื่อ-สกุล	กาญจนมาส เปลี่ยนสกุล
วัน เดือน ปี เกิด	14 กุมภาพันธ์ 2540
สถานที่เกิด	ปทุมธานี
วุฒิการศึกษา	พ.ศ. 2562 วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ จาก มหาวิทยาลัยธรรมศาสตร์
ที่อยู่ปัจจุบัน	200/240 หมู่1 ตำบลหลักหก อำเภอเมือง จังหวัดปทุมธานี 12000