การประยุกต์ใช้การเรียนรู้ของเครื่องในการส่งแคมเปญบัญชีเงินฝากดิจิทัล : กรณีศึกษาธนาคารพาณิชย์แห่งหนึ่ง

APPLYING MACHINE LEARNING MODEL IN DIGITAL DEPOSIT ACCOUNT CAMPAIGN : A CASE STUDY OF A COMMERCIAL BANK

กฤษฎ์พล กิตตินาถไกรวัฒน์

สหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชาสถิติ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2565

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

APPLYING MACHINE LEARNING MODEL IN DIGITAL DEPOSIT ACCOUNT CAMPAIGN : A CASE STUDY OF A COMMERCIAL BANK

KRISPOL KITTINARTKRAIWAT

A COOPERATIVE EDUCATION SUBMITTED IN PARTIAL
FULFILLMENT OF THE REQUIREMENT FOR
THE DEGREE OF BACHELOR OF SCIENCE (APPLIED STATISTICS)
DEPARTMENT OF STATISTICS, SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2022

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อสหกิจศึกษา	การประยุกต์ใช้การเรียนรู้ของเครื่องในการส่งแคมเปญบัญชีเงินฝากดิจิทัล : กรณีศึกษาธนาคารพาณิชย์แห่งหนึ่ง Applying Machine Learning Model in Digital Deposit Account Campaign : A Case Study of A Commercial Bank
ชื่อนักศึกษา	นาย กฤษฎิ์พล กิตตินาถไกรวัฒน์ รหัสนักศึกษา 62050743
ปริญญา	วิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชา	สถิติ
ปีการศึกษา	2565
อาจารย์ที่ปรึกษา	ดร. สกุณา ศรีอโนมัย

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) อนุมัติให้สหกิจศึกษานี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (สถิติประยุกต์) ประจำปีการศึกษา 2565

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.ดร.กนกกรรณ์ ลิ้โรจนาประภา ประธานกรรมการ	
คุณรัฐชัย จิระรัตนานนท์ กรรมการ	
ดร.สกุณา ศรีอโนมัย กรรมการและอาจารย์ที่ปรึกษา	

ลิขสิทธิ์ของคณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อสหกิจศึกษา การประยุกต์ใช้การเรียนรู้ของเครื่องในการส่งแคมเปญบัญชีเงินฝากดิจิทัล : กรณีศึกษาธนาคารพาณิชย์แห่งหนึ่ง
Applying Machine Learning Model in Digital Deposit Account Campaign : A Case Study of A Commercial Bank

ชื่อนักศึกษา นาย กฤษฎ์พล กิตตินาถไกรวัฒน์ รหัสนักศึกษา 62050743
ปริญญา วิทยาศาสตรบัณฑิต (สถิติประยุกต์)
ภาควิชา สถิติ
ปีการศึกษา 2565
อาจารย์ที่ปรึกษา ดร. สกุณา ศรีอโนมัย

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการประยุกต์ใช้การเรียนรู้ของเครื่องเข้ามาช่วยพยากรณ์ลูกค้าที่มีโอกาสเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ได้รับการอนุเคราะห์ข้อมูลจากธนาคารพาณิชย์แห่งหนึ่ง โดยผู้วิจัยได้ทำการจัดเตรียมข้อมูลและฝึกสอนแบบจำลองการเรียนรู้ของเครื่องทั้ง 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีการเรียนรู้นาอีฟเบย์ และวิธีป่าไม้สุ่ม เพื่อเปรียบเทียบประสิทธิภาพด้วยการพิจารณาจาก เมทริกซ์ความสับสน การตรวจสอบไขว้ และกราฟเส้นโค้ง ROC ผลการเปรียบเทียบพบว่า วิธีป่าไม้สุ่ม เป็นแบบจำลองที่ให้ประสิทธิภาพในการพยากรณ์ได้ดีที่สุด

จากนั้นทดลองนำไปใช้งาน ในการสร้างแคมเปญเสนอขายผลิตภัณฑ์บัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท โดยนำแบบจำลองไปพยากรณ์กลุ่มลูกค้า ณ วันที่ 1 เมษายน พ.ศ. 2566 ที่ตรงตาม 3 เงื่อนไขของงานวิจัยนี้ได้แก่ เป็นลูกค้าที่มีบัญชีออมทรัพย์แบบธรรมดา เป็นลูกค้าที่มีบัตรเครดิตอย่างน้อย 1 ประเภท และเป็นลูกค้าที่ไม่มีบัญชีเงินฝากประเภทดิจิทัล ได้ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ซึ่งมีลูกค้าที่ตรงตามเงื่อนไขทั้งสิ้นจำนวน 248,675 คน จากนั้นได้ทำการส่งแคมเปญให้แก่ลูกค้า 1 ครั้ง และพิจารณาผลการเข้ามาเปิดบัญชีหลังจากส่งแคมเปญ 21 วัน หรือ 3 สัปดาห์ โดยผลการเปิดบัญชีจะมีการแยกพิจารณาผลเป็น 2 มุมมอง ได้แก่ มุมมองผลการพยากรณ์ของแบบจำลอง พบว่า กลุ่มที่แบบจำลองทำนายว่ามีโอกาสเปิดบัญชี มีเปอร์เซ็นต์การเปิดบัญชีที่สูงกว่ากลุ่มที่แบบจำลองทำนายว่ามีโอกาสไม่เปิดบัญชี และมุมมองที่แยกตามคุณลักษณะของลูกค้า ได้แก่ อายุและพฤติกรรมการยอมรับความเสี่ยงในการลงทุน พบว่า กลุ่มลูกค้าที่มีอายุน้อยกว่า 41 ปี และมีพฤติกรรมการยอมรับความเสี่ยงในการลงทุนได้น้อย เป็นกลุ่มที่มีเปอร์เซ็นต์การเปิดบัญชีสูงที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

คำสำคัญ : การเรียนรู้ของเครื่อง, แคมเปญ, บัญชีเงินฝากดิจิทัล

Title	Applying Machine Learning Model in Digital Deposit Account Campaign : A Case Study of A Commercial Bank
Students	Mr. Krispol Kittinartkraiwat Student ID 62050743
Degree	Bachelor of Science (Applied Statistics)
Department	Statistics
School	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2022
Advisor	Dr. Sakuna Sriarnomai

Abstract

The objective of this research is to study the application of machine learning in predicting customers who have a chance of opening a digital high-interest deposit account with a deposit up to 100,000 baht. The data from a commercial bank was used to train the model. Three methods were employed: the k-Nearest Neighbors, the Naive Bayes, and the Random Forest. The effectiveness of these methods was compared by using Confusion Matrices, Cross-Validation, and ROC Curve. Random Forest method indicated the best accuracy.

The best model was deployed to create a campaign for offering digital high-interest deposit account products with a deposit up to 100,000 baht. The model was used to predict customers who met three conditions on April 1, 2023: having a regular savings account, having at least one type of credit card, and not having a digital high-interest deposit account. One campaign message was sent to 247,675 customers. After 3 weeks, (21 days) who opened the campaign account was used as respondent data. This research analyzed final results into two perspectives. First, comparing result from predictive model, customers who predicted to have a chance to open deposit account have higher percentage of respondence. Second, comparing with

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

customer characteristics including age and risk acceptance, customers under 41 years old with lower risk acceptance level have the highest percentage of respondence.

Keywords : Machine Learning, Campaign, Digital Deposit Account

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

การทำสหกิจศึกษาฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีเนื่องด้วยได้รับความกรุณาจาก ดร.สกุณา ศรีอโนมัย อาจารย์ที่ปรึกษาที่กรุณาสละเวลาอันมีค่าให้คำแนะนำ ข้อเสนอแนะ คำปรึกษาและช่วยปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่อย่างดียิ่ง อีกทั้งแนะแนวความรู้ต่าง ๆ เอื้อเฟื้อเอกสารอ้างอิงในการค้นคว้าข้อมูลและติดตามความก้าวหน้าของงานทุกขั้นตอน จนทำให้สหกิจศึกษาฉบับนี้เสร็จสมบูรณ์ ผู้วิจัยตระหนักถึงความตั้งใจ ความเมตตาและความทุ่มเทของอาจารย์ จึงขอกราบขอบพระคุณด้วยความเคารพอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ ผศ.ดร.กนกกรรณ์ ลี้โรจนาประภา ที่ให้เกียรติเป็นคณะกรรมการสหกิจศึกษาที่กรุณาให้คำปรึกษา แนวคิดข้อเสนอแนะที่เป็นประโยชน์ต่อการจัดทำสหกิจศึกษาฉบับนี้ และสละเวลาตรวจทานชี้ให้เป็นถึงข้อบกพร่องต่าง ๆ ทำให้สหกิจศึกษาฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

รวมถึงขอกราบขอบพระคุณ คุณรัฐชัย จิระรัตนานนท์ ที่เมตตาให้คำแนะนำ รวมทั้งให้คำปรึกษาเพื่อแก้ไขเครื่องมือที่ใช้ในการทำสหกิจศึกษาครั้งนี้ให้มีคุณภาพ และมอบความรู้ในการทำงานตลอดการทำสหกิจศึกษาครั้งนี้

สุดท้ายนี้ขอกราบขอบพระคุณ บิดา มารดา ที่สนับสนุนและส่งเสริมกำลังใจเสมอมาและขอขอบคุณเพื่อน ๆ ที่ให้คำปรึกษาและช่วยเหลือตลอดจนทำให้สหกิจศึกษาฉบับนี้สำเร็จตามที่ได้ตั้งใจ และคณะผู้วิจัยหวังอย่างยิ่งว่าสหกิจศึกษาฉบับนี้จะเป็นประโยชน์และเป็นแนวทางในการศึกษาต่อไป

กฤษฎ์พล กิตตินาถไกรวัฒน์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่ **หน้า**

เอกสารนี้สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่ **หน้า**

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

ธนาคารเป็นองค์กรทางการเงินที่มีผลิตภัณฑ์และบริการหลากหลาย เพื่อตอบสนองความต้องการและความพึงพอใจของลูกค้า เนื่องจากธนาคารมีผลิตภัณฑ์ที่หลากหลาย เช่น กองทุน ประกันชีวิต บัญชีเงินฝาก ฯลฯ จึงพยายามหาลูกค้าเพื่อเข้ามาใช้บริการของธนาคารที่มีอยู่ ด้วยการทำแคมเปญ (Campaign) ที่เป็นหนึ่งในวิธีการกระตุ้นยอดขายผลิตภัณฑ์ของธนาคาร เป็นกิจกรรมที่มุ่งเน้นการนำเสนอผลิตภัณฑ์ หรือ บริการของธนาคารให้แก่ลูกค้าผ่านการใช้ช่องทางต่าง ๆ โดยในอดีตที่ผ่านมา การทำแคมเปญเสนอขายผลิตภัณฑ์ของธนาคาร เป็นการส่งแคมเปญผ่านทางข้อความสั้น (SMS) เพราะเป็นวิธีที่สะดวก และง่ายที่สุดในการสื่อสารกับลูกค้า มีระบบการส่งข้อความมือถือที่เข้าถึงผู้ใช้งานโทรศัพท์มือถือโดยตรง แต่เป็นช่องทางที่มีค่าใช้จ่ายให้แก่บริษัทที่เป็นตัวกลาง ในการกระจายข้อความให้แก่ลูกค้าของธนาคาร ปัจจุบันการส่งแคมเปญเพื่อให้ลูกค้ารับรู้ถึงข้อมูลข่าวสารหรือผลิตภัณฑ์ของธนาคาร มีหลากหลายช่องทางมากยิ่งขึ้น เช่น ช่องทางออนไลน์ (Online) หรือช่องทางผ่านแอปพลิเคชันของธนาคาร (Application) จากกรณีศึกษาการส่งแคมเปญผ่านช่องข้อความสั้น (SMS) ของธนาคารพาณิชย์แห่งหนึ่ง พบว่าอัตราที่ลูกค้าเข้ามาใช้ผลิตภัณฑ์ของธนาคารน้อยกว่า 5 เปอร์เซ็นต์ ดังนั้นการส่งแคมเปญผ่านช่องข้อความสั้น (SMS) จะต้องมีการวิเคราะห์ข้อมูลเลือกกลุ่มลูกค้าเป้าหมายอย่างรอบคอบ เพื่อให้ได้ลูกค้าที่มีแนวโน้มเข้ามาใช้บริการหรือผลิตภัณฑ์ของธนาคารให้สอดคล้องกับงบประมาณที่เสียไป

จากการสำรวจรวบรวมข้อมูลการใช้อินเทอร์เน็ต (Internet) กับการใช้แอปพลิเคชันของธนาคาร เมื่อวันที่ 31 ม.ค. 2562 พบว่าจากผู้ใช้งานอินเทอร์เน็ต 43.8 ล้านคน คิดเป็นจำนวนคนใช้งาน แอปพลิเคชันธนาคารประมาณถึง 32.4 ล้านราย ซึ่งคิดเป็น 74 เปอร์เซ็นต์จากผู้ใช้งานอินเทอร์เน็ตทั้งหมด (กสทช., 2562) และปัจจุบันเมื่อวันที่ 1 ก.พ. 2566 ธนาคารออกนโยบายการส่งแคมเปญในช่องทางข้อความสั้น (SMS) ไว้ว่า การส่งแคมเปญในช่องทางสั้นนี้ห้ามมีการใส่ลิงก์เชื่อมโยง (Deep Link) หมายถึงปุ่มข้อความที่เชื่อมโยงไปยังเนื้อหา หรือระบบในแอปพลิเคชันหรือเว็บไซต์โดยตรง เพราะปัจจุบันได้มีการฉ้อโกงผ่านทางข้อความสั้น (SMS) โดยเมื่อลูกค้ากดเข้าไปผ่านลิงก์แล้ว จะทำให้สามารถสืบข้อมูล หรือโจรกรรมทรัพย์สินของลูกค้าให้เกิดความเสียหาย ดังนั้นเนื่องจากไม่สามารถใส่ลิงก์เชื่อมโยงไปยังแอปพลิเคชันธนาคารได้ ทำให้ความสะดวกของลูกค้าลดลง ซึ่งอาจจะลดประสิทธิภาพของการเข้ามาใช้ผลิตภัณฑ์ของธนาคาร และจากการที่มีผู้ใช้แอปพลิเคชันธนาคารที่เพิ่มมากขึ้น ทางธนาคารจึงได้หันมาใช้การส่งข้อความผ่านทางแอปพลิเคชัน (Application) มากขึ้น เพราะ เนื่องจากเป็นช่องทางการโฆษณาแคมเปญส่วนตัวของธนาคาร ที่สร้างขึ้นมาเองจึงทำให้การส่ง

เอกสารนี้ [illegible] ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แคมเปญหาลูกค้าเป็นจำนวนมากได้โดยไม่ต้องคำนึงถึงค่าใช้จ่าย แต่อย่างไรก็ตามการส่งแคมเปญให้กับลูกค้าจำนวนมากที่ไม่ใช่กลุ่มเป้าหมายของผลิตภัณฑ์ จะสร้างความไม่พึงพอใจให้แก่ลูกค้ามากเกินไปเพราะอาจจะไม่ใช่ผลิตภัณฑ์ที่ลูกค้าต้องการ ซึ่งหากนำการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาประยุกต์ใช้ในการส่งแคมเปญ เพื่อจำแนกกลุ่มลูกค้าที่เป็นกลุ่มเป้าหมายของผลิตภัณฑ์ จะช่วยให้แคมเปญที่ส่งไปเสนอผลิตภัณฑ์ของธนาคารตรงตามกลุ่มเป้าหมายมากยิ่งขึ้น เพื่อให้ลูกค้าเข้ามาใช้ผลิตภัณฑ์ของธนาคารมากขึ้น

ดังนั้นผู้วิจัยจึงสนใจการนำการเรียนรู้ของเครื่อง มาใช้ในการจำแนกกลุ่มลูกค้าที่มีเป็นกลุ่มลูกค้าเป้าหมายของแคมเปญ เพื่อการส่งแคมเปญที่ได้ลูกค้าเข้ามาใช้ผลิตภัณฑ์มากขึ้น โดยสร้างการเรียนรู้ของเครื่องด้วยกัน 3 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) วิธีนาอีฟเบย์ (Naïve Bayes) และวิธีป่าไม้สุ่ม (Random Forest) มาเปรียบเทียบประสิทธิภาพโดยพิจารณาจาก 1. เมทริกซ์ความสับสน (Confusion Matrix) 2. การวิเคราะห์ความแม่นตรงของแบบจำลองด้วยวิธีตรวจสอบไขว้แบบแบ่งชั้น (Stratified K-fold Cross Validation) 3. กราฟเส้นโค้ง ROC Curve (Receiver Operating Characteristic Curve) เพื่อนำการเรียนรู้ของเครื่องที่ให้ประสิทธิภาพจำแนกดีที่สุด ไปพยากรณ์กลุ่มลูกค้าที่ทางธนาคารสนใจ และทดลองการนำไปใช้งานสร้างแคมเปญเสนอขายผลิตภัณฑ์ของธนาคาร

1.2 วัตถุประสงค์

1) เพื่อศึกษาและพัฒนาแบบจำลองการเรียนรู้ของเครื่อง สำหรับการส่งแคมเปญเสนอขายผลิตภัณฑ์บัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท
2) เพื่อเปรียบเทียบวิธีการเรียนรู้ของเครื่องที่เหมาะสม สำหรับการนำไปใช้พยากรณ์กลุ่มลูกค้าที่มีโอกาสจะเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท
3) เพื่อทดสอบการนำแบบจำลองไปใช้งาน ในการส่งแคมเปญเสนอขายบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท

1.3 ขอบเขตของงานวิจัย

1) ขอบเขตด้านข้อมูล

รวบรวมข้อมูลลูกค้าที่ตรงตาม 3 เงื่อนไข ได้แก่ มีบัญชีออมทรัพย์แบบธรรมดา มีบัตรเครดิตอย่างน้อย 1 ประเภท และไม่มีบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ย้อนหลัง 1 ปี ตั้งแต่วันที่ 1 มกราคม พ.ศ. 2565 ถึง 31 ธันวาคม พ.ศ. 2565 มีข้อมูลทั้งสิ้น 253,594 รายการ แบ่งออกเป็นชุดที่ 1. ข้อมูลของลูกค้าที่เปิดใช้ผลิตภัณฑ์ภายใน 1 ปีเท่ากับ 4,312 รายการ คิดเป็นสัดส่วนประมาณ 1.7 เปอร์เซ็นต์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

และชุดที่ 2. ข้อมูลของลูกค้าที่ไม่เปิดใช้ผลิตภัณฑ์ภายใน 1 ปีเท่ากับ 249,282 รายการ คิดเป็นสัดส่วน 98.3 เปอร์เซ็นต์

2) ขอบเขตด้านเทคนิค

2.1 การกำจัดค่านอกเกณฑ์ด้วยค่าพิสัยควอร์ไทล์ (Interquartile Range)

2.2 การแปลงข้อมูล ด้วยวิธีการเข้ารหัสป้ายกำกับ (Label Encoding) และการปรับค่าให้อยู่ในช่วงสูงสุดและต่ำสุด (Min-Max Normalization)

2.3 การเลือกคุณลักษณะ ด้วยวิธีวิเคราะห์ความแปรปรวนทางเดียว (One way ANOVA) และสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient)

2.4 การจัดการความไม่สมดุลของข้อมูล ด้วยวิธีการสุ่มเพิ่ม Synthetic Minority Oversampling Technique และการสุ่มลด Edited Nearest Neighbor

2.5 วิธีการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) วิธีนาอีฟเบย์ (Naïve Bayes) และวิธีป่าไม้สุ่ม (Random Forest)

2.6 พิจารณาประสิทธิภาพโมเดลด้วยวิธี และเมทริกซ์ความสับสน (Confusion Matrix) วิธีตรวจสอบไขว้แบบแบ่งชั้น (Stratified K-Fold Cross Validation) และกราฟเส้นโค้ง ROC Curve (Receiver Operating Characteristic Curve)

3) ขอบเขตด้านระยะเวลา

การส่งแคมเปญข้อความเสนอขายผลิตภัณฑ์บัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ในวันที่ 20 เมษายน พ.ศ. 2566 และเก็บผลการเข้ามาเปิดบัญชีหลังจากการส่งแคมเปญ ในวันที่ 11 พฤษภาคม พ.ศ. 2566 ระยะเวลารวม 21 วัน

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ทราบถึงแนวทางในการพัฒนาแบบจำลอง สำหรับการพยากรณ์ลูกค้าที่มีโอกาสเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท
2) ธนาคารได้ลูกค้าเข้ามาใช้ผลิตภัณฑ์เพิ่มมากขึ้น หลังจากการส่งแคมเปญเสนอขายบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท

1.5 นิยามศัพท์เฉพาะ

ในงานวิจัยครั้งนี้ ผู้วิจัยได้กำหนดความหมายของคำศัพท์ที่เกี่ยวข้องไว้ เพื่อทำให้ผู้ที่จะนำงานวิจัยนี้ไปศึกษาต่อได้เกิดความเข้าใจในแนวทางเดียวกัน ดังนี้

1) ผลิตภัณฑ์ (Product) หมายถึง สิ่งที่จับต้องได้และสิ่งที่จับต้องไม่ได้ ซึ่งอยู่ในรูปแบบผลิตภัณฑ์และการบริการแตกต่างกันไปตามธุรกิจ เพื่อเป็นการตอบสนองต่อความจำเป็นและความต้องการของผู้บริโภค ในธุรกิจธนาคารผลิตภัณฑ์ของธนาคารมักจะ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รวมถึงบัญชีเงินฝาก บัตรเครดิต สินเชื่อส่วนบุคคล สินเชื่อเพื่อธุรกิจ การลงทุน การซื้อขายหลักทรัพย์ และบริการทางการเงินอื่นๆ (จิตพนธ์, 2560)

2) การเรียนรู้ของเครื่อง (Machine Learning) หมายถึง การทำให้ระบบคอมพิวเตอร์เรียนรู้และสร้างขั้นตอนวิธี (Algorithm) ที่สามารถเรียนรู้ข้อมูลและพยากรณ์ข้อมูลได้ (วีระพันธ์, 2564)
3) แคมเปญ (Campaign) หมายถึง ชุดข้อความโฆษณาที่ใช้วิธีคิดหรือแนวคิดที่กำหนดวัตถุประสงค์ของการทำโฆษณาเพื่อสร้างความน่าสนใจและเชื่อมโยงกับกลุ่มเป้าหมายและเพิ่มยอดขายของสินค้าหรือบริการ (พงศกร, 2565)
4) การเรียนรู้แบบมีผู้สอน (Supervised Learning) หมายถึง การเรียนรู้แบบมีผู้สอน หรือจากข้อมูลตัวอย่างในอดีตที่เฉลยผลลัพธ์ที่ควรจะเป็นแสดง เป็นลาเบล (Label) ไว้นำมาสอนเครื่องให้ค้นหาความสัมพันธ์ และสร้างกฎทั่วไปไว้ เพื่อพยากรณ์ว่าข้อมูลนำเข้าแล้วจะทำให้ได้ข้อมูลส่งออกแบบใด (มาณวิกา, 2562)
5) วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) จะใช้หลักการเปรียบเทียบความคล้ายคลึงกันของข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายคลึงหรืออยู่ใกล้กับข้อมูลใดมากที่สุด k ตัว จากนั้นจะทำการตัดสินใจว่าคำตอบของข้อมูลที่สนใจนั้นควรเป็นคำตอบเดียวกับข้อมูลที่อยู่ใกล้ที่สุด k ตัวนั้น โดยที่ k คือความถี่ของข้อมูลที่อยู่ใกล้กับข้อมูลที่สนใจ (พัชนา, 2561)
6) วิธีนาอีฟเบย์ (Naïve Bayes) คือ การเรียนรู้ของเครื่องที่อาศัยหลักการความน่าจะเป็นตามทฤษฎีเบส์ (Bayes Theorem) ซึ่งมีขั้นตอนวิธี (Algorithm) ที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูลโดยการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำมาสร้างเงื่อนไขการจำแนกข้อมูลใหม่ (อนัตต์ชัย และจรัญ, 2561)
7) วิธีป่าไม้สุ่ม (Random forest) คือ แบบจำลองเชิงตัวเลขที่ใช้หลายต้นไม้เป็นส่วนประกอบของการตัดสินใจ แต่ละต้นไม้จะถูกสร้างโดยการสุ่มตัวอย่างของข้อมูลและคุณสมบัติเพื่อสร้างโมเดลเบื้องต้น พยากรณ์ผลจากวิธีโหวตของต้นไม้แต่ละต้นเพื่อให้ส่งผลลัพธ์การตัดสินใจที่แม่นยำขึ้น (Petkovic et al., 2018)
8) คุณลักษณะ (Features) หมายถึง ตัวแปรที่แยกต่างหากที่มีบทบาทเป็นข้อมูลที่เข้าไปในแบบจำลอง โดยแบบจำลองจะใช้คุณลักษณะเหล่านั้นในการพยากรณ์ผลลัพธ์สุดท้ายหรือตัวแปรตาม (Brown, 2019)
9) บัญชีเงินฝากดิจิทัล (Digital Deposit Account) หรือบัญชีเงินฝากออนไลน์ เป็นรูปแบบของบัญชีเงินฝากที่ให้บริการผ่านช่องทางต่าง ๆ เช่น เว็บไซต์ หรือ แอปพลิเคชันบัญชีธนาคาร โดยทั่วไปแล้ว เงินฝากดิจิทัลมีลักษณะเดียวกับบัญชีธนาคารทั่วไป โดยสามารถฝากเงินเข้าบัญชีเงินฝาก ซึ่งจะถูกจัดเก็บในรูปแบบดิจิทัลแทนที่จะเป็นเงินสด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สามารถทำธุรกรรมต่าง ๆ เช่น ถอนเงิน โอนเงิน ตรวจสอบยอดคงเหลือ และตรวจสอบรายการธุรกรรมได้ผ่านทางอินเทอร์เน็ตได้ตลอดเวลา การทำธุรกรรมเหล่านี้สามารถทำได้ผ่านอุปกรณ์ที่เชื่อมต่ออินเทอร์เน็ต เช่น คอมพิวเตอร์ สมาร์ทโฟน แท็บเล็ต หรือเครื่องมืออื่น ๆ ที่สามารถเข้าถึงแพลตฟอร์มดิจิทัลได้

10) คลาส (Class) หมายถึง เป้าหมายที่ต้องการให้ระบบการเรียนรู้ของเครื่องมีความสามารถพยากรณ์จำแนกแยกแยะโดยมีทั้งสิ้น 2 คลาส 1. กลุ่มลูกค้าที่เปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท 2. กลุ่มลูกค้าไม่เปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท
11) คลาส DEPOSIT หมายถึง เป้าหมายที่ทางผู้จัดทำสนใจเป็นเป้าหมายหลักหรือเป้าหมายเชิงบวก (Positive) ในการพยากรณ์ คือ กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มลูกค้าเหล่านี้มีโอกาสที่จะเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท
12) คลาส NON_DEPOSIT หมายถึง เป้าหมายที่ทางผู้จัดทำสนใจเป็นหมายเชิงลบ (Negative) ในการพยากรณ์ คือ กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มลูกค้าเหล่านี้มีโอกาสไม่เปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การศึกษาและพัฒนาการเรียนรู้ของเครื่องในอุตสาหกรรมธนาคาร เพื่อการส่งขายผลิตภัณฑ์ของธนาคารจะมีกระบวนการสร้างแบบจำลองเพื่อให้เหมาะสมสำหรับการประยุกต์ใช้งาน ผู้วิจัยได้ทำการศึกษาค้นคว้าแนวคิดทฤษฎีและงานวิจัย ที่เกี่ยวข้องเพื่อเป็นพื้นฐานในการปฏิบัติงานและการวิเคราะห์ โดยนำเสนอดังหัวข้อต่อไปนี้

2.1 กระบวนการ CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM เป็นกระบวนการทำงานที่เป็นมาตรฐานสากลสำหรับการดำเนินการวิเคราะห์และประมวลผลข้อมูล ที่ช่วยให้ผู้ทำงานสามารถมีกรอบแนวทางที่ชัดเจนและโครงสร้างที่สามารถนำไปประยุกต์ใช้ในงานวิจัยอย่างมีประสิทธิภาพ ทำให้ผู้วิจัยสามารถวางแผนและดำเนินงานวิจัยในลักษณะที่มีขั้นตอนที่ชัดเจนและช่วยให้ผู้ทำวิจัยสามารถจัดทำเอกสารรายงานและแบ่งแยกหน้าที่ในการทำงานได้อย่างมีระบบ โดยกระบวนการ CRISP-DM (Cross Industry Standard Process for Data- Mining) เป็นกระบวนการทำงานที่มีขั้นตอนแบ่งออกเป็น 6 ขั้นตอนหลักที่ใช้ในการวิเคราะห์และประมวลผลข้อมูลในงานวิจัย ดังนี้ (Shearer, 2000)

1) การเข้าใจและการกำหนดขอบเขตของงาน (Business Understanding)

ขั้นตอนนี้เน้นในการเข้าใจและกำหนดเป้าหมายที่ต้องการให้การวิจัยเกิดขึ้น โดยการระบุคำถามวิจัยที่ต้องการตอบคำถามหรือปัญหาที่ต้องการแก้ไข และสำรวจข้อมูลที่มีอยู่เพื่อเข้าใจแนวทางของงานวิจัย

2) การทำกิจกรรมและวางแผน (Data Understanding)

ขั้นตอนนี้เน้นในการสำรวจและเข้าใจเกี่ยวกับข้อมูลที่ใช้ในงานวิจัย รวมถึงการเก็บรวบรวมข้อมูลและการทำความเข้าใจเกี่ยวกับคุณภาพและความเหมาะสมของข้อมูลที่ใช้ในการวิจัย

3) การเตรียมข้อมูล (Data Preparation)

ขั้นตอนนี้เน้นในการทำความสะอาดข้อมูลที่เกี่ยวข้องกับงานวิจัย รวมถึงการตรวจสอบความสมบูรณ์และความถูกต้องของข้อมูล การจัดรูปแบบข้อมูล และการนำข้อมูลมาเตรียมพร้อมใช้ในขั้นตอนถัดไปของกระบวนการวิจัย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4) การวิเคราะห์และการประมวลผล (Data Analysis and Modeling)

ขั้นตอนนี้เน้นในการนำข้อมูลที่เตรียมพร้อมใช้ มาวิเคราะห์และประมวลผลเพื่อหาคำตอบหรือแนวทางในการแก้ไขปัญหาวิจัย โดยใช้เทคนิควิเคราะห์ต่างๆ เช่น การสร้างแนวโน้ม (Trend Analysis) การสร้างโมเดลทางสถิติ (Statistical Modeling) การสร้างโมเดลทางเครื่องควบคุม (Machine learning- modeling) หรือการใช้เทคนิคการวิเคราะห์อื่นๆ เพื่อหาคำตอบหรือแนวทางที่เหมาะสมในงานวิจัย

5) การตรวจสอบและการประเมิน (Evaluation)

ขั้นตอนนี้เน้นในการประเมินผลลัพธ์ที่ได้จากการวิเคราะห์และประมวลผล โดยใช้เกณฑ์หรือตัวชี้วัดที่กำหนดไว้ก่อนหน้านี้ และตรวจสอบความถูกต้องและเปรียบเทียบกับเป้าหมายที่ตั้งไว้ในขั้นตอนแรกของงานวิจัย

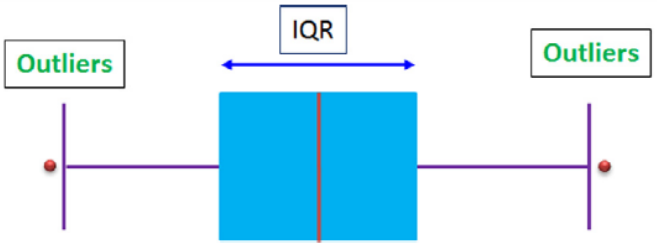
6) การนำเสนอผลและการนำไปใช้ (Deployment)

ขั้นตอนสุดท้ายเน้นในการนำเสนอผลการวิจัยและการนำผลลัพธ์ไปใช้ในการแก้ไขปัญหาหรือการตัดสินใจทางธุรกิจ รวมถึงการจัดทำรายงานและการนำผลลัพธ์ไปใช้ในระบบหรือกระบวนการที่เกี่ยวข้อง

ความสำคัญของ CRISP-DM ในงานวิจัยอยู่ที่การมีกระบวนการทำงานที่เป็นระเบียบและมีขั้นตอนที่ชัดเจนในการวิเคราะห์และประมวลผลข้อมูล ช่วยให้นักวิจัยสามารถมีการปรับปรุงและส่งเสริมประสิทธิภาพของงานวิจัยได้อย่างมีประสิทธิภาพและนำผลลัพธ์ไปใช้ให้เกิดประโยชน์จริงในองค์กรหรือองค์กรอื่น ๆ นอกจากนี้ยังช่วยให้งานวิจัยมีความสามารถในการทำซ้ำและทดสอบความถูกต้องของผลลัพธ์ และช่วยให้ผู้ที่ไม่เชี่ยวชาญในงานวิจัยสามารถเข้าใจกระบวนการและผลลัพธ์ที่ได้จากงานวิจัยนั้น ๆ ได้อย่างชัดเจน

2.2 การกำจัดค่านอกเกณฑ์ด้วยพิสัยควอไทล์ (Interquartile Range)

Interquartile Range หรือ IQR เป็นเครื่องมือทางสถิติที่ใช้ในการวัดการกระจายของข้อมูลในช่วงควอร์ไทล์ (Quartile) 25% ถึง 75% ของข้อมูล โดยค่าที่อยู่นอกช่วง IQR อาจถือว่าเป็นค่านอกเกณฑ์ (Kumar, 2023)



รูปที่ 2.1 ช่วงของค่านอกเกณฑ์ในแผนภาพรูปกล่อง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 2.1 แสดงช่วงของค่านอกเกณฑ์ในกราฟรูปกล่อง พบว่าค่าที่ถูกกำหนดให้เป็นค่านอกเกณฑ์คือค่าที่อยู่นอกแผนภาพรูปกล่อง (Box-Plot) จุดสีแดง ซึ่งสามารถหาได้จากสมการดังนี้

$$\text{Lower Limit} = \text{Q1} - 1.5\ \text{IQR} = \text{Q1} - 1.5(\text{Q3} - \text{Q1}) \quad (2.1)$$

$$\text{Upper Limit} = \text{Q3} - 1.5\ \text{IQR} = Q1 - 1.5(Q3 - Q1) \quad (2.2)$$

โดยกำหนดให้

Q1 คือ ค่าที่อยู่ควอร์ไทล์ที่ 25%

Q2 คือ ค่าที่อยู่ควอร์ไทล์ที่ 50%

Q3 คือ ค่าที่อยู่ควอร์ไทล์ที่ 75%

IQR คือ ช่วงระหว่าง *Q1* ถึง *Q3* หรือ ควอร์ไทล์ที่ 25% ถึง ควอร์ไทล์ที่ 75%

2.3 การเข้ารหัสป้ายกำกับ (Label Encoding)

เป็นกระบวนการแปลงค่าของป้ายกำกับหรือคลาสในข้อมูลจากชนิดข้อมูลที่เป็นข้อความหรือประเภทอื่น ๆ เป็นตัวเลข ซึ่งจะทำการแปลงข้อมูลที่เป็นอันดับหรือข้อมูลที่ไม่ได้เป็นตัวเลข จะช่วยทำให้แบบจำลองใช้พื้นที่ในการเก็บข้อมูลน้อยลงและสามารถนำมาใช้กับอัลกอริทึมการเรียนรู้ของเครื่องได้ (จิราภรณ์ และคณะ, 2563)

ตารางที่ 2.1 ข้อมูลตัวอย่างที่ถูกแปลงด้วยการเข้ารหัสป้ายกำกับ

ตัวแปรที่ยังไม่ถูกแปลงค่า	ตัวแปรที่ถูกแปลงค่า
Dog	1
Cat	2
Dog	1
Mouse	3

จากตารางที่ 2.1 แสดงข้อมูลตัวอย่างที่ถูกแปลงด้วยการเข้ารหัสป้ายกำกับ ตัวแปรที่ยังไม่ถูกแปลงค่าคือ Dog, Cat, และ Mouse ที่เป็นตัวแปรประเภทข้อความ หลังจากแปลงค่าเข้ารหัสป้ายกำกับจะได้เป็น Dog แทนด้วยเลข 1 Cat แทนด้วยแลข 2 และMouse แทนด้วยเลข 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.4 การปรับค่าให้อยู่ในช่วงสูงสุดและต่ำสุด (Min-Max Normalization)

การปรับค่าให้อยู่ในช่วงสูงสุดและต่ำสุด (Min-Max Normalization) เป็นกระบวนการปรับค่าข้อมูลให้อยู่ในช่วงที่กำหนดไว้ เพื่อทำให้ข้อมูลให้อยู่ในรูปคำแนนปรกติมาตรฐานน้อยที่สุดและมากที่สุด ซึ่งส่วนมากจะเป็นช่วงระหว่าง 0 ถึง 1 (ปิยวรรณ และคณะ, 2564)

$$X_{new} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2.3}$$

โดย

X_{new} หมายถึง ค่าใหม่หลังจากที่ทำการแปลงข้อมูลแล้ว

$X_{\min}$ หมายถึง ค่าที่ต่ำสุดของคุณลักษณะนั้นในข้อมูลชุดฝึกสอน

$X_{\max}$ หมายถึง ค่าที่มากที่สุดของคุณลักษณะนั้นในข้อมูลชุดฝึกสอน

X หมายถึง ตัวแปรที่ต้องการจะแปลงค่า

2.5 สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient)

การวัดความเหมือนหรือความสัมพันธ์ของคุณสมบัติสองอย่างจะใช้วัดความสัมพันธ์ระหว่างกันด้วย Correlation Measures ซึ่งเป็นวิธีการที่นิยมใช้งานอย่างแพร่หลาย โดยเมื่อสองคุณสมบัติเป็นฟังก์ชันเชิงเส้นตรงค่าสัมประสิทธิ์สหสัมพันธ์จะเป็นบวก 1 หรือลบ 1 หากไม่มีความสัมพันธ์กัน ค่าค่าสัมประสิทธิ์สหสัมพันธ์จะเป็น 0 การหาความสัมพันธ์ระหว่างคุณสมบัติสองอย่างสามารถใช้วิธีการ Correlation ได้โดยแบ่งออกเป็นสองประเภท คือ การวัดความสัมพันธ์เชิงเส้นแบบคลาสสิก และการวัดความสัมพันธ์โดยใช้ทฤษฎีสารสนเทศ จากทั้งสองวิธีนี้ วิธีการวัดความสัมพันธ์แบบเชิงเส้นเป็นวิธีที่นิยมใช้งานมากที่สุด โดยตามวิทยาการมาตรฐาน สำหรับคู่ของตัวแปร (X, Y) สมการ Correlation Coefficient ‘r’ จะเป็นดังสมการที่ 2.3 (Blessie and Karthikeyan, 2012)

$$r = \frac{\sum (X_i - \overline{X}_i)(Y_i - \overline{Y}_i)}{\sqrt{\sum (X_i - \overline{X}_i)^2}\sqrt{\sum (Y_i - \overline{Y}_i)^2}} \tag{2.4}$$

โดยกำหนดให้

r คือ ค่าสัมประสิทธิ์สหสัมพันธ์

X_i คือ ค่าตัวแปร X ณ ชุดข้อมูลที่ i

$\overline{X}_i$ คือ ค่าเฉลี่ยของตัวแปร X

Y_i คือ ค่าตัวแปร Y ณ ชุดข้อมูลที่ i

$\overline{Y}_i$ คือ ค่าเฉลี่ยของตัวแปร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การแปลความหมายค่าสัมประสิทธิ์สหสัมพันธ์ (สุจิตรา, 2563)

$0 \le r < 0.3$	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในระดับต่ำมาก
$0.3 \le r < 0.5$	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในระดับต่ำ
$0.5 \le r < 0.7$	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในระดับปานกลาง
$0.7 \le r < 0.9$	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในระดับสูง
$0.9 \le r < 1$	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในระดับสูงมาก
r เป็นบวก	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในทิศทางเดียวกัน (ตัวแปรหนึ่งมีค่าสูงตัวแปรอีกตัวหนึ่งจะมีค่าสูงไปด้วย)
r เป็นลบ	แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในทิศทางตรงกันข้าม (ตัวแปรหนึ่งมีค่าสูงตัวแปรอีกตัวหนึ่งจะมีค่าต่ำไปด้วย)

2.6 วิเคราะห์ความแปรปรวน (Analysis of Variance: ANOVA)

เป็นเทคนิคการวิเคราะห์ความแปรปรวนซึ่งเป็นการวิเคราะห์ข้อมูลที่ใช้ในการทดสอบ สมมติฐานเกี่ยวกับความแตกต่างของค่าเฉลี่ยกรณีประชากรมากกว่า 3 กลุ่ม (k กลุ่ม) ขึ้นไป กำหนดสมมติฐานของการทดสอบดังนี้

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: \mu_i \neq \mu_j \text{ อย่างน้อย } 1 \text{ คู่ โดยที่ } i \neq j$$

2.6.1 การทดสอบเอฟ (F-test)

การทดสอบเอฟจะใช้การวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA) ในการเปรียบเทียบความแตกต่างของค่าเฉลี่ยตั้งแต่ 3 กลุ่มขึ้นไปและการวิเคราะห์ต้องเป็นไปตามข้อกำหนดเบื้องต้น 3 ข้อดังนี้ (จิรภา, 2560)

1) ข้อมูลต้องมีการแจกแจงปรกติ
2) ข้อมูลแต่ละกลุ่มต้องมีความแปรปรวนเท่ากัน
3) ข้อมูลแต่ละกลุ่มต้องอิสระกัน

ตัวสถิติทดสอบ

$$F_{cal} = \frac{(\sum_{i=1}^{k} \frac{x_{i\bullet}^2}{n_i} - \frac{x_{\bullet\bullet}^2}{n}) / (k-1)}{(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} \frac{x_{i\bullet}^2}{n_i}) / (n-k)} = \frac{MSB}{MSW} \qquad (2.5)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แสดงในรูปแบบตารางได้ดังนี้

ตารางที่ 2.2 ตารางวิเคราะห์ความแปรปรวน

แหล่งความแปรปรวน	**SS**	**df**	**MS**	F_{cal}
ระหว่างกลุ่ม	SSB	k-1	MSB	$\frac{MSB}{MSW}$
ภายในกลุ่ม	SSW	n-k	MSW	
รวม	SST	n-1		

จากตารางที่ 2.2 แสดงตารางวิเคราะห์ความแปรปรวนโดยแต่ละค่าในตารางสามารถคำนวณได้ดังนี้

1) ผลรวมของจำนวนค่าสังเกตทั้งหมด

$$n = \sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k \tag{2.6}$$

2) ผลบวกกำลังสองของยอดรวม (Total Sum of Square, SST)

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{x_{\bullet\bullet}^{\ 2}}{n} \tag{2.7}$$

3) ผลบวกกำลังสองระหว่างกลุ่ม (Between Group Sum of Square, SSB)

$$SSB = \sum_{i=1}^{k} \frac{x_{i\bullet}^2}{n_i} - \frac{x_{\bullet\bullet}^2}{n} \tag{2.8}$$

4) ผลบวกกำลังสองภายในกลุ่ม (Within Groups Sum of Square, SSW)

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} \frac{x_{i\bullet}^2}{n_i} \tag{2.9}$$

5) กำลังสองเฉลี่ยระหว่างกลุ่ม (Mean Square Between Groups)

$$MSB = \frac{SSB}{k-1} \tag{2.10}$$

6) กำลังสองเฉลี่ยภายในกลุ่ม (Mean Square Within Groups)

$$MSW = \frac{SSW}{n-k} \tag{2.11}$$

จะได้สถิติทดสอบ $F_{cal} = \frac{MSB}{MSW}$ (2.12)

จะได้เขตวิกฤตดังนี้ หาก $\mathrm{F_{cal}}$ มากกว่า $F_{\alpha,df1,df2}$ จะทำให้ปฏิเสธสมมติฐานว่าง โดยมี $\mathrm{df_1 = k - 1}$ และ $\mathrm{df_2 = n - k}$ หรือ p-value $\leq \alpha\ (0.05)$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

n_i คือ ค่าสังเกตของแต่ละกลุ่มตัวอย่าง i=1,2,...,k

k คือ จำนวนกลุ่มตัวอย่าง

n คือ ผลรวมของจำนวนของค่าสังเกตทั้งหมด

α คือ ระดับนัยสำคัญ

$x_{i.}$ คือ ผลรวมของค่าสังเกตของแต่ละกลุ่มตัวอย่างที่ i=1,2,...,k

$x_{..}$ คือ ผลรวมของค่าสังเกตทุกค่า

2.7 ข้อมูลที่ไม่สมดุล

ข้อมูลที่ไม่สมดุล คือ ข้อมูลที่มีจำนวนข้อมูลของกลุ่มหนึ่งมากกว่าจำนวนข้อมูลของกลุ่มที่เหลือเป็นจำนวนมาก ซึ่งข้อมูลที่ไม่สมดุลนี้อาจจะส่งผลกระทบต่อแบบจำลองที่เราสร้างขึ้น โดยเฉพาะแบบจำลองในงานจำแนกข้อมูลหรือการจำแนกหมวดหมู่ โดยแบบจำลองจะจำแนกกลุ่มข้อมูลจำนวนมากได้แม่นยำแต่ในขณะเดียวกันจะจำแนกกลุ่มข้อมูลจำนวนน้อยได้ไม่แม่นยำเท่าที่ควร โดยทั่วไปแล้วกลุ่มที่มีจำนวนน้อยจะถูกเรียกว่า คลาสส่วนน้อย (Minority Class) และกลุ่มข้อมูลที่มีจำนวนมากนั้นจะถูกเรียกว่า คลาสส่วนมาก (Majority Class) (ภาสพิชญ์, 2557)

ข้อมูลที่ไม่สมดุลนั้นมันเกิดจากหลายสาเหตุหลายปัจจัยที่เกิดขึ้น เช่น ข้อมูลที่ไม่สมดุลอาจจะเกิดจากธรรมชาติของข้อมูลอยู่แล้วซึ่งอาจจะพบเจอได้ในข้อมูลการวินิจฉัยทางการแพทย์ต่าง ๆ ที่มีข้อมูลเกี่ยวกับผู้ป่วยด้วยโรคร้ายแรงซึ่งแน่นอนว่าโรคร้ายแรงนั้นย่อมน้อยกว่าผู้ป่วยที่สุขภาพดีหรือโรคไม่ร้ายแรงจำนวนมาก ข้อมูลบัตรเครดิตที่มีข้อมูลของลูกค้าปกติซึ่งมากกว่าข้อมูลลูกค้าที่ผิดปกติหรือว่าข้อมูลที่ไม่สมดุลอาจจะเกิดจากข้อจำกัดต่าง ๆ ในการเก็บข้อมูลเป็นต้น

รูปที่ 2.2 การกระจายของข้อมูลที่มีปัญหาความไม่สมดุลของข้อมูล

จากรูปที่ 2.2 แสดงการกระจายของข้อมูลที่มีปัญหาความไม่สมดุลของข้อมูล มีข้อมูลจำนวน 27 ตัวอย่างซึ่งแบ่งออกเป็น 2 กลุ่ม กลุ่มแรกมีจำนวน 19 ตัวอย่าง กลุ่มที่ 2 มีเพียง 8 ตัวอย่าง ซึ่งหาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เราสร้างกราฟดูการกระจายจากรูปที่ 2.2 จะพบว่าข้อมูลส่วนมากเป็นจุดสีเขียวมากกว่าข้อมูลที่เป็นจุดสีส้ม เมื่อนำข้อมูลชุดนี้เข้าแบบจำลองที่สร้างขึ้นผลลัพธ์อาจจะพบว่าความถูกต้องในการจำแนกข้อมูลมีความเอนเอียงกล่าวคือ ความสามารถในการจำแนกข้อมูลส่วนน้อยไม่สามารถจำแนกประเภทข้อมูลได้ดีแต่ในขณะเดียวกันแบบจำลองสามารถจำแนกประเภทข้อมูลที่เป็นกลุ่มข้อมูลที่อยู่ในคลาสส่วนมากได้อย่างถูกต้องแม่นยำกว่า ทั้งนี้หากนำข้อมูลที่ไม่เคยผ่านการเรียนรู้เข้าไปทดสอบจะพบว่าความน่าจะเป็นของการจำแนกข้อมูลจะเกิดความเอนเอียงไปยังกลุ่มคลาสมากส่วนใหญ่ส่งผลให้ข้อมูลกลุ่มที่เป็นคลาสส่วนน้อยเกิดการจำแนกผิดกลุ่มได้

2.8 วิธีการแก้ปัญหาข้อมูลไม่สมดุล

นักวิจัยต่าง ๆ ได้เสนอวิธีการต่าง ๆ เพื่อมาแก้ไขปัญหาข้อมูลที่ไม่สมดุล โดยวิธีการนั้นได้ทำการแบ่งออกเป็น 3 ระดับคือ 1.การแก้ปัญหาข้อมูลไม่สมดุลที่ระดับข้อมูล (Data Level Solutions) 2. การแก้ไขปัญหาข้อมูลที่ไม่สมดุลที่ระดับขั้นตอนวิธีการ (Algorithmic Level Solutions) 3.การแก้ปัญหาข้อมูลที่ไม่สมดุลด้วยการเรียนรู้แบบมีค่าใช้จ่าย โดยทั้ง 3 ระดับมีเป้าหมายตรงกันคือการเพิ่มประสิทธิภาพและความแม่นยำในการจำแนกประเภทของแบบจำลองที่สร้าง (López et al., 2012)

2.8.1 การแก้ปัญหาระดับข้อมูล

เป็นการแก้ปัญหาก่อนที่จะสร้างแบบจำลองขึ้นมาซึ่งอยู่ในขั้นตอนเตรียมข้อมูล (Preprocessing) ซึ่งจะเกี่ยวข้องกับข้อมูลโดยตรงโดยการทำให้ข้อมูลที่ไม่สมดุลให้กลายเป็นข้อมูลที่สมดุลกันมากขึ้นด้วยเทคนิคต่างๆ ซึ่งมีเทคนิคการสุ่มเลือกข้อมูลที่ได้รับความนิยมจะแบ่งออกเป็น 3 กลุ่มดังต่อไปนี้

1) วิธีสุ่มเกิน (Oversampling) เป็นวิธีการเพิ่มจำนวนที่อยู่ในคลาสน้อยให้มีจำนวนใกล้เคียงกับคลาสมากซึ่งการเพิ่มข้อมูลนั้นจะเพิ่มโดยการสุ่มจากข้อมูลเดิมที่มีอยู่ หรือสร้างขึ้นมาใหม่จากตัวอย่างที่มีอยู่นั้น โดยวิธีการสุ่มเกินที่ได้รับความนิยม เช่น Synthetic Minority Oversampling Technique (SMOTE)

2) วิธีสุ่มลด (Undersampling) เป็นวิธีการที่ลดจำนวนข้อมูลที่อยู่ในคลาสมากให้มีจำนวนข้อมูลใกล้เคียงกับคลาสน้อยโดยวิธีสุ่มลดที่ได้รับความนิยม เช่น Wilson’s edited nearest neighbor (ENN)

3) วิธีผสมผสาน (Combination of Oversampling and Undersampling) เป็นวิธีการที่นำเทคนิควิธีสุ่มเกินและวิธีส่วนลดมาทำงานร่วมกัน เช่น การเอาเทคนิค SMOTE มารวมกับเทคนิค ENN จะได้เทคนิคใหม่ที่ชื่อว่า “SMOTE-ENN”

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.8.2 การแก้ปัญหาระดับขั้นตอนวิธีการ

เป็นการแก้ปัญหาโดยการปรับการเรียนรู้ของแบบจำลองสำหรับการจำแนกข้อมูลเดิมที่มีอยู่ให้สามารถเรียนรู้ข้อมูลที่ไม่สมดุลได้โดยการเอนเอียงไปทางข้อมูลส่วนน้อยมากกว่า เช่น การปรับ นำหนักคลาส (Weight Class) ในอัลกอริทึมป่าไม้สุ่มหรือซัพพอร์ตเวกเตอร์แมชชีน

2.8.3 การแก้ปัญหาด้วยวิธีการเรียนรู้แบบมีค่าใช้จ่าย

เป็นวิธีการแก้ปัญหาที่นำทั้งการแก้ปัญหาที่ระดับข้อมูลและการแก้ปัญหาที่ระดับขั้นตอนและวิธีการมาทำงานร่วมกัน โดยที่ระดับข้อมูลจะทำการเพิ่มค่าใช้จ่าย (Cost) ที่พิเศษสำหรับกรณีที่มีการจำแนกผิดพลาดและระดับขั้นตอนวิธีการจะทำการปรับปรุงการเรียนรู้ของอัลกอริทึมมาตรฐานให้สอดคล้องกับการจำแนกข้อมูลที่ผิดพลาด

2.9 การผสมผสานเทคนิควิธีสุ่มเกินและวิธีสุ่มลด

การผสมผสานเทคนิควิธีสุ่มเกินและวิธีสุ่มลด หรือ Combination of Oversampling and Undersampling เป็นการปรับปรุงประสิทธิภาพของแบบจำลองที่ถูกฝึกสอนด้วยข้อมูลไม่สมดุล โดยทำการเพิ่มจำนวนตัวอย่างของคลาสน้อยและลดจำนวนตัวอย่างของคลาสมาก เพื่อสร้างความสมดุลให้กับข้อมูลโดยมีเทคนิคมากมายที่สามารถปรับความสมดุลให้แก้ข้อมูลได้ เช่น การผสมผสานระหว่างเทคนิค Synthetic Minority Oversampling Technique (SMOTE) กับเทคนิค Edited Nearest Neighbor (ENN) (ภาสพิชญ์, 2557)

2.9.1 เทคนิค Synthetic Minority Oversampling Technique (SMOTE)

เทคนิค Synthetic Minority Over-sampling Technique (SMOTE) เป็นเทคนิคที่ใช้ในการจัดการกับปัญหาข้อมูลไม่สมดุลในชุดข้อมูลที่มีคลาสน้อย (Minority Class) โดยการสร้างตัวอย่างสามัญใหม่ของคลาสน้อย วิธีการของ SMOTE คือการสุ่มตัวอย่างของคลาสน้อยแล้วนำตัวอย่างที่สุ่มมาไปสร้างตัวอย่างใหม่ (Feng and Hang, 2013) ตามสมการที่ 2.13

$$X_{new} = X_i + \left|X_i' - X_i\right| \times \delta \tag{2.13}$$

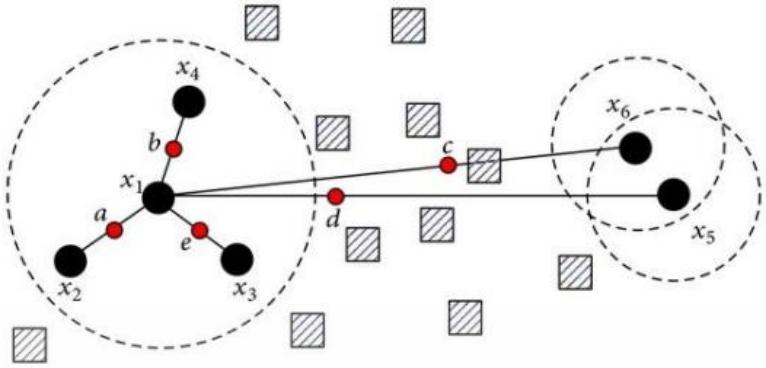
$\mathbf{x_{new}}$ คือ ตัวอย่างใหม่

$\mathbf{x_i}$ คือ ตัวอย่างในคลาสน้อย

$\mathbf{x_i'}$ คือ ตัวอย่างที่เป็นเพื่อนบ้านใกล้ที่สุดของ x_i

δ คือ ตัวเลขที่สุ่มขึ้น โดยที่ $\delta \in [0,1]$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



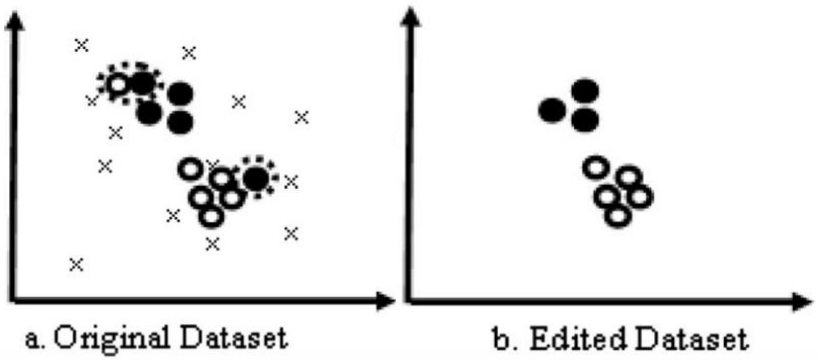
รูปที่ 2.3 หลักการทำงานของ SMOTE (Synthetic Minority Oversampling Technique)

จากรูปที่ 2.3 แสดงหลักการทำงานของ SMOTE โดยใช้หลักการทำงานของเพื่อนบ้านใกล้สุด k ตัวเพื่อสร้างข้อมูลใหม่ โดยเริ่มต้นจากการเลือกข้อมูลแบบสุ่มจากคลาสของข้อมูลชนกลุ่มน้อย หลังจากนั้นตั้งค่า k-Nearest Neighbors หรือเพื่อนบ้านใกล้เคียงเพื่อหาเพื่อนบ้านที่ใกล้ที่สุดจากข้อมูล จากนั้นข้อมูลสังเคราะห์จะสร้างขึ้นระหว่างข้อมูลสุ่มและเพื่อนบ้านที่ใกล้ที่สุดจากจำนวน k ที่เลือกแบบสุ่ม ขั้นตอนนี้จะทำซ้ำหลายครั้งมากพอจนกระทั่งคลาสของข้อมูลชนกลุ่มน้อยมีสัดส่วนเดียวกับคลาสของข้อมูลชนกลุ่มใหญ่

2.9.2 เทคนิค Edited Nearest Neighbor (ENN)

เป็นอัลกอริทึมที่ใช้ในการแก้ไขข้อมูลโดยการลดจำนวนข้อมูลที่อยู่ในคลาสมากให้มีจำนวนข้อมูลใกล้เคียงกับคลาสน้อยเพื่อลดความซับซ้อนของข้อมูล โดยอัลกอริทึม ENN จะคำนวณค่าระยะทาง (Distance) ระหว่างตัวอย่างข้อมูลทั้งหมดในชุดข้อมูลการฝึกสอนโดยใช้กฎของ k-NN ซึ่งกำหนดให้กับตัวอย่างที่ต้องการแก้ไข จากนั้นจำแนกตัวอย่างที่ถูกจำแนกผิดพลาดโดย k-NN และทำเครื่องหมาย (Mark) ตัวอย่างนั้นในชุดข้อมูลการฝึกสอน จากนั้นทำการลบตัวอย่างที่ถูกจำแนกผิดพลาดออกจากชุดข้อมูลการฝึกสอน ที่ได้รับเครื่องหมายในขั้นตอนก่อนหน้า ให้เหลือเฉพาะตัวอย่างที่ถูกจำแนกถูกต้อง (Donghai Guan et al., 2009) ดังรูปที่ 2.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

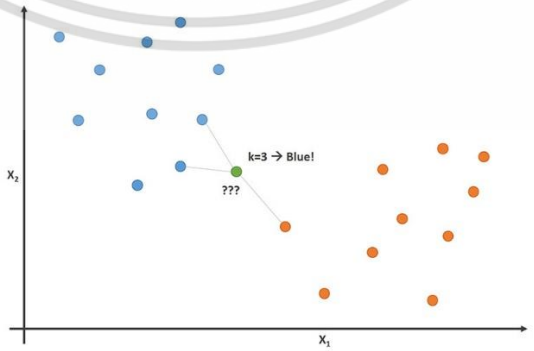


รูปที่ 2.4 หลักการทำงานของ ENN (Edited Nearest Neighbor)

จากรูปที่ 2.4 แสดงข้อมูลต้นฉบับก่อนและหลังจากการทำ Edited Nearest Neighbor โดยจะเห็นว่า ข้อมูลต้นฉบับ (Original Dataset) ที่ได้ทำเครื่องหมาย (Mark) หลังจากการจำแนกตัวอย่างที่ถูกจำแนกผิดพลาดโดย k-NN หรือหมายความว่าไม่เป็นไปตามหลักการเพื่อนบ้าน k-NN ดังนั้นข้อมูลที่ถูกปรับแต่ง (Edited Dataset) จึงได้ทำการลบข้อมูลที่ถูกจำแนกผิดพลาดนั้นออกเพื่อลดความซับซ้อนของข้อมูล

2.10 วิธีเพื่อนบ้าน k ตัว (K-Nearest Neighbors)

วิธีเพื่อนบ้านใกล้สุด k ตัว จะใช้หลักการเปรียบเทียบความคล้ายคลึงกันของข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายคลึงหรืออยู่ใกล้กับข้อมูลใดมากที่สุด k ตัว จากนั้นจะทำการตัดสินใจว่าคำตอบของข้อมูลที่สนใจนั้นควรเป็นคำตอบเดียวกับข้อมูลที่อยู่ใกล้ที่สุด k ตัวนั้น โดยที่ k คือความถี่ของข้อมูลที่อยู่ใกล้กับข้อมูลที่สนใจ สามารถทำได้โดยกำหนดค่า k จากนั้นคำนวณหาระยะห่างระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่นๆ ทุกตัวด้วยวิธีระยะห่างยูคลิเดียน (Euclidian Distance) (พัชนา, 2561) ดังสมการที่ 2.14



รูปที่ 2.5 วิธีเพื่อนบ้านใกล้สุด k ตัว โดยค่า k เท่ากับ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$dist(X_i, X_j) = \sqrt{\sum_{k=1}^{n}(X_{i,k} - X_{j,k})^2} \quad (2.14)$$

โดยที่ $dist(X_i, X_j)$ คือ ระยะห่างระหว่างตัวอย่าง X_i กับตัวอย่าง X_j

n คือ จำนวนข้อมูล

X_i คือ สมบัติทั้งหมดของตัวอย่าง X_i

$X_{i,k}$ คือ สมบัติตัวที่ k ของตัวอย่าง X_i

X_j คือ สมบัติทั้งหมดของตัวอย่าง X_j

$X_{j,k}$ คือ สมบัติตัวที่ k ของตัวอย่าง X_j

2.11 วิธีนาอีฟเบย์ (Naïve Bayes)

เป็นวิธีที่ให้ผลการจำแนกได้ดีไม่แตกต่างวิธีการอื่นโดยมีขั้นตอนวิธีการทำงานที่ไม่ซับซ้อน การเรียนรู้ของนาอีฟเบย์จะเป็นการเรียนรู้โดยใช้หลักการของความน่าจะเป็น (Probability) ซึ่งมีพื้นฐานมาจากทฤษฎีเบส์ (Bayes Theorem) หรือทฤษฎีว่าด้วยโอกาสที่จะเกิดของเหตุการณ์ต่างๆ ซึ่งจะใช้การคำนวณความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability) ดังสมการที่ 2.15 (Dietrich et al., 2015)

$$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \quad (2.15)$$

D แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็นภายหลัง (Posteriori Probability) ของการเกิดเหตุการณ์ h คือ $P(h|D)$

โดยที่ $P(h)$ คือ ค่าความน่าจะเป็นก่อน (Prior probability) ของการเกิดเหตุการณ์ h

$P(D)$ คือ ค่าความน่าจะเป็นก่อนของชุดข้อมูลตัวอย่าง D

$P(h|D)$ คือ ค่าความน่าจะเป็นของ h เมื่อรู้ D

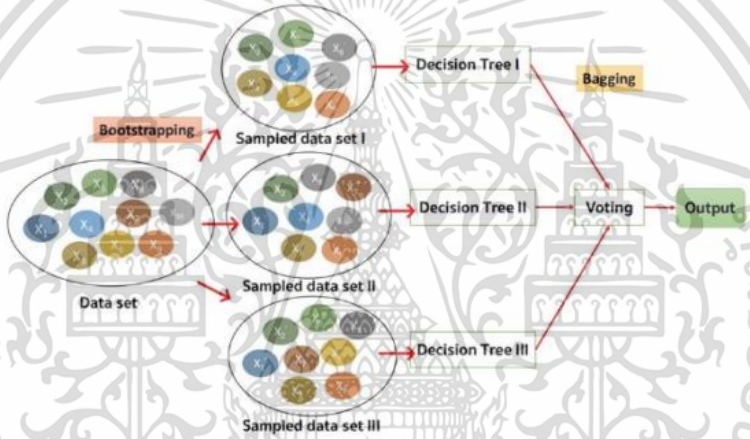
$P(D|h)$ คือ ค่าความน่าจะเป็นของ D เมื่อรู้ h

กำหนดให้ $P(h)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h และ $P(h|D)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D แล้วจากตัวแปรที่กำหนด และแนวคิดของเบส์นั้นเราสามารถพยากรณ์เหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่าง ๆ ได้ดังสมการที่ 4 ข้างต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.12 วิธีป่าไม้สุ่ม (Random Forest)

วิธีป่าไม้สุ่ม เป็นแบบจำลองการเรียนรู้แบบมีผู้สอนที่ถูกพัฒนาขึ้นจาก วิธีต้นไม้ตัดสินใจ (Decision Tree) โดยที่วิธีการป่าไม้สุ่มนั้นเป็นการเพิ่มจำนวนต้นไม้เป็นต้นไม้หลาย ๆ ต้นทำให้ประสิทธิภาพในการทำงานสูงขึ้นแม่นยำมากขึ้น ซึ่งวิธีป่าไม้สุ่มนั้นเป็นแบบจำลองที่ได้รับความนิยมไปอย่างมากโดยหลักการของโมเดลป่าไม้สุ่มคือการสร้างโมเดลจากวิธีการต้นไม้ตัดสินใจหลาย ๆ โมเดล โดยแต่ละโมเดลจะได้รับชุดข้อมูลไม่เหมือนกันซึ่งเป็นข้อมูลชุดย่อยของชุดข้อมูลทั้งหมดตอนการฝึกสอนแบบจำลอง ในการพยากรณ์ก็จะให้แต่ละต้นไม้ตัดสินใจและพยากรณ์ผลของตัวเองจากนั้นเมื่อได้ผลแต่ละต้นไม้แล้วจะทำการโหวต (Vote) หรือทำการตัดสินใจเลือกค่าที่ดีที่สุด (จิรวัฒน์ และคณะ, 2565)



รูปที่ 2.6 กระบวนการทำงานของวิธีป่าไม้สุ่ม

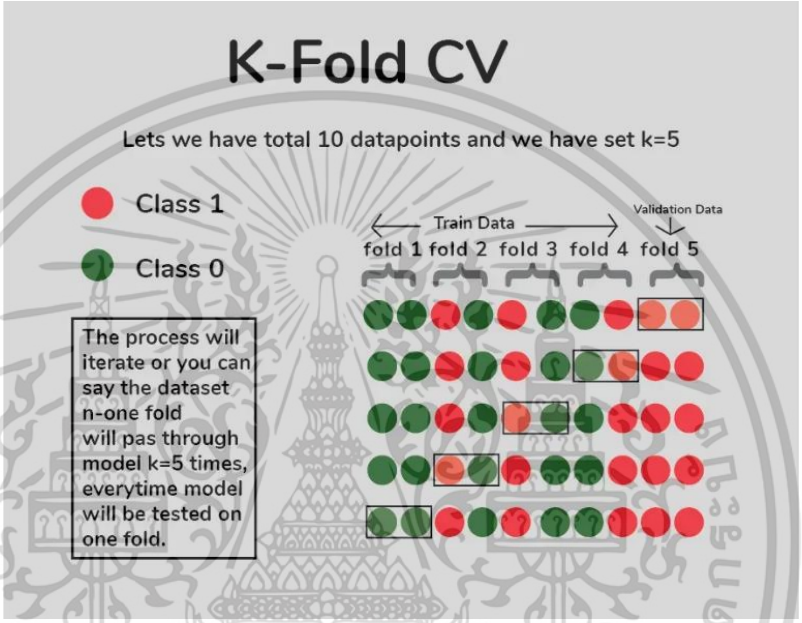
จากรูป 2.6 จะแสดงให้เห็นถึงการสร้างต้นไม้และเป็นประเภทข้อมูลแบบ Bagging (Bootstrap Aggregation) โดยต้นไม้แต่ละต้นที่นำมาฝึกสอนในแบบจำลอง จะมีตัวแปรแต่ละตัวเป็นส่วนหนึ่งของ คุณลักษณะ (Feature) ซึ่งจะนำมาฝึกสอนในรูปแบบสุ่มและในส่วนขั้นตอนการพยากรณ์ข้อมูล จะกำหนดให้ต้นไม้แต่ละต้นพยากรณ์ในต้นของตัวเองและคัดเลือกผลพยากรณ์สุดท้ายจากค่าพยากรณ์ที่ได้รับการโหวตมากที่สุดเทคนิคดังกล่าวเรียกว่า "การสุ่มตัวอย่างข้อมูล" และเทคนิคการจำแนกประเภทข้อมูลแบบ Bagging (สุภาภรณ์, 2564)

2.13 วิธีการตรวจสอบไขว้ (K-Fold Cross-Validation)

วิธี K-fold Cross Validation เป็นเทคนิคที่ใช้ในการประเมินประสิทธิภาพของแบบจำลองเพื่อวัดความสามารถในการพยากรณ์ผลลัพธ์ของแบบจำลอง โดยการแบ่งข้อมูลออกเป็น K ส่วนเท่า ๆ กัน (folds) โดยแต่ละ Fold จะมีขนาดเท่า ๆ กันหรือใกล้เคียงกัน จากนั้นจะทำการทดสอบโมเดล

เอกสารนี้ ... ไม่ว่ากรณีใด ...

K ครั้ง โดยใช้แต่ละ Fold ในการทดสอบโมเดลคนละครั้งและใช้ Folds ที่เหลือในการฝึกโมเดล กล่าวคือจะใช้ K-1 Folds ในการฝึกโมเดลและ Fold ที่เหลือในการทดสอบโมเดลจนครบ K รอบ เมื่อทำการทดสอบโมเดล K ครั้งเสร็จแล้วจะนำผลลัพธ์ที่ได้มาเฉลี่ยกันเพื่อประเมินประสิทธิภาพของโมเดล ซึ่งการใช้ K-fold Cross Validation จะช่วยประเมินประสิทธิภาพในการจำแนกและช่วยลดความผิดพลาดจากการใช้ข้อมูลฝึกสอนและข้อมูลทดสอบเดียวกันโดยตรงและช่วยป้องกันการ Overfitting ของโมเดลในการฝึกโดยใช้ข้อมูลฝึกสอนหนึ่งเดียว (รัชพล และจรัญ, 2561)



รูปที่ 2.7 วิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ด้วยการตรวจสอบไขว้

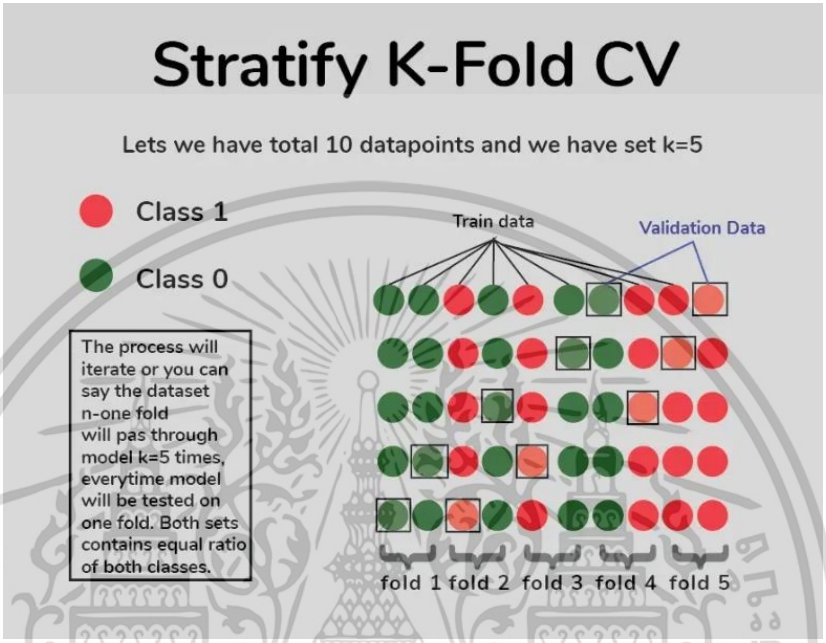
จากรูปที่ 2.7 เป็นการแสดงการวิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ ที่กำหนดให้ k เท่ากับ 5 หรือ 5 ชุดข้อมูลจากนั้นทำการแบ่งข้อมูลของแต่ละชุดให้เป็นข้อมูลชุดทดสอบ 1 Fold และข้อมูลชุดฝึกสอน 4 Fold จากนั้นทำการสลับข้อมูลชุดทดสอบกับข้อมูลชุดฝึกสอนไปเรื่อย ๆ จนครบทุกช่องแต่ข้อเสียมีดังนี้

1) อาจเกิดความเบี่ยงเบน เนื่องจากไม่ทราบว่าจำนวนตัวอย่างของคลาสบวก (Positive Class) และคลาสลบ (Negative Class) ที่ใช้ในการฝึกและการตรวจสอบมีเท่าไรอาจเกิดการเอาตัวอย่างของคลาสเดียวกันมาใช้หลายครั้งในการฝึกหรือการตรวจสอบ ซึ่งอาจทำให้โมเดลมีความแม่นยำต่ำกว่าที่ควร

2) ถ้ากำหนดจำนวน Fold น้อยเกินไปอาจทำให้โมเดลไม่สามารถเข้าใจและจำกันข้อมูลได้อย่างเพียงพอ หรือถ้ากำหนดจำนวน Fold มากเกินไปอาจทำให้มีเวลาในการฝึกและการตรวจสอบที่นานขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เพื่อแก้ไขปัญหาที่หนึ่งของการใช้ K-Fold Validation จะต้องรักษาอัตราส่วนของจำนวนข้อมูลจากทั้งสองคลาสให้เท่ากัน โดยจะต้องทำการจัด Stratification ให้กับชุดข้อมูลก่อนแล้วจึงแบ่งเป็นส่วนต่าง ๆ เพื่อให้แน่ใจว่าจะมีอัตราส่วนของข้อมูลจากทั้งสองคลาสที่เท่ากันเหมือนกันในการฝึกและการทดสอบ



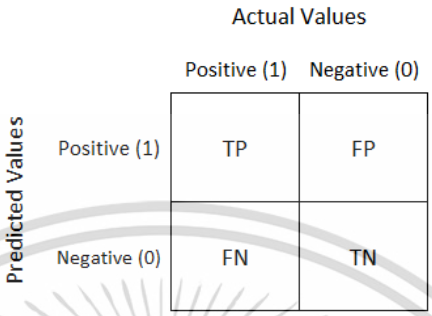
รูปที่ 2.8 วิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ด้วยการตรวจสอบไขว้แบบแบ่งชั้น

จากรูปที่ 2.8 เป็นการแสดงการวิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ด้วยการตรวจสอบไขว้แบบแบ่งชั้น k = 5 หรือ 5 ชุดข้อมูลจากนั้นทำการแบ่งข้อมูลของแต่ละชุดให้เป็นข้อมูลชุดทดสอบ 1 Fold และข้อมูลชุดฝึกสอน 4 Fold แต่การแบ่งชุดข้อมูลทดสอบ หรือ Validation Data ครั้งนี้จะทำการแบ่งข้อมูลให้มีคลาสที่เท่ากันทุกครั้งในการทดสอบ เพื่อลดความเบี่ยงเบนที่อาจเกิดการเอาตัวอย่างของคลาสเดียวกันมาใช้หลายครั้งในการฝึกหรือการตรวจสอบ (Prusty et al., 2022)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.14 เมทริกซ์ความสับสน (Confusion Matrix)

เป็นเครื่องมือที่สำคัญในการประเมินวัดความแม่นยำและประสิทธิภาพของแบบจำลองที่พยากรณ์กับข้อมูลที่เกิดขึ้นจริง (เอกพันธ์, 2563) ดังรูปที่ 2.9



รูปที่ 2.9 เมทริกซ์ความสับสน

บวกจริง (True Positive : TP) คือ สิ่งที่พยากรณ์ตรงกับสิ่งที่เกิดขึ้นจริง
ลบจริง (True Negative : TN) คือ สิ่งที่พยากรณ์ตรงกับสิ่งที่เกิดขึ้นจริง
บวกเท็จ (False Positive :FP) คือ สิ่งที่พยากรณ์ไม่ตรงกับสิ่งที่เกิดขึ้น
ลบเท็จ (False Negative : FN) คือ สิ่งที่พยากรณ์ไม่ตรงกับที่เกิดขึ้นจริง

ค่าความแม่น (Accuracy) เป็นการวัดความแม่นยำของแบบจำลองโดยรวม กล่าวคือ แบบจำลองพยากรณ์ถูกกี่ครั้งจากจำนวนการพยากรณ์ทั้งหมด สามารถหาได้ดังสมการที่ 2.16 (ธนาภัทร, 2563)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2.16)$$

ค่าความเที่ยง (Precision) เป็นค่าที่แบบจำลองพยากรณ์เป็นคลาสที่กำลังพิจารณาและถูกต้องต่อค่าที่แบบจำลองพยากรณ์ว่าเป็นคลาสที่กำลังพิจารณาทั้งถูกและผิด สามารถหาได้ดังสมการที่ 2.17 (ธนาภัทร, 2563)

$$Precision = \frac{TP}{TP + FP} \quad (2.17)$$

ค่าระลึก (Recall) เป็นค่าที่แบบจำลองพยากรณ์เป็นคลาสที่กำลังพิจารณาและถูกต้องต่อคลาสที่สนใจทั้งหมด สามารถหาได้ดังสมการที่ 2.18 (ธนาภัทร, 2563)

$$Recall = \frac{TP}{TP + FN} \quad (2.18)$$

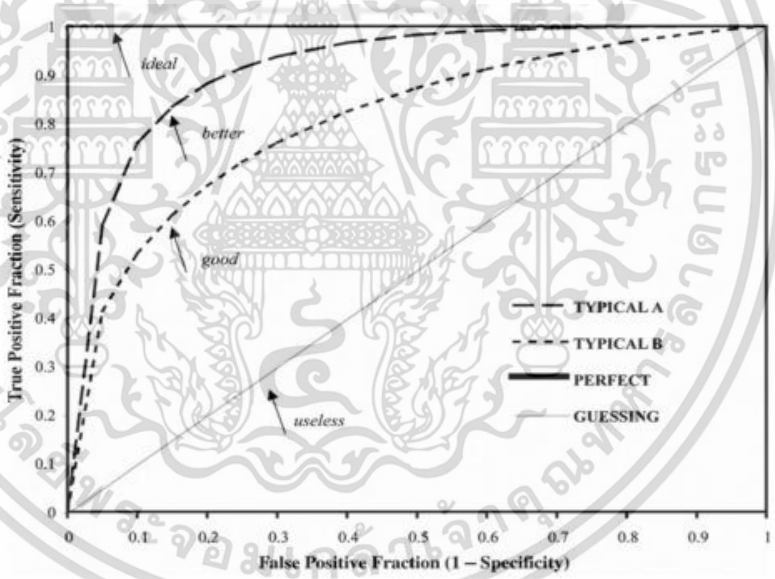
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค่าประสิทธิภาพโดยรวม (F-Measure) เป็นการวัดความเที่ยงและค่าระลึกของแบบจำลองไปพร้อม ๆ กัน สามารถหาได้ดังสมการที่ 2.19 (ธนาภัทร, 2563)

$$F-Measure = \frac{2\times precision\times recall}{precision+recall} \quad (2.19)$$

2.15 เส้นโค้ง ROC Curve (Receiver Operating Characteristic Curve)

เป็นการวัดประสิทธิภาพความถูกต้องของการพยากรณ์ที่ได้รับความนิยม โดยพื้นที่ใต้กราฟ ROC ที่มีค่ามากแสดงว่า แบบจำลองนั้นมีผลการพยากรณ์ได้อย่างถูกต้องมาก ROC Curve เป็นกราฟที่มีความสัมพันธ์ระหว่างแกน Y แทน Sensitivity (True Positive Rate) กับแกน X แทน 1-Specificity (False Positive Rate) โดยค่า Sensitivity หรือ "ความไวในการเจาะจง" ซึ่งหมายถึงสัดส่วนของผลลัพธ์ที่ถูกตรวจพบที่เป็นบวก (Positive) ต่อทั้งหมดของผลลัพธ์ที่เป็นบวก และค่า Specificity หรือ "ความสามารถในการระบุ" หรือ ซึ่งหมายถึงสัดส่วนของผลลัพธ์ที่ถูกตรวจพบที่เป็นลบ (Negative) ต่อทั้งหมดของผลลัพธ์ที่เป็นลบ



รูปที่ 2.10 แนวคิดของ Receiver Operating Characteristic Curve (ROC Curve)

จากรูปที่ 2.10 แสดงแนวคิดของ Receiver Operating Characteristic Curve (ROC Curve) จะเห็นได้ว่า ถ้าค่าของ Sensitivity และ 1-Specificity มีค่าสูงกราฟ ROC Curve จะโค้งเข้าหามุมทางด้านซ้ายบน หรือพื้นที่ใต้กราฟ ROC Curve เป็นการบ่งบอกถึงความถูกต้องหรือความน่าเชื่อถือของแบบจำลอง ถ้าแบบจำลองใดที่มีพื้นที่ใต้กราฟ ROC Curve (ROC AUC) สูงถือว่ามีประสิทธิภาพมากโดยอ้างอิงจากตารางที่ 2.2 (พงษเดช, 2564)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.3 เกณฑ์การตัดสินใจทั่วไปของค่าประมาณพื้นที่ใต้โค้ง ROC หรือ AUC

ค่าพื้นที่ใต้โค้ง ROC หรือ AUC	ความหมายของการจำแนกได้ถูกต้อง
มากกว่า 0.90	ดีมาก
0.75 – 0.90	ดี
0.50 – 0.75	ค่อนข้างต่ำ
ต่ำกว่า 0.5	ไม่สามารถนำมาใช้ได้

จากตารางที่ 2.3 แสดงเกณฑ์การตัดสินใจทั่วไปของค่าประมาณพื้นที่ใต้โค้ง ROC หรือ AUC จะพบว่าหากพื้นที่ใต้โค้ง AUC มีค่าสูงกว่า 0.90 หมายความว่าแบบจำลองที่สร้างขึ้นมีความแม่นยำสูงจำแนกข้อมูลได้อย่างถูกต้อง ในขณะเดียวกันหากมีค่าต่ำกว่า 0.5 หมายความว่าแบบจำลองที่สร้างไม่สามารถพยากรณ์ข้อมูลทดสอบได้เลยจึงทำให้ไม่สามารถนำไปใช้งานได้

2.16 การตลาดแบบบุคคลส่วนตัว (Personalized Marketing)

การตลาดแบบบุคคลส่วนตัวเป็นแนวคิดที่เกิดขึ้นเมื่อข้อมูลเกี่ยวกับผู้บริโภคถูกผสานกับเทคโนโลยีเพื่อสร้างปฏิสัมพันธ์ระหว่างบริษัทและผู้บริโภคแต่ละคน แนวคิดนี้ถูกอธิบายว่าเป็นการสื่อสารที่เน้นไปที่ผู้บริโภคแต่ละคนโดยที่ถูกปรับแต่งขึ้นจากข้อมูลส่วนบุคคล เช่น อายุ เชื้อชาติ เพศ ความต้องการ และความชอบ การตลาดแบบบุคคลส่วนบุคคลจะใช้ข้อมูลส่วนบุคคลเพื่อจับคู่ความต้องการหรือความชอบของผู้บริโภคกับผลิตภัณฑ์ที่เฉพาะเจาะจง เนื่องจากการพัฒนาเทคโนโลยีและความสามารถในการเก็บข้อมูล การตลาดได้รับการปรับแต่งให้เป็นรูปแบบที่เข้ากับผู้บริโภคได้มากกว่าแต่ก่อน นอกจากนี้การใช้อินเทอร์เน็ตมากขึ้นทำให้ธุรกิจสามารถสื่อสารโดยตรงกับผู้บริโภคได้ง่ายขึ้นทั้งบริษัทขนาดใหญ่และขนาดเล็กไม่ว่าจะอยู่ในอุตสาหกรรมใด เป็นสิ่งที่สร้างการแข่งขันในธุรกิจเนื่องจากการเข้าถึงข้อมูลส่วนบุคคลที่ชัดเจน ซึ่งช่วยให้บริษัทสามารถระบุความต้องการของผู้บริโภคได้อย่างแม่นยำ พร้อมทั้งสร้างความรู้สำหรับผู้บริโภค แนวคิดการตลาดแบบบุคคลส่วนตัวสามารถอธิบายได้ผ่านสองคำศัพท์ที่แตกต่างกัน คือ ส่วนบุคคล (Personalization) และ การปรับแต่ง (Customization) โดยส่วนบุคคล (Personalization) เป็นกระบวนการที่พิจารณาจากข้อมูลที่เกี่ยวข้องกับพฤติกรรมก่อนหน้านี้ ความชื่นชอบ ประวัติการซื้อขาย และข้อมูลส่วนบุคคลอื่น ๆ กิจกรรมทางออนไลน์ที่ทิ้งร่องรอยการใช้งานที่ช่วยให้บริษัทสร้างโปรไฟล์สำหรับการตลาดบุคคลที่ปรับแต่ง ตามพฤติกรรมประกอบด้วยข้อมูล เช่น ชื่อ สินค้าที่สนใจ ตำแหน่งปัจจุบันและข้อมูลประชากรศาสตร์ ฐานข้อมูลที่ใหญ่มหาศาลและการเก็บรวบรวมเกี่ยวกับพฤติกรรมของผู้บริโภคบนเว็บไซต์ต่าง ๆ ทำให้บริษัทสามารถสื่อสารได้อย่างแม่นยำกับผู้บริโภคอย่างเฉพาะเจาะจง โอกาสสำหรับการสื่อสารที่แม่นยำมากขึ้นเพิ่มโอกาสในการสร้างความสัมพันธ์ระยะยาวกับผู้บริโภคโดยแสดงความเข้าใจในกิจกรรมและความต้องการของลูกค้า

เอกสารนี้ ... ไม่ว่ากรณีใด ๆ ... อ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

อีกหนึ่งคำคือ การปรับแต่ง (Customization) ซึ่งหมายถึงว่าผู้ใช้สามารถกำหนดประเภทของโฆษณาที่ต้องการได้เอง บริษัทจะติดตามความชื่นชอบส่วนตัวของพวกเขาอย่างต่อเนื่องและปรับโฆษณาตามนั้น ทำให้บริษัทได้ประโยชน์จากความคิดเห็นของลูกค้า ตัวอย่างที่ดีของบริษัทที่ใช้การกำหนดเองคือ Twitter ที่ให้ผู้บริโภคตัดสินใจว่าจะถูกโฆษณาประเภทใด สามารถทำได้โดยการเผยแพร่ข้อมูลเองและแชร์โพสต์ของผู้อื่นที่พิจารณาว่าเกี่ยวข้อง แม้ว่าการตลาดแบบบุคคลส่วนตัวสามารถนำมาใช้ได้ในวิธีการอย่างน้อยสองวิธี แต่สิ่งสำคัญที่สุดคือความเกี่ยวข้องของข้อความ ในข้อความนั้นต้องสื่อสารให้แม่นยำ ต้องสร้างความรู้สึกว่าผู้บริโภคได้รับการเลือกเฉพาะสำหรับตัวเองโดยเฉพาะแทนการได้ข้อความที่ไม่เจาะจงต่อบุคคล (Fridh and Dahl, 2019)

2.17 งานวิจัยที่เกี่ยวข้อง

Ibrahim and Osman (2014) งานวิจัยนี้เป็นการศึกษาเกี่ยวกับการใช้เทคนิคการเลือกคุณลักษณะ (Feature Selection) ในงานจัดประเภทสแปมอีเมล (E-Mail Spam Classification) โดยใช้เทคนิค การวิเคราะห์ความแปรปรวน (One-Way ANOVA) ซึ่งเป็นเทคนิคทางสถิติที่ใช้ในการทดสอบความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มตัวอย่าง ในงานวิจัยนี้ผู้จัดทำได้พัฒนาเทคนิคใหม่ในการเลือกคุณลักษณะที่เหมาะสมในงานจัดประเภทสแปมอีเมล โดยใช้ การวิเคราะห์ความแปรปรวนเพื่อตรวจสอบความสามารถในการแยกแยะสถานะสแปมและสถานะไม่ใช่สแปมของอีเมล ผลการวิจัยนี้อาจมีประโยชน์ในการพัฒนาระบบกรองสแปมอีเมลที่มีประสิทธิภาพสูงขึ้น งานวิจัยนี้ได้รับการตีพิมพ์ในวารสารทางวิชาการและถือเป็นแหล่งข้อมูลที่น่าสนใจสำหรับผู้ที่สนใจในงานวิจัยที่เกี่ยวกับหลักการเลือกคุณลักษณะในงานจัดประเภทสแปมอีเมลโดยใช้เทคนิคการวิเคราะห์ความแปรปรวน

Yu and Liu (2003) งานวิจัยนี้ได้นำเสนอวิธีการเลือกคุณลักษณะสำหรับข้อมูลที่มีมิติสูง โดยใช้วิธีกรองที่ใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ซึ่งเป็นวิธีที่รวดเร็วและมีประสิทธิภาพในการคัดเลือกคุณลักษณะที่สำคัญ วิธีการทำงานคือการคำนวณค่าสัมพันธ์ระหว่างคุณลักษณะแต่ละคู่และตัวแปรเป้าหมาย โดยคัดเลือกคุณลักษณะที่มีค่าสัมพันธ์สูงกับตัวแปรเป้าหมายและค่าสัมพันธ์ต่ำกับคุณลักษณะอื่น ๆ โดยการลดมิติข้อมูลและลบคุณลักษณะที่ไม่สำคัญออกจากชุดข้อมูล ผลการทดลองแสดงให้เห็นถึงประสิทธิภาพของวิธีที่เสนอในการเลือกคุณลักษณะ วิธีนี้มีความเร็วและมีประสิทธิภาพในการประมวลผลสูงและสามารถจัดการกับข้อมูลที่มีมิติสูงได้ดี

Wang et al. (2021) งานวิจัยนี้เป็นการศึกษาเกี่ยวกับการนำเทคนิคการ การสุ่มเพิ่ม (Oversampling) และ (Undersampling) โดยใช้วิธี SMOTE และENN มาช่วยในการพยากรณ์ความเสี่ยงในผู้ป่วยโรคหัวใจเรื้อรัง (Chronic Heart Failure) โดยใช้การเรียนรู้ของเครื่องเป็นเครื่องมือในการพยากรณ์ งานวิจัยนี้ได้นำข้อมูลจาก Electronic Medical Record (EMR) ของผู้ป่วยโรคหัวใจเรื้อรังที่รักษาอยู่ในโรงพยาบาลเฉพาะทางมาใช้ในการศึกษา โดยแบ่งข้อมูลออกเป็นชุดข้อมูลสำหรับการฝึกสอนและชุดข้อมูลสำหรับการทดสอบ วิธีการพยากรณ์ที่ใช้ในงานวิจัยนี้เป็นการนำข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จำนวนมากที่มีความไม่สมดุล (Imbalanced Data) มาใช้เพื่อพยากรณ์ความเสี่ยงในการเกิดผลเสียจากโรคหัวใจเรื้อรัง เช่น การเสียชีวิต การเจ็บป่วยรุนแรงและการเจ็บป่วยที่ต้องนอนโรงพยาบาลนาน ผลการวิจัยพบว่าการนำเอาเทคนิคการสุ่มเพิ่มและการสุ่มลดมารวมกันโดยใช้วิธี SMOTE และENN นั้นสามารถช่วยปรับปรุงการจำแนกความเสี่ยงของผู้ป่วยโรคหัวใจเป็นเรื่องราวสำคัญที่มีผลกับการบริหารจัดการโรคเป็นอย่างมาก โดยค่าความแม่นของโมเดลที่ใช้วิธี SMOTEและENN นั้นสูงกว่าโมเดลที่ใช้เทคนิค การสุ่มเพิ่ม หรือ การสุ่มลดแบบเดียวเท่านั้น

Waritpon et al. (2022) งานวิจัยนี้มีการทำการศึกษาเปรียบเทียบกันระหว่างอัลกอริทึมการเรียนรู้ของเครื่อง ทั้งหมด 5 วิธี วิธีต้นไม้ตัดสินใจ (Decision Tree) วิธีต้นไม้สุ่ม (Random Tree) วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) วิธีนาอีฟเบย์ (Naïve Bayes) และวิธีป่าไม้สุ่ม (Random Forest) เพื่อจำแนกว่าลูกค้าจะสมัครสมาชิกเงินฝากระยะสั้นหรือไม่ ข้อมูลที่ใช้ในงานวิจัยเป็นชุดข้อมูล Bank Marketing ที่เกี่ยวข้องกับแคมเปญการตลาดโดยตรงของสถาบันการเงินในประเทศโปรตุเกส โดยเนื่องจากชุดข้อมูลเดิมมีความไม่สมดุลกัน ในงานวิจัยนี้ได้ใช้เทคนิคการลดจำนวนตัวอย่างในกลุ่มส่วนมาก (คลาส 'no') เพื่อทำให้ได้ชุดข้อมูลที่สมดุลเพื่อสร้างแบบจำลองและเปรียบเทียบประสิทธิภาพของแต่ละเทคนิคการเรียนรู้ของเครื่อง ผลลัพธ์แสดงให้เห็นว่า Random Forest (RF) มีความแม่นยำสูงสุด ตามด้วย Decision Tree (J48) ที่มีความแม่นยำใกล้เคียงกับ RF โดยมีความแตกต่างเพียง 0.01 เท่านั้น ในสรุปงานวิจัยนี้, การใช้เทคนิค Machine Learning ในการจำแนกลูกค้าธนาคารสามารถช่วยให้เราสามารถวิเคราะห์และพยากรณ์พฤติกรรมของลูกค้าในการตลาดได้อย่างมีประสิทธิภาพ โดยการลดจำนวนตัวแปรที่ใช้ในโมเดล J48 เป็นต้น ซึ่งสามารถช่วยให้ธนาคารพัฒนากลยุทธ์การตลาดและการจัดการลูกค้าได้อย่างมีประสิทธิภาพและเหมาะสม

ดังนั้นจากงานวิจัยนี้ได้เลือกใช้เทคนิคการคัดเลือกคุณลักษณะ (Feature Selection) ด้วยเทคนิคการวิเคราะห์ความแปรปรวน (One-Way ANOVA) และ เทคนิคสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) เลือกคุณลักษณะที่มีผลต่อตัวแปรตามของงานวิจัย เพื่อลดความไม่เกี่ยวข้องของตัวแปร และใช้เทคนิคการผสมผสานการสุ่มเพิ่ม (Oversampling) ด้วยวิธี Synthetic Minority Oversampling Technique (SMOTE) และการสุ่มลด (Undersampling) ด้วยวิธี Edited Nearest Neighbor (ENN) เพื่อปรับปรุงข้อมูลให้มีความสมดุลกันมากขึ้น โดยหลังจากจัดเตรียมข้อมูลด้วยเทคนิคดังกล่าว งานวิจัยนี้ได้เลือกใช้วิธีการเรียนรู้ของเครื่อง 3 วิธีได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และ วิธีป่าไม้สุ่ม ซึ่งแต่ละวิธีมีความสามารถและจุดเด่นในการพยากรณ์ข้อมูลที่แตกต่างกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
วิธีดำเนินงานวิจัย

การวิจัยครั้งนี้เป็นการศึกษาการประยุกต์ใช้การเรียนรู้ของเครื่อง มาใช้ในการจำแนกกลุ่มลูกค้าที่เป็นกลุ่มลูกค้าเป้าหมายของแคมเปญ เพื่อการส่งแคมเปญที่ได้ลูกค้าเข้ามาใช้ผลิตภัณฑ์อย่างมีประสิทธิภาพ โดยจะอธิบายการดำเนินงานวิจัยดังหัวข้อต่อไปนี้

3.1 เครื่องมือที่ใช้ในงานวิจัย

ผู้วิจัยทำการวิจัยโดยใช้เครื่องมือดึงข้อมูลตัวอย่างมาจากคลังข้อมูลของธนาคาร นำออกในรูปแบบแฟ้มข้อมูล CSV (Comma Separated Value) มีจำนวนข้อมูลมาก ดังนั้นการใช้ระบบปฏิบัติการควรใช้ตั้งแต่วินโดวส์ 10 ขึ้นไป ซึ่งเครื่องมือผู้วิจัยใช้มีดังนี้

3.1.1 ฮาร์ดแวร์ (Hardware)

เครื่องคอมพิวเตอร์พกพาที่ใช้ระบบปฏิบัติการวินโดว์ 10

1. Processor: Intel Core i5-1335U 3.0GHz
2. Graphic Card: Intel Iris Xe Graphics
3. Memory: 16 GB
4. HDD: 1 TB

3.1.2 ซอฟต์แวร์ (Software)

ซอฟต์แวร์ คือโปรแกรมคอมพิวเตอร์หรือชุดคำสั่งที่ถูกสร้างขึ้นเพื่อทำงานบนเครื่องคอมพิวเตอร์ ซึ่งซอฟต์แวร์ที่ผู้วิจัยใช้มีดังนี้

- Jupyter Notebook เป็นแอปพลิเคชันที่ใช้ในการสร้างและแชร์เอกสารที่ประกอบด้วยโค้ดและข้อความในรูปแบบที่เรียกว่า "โน้ตบุ๊ค" โดยสนับสนุนหลายภาษาโปรแกรม เช่น C++, Java, Python ฯลฯ
- Microsoft Excel เป็นโปรแกรมสำหรับการจัดการข้อมูลที่เป็นส่วนหนึ่งของชุดโปรแกรม Microsoft Office ที่พัฒนาโดยบริษัท Microsoft โดย Excel มีความสามารถในการจัดเก็บข้อมูลที่เกี่ยวข้องกันในรูปแบบตารางสามารถนำออกเป็นรูปแบบแฟ้มข้อมูลประเภทต่าง ๆ ได้

3.1.3 ชุดคำสั่งในงานวิจัย

โดยชุดคำสั่งในงานวิจัยครั้งนี้เป็นชุดคำสั่งของโปรแกรมภาษา Python ที่ใช้ในการเตรียมข้อมูลสร้างแบบจำลองการเรียนรู้ของเครื่องและวัดประสิทธิภาพแบบจำลอง โดยชุดคำสั่งที่ใช้ในงานวิจัยมีดังนี้

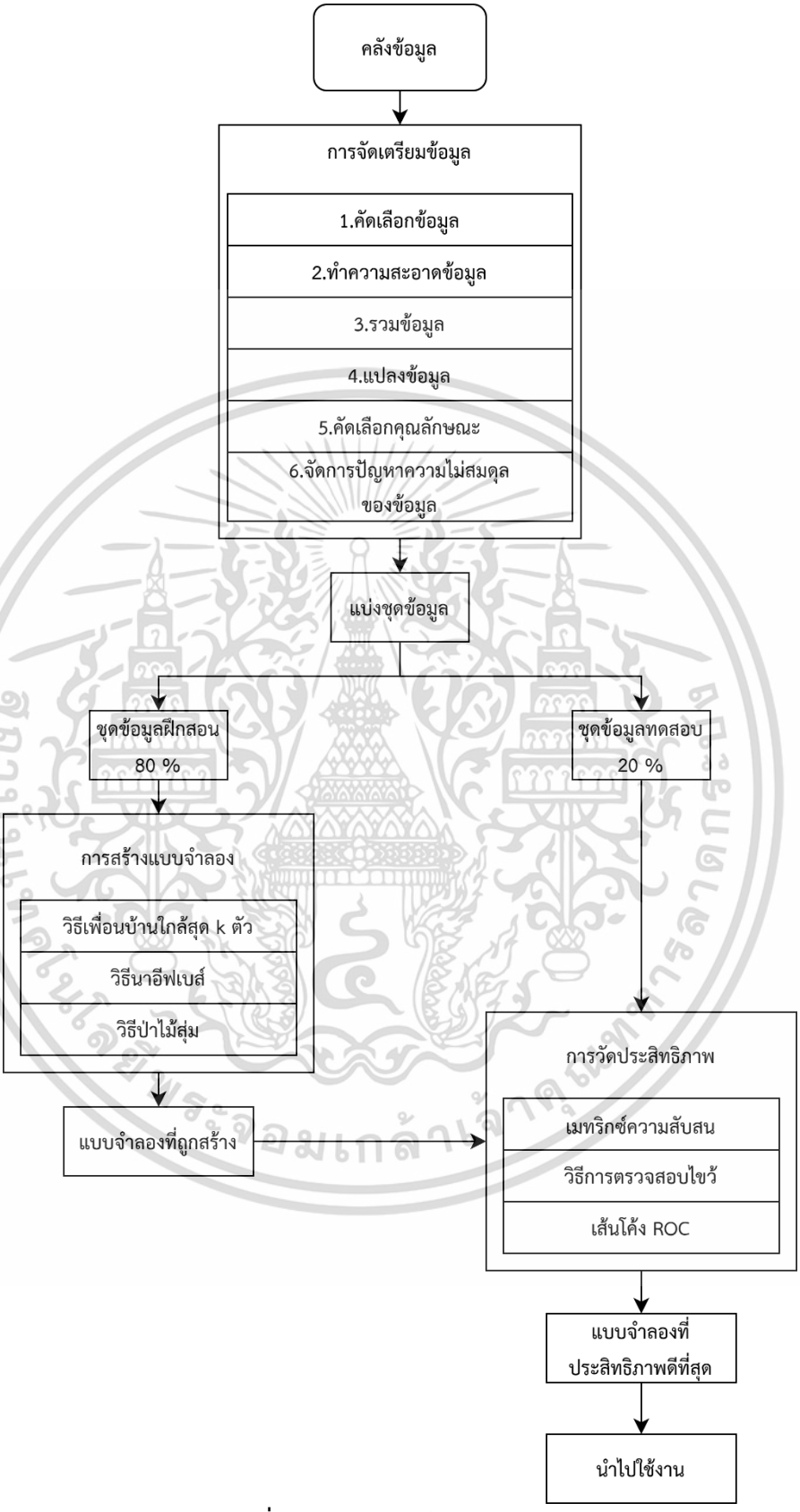
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.1 ชุดคำสั่งที่ใช้งานวิจัย

คำสั่งไลบรารี่	คำอธิบาย
	Scikit-learn เป็นชุดคำสั่งของภาษา Python ใช้สำหรับการเรียนรู้ของเครื่องและสร้างแบบจำลองทางสถิติ การจำแนกประเภท เช่น Regression, Classification และ Clustering
	เป็นชุดเครื่องมือในภาษา Python ที่ใช้สำหรับการจัดการกับข้อมูลที่มีความไม่สมดุลกัน (imbalanced data) ในงานที่เกี่ยวข้องกับการเรียนรู้เครื่องจักร (machine learning) หรือการประมวลผลทางสถิติ (statistical processing)
	แพ็กเกจสำหรับการสร้างกราฟและการวิเคราะห์ข้อมูลที่มีความสวยงามและง่ายต่อการใช้งานในภาษา Python ซึ่งช่วยให้นักวิเคราะห์ข้อมูลและนักพัฒนาทำงานกับข้อมูลและสร้างกราฟได้อย่างสะดวกมากขึ้น
	NumPy เป็นชุดคำสั่งพื้นฐานที่ใช้คำนวณทางคณิตศาสตร์ด้วยภาษา Python สามารถคำนวณหรือ ดำเนินการทางตรรกะใน Array หลายมิติ หรือ Matrix ได้อย่างรวดเร็ว
	pandas คือหนึ่งในชุดคำสั่งสำคัญของภาษา Python มีความสามารถในการจัดการ และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพตั้งแต่ข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดใหญ่สามารถใช้การเขียนโค้ด เพื่อปรับแต่ง หรือเชื่อมต่อกับโปรแกรมอื่นๆเพื่อดู Data set
	Matplotlib เป็นชุดคำสั่งของภาษา Python เพื่อใช้ในการสร้างหรือแสดงผล Data visualization ช่วยในการสร้างแผนภูมิและกราฟต่างๆเพื่อช่วยในการวิเคราะห์ที่ทำให้ดูง่ายขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 ขั้นตอนการดำเนินงาน



รูปที่ 3.1 กระบวนสร้างแบบจำลอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.1 แสดงกระบวนสร้างแบบจำลอง โดยเริ่มจากการนำข้อมูลจากคลังข้อมูลธนาคารเข้ากระบวนการจัดเตรียมข้อมูลเพื่อช่วยลดความผิดพลาดที่อาจเกิดขึ้นในกระบวนการสร้างแบบจำลองโดยจะมีขั้นตอน เช่น คัดเลือกข้อมูล ทำความสะอาดข้อมูล รวมข้อมูล แปลงข้อมูล คัดเลือกคุณลักษณะที่สำคัญ จัดการข้อมูลที่ไม่สมดุล

เพื่อนำข้อมูลที่ผ่านกระบวนการจัดเตรียมข้อมูลเสร็จสิ้นสร้างแบบจำลอง โดยจะแบ่งข้อมูลออกเป็น 2 ส่วน คือ ชุดข้อมูลฝึกสอนที่มีสัดส่วน 80 เปอร์เซ็นต์ของข้อมูลทั้งหมด และชุดข้อมูลทดสอบที่มีสัดส่วน 20 เปอร์เซ็นต์ของข้อมูลทั้งหมด นำชุดข้อมูลฝึกสอนสร้างแบบจำลองทั้ง 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และวิธีป่าไม้สุ่ม โดยผ่านการตั้งค่าพารามิเตอร์ที่เหมาะสม นำแบบจำลองที่ถูกสร้างเสร็จสิ้นมาประเมินประสิทธิภาพด้วยชุดข้อมูลทดสอบที่แบ่งไว้ 20 เปอร์เซ็นต์โดยพิจารณาจาก 3 เทคนิค ดังนี้

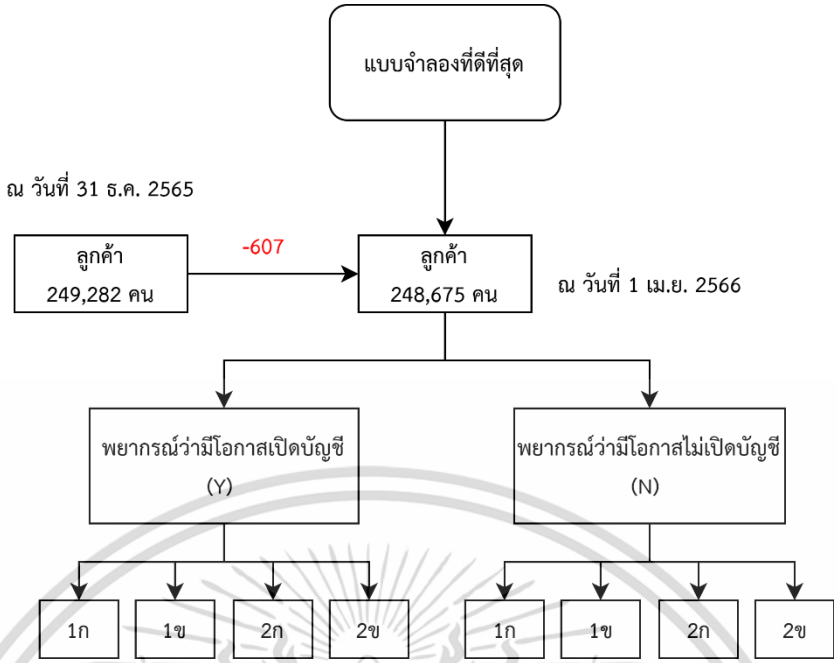
1) เมทริกซ์ความสับสน (Confusion Matrix) ที่ดูจากค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure)

2) การตรวจสอบไขว้ (K-Fold Cross-Validation) ที่จะช่วยให้เห็นถึงความเสถียรภาพของแบบจำลองว่ามีความแม่นยำและสมบูรณ์แบบในการพยากรณ์ข้อมูลที่ไม่เคยเห็นมาก่อนในชุดข้อมูล

3) เส้นโค้ง ROC Curve (Receiver Operating Characteristic Curve) ช่วยประเมินแบบจำลองที่สร้างขึ้นโดยดูจากพื้นที่ใต้เส้นโค้งเพื่อประเมินประสิทธิภาพการจำแนกของแบบจำลอง

เปรียบเทียบแบบจำลองทั้ง 3 วิธี เพื่อหาวิธีที่ให้ประสิทธิภาพในการจำแนกข้อมูลดีที่สุด จากนั้นนำแบบจำลองที่มีประสิทธิภาพดีที่สุดไปใช้งาน โดยจะเป็นการทดสอบการนำแบบจำลองใช้งาน ในการส่งแคมเปญข้อเสนอบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.2 กระบวนการสร้างแคมเปญ

จากรูปที่ 3.2 แสดงกระบวนการสร้างแคมเปญหลังจากได้แบบจำลองที่ดีที่สุดแล้ว นำมาพยากรณ์กลุ่มลูกค้าที่ตรงตาม 3 เงื่อนไขของงานวิจัยนี้ได้ตั้งไว้ได้แก่ เป็นลูกค้าที่มีบัญชีออมทรัพย์แบบธรรมดา เป็นลูกค้าที่มีบัตรเครดิตอย่างน้อย 1 ประเภท และเป็นลูกค้าที่ไม่มีบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ซึ่งกลุ่มลูกค้าเหล่านี้มาจากข้อมูลตัวอย่างของลูกค้าที่ไม่มีการเปิดบัญชี ณ วันที่ 31 ธ.ค. 2565 มีจำนวน 249,282 คน หลังจากเวลาผ่านไป 4 เดือนมีลูกค้าเปิดบัญชีไปแล้วทั้งสิ้น 607 คน ทำให้เหลือลูกค้า จำนวน 248,675 คน ณ วันที่ 1 เม.ย. 2566

โดยการพยากรณ์กลุ่มลูกค้า แบบจำลองจะจำแนกลูกค้าออกเป็น 2 กลุ่ม

1) กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มคนเหล่านี้มีโอกาสเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม้เกิน 1 แสนบาท หรือกลุ่ม ‘Y’

2) กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มคนเหล่านี้มีโอกาสที่จะไม่เปิดบัญชีเงินเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม้เกิน 1 แสนบาท หรือกลุ่ม ‘N’

แคมเปญที่สร้างจะจำแนกลูกค้าเหล่านั้นออกตามคุณลักษณะส่วนบุคคล ได้แก่ อายุ และพฤติกรรมการยอมรับความเสี่ยง เพื่อสร้างข้อความเฉพาะบุคคล (Personalized Message) ที่มีเนื้อหาแตกต่างกัน โดยขั้นต้นจะแบ่งออกเป็น 2 ระดับตามลักษณะส่วนบุคคล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ระดับที่ 1 กลุ่มข้อความที่จำแนกตาม อายุ แบ่งออกเป็น 2 กลุ่ม ได้แก่

กลุ่มที่ ‘1’ จำแนกตามอายุของลูกค้าที่อายุ 18 ถึง 40 ปี

กลุ่มที่ ‘2’ จำแนกตามอายุของลูกค้าที่อายุ 41 ถึง 90 ปี

ระดับที่ 2 กลุ่มข้อความที่จำแนกตาม การใช้บัตรเครดิต แบ่งออกเป็น 2 กลุ่ม ได้แก่

กลุ่ม ‘ก’ จำแนกตามการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

กลุ่ม ‘ข’ จำแนกตามการที่ไม่ใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

ซึ่งการสร้างข้อความเฉพาะบุคคล จะทำการรวมกลุ่มทั้ง 2 ระดับเข้าด้วยกัน เพื่อให้ได้ข้อความที่จำแนกให้แก่ลูกค้า 4 กลุ่มที่แตกต่างกัน ดังต่อไปนี้

กลุ่ม 1ก) เป็นข้อความที่ถูกส่งให้แก่กลุ่มลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

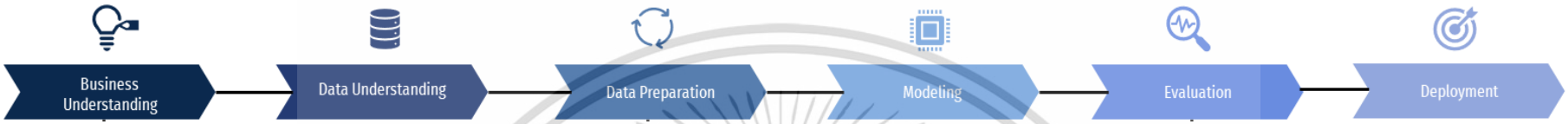
กลุ่ม 1ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

กลุ่ม 2ก) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

กลุ่ม 2ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

โดยการส่งแคมเปญให้ลูกค้าจะเกิดขึ้นเพียงครั้งเดียว วันที่ 20 เมษายน พ.ศ. 2566 และเก็บผลการเปิดบัญชีในวันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นระยะเวลา 21 วัน หรือ 3 สัปดาห์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



3.3 การทำความเข้าใจธุรกิจ

3.4 การทำความเข้าใจข้อมูล
- 3.4.1 การเก็บข้อมูล
- 3.4.2 การอธิบายข้อมูล
- 3.4.3 การตรวจสอบคุณภาพของข้อมูล

3.5 การจัดเตรียมข้อมูล
- 3.5.1 การคัดเลือกข้อมูล
- 3.5.2 การทำความสะอาดข้อมูล
- 3.5.3 การรวมข้อมูล
- 3.5.4 การแปลงข้อมูล
- 3.5.5 การเลือกคุณลักษณะที่สำคัญ
- 3.5.6 การจัดการความไม่สมดุลของข้อมูล

3.6 การสร้างแบบจำลอง
- 3.6.1 วิธีเพื่อนบ้านใกล้สุด k ตัว
- 3.6.2 วิธีนาอีฟเบส์
- 3.6.3 วิธีป่าไม้สุ่ม

3.7 การวัดประสิทธิภาพ
- 3.7.1 การวัดประสิทธิภาพวิธีเพื่อนบ้านใกล้สุด k ตัว
- 3.7.2 การวัดประสิทธิภาพวิธีนาอีฟเบส์
- 3.7.3 การวัดประสิทธิภาพวิธีป่าไม้สุ่ม
- 3.7.4 ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง

3.8 การนำแบบจำลองที่เหมาะสมไปใช้งาน

รูปที่ 3.3 กระบวนการวิเคราะห์ข้อมูลโดยใช้โมเดล CRISP-DM

โดยบทที่ 3 จะเป็นการอธิบายขั้นตอนการดำเนินงานตามกระบวนการ CRISP-DM ทั้งสิ้น 6 ขั้นตอนตามรูปที่ 3.3 เพื่อเป็นไปตามหลักการทำเหมืองข้อมูล ดังนี้

ขั้นตอนที่ 1) การทำความเข้าใจธุรกิจ (Business Understanding) เป็นขั้นตอนการทำความเข้าใจปัญหาของการส่งแคมเปญของธนาคาร และเป็นขั้นตอนสำคัญในการกำหนดวัตถุประสงค์ของการทำงานวิจัยครั้งนี้

ขั้นตอนที่ 2) การทำความเข้าใจข้อมูล (Data Understanding) เป็นขั้นตอนในการทำความเข้าใจข้อมูลที่กำลังจะถูกนำไปใช้งาน โดยมีหัวข้อย่อย เช่น

3.4.1 การเก็บข้อมูล

3.4.2 การอธิบายข้อมูล

3.4.3 การตรวจสอบคุณภาพของข้อมูล

ขั้นตอนที่ 3) การเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนในการเตรียมข้อมูลก่อนการสร้างแบบจำลอง โดยจะมีหัวข้อย่อย เช่น

3.5.1 การคัดเลือกข้อมูล

3.5.2 ทำความสะอาดข้อมูล

3.5.3 การรวมข้อมูล

3.5.4 การแปลงข้อมูล

3.5.5 การเลือกคุณลักษณะที่สำคัญ

3.5.6 การจัดการความไม่สมดุลของข้อมูล

ขั้นตอนที่ 4) การสร้างแบบจำลอง (Modeling) เป็นขั้นตอนหลังจากเตรียมข้อมูลเสร็จสิ้น นำข้อมูลชุดฝึกสอนที่ถูกแบ่งมา 80 เปอร์เซ็นต์ของข้อมูลเข้าแบบจำลองให้เกิดการเรียนรู้ โดยงานวิจัยนี้ได้สร้างแบบจำลอง 3 วิธี

3.6.1 วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors)

3.6.2 วิธีนาอีฟเบย์ (Naïve Bayes)

3.6.3 วิธีป่าไม้สุ่ม (Random Forest)

และจะมีการกำหนดพารามิเตอร์ที่เหมาะสมสำหรับแบบจำลองนั้น ๆ

ขั้นตอนที่ 5) การวัดประสิทธิภาพ (Evaluation) เป็นขั้นตอนหลังจากการสร้างแบบจำลองเสร็จสิ้น นำชุดข้อมูลทดสอบที่ถูกแบ่งจากชุดข้อมูลหลัก 20 เปอร์เซ็นต์มาวัดประสิทธิภาพแบบจำลองด้วยการพิจารณาต่าง ๆ เช่น เมทริกซ์ความสับสน การวิเคราะห์ความแม่นตรงของแบบจำลองด้วยการตรวจสอบไขว้ และการตรวจสอบกราฟเส้นโค้ง ROC เพื่อหาแบบจำลองที่ถูกประเมินว่าให้ประสิทธิภาพดีที่สุดในการพยากรณ์นำไปใช้งานในการส่งแคมเปญ โดยจะมีหัวข้อ เช่น

3.7.1 การวัดประสิทธิภาพวิธีเพื่อนบ้านใกล้สุด k ตัว

3.7.2 การวัดประสิทธิภาพวิธีนาอีฟเบย์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.7.3 การวัดประสิทธิภาพวิธีป่าไม้สุ่ม

3.7.4 ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง

ขั้นตอนที่ 6) การนำแบบจำลองไปใช้งาน (Deployment) เป็นการนำแบบจำลองที่เหมาะสมไปใช้ในการส่งแคมเปญ โดยจะมีการพยากรณ์กลุ่มลูกค้าและมีการสร้างข้อความเฉพาะบุคคลให้แก่ลูกค้าแต่ละกลุ่มที่แตกต่างกัน โดยจะอธิบายรายละเอียดในลำดับถัดไป

3.3 การทำความเข้าใจธุรกิจ (Business Understanding)

ขั้นตอนที่ 1 ทำความเข้าใจธุรกิจ ปัจจุบันทางธนาคารได้มีการสร้างแคมเปญขึ้นมาอย่างต่อเนื่อง โดย ผู้สร้างแคมเปญจะวิเคราะห์ข้อมูลลูกค้าจากคลังข้อมูล เพื่อกำหนดกลุ่มเป้าหมายและสร้างข้อความเชิญชวนให้เป็นข้อความเฉพาะบุคคล เพื่อสร้างความน่าสนใจและสร้างความสัมพันธ์ระหว่างลูกค้ากับธนาคาร โดยในแต่ละปีจะมีการส่งเสริมการกระตุ้นแคมเปญที่แตกต่างกันเพื่อสร้างผลกำไรและสร้างเสถียรภาพแก่ธนาคาร

การส่งแคมเปญเพื่อเสนอขายผลิตภัณฑ์ของธนาคารมีหลายประเภท เช่น บัญชีฝากประจำ (Term Deposit) บัญชีเงินฝาก (Saving Deposit) กองทุน (Mutual Funds) บัตรเครดิต (Credit Card) ประกันชีวิต (Bancassurance) ฯลฯ โดยในแต่ละแคมเปญจะมีการส่งข้อความเสนอขายผลิตภัณฑ์ในหลายช่องทางที่แตกต่างกัน เช่น ช่องทางข้อความสั้น (SMS) ช่องทางออนไลน์ (Online) ช่องทางแอปพลิเคชันธนาคาร (Application) ที่ส่งไปเสนอขายผลิตภัณฑ์ให้แก่ลูกค้าแต่ละกลุ่มที่ต่างกันตามเงื่อนไขและข้อจำกัดที่กำหนด

ตารางที่ 3.2 ข้อมูลการส่งแคมเปญเพื่อเสนอขายผลิตภัณฑ์ 1 ม.ค 2565 – 31 ธ.ค 2565

	ประเภทผลิตภัณฑ์			
	กองทุน	**บัตรเครดิต**	**ประกันชีวิต**	**บัญชีเงินฝาก**
จำนวนลูกค้าที่ส่งข้อเสนอ (คน)	276K	595K	617K	620K
การเข้ามาใช้ผลิตภัณฑ์ (คน)	13K	19K	8.5K	2.5K
การเข้ามาใช้ผลิตภัณฑ์ (%)	4.9	2.7	1.4	0.42

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้ง

จากตารางที่ 3.2 ข้อมูลการส่งแคมเปญเพื่อเสนอขายผลิตภัณฑ์เมื่อวันที่ 1 ม.ค 2565 – 31 ธ.ค 2565 ผ่านช่องทางข้อความสั้น (SMS) โดยยกตัวอย่างมา 4 แคมเปญคือ ข้อเสนอเกี่ยวกับการเปิด

พอร์ตกองทุน ข้อเสนอเกี่ยวกับการเปิดบัตรเครดิต ข้อเสนอที่เกี่ยวกับการซื้อประกันชีวิต และข้อเสนอที่เกี่ยวกับการเปิดบัญชีเงินฝาก โดยจำนวนลูกค้าที่ได้รับข้อความจากแคมเปญต่าง ๆ เป็นจำนวนของลูกค้าทั้งหมดที่แตกต่างกัน (Distinct) ตั้งแต่ 1 ม.ค 2565 – 31 ธ.ค 2565 จากข้างต้นพบว่า ภายในระยะเวลา 1 ปี มีจำนวนลูกค้าที่ได้รับการส่งแคมเปญขายผลิตภัณฑ์ของธนาคารที่เกี่ยวกับบัญชีเงินฝากมากที่สุด ประมาณ 620,000 คน แต่มีลูกค้าเข้ามาใช้ผลิตภัณฑ์เพียง ประมาณ 2,500 คน ซึ่งหากคิดเป็นเปอร์เซ็นต์การเข้ามาใช้น้อยกว่า 1 เปอร์เซ็นต์ เป็นจำนวนน้อยที่สุดเมื่อเทียบกับ 3 ผลิตภัณฑ์ที่เหลือ จึงเป็นสาเหตุของงานวิจัยนี้ที่พยายามจะเพิ่มยอดการเข้ามาใช้ผลิตภัณฑ์ โดยการประยุกต์ใช้การเรียนรู้ของเครื่องสำหรับการส่งแคมเปญที่เกี่ยวกับบัญชีเงินฝากเพื่อการส่งแคมเปญที่ตรงตามกลุ่มเป้าหมายมากขึ้น

ตารางที่ 3.3 ข้อมูลผลิตภัณฑ์บัญชีเงินฝากของธนาคาร

	บัญชีเงินฝาก ประเภท A	**บัญชีเงินฝาก ประเภท B**	**บัญชีเงินฝาก ประเภท C**	**บัญชีเงินฝาก ประเภท D**
ดอกเบี้ย	เป็นไปตามประกาศของธนาคาร	เป็นไปตามประกาศของธนาคาร	0.8 %	0.2 %
เงื่อนไขการได้ดอกเบี้ย	ฝากเงินไว้ในบัญชี	ฝากเงินไว้ในบัญชี	ทำรายการกับบัญชีออมทรัพย์ 5 ครั้งต่อเดือน	ฝากไม่เกิน 1 แสนบาท ยอดฝากมากกว่าถอนแต่ละเดือน
การจ่ายดอกเบี้ย	คำนวณดอกเบี้ยทุกวันและจ่ายให้ปีละ 2 ครั้ง มิถุนายนและธันวาคม	คำนวณดอกเบี้ยทุกวันและจ่ายให้ปีละ 2 ครั้ง มิถุนายนและธันวาคม	คำนวณดอกเบี้ยทุกวันและจ่ายทุกเดือน	คำนวณดอกเบี้ยทุกวันและจ่ายทุกเดือน
ประกัน	ไม่มี	เปิดบัญชีขั้นต่ำ 5,000 บาท เพื่อรับความคุ้มครองอุบัติเหตุ	ไม่มี	ไม่มี
ประเภทบัญชี	บัญชีปกติ	บัญชีปกติ	บัญชีปกติ	บัญชีดิจิทัล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 3.3 แสดงข้อมูลผลิตภัณฑ์บัญชีเงินฝากของธนาคาร ที่มีเงื่อนไขในการได้รับดอกเบี้ยและสิทธิประโยชน์ที่แตกต่างกัน ซึ่งในปี พ.ศ 2566 ทางธนาคารได้มีการวางแผนการส่งเสริมแคมเปญที่เกี่ยวกับบัญชีเงินฝากประเภท D มากกว่าปีที่ผ่านมาเพราะทางธนาคารเล็งเห็นว่า บัญชีเงินฝากประเภท D เป็นบัญชีดิจิทัล (Digital) ที่ลูกค้าสามารถเปิดบัญชีผ่านแอปพลิเคชันธนาคารได้โดยตรง ไม่จำเป็นต้องมาเปิดบัญชีที่สาขาของธนาคาร ทำให้สร้างความสะดวกสบายให้แก่ลูกค้าสามารถฝากเงินหรือถอนเงินตอนไหนก็ได้ และบัญชีประเภทนี้ยังส่งเสริมเรื่องการออมเงินเพราะ การที่จะได้ดอกเบี้ยรวมโบนัสสูงถึง 2 เปอร์เซ็นต์ ต้องฝากเงินไม่เกิน 1 แสนบาท มียอดฝากเงินมากกว่ายอดถอนเงินในแต่ละเดือน ทำให้ลูกค้ามีการปรับเปลี่ยนพฤติกรรมการออมเงินให้มีวินัยในการออมทุกเดือน เพื่อรับดอกเบี้ยจากแต่ก่อนอาจจะออมเงินไม่สม่ำเสมอ โดยแคมเปญบัญชีเงินฝากประเภท D มีการส่งแคมเปญเพียง 1 แคมเปญเมื่อปี พ.ศ. 2565 ที่ถูกส่งผ่านช่องทางข้อความสั้น (SMS)

ตารางที่ 3.4 ข้อมูลการส่งแคมเปญบัญชีเงินฝากประเภท D ในอดีต

ชื่อแคมเปญ	**รหัสแคมเปญ**	**จำนวนลูกค้าที่ส่งข้อเสนอ (คน)**	**การเข้ามาใช้ผลิตภัณฑ์ (คน)**	**การเข้ามาใช้ผลิตภัณฑ์ (%)**
X sell digital saving deposit type D	D9PBF*****	2,746	0	0

จากตารางที่ 3.4 แสดงข้อมูลการส่งแคมเปญบัญชีเงินฝากประเภท D โดยแคมเปญนี้จะถูกเสนอให้กับลูกค้า 2,746 คน ผ่านทางช่องทางข้อความสั้นในระยะเวลาตั้งแต่ 1 ธ.ค 2565 ถึง 1 ม.ค 2566 พบว่าภายในระยะเวลาแคมเปญ 31 วัน การเข้ามาใช้ผลิตภัณฑ์ของลูกค้าภายใต้ข้อเสนอรหัสแคมเปญ D9PBF***** มีจำนวน 0 คน คิดเป็น 0 เปอร์เซ็นต์ของการเข้ามาใช้

ปัญหาทั้งหมดที่เกิดขึ้นจะพบว่า การสร้างแคมเปญที่เกี่ยวกับบัญชีเงินฝากทั้งหมด ตั้งแต่วันที่ 1 ม.ค 2565 – วันที่ 31 ธ.ค. 2565 ไม่มีประสิทธิภาพเท่าที่ควรเมื่อเทียบกับผลิตภัณฑ์อื่น ๆ แสดงให้เห็นถึงเปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์ของแคมเปญน้อยกว่า 1 เปอร์เซ็นต์ ซึ่งปัจจุบันปี พ.ศ 2566 ทางธนาคารต้องการส่งเสริม การทำแคมเปญที่เกี่ยวกับบัญชีเงินฝากประเภท D ให้มากกว่าปีก่อนที่ผ่านมา และหากย้อนดูข้อมูลแคมเปญที่เกี่ยวกับบัญชีเงินฝากประเภท D เมื่อปีที่ผ่านมาปรากฏว่า ไม่มีลูกค้าเข้ามาใช้ผลิตภัณฑ์เลยอาจจะเกิดจากสาเหตุ เช่น การเลือกกลุ่มลูกค้าที่ยังไม่ใช้ลูกค้าเป้าหมายของผลิตภัณฑ์ ดังนั้นงานวิจัยนี้จึงเสนอการนำการเรียนรู้ของเครื่องเข้ามาจำแนกกลุ่มลูกค้าเป้าหมาย ที่ต้องการจะส่งแคมเปญข้อเสนอบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท หรือในหัวข้อต่อไปงานวิจัยนี้จะใช้ชื่อว่า “บัญชีเงินฝากประเภท D” ซึ่งหากสร้าง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

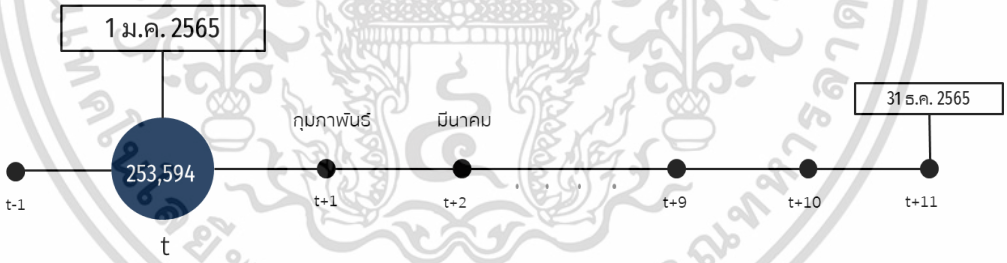
แบบจำลองที่สามารถจำแนกกลุ่มเป้าหมายของผลิตภัณฑ์ให้กับแคมเปญนี้ได้ อาจจะทำให้การส่งแคมเปญมีประสิทธิภาพในการที่ลูกค้าสนใจเข้ามาใช้ผลิตภัณฑ์มากขึ้น

3.4 การทำความเข้าใจข้อมูล (Data Understanding)

ขั้นตอนที่ 2 ทำความเข้าใจข้อมูล เป็นกระบวนการทำความเข้าใจบริบทของข้อมูลที่กำลังจะนำเข้าสู่การฝึกสอนแบบจำลองโดยหัวข้อนี้จะกล่าวถึง การเก็บข้อมูลขั้นต้น อธิบายข้อมูล และการตรวจสอบคุณภาพข้อมูล ดังต่อไปนี้

3.4.1 การเก็บข้อมูล (Collect Initial Data)

การเก็บข้อมูลเป็นกระบวนการที่เริ่มต้นในการรวบรวมข้อมูลที่จำเป็นสำหรับการสร้างแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ขั้นตอนนี้เป็นการรวบรวมข้อมูลที่ใช้เป็นตัวคุณลักษณะ (Features) และตัวเป้าหมาย (Target) ที่เราต้องการพยากรณ์หรือวิเคราะห์โดยงานวิจัยนี้ได้นำข้อมูลมาจากฐานข้อมูลของธนาคาร โดยนำออกมาในรูปแบบแฟ้มข้อมูล CSV (Comma Separated Value) เป็นแฟ้มข้อมูลข้อความประเภทหนึ่งที่ใช้สำหรับเก็บข้อมูลในรูปแบบตารางซึ่งที่มีจำนวนข้อมูลทั้งสิ้น 12 แฟ้มข้อมูล โดยแฟ้มข้อมูลจะเป็นข้อมูลของในแต่ละเดือนที่เก็บรวบรวมข้อมูลตัวอย่างลูกค้าที่มีการเปิดบัญชีในช่วงระยะเวลา 1 ปี และรวบรวมข้อมูลตัวอย่างลูกค้าที่ไม่มีการเปิดบัญชีในช่วงระยะเวลา 1 ปี



รูปที่ 3.4 แผนการเก็บข้อมูลช่วงระยะเวลา t ถึง t+11

จากรูปที่ 3.4 แสดงแผนการเก็บข้อมูลช่วงระยะเวลา t ถึง t+11 หรือในช่วงระยะเวลาวันที่ 1 ม.ค 2565 ถึง 31 ธ.ค 2565 ทางผู้วิจัยได้กำหนดเงื่อนไขของกลุ่มข้อมูลที่จะศึกษา มีอยู่ด้วยกัน 3 เงื่อนไข

1) ต้องเป็นข้อมูลลูกค้าที่มีผลิตภัณฑ์บัญชีออมทรัพย์แบบธรรมดา เพราะการที่จะเปิดบัญชีเงินฝากประเภท D ได้ต้องมีบัญชีออมทรัพย์แบบธรรมดา

2) ต้องเป็นข้อมูลลูกค้าที่มีผลิตภัณฑ์บัตรเครดิตประเภทใดก็ได้ เพราะงานวิจัยนี้ต้องการที่จะนำตัวแปรพฤติกรรมในการใช้จ่ายบัตรเครดิตมาฝึกสอนแบบจำลอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งไม่อนุญาตให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3) ต้องเป็นข้อมูลลูกค้าที่ยังไม่มีผลิตภัณฑ์บัญชีเงินฝากประเภท D

โดยจากเงื่อนไข พบว่า วันที่ 1 ม.ค. 2565 มีจำนวนลูกค้าที่ยังไม่มีบัญชีเงินฝากประเภท D และตรงตามเงื่อนไขข้อ 1 2 และ3 มีจำนวนข้อมูลอยู่ 253,594 คน จากนั้นสำรวจเป็นรายเดือนว่าลูกค้าที่อยู่ภายใต้กลุ่ม 253,594 คนนั้นมีคนใดบ้างที่มีการเปิดบัญชีเงินฝากประเภท D หากลูกค้ามีการเปิดบัญชีเดือนใดผู้วิจัยจะนำข้อมูลเดือน t-1 หรือย้อนหลัง 1 เดือนมาเป็นข้อมูลที่จะฝึกสอนแบบจำลอง เช่น หากลูกค้าเปิดบัญชีในเดือน t+2 หรือเดือน มีนาคม 2565 ทางผู้วิจัยจะใช้ข้อมูลลูกค้าเดือนที่ t+1 หรือเดือน กุมภาพันธ์ 2565 เพราะการดึงข้อมูลย้อนหลังจะได้ข้อมูลลูกค้าหรือพฤติกรรมก่อนที่ลูกค้าจะเปิดบัญชี ว่ามีพฤติกรรมการใช้เงินเป็นอย่างไร มีลักษณะส่วนบุคคลเป็นอย่างไร หากลูกค้าไม่มีการเปิดบัญชีภายในระยะเวลา 1 ปี ทางผู้วิจัยจะใช้ข้อมูลลูกค้าเดือนสุดท้ายหรือเดือนธันวาคม ดังนั้นข้อมูลลูกค้าที่เปิดบัญชีจะมีการนำข้อมูลมาจากเดือนที่แตกต่างกัน ส่วนข้อมูลลูกค้าที่ไม่เปิดบัญชีจะเป็นข้อมูลเดือนสุดท้ายหรือเดือนธันวาคมทั้งหมดเหมือนกัน

ตารางที่ 3.5 จำนวนข้อมูลของตัวแปรเป้าหมายและคำอธิบาย

ข้อมูล	จำนวนข้อมูล	คำอธิบาย
DEPOSIT	4,312	ข้อมูลลูกค้าที่มีการเปิดบัญชีภายในระยะเวลา วันที่ 1 ม.ค พ.ศ 2565 ถึง 31 ธ.ค พ.ศ. 2565
NON_DEPOSIT	249,282	ข้อมูลลูกค้าที่ไม่มีการเปิดบัญชีภายในระยะเวลา วันที่ 1 ม.ค พ.ศ 2565 ถึง 31 ธ.ค พ.ศ. 2565

จากตารางที่ 3.5 แสดงจำนวนข้อมูลของตัวแปรเป้าหมายและคำอธิบาย พบว่าจำนวนข้อมูลทั้งหมดที่งานวิจัยนี้ได้รวบรวมเพื่อฝึกสอนแบบจำลองมีจำนวนทั้งสิ้น 253,594 ข้อมูล แบ่งออกเป็นข้อมูลของกลุ่มลูกค้าที่มีการเปิดบัญชีภายในระยะเวลา 1 ม.ค 2565 ถึง 31 ธ.ค 2565 มีจำนวน 4,312 ข้อมูล ให้ชื่อว่าตัวแปรเป้าหมายว่า ‘DEPOSIT’ และข้อมูลของกลุ่มลูกค้าที่ไม่มีการเปิดบัญชีภายในระยะเวลา 1 ม.ค 2565 ถึง 31 ธ.ค 2565 มีจำนวน 249,282 ข้อมูล ให้ชื่อตัวแปรเป้าหมายว่า ‘NON_DEPOSIT’

โดยงานวิจัยนี้ได้ดึงคุณลักษณะ (Feature) ที่เกี่ยวข้องกับตัวแปรเป้าหมาย มาทั้งสิ้น 38 คุณลักษณะที่เกี่ยวข้องแบ่งออกเป็น 4 ประเภท

1. ตัวแปรข้อมูลส่วนบุคคลทั่วไปในธนาคาร
2. ตัวแปรข้อมูลถือครองผลิตภัณฑ์ในธนาคาร
3. ตัวแปรข้อมูลการใช้จ่ายบัตรเครดิตย้อนหลัง
4. ตัวแปรเป้าหมาย

และจะแสดงดังตารางที่ 3.6 3.7 3.8 3.9 ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.6 ตัวแปรคุณลักษณะส่วนบุคคลทั่วไปของลูกค้าและคำอธิบาย

ชื่อตัวแปร	คำอธิบาย	ประเภทของตัวแปร
AGE	อายุ (ปี)	ตัวแปรเชิงปริมาณ
CUST_SEGMENT_RANGE	กลุ่มลูกค้า	ตัวแปรเชิงคุณภาพ
OCCUPATION_GRP	อาชีพ	ตัวแปรเชิงคุณภาพ
GENDER_GRP	เพศ	ตัวแปรเชิงคุณภาพ
MARITAL_STATUS_GRP	สถานะสมรส	ตัวแปรเชิงคุณภาพ
EDUCATION	ระดับการศึกษา	ตัวแปรเชิงคุณภาพ
REGION	ภูมิภาค	ตัวแปรเชิงคุณภาพ
INCOME	รายได้ (บาท)	ตัวแปรเชิงปริมาณ
MOBILE	มีแอปพลิเคชันธนาคาร	ตัวแปรเชิงคุณภาพ
PAYROLL	ลูกค้าที่รับเงินเดือนผ่านธนาคาร	ตัวแปรเชิงคุณภาพ
DURATION	ระยะเวลาที่เป็นลูกค้าของธนาคาร (เดือน)	ตัวแปรเชิงปริมาณ
MAIN_BANK	เป็นลูกค้าประจำของธนาคาร	ตัวแปรเชิงคุณภาพ
CAMPAIGN_CONTACT	การถูกเสนอขายผลิตภัณฑ์	ตัวแปรเชิงคุณภาพ
AUM	สินทรัพย์รวมในธนาคาร (บาท)	ตัวแปรเชิงปริมาณ
WM	กลุ่มลูกค้าที่ขึ้นอยู่กับทรัพย์สินในธนาคาร	ตัวแปรเชิงคุณภาพ

จากตารางที่ 3.6 แสดงตัวแปรคุณลักษณะส่วนบุคคลทั่วไปของลูกค้าและคำอธิบายเบื้องต้น โดยตัวแปรคุณลักษณะทั่วไปของลูกค้าในธนาคารเป็นตัวแปรที่สำคัญในการสร้างแบบจำลองพยากรณ์ลูกค้าที่มีแนวโน้มในการเปิดบัญชี เช่น อายุ กลุ่มลูกค้า อาชีพ เพศ สถานะสมรส และระดับการศึกษา เป็นปัจจัยที่สามารถพยากรณ์การเปิดบัญชีฝากเงินได้เนื่องจากมีความสัมพันธ์กับพฤติกรรมของลูกค้าในการเลือกใช้บริการธนาคาร หรือ รายได้ กลุ่มลูกค้าที่ขึ้นอยู่กับทรัพย์สินในธนาคาร และสินทรัพย์รวมในธนาคาร อาจจะเป็นตัวชี้วัดที่แสดงถึงความสามารถในการชำระเงินของลูกค้าซึ่งอาจมีผลต่อการเปิดบัญชีฝากเงินกับธนาคาร หรือแม้แต่ตัวแปรลูกค้าที่รับเงินเดือนผ่าน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ธนาคาร ระยะเวลาที่อยู่ในธนาคาร เป็นลูกค้าประจำของธนาคาร ก็เป็นตัวชี้วัดอย่างหนึ่งที่สามารถแสดงถึงความสัมพันธ์ และความน่าเชื่อถือของลูกค้ากับธนาคาร ลูกค้าที่มีความสัมพันธ์ใกล้ชิดกับธนาคารอาจจะมีโอกาสเปิดบัญชีฝากเงินสูงขึ้น เป็นต้น

ตารางที่ 3.7 ตัวแปรข้อมูลถือครองผลิตภัณฑ์ในธนาคารและคำอธิบาย

ชื่อตัวแปร	คำอธิบาย	ประเภทของตัวแปร
SUM_OS_SAVING_DEP	ยอดเงินรวมในบัญชีออมทรัพย์	ตัวแปรเชิงปริมาณ
MF (Mutual Funds)	ผลิตภัณฑ์เกี่ยวกับกองทุน	ตัวแปรเชิงคุณภาพ
BA (Bancassurance)	ผลิตภัณฑ์เกี่ยวกับประกัน	ตัวแปรเชิงคุณภาพ
SL (Secured Loan)	ผลิตภัณฑ์เกี่ยวกับสินเชื่อค้ำประกัน	ตัวแปรเชิงคุณภาพ
UL (Unsecured Loan)	ผลิตภัณฑ์เกี่ยวกับสินเชื่อไม่ค้ำประกัน	ตัวแปรเชิงคุณภาพ
HP (Hire Purchase)	ผลิตภัณฑ์เกี่ยวกับสินเชื่อเช่าซื้อ	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_A	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 1	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_B	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 2	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_C	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 3	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_D	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 4	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_E	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 5	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_F	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 6	ตัวแปรเชิงคุณภาพ
NO_ACCT_ TYPE_G	ผลิตภัณฑ์บัตรเครดิตประเภทที่ 7	ตัวแปรเชิงคุณภาพ
TOTAL_MF_OS_01	มูลค่ารวมของกองทุนที่ครอบครองย้อนหลัง 1 เดือน (บาท)	ตัวแปรเชิงปริมาณ
TOTAL_MF_OS_02	มูลค่ารวมของกองทุนที่ครอบครองย้อนหลัง 2 เดือน (บาท)	ตัวแปรเชิงปริมาณ
TOTAL_MF_OS_03	มูลค่ารวมของกองทุนที่ครอบครองย้อนหลัง 3 เดือน (บาท)	ตัวแปรเชิงปริมาณ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตาราที่ 3.7 (ต่อ) ตัวแปรข้อมูลถือครองผลิตภัณฑ์ในธนาคารและคำอธิบาย

ชื่อตัวแปร	คำอธิบาย	ประเภทของตัวแปร
SUM_OS_BA01	มูลค่ารวมของประกันที่ครอบครองย้อนหลัง 1 เดือน (บาท)	ตัวแปรเชิงปริมาณ
SUM_OS_BA02	มูลค่ารวมของประกันที่ครอบครองย้อนหลัง 2 เดือน (บาท)	ตัวแปรเชิงปริมาณ
SUM_OS_BA03	มูลค่ารวมของประกันที่ครอบครองย้อนหลัง 3 เดือน (บาท)	ตัวแปรเชิงปริมาณ

จากตารางที่ 3.7 แสดงตัวแปรข้อมูลถือครองผลิตภัณฑ์ในธนาคารและคำอธิบายการเลือกคุณลักษณะเบื้องต้น ข้อมูลถือครองผลิตภัณฑ์ของธนาคารเป็นสิ่งสำคัญ เพื่อพยากรณ์ลูกค้าที่มีแนวโน้มในการเปิดบัญชีฝากเงินกับธนาคาร เนื่องจากผลิตภัณฑ์ธนาคารที่ลูกค้าถือครองและใช้บ่อย ๆ อาจจะสามารถสะท้อนพฤติกรรมและแนวโน้มของลูกค้าได้ อาจส่งผลต่อความน่าจะเป็นในการเปิดบัญชีใหม่ ยกตัวอย่างเช่น ยอดเงินรวมในบัญชีออมทรัพย์ สามารถแสดงถึงความสามารถในการออมเงินและเก็บเงินของลูกค้า ลูกค้าที่มียอดเงินสะสมมากอาจมีความเป็นไปได้ที่จะเปิดบัญชีฝากเงินเพิ่มขึ้น หรือคุณลักษณะที่เกี่ยวกับการมีผลิตภัณฑ์ กองทุน ประกัน สินเชื่อค้ำประกัน สินเชื่อไม่ค้ำประกัน สินเชื่อเช่าซื้อ ผลิตภัณฑ์บัตรเครดิตประเภทต่าง ๆ อาจจะมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน และคุณลักษณะที่เกี่ยวกับมูลค่ารวมของกองทุนและประกัน แสดงถึงมูลค่าของกองทุนหรือผลประโยชน์ที่ลูกค้าได้รับจากการถือครองประกันในช่วงเวลาที่แตกต่างกัน ลูกค้าที่มีมูลค่าในการถือครองผลิตภัณฑ์มากขึ้นอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงินมากขึ้น

ตารางที่ 3.8 ตัวแปรข้อมูลการใช้จ่ายบัตรเครดิตย้อนหลังและคำอธิบาย

ชื่อตัวแปร	คำอธิบาย	ประเภทของตัวแปร
CC_Spending_1	การใช้จ่ายบัตรเครดิตย้อนหลัง 1 เดือน (บาท)	ตัวแปรเชิงปริมาณ
CC_Spending_2	การใช้จ่ายบัตรเครดิตย้อนหลัง 2 เดือน (บาท)	ตัวแปรเชิงปริมาณ
CC_Spending_3	การใช้จ่ายบัตรเครดิตย้อนหลัง 3 เดือน (บาท)	ตัวแปรเชิงปริมาณ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 3.8 แสดงตัวแปรข้อมูลการใช้จ่ายบัตรเครดิตย้อนหลัง โดยการเลือกคุณลักษณะที่เป็นตัวแปรข้อมูลการใช้จ่ายบัตรเครดิตย้อนหลัง เป็นข้อมูลที่แสดงถึงรูปแบบพฤติกรรมการใช้เงินของลูกค้า อาจช่วยให้เข้าใจถึงลักษณะการใช้เงินของลูกค้าว่าเป็นคนที่ใช้เงินอย่างระมัดระวังหรือใช้เงินอย่างสม่ำเสมอ หรืออาจแสดงถึงระดับการเงินและความสามารถในการออกจ่ายของลูกค้า ลูกค้าที่มีรายได้สูงหรือใช้จ่ายมากอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน เป็นต้น

ตารางที่ 3.9 ตัวแปรเป้าหมายและคำอธิบาย

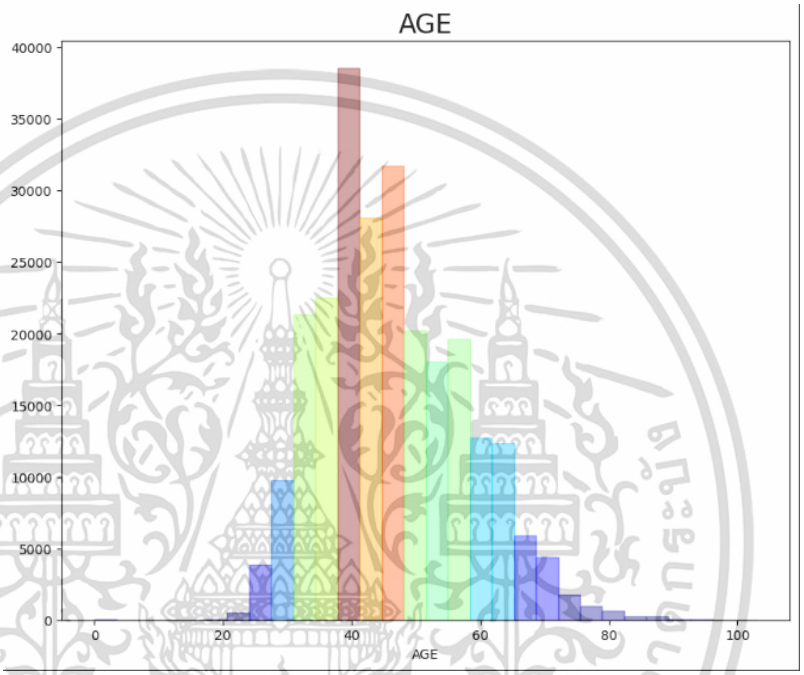
ชื่อตัวแปร	คำอธิบาย	ประเภทของตัวแปร
Class	ตัวแปรตามหรือตัวแปรเป้าหมายของข้อมูล	ตัวแปรเชิงคุณภาพ

ตารางที่ 3.9 แสดงตัวแปรเป้าหมายของข้อมูลและคำอธิบาย โดยตัวแปรเป้าหมายของงานวิจัยนี้คือตัวแปรที่บอกถึงลักษณะของข้อมูลตัวอย่างของลูกค้าที่มีการเปิดบัญชีเงินฝากประเภท D ตัวแปรเป้าหมายมีค่าเป็นไปได้อยู่ 2 ค่าคือ ข้อมูลลูกค้าที่ภายในระยะเวลา 1 ปีตั้งแต่ 1 ม.ค 2565 ถึง 31 ธ.ค 2565 มีการเปิดบัญชีเงินฝากประเภท D จะเป็นตัวแปรตามที่ชื่อว่า ‘DEPOSIT’ และ ข้อมูลลูกค้าที่ภายในระยะเวลา 1 ปี ตั้งแต่ 1 ม.ค 2565 ถึง 31 ธ.ค 2565 ไม่มีการเปิดบัญชีเงินฝากประเภท D ให้เป็นตัวแปรที่ชื่อว่า ‘NON_DEPOSIT’

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.4.2 อธิบายข้อมูล (Describe Data)

การอธิบายข้อมูลเป็นกระบวนการในการสรุปและสกัดข้อมูลที่มีอยู่ในชุดข้อมูล เป็นขั้นตอนสำคัญในการวิเคราะห์ข้อมูล ซึ่งจะช่วยให้เข้าใจเกี่ยวกับข้อมูลที่จะนำเข้าสู่การฝึกสอนแบบจำลอง ในงานวิจัยนี้ได้เก็บข้อมูลมาจากคลังข้อมูลของธนาคารโดยมีจำนวนทั้งสิ้น 253,594 ข้อมูล โดยเลือกตัวแปรอายุ (Age) มาอธิบายข้อมูล โดยใช้การสร้างกราฟและสถิติพรรณนา (Descriptive Statistics) เพื่ออธิบายข้อมูล



รูปที่ 3.5 กราฟแท่งกลุ่มช่วงอายุลูกค้า

จากรูปที่ 3.5 กราฟแท่งแสดงกลุ่มช่วงอายุลูกค้าของข้อมูลที่นำเข้าแบบจำลอง โดยค่าทางสถิติเป็นดังนี้ ค่าเฉลี่ย (Mean) เท่ากับ 46.287 ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) เท่ากับ 11.136 ค่าน้อยที่สุด (Min) เท่ากับ 0 ค่ามากที่สุด (Max) เท่ากับ 103 จากข้อมูลทางสถิติข้างต้นพบว่าข้อมูลส่วนใหญ่ที่นำเข้าสู่แบบจำลองเพื่อรับการฝึกสอนส่วนใหญ่เป็นลูกค้าอายุ 46 ปี และจะพบว่ามีค่าที่น้อยที่สุดคือ 0 และมีค่ามากที่สุดคือ 103 เมื่อพิจารณาตามหลักความเป็นจริงบุคคลที่อายุเท่ากับ 0 ปีหรือเท่ากับ 103 ปีแทบจะเป็นไปไม่ได้ ดังนั้นจึงสรุปขั้นต้นว่าข้อมูลดังกล่าวอาจจะมีค่านอกเกณฑ์ที่ไม่สมเหตุสมผลอยู่บางส่วน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.10 ประเภทของตัวแปรและค่าที่เป็นไปได้

ชื่อตัวแปร	ค่าที่เป็นไปได้
AGE	ช่วง 0 ถึง 103
CUST_SEGMENT_RANGE	‘Midincome’,’Affluent’,’Non_Retail’,’Mass’
OCCUPATION_GRP	‘Private Employee’,’Owner Operator’,’State Enterprise’,’Military’,’Housewife’,’Student’,’Professional’,’Police’
GENDER_GRP	‘Male’,’Female’,’NA’
MARITAL_STATUS_GRP	‘Single’,’Married’,’Divorce’, ’Widow’,’Undeffined’
EDUCATION	‘Bachelor’s degree’,’Secondary School’,’Master’s degree’,’High Vocational Certificate’,’Vocational Certificate’,’Primary School’,’Doctor of Philosophy’,’Technical Certificate’
REGION	‘Bangkok’,’Vincity’,’East’,’South’,’North’,’North_East’,’West’,’Central’,’Other’
INCOME	ช่วง 0 ถึง 3.33×10^{10}
MOBILE	‘Yes’,’No’
PAYROLL	‘Yes’,’No’
DURATION	อยู่ในช่วง 2 ถึง 295
MAIN_BANK	‘Yes’,’No’
CAMPAIGN_CONTRACT	‘Yes’,’No’
AUM	ช่วง 0 ถึง 1×10^{7}
WM	‘Non’,’SB’,’TB’,’WB’,’PB’
SUM_OS_SAVING_DEP	ช่วง 0 ถึง 3.21×10^{8}
MF (Mutual Funds)	‘Yes’,’No’
BA (Bancassurance)	‘Yes’,’No’

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.10 (ต่อ) ประเภทของตัวแปรและค่าที่เป็นไปได้

ชื่อตัวแปร	ค่าที่เป็นไปได้
SL (Secured Loan)	‘Yes’,’No’
UL (Unsecured Loan)	‘Yes’,’No’
HP (Hire Purchase)	‘Yes’,’No’
NO_ACCT_ TYPE_A	‘Yes’,’No’
NO_ACCT_ TYPE_B	‘Yes’,’No’
NO_ACCT_ TYPE_C	‘Yes’,’No’
NO_ACCT_ TYPE_D	‘Yes’,’No’
NO_ACCT_ TYPE_E	‘Yes’,’No’
NO_ACCT_ TYPE_F	‘Yes’,’No’
NO_ACCT_ TYPE_G	‘Yes’,’No’
TOTAL_MF_OS_01	ช่วง 0 ถึง 9.885×10^8
TOTAL_MF_OS_02	ช่วง 0 ถึง 1.019×10^9
TOTAL_MF_OS_03	ช่วง 0 ถึง 1.018×10^9
SUM_OS_BA01	ช่วง -1.17 ถึง 1.081×10^7
SUM_OS_BA02	ช่วง -1.17 ถึง 1.081×10^7
SUM_OS_BA03	ช่วง -1.17 ถึง 1.081×10^7
CC_Spending_1	ช่วง 0 ถึง 4.737×10^6
CC_Spending_2	ช่วง 0 ถึง 3.230×10^6
CC_Spending_3	ช่วง 0 ถึง 4.453×10^6
Class	‘DEPOSIT’,’NON_DEPOSIT’

จากตารางที่ 3.10 แสดงตารางที่บอกประเภทของตัวแปรและค่าที่เป็นไปได้พบว่า มีตัวแปรทั้งหมด 38 ตัวแปร ส่วนใหญ่เป็นตัวแปรเชิงคุณภาพ (Qualitative Variable) หมายถึงตัวแปรที่อธิบายลักษณะเฉพาะของกลุ่มข้อมูลโดยค่าที่เป็นไปได้ส่วนใหญ่จะอยู่ในรูปข้อความที่แสดงถึงลักษณะกลุ่มมีจำนวน 24 ตัวแปร ยกตัวอย่างเช่น MARITAL_STATUS_GRP มีค่าที่เป็นไปได้สำหรับของข้อมูลชุดนี้ คือ ‘Single’,’Married’,’Divorce’,’Widow’,’Undeffined’ ซึ่งแสดงลักษณะของลูกค้าที่มีสถานภาพสมรส และตัวแปรเชิงปริมาณ (Quantitative Variable) หมายถึงตัวแปรที่มีค่าที่สามารถนับหรือวัดได้เป็นตัวเลขและมีลักษณะที่สามารถใช้ในการพยากรณ์หรือคำนวณค่าต่าง ๆ ได้ มีจำนวน 14 ตัวแปร ยกตัวอย่างเช่น AGE มีค่าที่เป็นไปได้สำหรับข้อมูลชุดนี้ อยู่ในช่วง 0 ถึง 103 ซึ่งแสดงอายุของลูกค้า เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.4.3 การตรวจสอบคุณภาพของข้อมูล (Verify Data Quality)

การตรวจสอบคุณภาพข้อมูลเป็นขั้นตอนที่สำคัญในการทำงานกับข้อมูล เนื่องจากข้อมูลที่ไม่สมบูรณ์ไม่ถูกต้องหรือไม่น่าเชื่อถือ อาจทำให้เกิดผลลัพธ์ที่ไม่แม่นยำไม่สามารถนำไปใช้ประโยชน์ได้อย่างเต็มที่ได้ จากการตรวจสอบพบว่ามีข้อมูลสูญหาย (Missing Value) และค่านอกเกณฑ์ (Outlier) เป็นจำนวนมากที่อาจจะมาจากความผิดพลาดของระบบฐานข้อมูล และการดึงข้อมูลจากแหล่งต่าง ๆ โดยฐานข้อมูลของธนาคารมีการเก็บข้อมูลไว้หลายแหล่งจึงมีข้อจำกัดในด้านการค้นหาและรวบรวมเข้ามาไว้ในที่เดียว

```
5   REGION          253594 non-null  object
6   INCOME          253594 non-null  int64
7   MOBILE          253594 non-null  object
8   PAYROLL         253205 non-null  object
9   DURATION        253205 non-null  float64
10  EDUCATION       253205 non-null  object
11  V3              253594 non-null  float64
12  V2              253594 non-null  float64
13  V1              253594 non-null  float64
14  MAIN_BANK       253594 non-null  object
15  SUM_OS_SAVING_DEP  253594 non-null  float64
```

รูปที่ 3.6 ค่าสูญหายจากข้อมูลที่มาจากฐานข้อมูลธนาคาร

หลังจากที่นำข้อมูลมาจากฐานข้อมูลธนาคารเพื่อเข้าสู่โปรแกรมจำลองฐานข้อมูล Jupyter Notebook เสร็จสิ้น จากรูปที่ 3.6 แสดงค่าสูญหายจากข้อมูลที่มาจากฐานข้อมูลธนาคาร โดยดูจากตัวเลขที่อยู่นอกกรอบสีแดงและในกรอบพบว่า ตัวเลขที่อยู่ในกรอบสีแดงมีจำนวนน้อยกว่าตัวเลขที่อยู่นอกกรอบแสดงถึงค่าที่สูญหายของตัวแปร Payroll, Duration และ Education มีค่าสูญหายจำนวน 389 จำนวน การสูญหายข้อมูลที่เกิดขึ้นอาจส่งผลต่อความถูกต้องในการจำแนก ดังนั้นจึงพิจารณาสาเหตุของค่าสูญหายข้อมูลและผลกระทบที่อาจเกิดขึ้น

1) การสูญหายอาจจะเกิดจากข้อผิดพลาดของการเก็บข้อมูล เนื่องจากฐานข้อมูลของธนาคารมีการเก็บรวบรวมจากหลายแหล่งและมีการทดสอบระบบฐานข้อมูลอยู่บ่อยครั้ง บางครั้งผู้พัฒนาอาจจะสร้างข้อมูลเสมือนเพื่อทดสอบระบบและอาจจะไม่ได้ลบข้อมูลดังกล่าวออกจึงทำให้มีข้อมูลค้างอยู่ในระบบฐานข้อมูล

2) การสูญหายข้อมูลอาจเกิดขึ้นในกระบวนการถ่ายโอนข้อมูลจากฐานข้อมูลไปยังโปรแกรม Jupyter Notebook หรือในกระบวนการจัดเก็บข้อมูลเอง สาเหตุอาจมาจากการขัดข้องในการเชื่อมต่อระบบหรือการดึงข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

	INCOME
count	2.535760e+05
mean	2.199385e+05
std	6.619609e+07
min	0.000000e+00
25%	2.080000e+04
50%	3.156000e+04
75%	6.000000e+04
max	3.333333e+10

รูปที่ 3.7 ตัวแปรรายได้ที่มีค่านอกเกณฑ์

จากรูปที่ 3.7 แสดงตัวแปรรายได้ที่มีค่านอกเกณฑ์จากการสังเกตพบว่าค่าที่มากที่สุดของตัวแปรรายได้ (Income) มีค่ามากที่สุดเท่ากับ 3.33×10^{10} หรือ 3,333,333,333 บาท ต่อเดือน เป็นค่าที่มากและไม่สมเหตุสมผล ซึ่งค่ารายได้สูงเกินไปนี้อาจมาจากความผิดพลาดของระบบฐานข้อมูล เช่น การทดสอบฐานข้อมูลจากผู้พัฒนาหรือเกิดจากความผิดพลาดของระบบจัดเก็บข้อมูล ดังนั้น ข้อมูลที่มีค่านอกเกณฑ์ที่ไม่สมเหตุสมผลดังกล่าวควรถูกกำจัดออกจากแบบจำลอง

3.5 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนที่ 3 การเตรียมข้อมูลเป็นขั้นตอนที่มีความสำคัญในกระบวนการวิเคราะห์ข้อมูลหรือสร้างแบบจำลองการเรียนรู้ของเครื่อง โดยการทำขั้นตอนนี้อย่างถูกต้องและเหมาะสมจะส่งผลสำคัญต่อผลลัพธ์ของแบบจำลองการเรียนรู้ของเครื่อง หากไม่ทำการเตรียมข้อมูลอย่างถูกต้องและเหมาะสมอาจจะเกิดผลที่ตามมา

1) การพยากรณ์หรือการวิเคราะห์ที่ไม่แม่นยำ เพราะข้อมูลที่ไม่ถูกต้องหรือไม่สมบูรณ์อาจทำให้แบบจำลองพยากรณ์ผลลัพธ์ที่ไม่ถูกต้องหรือให้ผลลัพธ์ที่ไม่แม่นยำเท่าที่ควร

2) เสียเวลาและทรัพยากร เพราะการวิเคราะห์หรือการสร้างแบบจำลองด้วยข้อมูลที่ไม่เตรียมพร้อมอาจเสียเวลาและทรัพยากรในการพัฒนาแบบจำลองที่ไม่เป็นประสิทธิภาพ

3) แบบจำลองให้ผลลัพธ์ที่ผิดพลาดในการตีความ เพราะข้อมูลที่ไม่ถูกต้องหรือไม่สมบูรณ์อาจทำให้ผิดพลาดในการตีความผลลัพธ์หรือทำให้ไม่สามารถตีความได้ถูกต้องและสอดคล้องกับความต้องการได้

ดังนั้นงานวิจัยนี้จึงได้กำหนดขั้นตอนวิธีการเตรียมข้อมูลขึ้นมาทั้งสิ้น 6 ขั้นตอน 1.การคัดเลือกข้อมูล (Data Selection) 2.ทำความสะอาดข้อมูล (Data Cleaning) 3.การรวมข้อมูล (Data Integration) 4.การแปลงข้อมูล (Data Transformation) 5.การเลือกคุณลักษณะที่สำคัญ (Feature Selection) 6.การจัดการความไม่สมดุลของคลาส (Data Imbalance Handling)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.5.1 การคัดเลือกข้อมูล (Data Selection)

การคัดเลือกข้อมูลคุณลักษณะ ที่มีความสัมพันธ์กับการเปิดบัญชีเงินฝากประเภท D มีความสำคัญในการสร้างแบบจำลองให้มีประสิทธิภาพ หากเลือกตัวแปรคุณลักษณะที่สร้างแบบจำลองมากเกินไป ทำให้เสียเวลาในการเรียนรู้ของแบบจำลอง และทำให้แบบจำลองมีการเรียนรู้ที่ซับซ้อน ดังนั้นควรเลือกตัวแปรที่สามารถบ่งบอก ลักษณะของกลุ่มลูกค้าที่เปิดบัญชีและไม่เปิดบัญชีอย่างชัดเจน เพื่อให้แบบจำลองมีประสิทธิภาพในการจำแนกกลุ่มข้อมูลทั้ง 2 กลุ่ม โดยเหตุผลการคัดเลือกคุณลักษณะที่มีความสัมพันธ์ดังนี้

1) อายุ (AGE): อายุของลูกค้าอาจเป็นตัวแปรที่สำคัญในการวิเคราะห์เนื่องจากอาจมีความสัมพันธ์กับพฤติกรรมการเปิดบัญชี ยกตัวอย่างเช่น กลุ่มลูกค้าที่อายุมากอาจมีแนวโน้มที่จะมีความสนใจในการเริ่มต้นการเปิดบัญชีฝากเงินมากกว่ากลุ่มลูกค้าที่มีอายุน้อย เนื่องจากการเปิดบัญชีเพื่อออมเงินเป็นการออมที่มีความเสี่ยงต่ำมาก ทำให้ลูกค้าที่มีอายุมากที่เป็นกลุ่มผู้สูงอายุที่ไม่ต้องการมีความเสี่ยงในการลงทุนเลือกใช้

2) รายได้ (INCOME): รายได้ของลูกค้าส่วนใหญ่อาจมีความสัมพันธ์กับความสามารถในการเปิดบัญชีฝากเงิน ลูกค้าที่มีรายได้สูงอาจมีแนวโน้มที่จะมีเงินเก็บออมหรือลงทุนในบัญชีธนาคารมากกว่าเพราะมีเงินเหลือต่อเดือนมาก

3) อาชีพ (OCCUPATION_GRP): อาชีพของลูกค้าอาจมีผลต่อแนวโน้มในการเปิดบัญชี กลุ่มลูกค้าที่มีอาชีพที่มั่นคงและรายได้สูงอาจมีแนวโน้มที่จะเปิดบัญชีฝากเงินกับธนาคารมากกว่า

4) เพศ (GENDER_GRP): เพศอาจมีความสัมพันธ์กับแนวโน้มในการเปิดบัญชี บางครั้งการบริหารเงินในบัญชีธนาคารอาจแตกต่างกันขึ้นอยู่กับเพศของลูกค้า

5) สถานะสมรส (MARITAL_STATUS_GRP): สถานะแต่งงานอาจมีผลต่อการตัดสินใจในการเปิดบัญชีฝากเงิน ลูกค้าที่แต่งงานแล้วอาจมีความเสถียรและความมั่นคงทางการเงินมากกว่าลูกค้าที่ยังไม่แต่งงาน

6) ระดับการศึกษา (EDUCATION): ระดับการศึกษาอาจสื่อถึงความรู้และความเข้าใจในเรื่องการเงิน ลูกค้าที่มีระดับการศึกษาสูงอาจมีแนวโน้มที่จะมีการทำกำไรหรือลงทุนในบัญชีธนาคารมากกว่า

7) ภูมิภาค (REGION): ภูมิภาคที่ลูกค้าอยู่อาจมีความสัมพันธ์กับพฤติกรรมการเปิดบัญชีฝากเงิน เช่น ภูมิภาคที่มีรายได้เฉลี่ยสูงอาจมีแนวโน้มที่จะมีการเปิดบัญชีฝากเงินมากกว่า

8) มีแอปพลิเคชันธนาคาร (MOBILE): การมีแอปพลิเคชันธนาคารอาจเป็นตัวบ่งชี้ว่าลูกค้ามีความสนใจและความสะดวกในการเปิดบัญชีผ่านแอปพลิเคชันธนาคาร ซึ่งอาจมีความเกี่ยวข้องกับการเปิดบัญชีฝากเงิน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

9) เป็นลูกค้าที่รับเงินเดือนผ่านธนาคาร (PAYROLL): ลูกค้าที่มีการรับเงินเดือนผ่านธนาคารอาจมีความเชื่อมั่นและความสัมพันธ์กับธนาคารมากขึ้นและมีแนวโน้มที่จะเปิดบัญชีฝากเงินกับธนาคาร

10) ระยะเวลาที่อยู่ในธนาคาร (DURATION): ระยะเวลาที่ลูกค้าเป็นลูกค้าของธนาคารอาจสื่อถึงความคงที่และความพึงพอใจในบริการของธนาคาร ลูกค้าที่ใช้บริการธนาคารเป็นเวลานานอาจมีแนวโน้มที่จะเปิดบัญชีฝากเงินมากกว่า

11) เป็นลูกค้าประจำของธนาคาร (MAIN_BANK): ลูกค้าที่เป็นลูกค้าประจำของธนาคารอาจมีความเชื่อมั่นและความสัมพันธ์กับธนาคารมากขึ้น และมีแนวโน้มที่จะเปิดบัญชีฝากเงินกับธนาคาร การเป็นลูกค้าประจำได้ต้องตรงตามเงื่อนไขข้อใดข้อหนึ่ง คือ 1. มีบัญชีออมทรัพย์ธรรมดา 2. มีผลิตภัณฑ์บัตรเครดิต 3. มีผลิตภัณฑ์กองทุนหรือประกัน

12) การถูกเสนอขายผลิตภัณฑ์ (CAMPAIGN_CONTACT): การถูกเสนอขายผลิตภัณฑ์ธนาคารอาจมีผลต่อการตัดสินใจในการเปิดบัญชีฝากเงิน ลูกค้าที่เคยได้รับข้อมูลเกี่ยวกับผลิตภัณฑ์และความประโยชน์ของธนาคารอาจมีแนวโน้มที่จะเปิดบัญชีฝากเงินกับธนาคาร

13) สินทรัพย์รวมในธนาคาร (AUM): สินทรัพย์รวมในธนาคารของลูกค้าอาจสื่อถึงความมั่งคั่งและความสามารถในการลงทุน ลูกค้าที่มีสินทรัพย์รวมสูงอาจมีแนวโน้มที่จะเปิดบัญชีฝากเงินกับธนาคารมากกว่า

14) กลุ่มลูกค้าที่ขึ้นอยู่กับทรัพย์สินในธนาคาร (WM): ลูกค้าที่มีทรัพย์สินในธนาคารสูงอาจมีความต้องการในการบริการทางการเงินที่ทันสมัยและแตกต่างจากกลุ่มลูกค้าอื่น

15) ยอดเงินรวมในบัญชีออมทรัพย์ (SUM_OS_SAVING_DEP): แสดงถึงความสามารถในการออมเงินและเก็บเงินของลูกค้า ลูกค้าที่มียอดเงินสะสมมากอาจมีความเป็นไปได้ที่จะเปิดบัญชีฝากเงินเพิ่มขึ้น

16) ผลิตภัณฑ์เกี่ยวกับกองทุน (Mutual Funds): แสดงถึงความสนใจและความเชื่อมั่นในการลงทุนของลูกค้า ลูกค้าที่มีการถือครองหรือลงทุนในกองทุนอาจมีแนวโน้มที่จะเปิดบัญชีฝากเงินเพิ่มขึ้น

17) ผลิตภัณฑ์เกี่ยวกับประกัน (Bancassurance): แสดงถึงความต้องการในการประกันภัยของลูกค้ามั่นใจในธนาคาร ลูกค้าที่มีการถือครองหรือสมัครประกันภัยอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

18) ผลิตภัณฑ์เกี่ยวกับสินเชื่อค้ำประกัน (Secured Loan): แสดงถึงความต้องการในการกู้ยืมเงินแบบมีค้ำประกันของลูกค้า ลูกค้าที่มีความสนใจในสินเชื่อค้ำประกันอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

19) ผลิตภัณฑ์เกี่ยวกับสินเชื่อไม่ค้ำประกัน (Unsecured Loan): แสดงถึงความสามารถในการกู้ยืมเงินของลูกค้าโดยไม่ต้องมีคนค้ำประกัน ลูกค้าที่สนใจในการกู้ยืมเงินแบบไม่ค้ำประกันอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

20) ผลิตภัณฑ์เกี่ยวกับสินเชื่อเช่าซื้อ (Hire Purchase): แสดงถึงความสนใจในการกู้ยืมเงินสำหรับซื้อรถ ลูกค้าที่สนใจในสินเชื่อนี้อาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

21) ผลิตภัณฑ์บัตรเครดิตประเภทที่ A ถึง G (NO_ACCT_TYPE): แสดงถึงความสนใจในการใช้บัตรเครดิตของลูกค้า ลูกค้าที่มีผลิตภัณฑ์บัตรเครดิตแบบต่าง ๆ อาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

22) มูลค่ารวมของกองทุนที่ครอบครองย้อนหลัง 3 เดือน (TOTAL_MF_OS): แสดงถึงการถือครองกองทุนในช่วงเวลาที่แตกต่างกัน โดยลูกค้าแต่ละคนจะมีมูลค่ารวมของกองทุนที่ครอบครองแต่ละเดือนที่ไม่เท่ากัน ดังนั้นการเลือกคุณลักษณะที่ดูย้อนหลัง 3 เดือนเพื่อเป็นการดูมูลค่าโดยเฉลี่ยของกองทุนที่ลูกค้ามี ลูกค้าที่มีมูลค่ากองทุนที่ครอบครองมากอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

23) มูลค่ารวมของประกันที่ครอบครองปัจจุบัน ย้อนหลัง 3 เดือน (SUM_OS_BA): แสดงถึงมูลค่าของผลประโยชน์ที่ลูกค้าได้รับจากการถือครองประกันในช่วงเวลาที่แตกต่างกัน โดยลูกค้าแต่ละคนจะมีมูลค่ารวมของประกันที่ครอบครองแตกต่างกัน ดังนั้นการเลือกคุณลักษณะที่ดูย้อนหลัง 3 เดือนเพื่อเป็นการดูมูลค่าโดยเฉลี่ยของประกันที่ลูกค้ามี ลูกค้าที่มีมูลค่าประกันที่ครอบครองมากอาจมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

27) การใช้จ่ายบัตรเครดิต ย้อนหลัง 3 เดือน (CC_Spending): แสดงถึงการใช้จ่ายบัตรเครดิตในแต่ละเดือนข้อมูลค้าในช่วงเวลาที่แตกต่างกัน โดยลูกค้าแต่ละคนจะมีการใช้จ่ายบัตรเครดิตที่ไม่เท่ากันในแต่ละเดือน ดังนั้นการเลือกคุณลักษณะที่ดูการใช้จ่ายย้อนหลัง 3 เดือนเป็นการดูพฤติกรรมของลูกค้าว่ามีการใช้จ่ายอย่างไร ลูกค้าที่มีพฤติกรรมใช้จ่ายมากอาจจะมีแนวโน้มที่จะสนใจในการเปิดบัญชีฝากเงิน

28) ตัวแปรเป้าหมาย (Class): แสดงถึงลักษณะของข้อมูลว่าเป็นข้อมูลตัวอย่างของลูกค้าที่เปิดบัญชีหรือไม่เปิดบัญชี เป็นตัวแปรที่สำคัญในฝึกสอนแบบจำลองให้เกิดการเรียนรู้และพยากรณ์ข้อมูลได้ถูกต้อง

ข้อมูลคุณลักษณะส่วนใหญ่เป็นตัวแปรเชิงคุณภาพ และเป็นข้อมูลคุณลักษณะส่วนบุคคล ซึ่งการนำคุณลักษณะส่วนบุคคลเข้าในแบบจำลองการเรียนรู้ของเครื่องมาก ๆ จะช่วยให้แบบจำลองสามารถรู้จักและจำแนกลูกค้าที่มีแนวโน้มในการเปิดบัญชีเงินฝาก ลักษณะคล้ายคลึงกันได้ดีขึ้น อาจจะช่วยเพิ่มโอกาสการได้ลูกค้าที่มีศักยภาพที่เหมาะสม สำหรับแคมเปญเสนอผลิตภัณฑ์บัญชีเงินฝากประเภท D ที่ส่งไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.5.2 การทำความสะอาดข้อมูล (Data Cleaning)

การทำความสะอาดข้อมูลเป็นขั้นตอนสำคัญในการเตรียมข้อมูล เพื่อให้ข้อมูลสมบูรณ์และน่าเชื่อถือก่อนที่จะนำมาใช้ในการสร้างแบบจำลอง เป็นขั้นตอนที่ช่วยลดความผิดพลาดที่อาจเกิดจากข้อมูลที่ไม่ถูกต้อง เช่น ข้อมูลที่ขาดหายไป (Missing Value) ข้อมูลที่มีค่านอกเกณฑ์ (Outlier) ที่จะส่งผลทำให้แบบจำลองมีประสิทธิภาพในการจำแนกน้อยลง แต่เนื่องจากงานวิจัยนี้มีข้อมูลตัวแปรเป้าหมายที่เป็นคลาสบวก คือ กลุ่มลูกค้าที่เปิดบัญชี หรือ ‘DEPOSIT’ มีจำนวนข้อมูล 4,312 ข้อมูล ซึ่งมีจำนวนน้อยกว่า กลุ่มลูกค้าที่ไม่เปิดบัญชี ‘NON_DEPOSIT’ ที่มีข้อมูลเท่ากับ 249,282 ข้อมูล ดังนั้นงานวิจัยนี้จึงต้องการที่จะคงสภาพข้อมูลให้เหลือจำนวนข้อมูลให้มีมากที่สุดเท่าที่เป็นไปได้จึงมีการผสมผสานระหว่างการกำจัดค่านอกเกณฑ์แบบใช้เทคนิคสถิติและการกำจัดค่านอกเกณฑ์ด้วยหลักเกณฑ์ของผู้วิจัย โดยการทำความสะอาดข้อมูลมี 3 ขั้นตอนดังนี้

ขั้นตอนที่ 1) ตรวจสอบข้อมูลที่ขาดหายไป (Missing Value) ผู้วิจัยได้ตรวจสอบค่าที่สูญหายพบว่ามีจำนวน 389 ข้อมูล และเป็นข้อมูลที่ไม่สามารถหาค่ามาแทนได้ เช่น ค่าเฉลี่ย หรืออื่น ๆ ดังนั้นผู้วิจัยจึงลบข้อมูลทั้ง 389 ข้อมูลออก

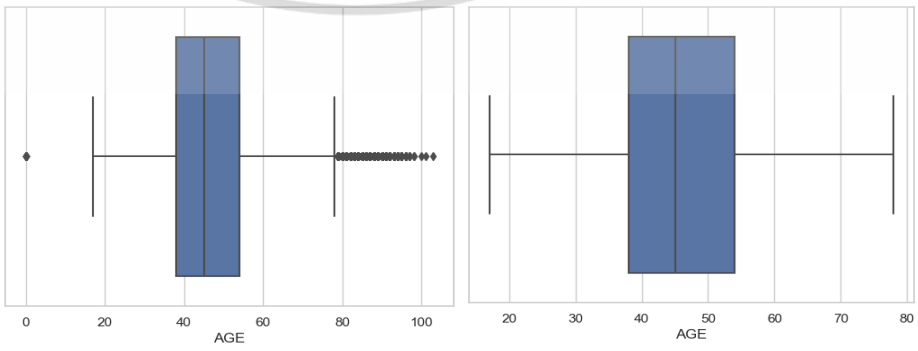
```
In [12]: df.dropna(inplace=True)
         print(df .isnull().sum().sum())

         0
```

รูปที่ 3.8 คำสั่งที่ใช้ในการลบข้อมูลสูญหาย

จากรูปที่ 3.8 แสดงคำสั่งที่ใช้ในการลบข้อมูลสูญหายของภาษาไพทอนที่ทำการลบแถว (Row) ออกจากข้อมูลหลังจากนั้นโชว์ตัวเลขค่าสูญหายจะเห็นว่าเท่ากับ 0

ขั้นตอนที่ 2) ตรวจสอบค่านอกเกณฑ์ของตัวแปรอายุ (Age) ด้วย แผนภาพกล่อง (Box Plot) เพื่อตรวจสอบว่ามีค่านอกเกณฑ์ที่อยู่นอกแผนภาพกล่องหรือไม่



Original Dataset Edited Dataset

รูปที่ 3.9 แผนภาพกล่องของข้อมูลอายุ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.9 แสดงแผนภาพกล่องของข้อมูลอายุที่ พบว่าค่าที่อยู่นอกแผนภาพกล่องทางซ้ายและทางขวาเป็นจำนวนมาก ดังนั้นจึงใช้การกำจัดค่านอกเกณฑ์เทคนิคที่ใช้คือการหาค่าพิสัยควอไทล์ (Interquartile Range) คือวิธีการในการตรวจจับข้อมูลที่มีค่าผิดปกติออกไปจากข้อมูล ค่าที่อยู่นอกช่วงพิสัยควอไทล์ อาจถือว่าเป็นค่านอกเกณฑ์ โดยช่วงขอบเขตต่ำสุด (Lower Limit) และขอบเขตสูงสุด (Upper Limit) คำนวณได้ดังสมการที่ 3.1 และ3.2

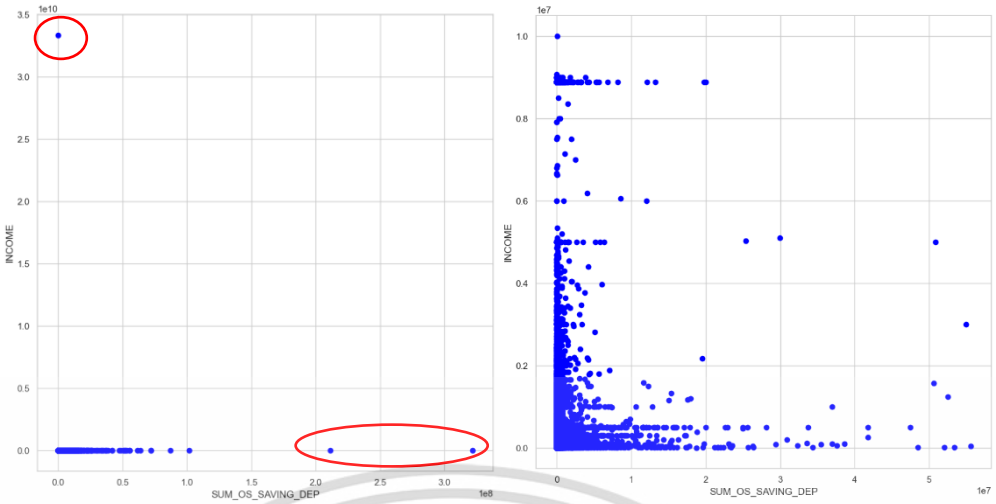
$$\text{Lower Limit} = Q1 - 1.5\ \text{IQR} = 38 - 1.5(16) = 14.0 \quad (3.1)$$

$$\text{Upper Limit} = Q3 - 1.5\ \text{IQR} = 54 - 1.5(16) = 78.0 \quad (3.2)$$

จากสมการที่ 3.1 และ 3.2 แสดงการคำนวณช่วงขอบเขตต่ำสุด (Lower Limit) และขอบเขตสูงสุด (Upper Limit) โดยช่วงเขตต่ำสุดมีค่าเท่ากับ 14.0 และช่วงเขตสูงสุดมีค่าเท่ากับ 78.0 หากข้อมูลอายุลูกค้าที่มีค่านอกช่วง 14.0 ถึง 78.0 ถือว่าเป็นค่าที่อยู่นอกเกณฑ์ดังนั้นจึงตัดข้อมูลดังกล่าวออก จากภาพด้านซ้ายมือ (Original Dataset) พบว่าหลังจากที่ตัดค่านอกเกณฑ์ที่อยู่นอกช่วงขอบเขตต่ำสุดและขอบเขตสูงสุดจะได้แผนภาพกล่องใหม่ด้านขวามือ (Edited Dataset) โดยจะได้แผนภาพที่จะไม่มีค่านอกเกณฑ์ที่อยู่นอกแผนภาพกล่องอีกต่อไป

ขั้นตอนที่ 3) ตรวจสอบค่านอกเกณฑ์ที่ไม่สมเหตุสมผล โดยการกำจัดค่านอกเกณฑ์ด้วยหลักเกณฑ์ของผู้วิจัยที่กำหนดขึ้น เนื่องจากข้อมูลคลาส (Class) ทั้งสองค่อนข้างไม่สมดุลกันเป็นอย่างมากระหว่างคลาสของข้อมูลกลุ่มคนที่เปิดบัญชี หรือ ‘DEPOSIT’ กับข้อมูลกลุ่มคนที่ไม่เปิดบัญชี หรือ ‘NON_DEPOSIT’ เพื่อให้เหลือจำนวนมากที่สุด จึงจำเป็นที่ต้องเลือกตัดข้อมูลที่เป็นค่านอกเกณฑ์ที่ไม่สมเหตุสมผลมากที่สุด ด้วยการสังเกตจากแผนภูมิการกระจายของข้อมูล หากค่าใดที่เป็นค่าอยู่นอกเกณฑ์ที่ทางผู้วิจัยมองแล้วว่าควรตัดออก จึงทำการตัดค่าดังกล่าวออกบางส่วน เพื่อให้เหลือข้อมูลมากที่สุดทำให้การฝึกสอนแบบจำลองมีประสิทธิภาพมากขึ้น ดังนั้นจึงสร้างแผนภูมิกระจายข้อมูลระหว่างตัวแปรเพื่อดูค่านอกเกณฑ์ดังนี้

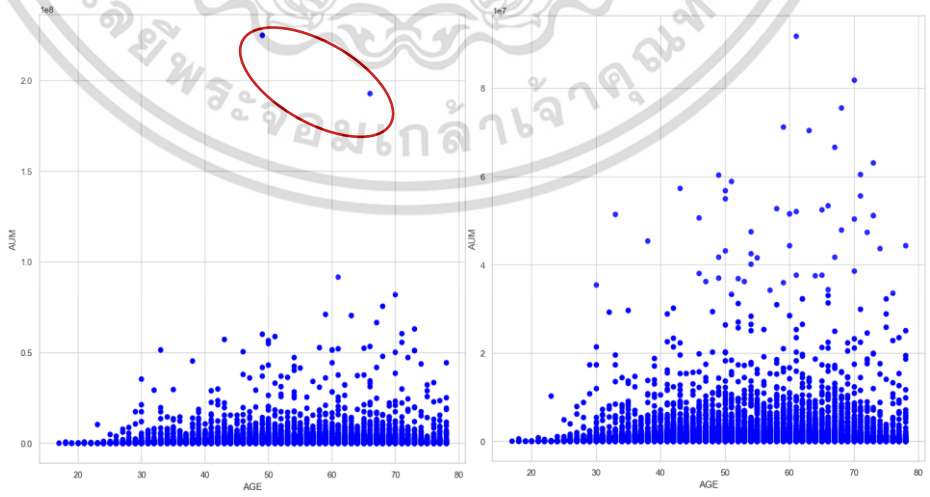
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



Original Dataset Edited Dataset

รูปที่ 3.10 แผนภูมิกระจายข้อมูลระหว่างยอดเงินรวมที่อยู่ในบัญชีออมทรัพย์กับรายได้

จากรูปที่ 3.10 จากแผนภูมิการกระจายข้อมูล (Scatter plot) ระหว่างยอดเงินรวมที่อยู่ในบัญชีออมทรัพย์ (SUM_OS_SAVING_DEP) กับรายได้ต่อเดือน (Income) เพื่อดูการกระจายและหาค่านอกเกณฑ์ที่ไม่สมเหตุสมผล พบว่าจะมีค่านอกเกณฑ์ที่แกน Y มีอยู่ 1 ค่าและแกน X มีอยู่ 2 ค่าดังภาพด้านซ้ายมือ (Original Dataset) จากการวิเคราะห์ของผู้วิจัยมีความเห็นว่า ค่าดังกล่าวคือค่านอกเกณฑ์ที่ไม่สมเหตุสมผลจึงตัดค่าออกบ้างส่วน เพื่อคงข้อมูลไว้ให้ได้มากที่สุดและจะได้แผนภูมิที่ถูกกำจัดค่านอกเกณฑ์เสร็จสิ้นที่เห็นถึงการกระจายข้อมูลมากขึ้นดังรูปด้านขวามือ (Edited Dataset)



Original Dataset Edited Dataset

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 3.11 แผนภูมิกระจายข้อมูลระหว่างสินทรัพย์รวมในธนาคารและอายุ

จากรูป 3.11 ตรวจสอบค่านอกเกณฑ์ที่ไม่สมเหตุสมผลโดยการสร้างแผนภูมิกระจายข้อมูลระหว่างข้อมูลสินทรัพย์รวมในธนาคาร (AUM) และอายุ (AGE) พบว่าจากรูปด้านซ้ายมือ (Original Dataset) มีค่านอกเกณฑ์ที่ไม่สมเหตุสมผลอยู่ 2 ค่า ดังวงกลมสีแดงผู้วิจัยจึงมองว่าเป็นค่านอกเกณฑ์ที่ไม่สมเหตุสมผล จึงทำการตัดค่าออกไปและจะเห็นการกระจายตัวของข้อมูลมากยิ่งขึ้นดังรูปด้านขวามือ (Edited Dataset)

ตารางที่ 3.11 เปรียบเทียบข้อมูลเดิมกับข้อมูลใหม่หลังจากกำจัดค่าสูญหายและค่านอกเกณฑ์

ตัวแปรเป้าหมาย	ข้อมูลดั้งเดิม	ข้อมูลใหม่
DEPOSIT	4,312	3,907
NON_DEPOSIT	249,282	248,023

จากตารางที่ 3.11 แสดงการเปรียบเทียบตัวแปรเป้าหมายเดิมกับตัวแปรเป้าหมายใหม่หลังจากทำขั้นตอนการกำจัดค่าสูญหายและค่านอกเกณฑ์แล้วจะพบว่าจากข้อมูลทั้งหมด 253,594 ข้อมูล หลังจากการกำจัดค่าสูญหายและค่านอกเกณฑ์ที่ไม่สมเหตุสมผลมีข้อมูลที่ถูกตัดออกไปจำนวน 1,664 ข้อมูล แบ่งเป็นข้อมูลของตัวแปรเป้าหมาย ‘DEPOSIT’ หายไป 405 ข้อมูลและ ‘NON_DEPOSIT’ หายไป 1,259 ข้อมูล อย่างไรก็ตามข้อมูลชุดนี้ยังมีค่านอกเกณฑ์อยู่บางส่วน แต่เพื่อวัตถุประสงค์ของผู้วิจัยที่ต้องการคงจำนวนข้อมูลให้ได้มากที่สุดเท่าที่เป็นไปได้ เพื่อให้การฝึกสอนแบบจำลองมีข้อมูลการเรียนรู้ที่มากขึ้น ดังนั้นจึงผสมผสานวิธีการกำจัดค่านอกเกณฑ์แบบใช้เทคนิคทางสถิติและกำจัดค่านอกเกณฑ์ด้วยหลักเกณฑ์ของผู้วิจัย

3.5.3 การรวมข้อมูล (Data Integration)

การรวมข้อมูลเป็นขั้นตอนสำคัญที่ช่วยเตรียมข้อมูลให้พร้อมใช้งาน ในการสร้างแบบจำลองการจำแนกหมวดหมู่ โดยเพิ่มความครอบคลุมและความถูกต้องของข้อมูล ลดความผิดพลาด ปรับสมดุลของข้อมูล และลดการซ้ำซ้อน เพื่อให้แบบจำลองมีประสิทธิภาพและมีความแม่นยำในการพยากรณ์สูงยิ่งขึ้น กระบวนการรวมข้อมูลเป็นกระบวนการที่นำข้อมูลมาเชื่อมต่อเข้าด้วยกันเพื่อสร้างข้อมูลที่ครอบคลุมและมีคุณภาพสูงขึ้น เนื่องจากข้อมูลคุณลักษณะบางคุณลักษณะมาจากแหล่งข้อมูลที่แตกต่างกัน ดังนั้นเพื่อลดความซ้ำซ้อนของการเรียนรู้แบบจำลองจึงได้มีการรวมคุณลักษณะต่าง ๆ เข้าด้วยกันดังขั้นตอนต่อไปนี้

1) การรวมคุณลักษณะ (Feature) การใช้จ่ายบัตรเครดิตย้อนหลัง 3 เดือน เนื่องจากคุณลักษณะการใช้จ่ายบัตรเครดิตของลูกค้าที่แต่ละคนแตกต่างกัน จึงต้องการรวมคุณลักษณะเข้าด้วยกันเป็นคุณลักษณะที่บอกการใช้จ่ายบัตรเครดิตย้อนหลัง เช่น ข้อมูลลูกค้าธนาคารรายหนึ่งได้เปิดบัญชีเงินฝากประเภท D เดือนที่ 5 ของปี พ.ศ. 2565 ดังนั้นข้อมูลการใช้จ่ายย้อนหลัง 3 เดือนจะเป็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การนำข้อมูลการใช้จ่ายบัตรเครดิตของเดือนที่ 4 3 2 ตามลำดับ หากเป็นข้อมูลที่เป็นลูกค้าที่ไม่เปิดบัญชีเลยภายใน 1 ปีจากที่กล่าวข้างต้นในขั้นตอนการเก็บข้อมูลงานวิจัยนี้จะใช้ข้อมูลสิ้นปีคือเดือนที่ 12 เป็นข้อมูลลูกค้าตัวอย่าง โดยคุณลักษณะการใช้จ่ายบัตรเครดิตย้อนหลังจะเหมือนกันทั้งหมดคือ เดือนที่ 12 11 10 ตามลำดับ การรวมคุณลักษณะเข้าด้วยกันเป็นคุณลักษณะเดียว จะแสดงดังสมการที่ 3.3

$$\text{CC\_EXPENSE\_AVG} = \frac{CC\_spending\_1+CC\_spending\_2+CC\_spending\_3}{3} \quad (3.3)$$

จากสมการที่ 3.3 แสดงการรวมคุณลักษณะการใช้จ่ายบัตรเครดิตย้อนหลัง 3 เดือน กลายเป็นคุณลักษณะใหม่ชื่อว่า CC_EXPENSE_AVG คือ คุณลักษณะการใช้จ่ายบัตรเครดิตโดยเฉลี่ยย้อนหลัง 3 เดือน ซึ่งบ่งบอกถึงพฤติกรรมการใช้เงินเฉลี่ยของลูกค้าก่อนที่จะเปิดบัญชีหรือพฤติกรรมการใช้เงินเฉลี่ยของลูกค้าที่ไม่เปิดบัญชี

2) การรวมคุณลักษณะ (Feature) มูลค่ารวมของกองทุนที่ครอบครองย้อนหลัง 3 เดือน: เนื่องจากคุณลักษณะมูลค่ารวมของกองทุนที่ครอบครองของลูกค้าที่แต่ละคนในแต่ละเดือนมีความแตกต่างกันดังคำอธิบายขั้นตอน 1 จึงรวมคุณลักษณะเข้าด้วยกันดังสมการที่ 3.4

$$\text{TOTAL\_MF\_AVG} = \frac{\text{TOTAL\_MF\_OS\_01}+\text{TOTAL\_MF\_OS\_01}+\text{TOTAL\_MF\_OS\_01}}{3} \quad (3.4)$$

จากสมการที่ 3.4 แสดงการรวมคุณลักษณะมูลค่ารวมของกองทุนที่ครอบครองย้อนหลัง 3 เดือน กลายเป็นคุณลักษณะใหม่ที่ชื่อว่า TOTAL_MF_AVG คือ คุณลักษณะมูลค่ารวมของกองทุนโดยเฉลี่ย 3 เดือนย้อนหลัง ซึ่งบ่งบอกถึงคุณลักษณะของลูกค้าว่ามีการถือครองผลิตภัณฑ์ของธนาคารที่เกี่ยวกับกองทุนเฉลี่ย 3 เดือนย้อนหลังมีมูลค่าอย่างไร

3) การรวมคุณลักษณะ (Feature) มูลค่ารวมของประกันที่ครอบครองย้อนหลัง 3 เดือน: เนื่องจากคุณลักษณะมูลค่ารวมของประกันที่ครอบครองของลูกค้าที่แต่ละคนในแต่ละเดือนมีความแตกต่างกันดังคำอธิบายขั้นตอน 1 จึงรวมคุณลักษณะเข้าด้วยกันดังสมการที่ 3.5

$$SUM\_OS\_BA\_AVG = \frac{SUM\_OS\_BA01+SUM\_OS\_BA02+SUM\_OS\_BA03}{3} \quad (3.5)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากสมการที่ 3.5 แสดงการรวมคุณลักษณะมูลค่ารวมของประกันที่ครอบครองย้อนหลัง 3 เดือน กลายเป็นคุณลักษณะใหม่ที่ชื่อว่า SUM_OS_BA_AVG คือ คุณลักษณะมูลค่ารวมของประกันโดยเฉลี่ย 3 เดือน ซึ่งบ่งบอกถึงคุณลักษณะของลูกค้าว่ามีการถือครองผลิตภัณฑ์ของธนาคารที่เกี่ยวกับประกันเฉลี่ย 3 เดือนย้อนหลังมีมูลค่าอย่างไร

ดังนั้นจากคุณลักษณะทั้งหมดที่จะนำเข้าสู่แบบจำลองมีเท่ากับ 38 ตัวแปร เมื่อมีการรวมคุณลักษณะเข้าด้วยกันให้กลายเป็นตัวแปรเดียวจะทำให้คุณลักษณะปัจจุบันที่นำเข้าสู่แบบจำลองเหลือเท่ากับ 32 คุณลักษณะที่สำคัญ ทำให้แบบจำลองที่ได้รับการฝึกสอนกับข้อมูลชุดนี้ที่มี 32 ตัวแปรอาจจะมีประสิทธิภาพในการจำแนกมากขึ้น เนื่องจากได้รวมคุณลักษณะที่คล้ายกันเข้าด้วยกันทำให้ลดความซ้ำซ้อนของข้อมูล ทำให้แบบจำลองอาจมีประสิทธิภาพดีขึ้น โดยขั้นตอนต่อไปในการเตรียมแบบจำลองจะเป็นการนำตัวแปรที่เหลืออยู่ 32 ตัวแปรไปใช้งานในลำดับต่อไป

3.5.3 การแปลงข้อมูล (Data Transformation)

นามธรรมของข้อมูลมักแตกต่างกันและอยู่ในรูปแบบที่ไม่เหมือนกัน และในบางครั้งอาจจะมีขอบเขตหรือข้อจำกัดในการนำข้อมูลไปใช้ในกระบวนการวิเคราะห์ ดังนั้นการแปลงข้อมูลสามารถทำได้หลากหลายวิธีเพื่อให้ข้อมูลเข้ามาใช้ในกระบวนการสร้างแบบจำลองได้ถูกต้องและเหมาะสมกับแบบจำลอง โดยทางผู้จัดทำได้แบ่งการแปลงข้อมูลออกเป็น 2 วิธีที่ใช้กับตัวแปรที่แตกต่างกันดังนี้

1) การเข้ารหัสป้ายกำกับ (Label Encoding) ใช้กับข้อมูลเชิงคุณภาพ (Qualitative Data) เพื่อกำหนดค่าตัวเลขหรือรหัสสำหรับข้อมูลเชิงคุณภาพทั้งหมด 24 ตัว เช่น การแปลงจาก ‘Yes’ เป็น 0 ‘No’ เป็น 1 หรือ Marital Status โดยที่ ‘Single’ เป็น 0 ‘Married’ เป็น 1 ’Divorce’ เป็น 2 ‘Widow’ เป็น 3 ‘Undefined’ เป็น 4 เป็นต้น การเข้ารหัสป้ายกำกับช่วยในการแปลงตัวแปรประเภทข้อมูลที่เป็นข้อความหรือป้ายกำกับให้เป็นตัวเลข ซึ่งเป็นรูปแบบที่แบบจำลองจะเรียนรู้และพยากรณ์ได้ง่ายขึ้น

2) การปรับค่าให้อยู่ในช่วงสูงสุดและต่ำสุด (Min-Max Normalization) ใช้กับข้อมูลเชิงปริมาณ (Quantitative Data) เป็นวิธีการปรับค่าข้อมูลให้อยู่ในช่วงระหว่างค่าต่ำสุด (min) ถึงค่าสูงสุด (max) ที่กำหนดไว้ เพื่อให้ข้อมูลมีช่วงค่าที่เท่าเดิมและอยู่ในช่วงที่เซตค่าสามารถประมวลผลได้ง่ายและมีความสมดุล โดยทำการปรับค่าของข้อมูลให้อยู่ในช่วงระหว่าง 0 ถึง 1 โดยตัวแปรที่ถูกแปลงค่าคือตัวแปรที่เป็นเชิงปริมาณ เช่น INCOME, AGE, SUM_OS_SAVING_DEP, CC_EXPENSE_AVG, DURATION, TOTAL_MF_AVG, SUM_OS_BA และ AUM

ขั้นตอนการแปลงข้อมูลสำคัญเป็นอย่างมากเนื่องจากบางครั้งข้อมูลอาจอยู่ในรูปแบบที่ไม่เหมาะสมสำหรับการนำเข้าแบบจำลอง เช่น ข้อมูลที่เป็นข้อความหรือสัญลักษณ์ต้องถูกแปลงเป็นตัวเลข หรือข้อมูลที่มีหน่วยวัดที่แตกต่างกันต้องถูกปรับให้เหมือนกัน เพื่อให้แบบจำลองสามารถ

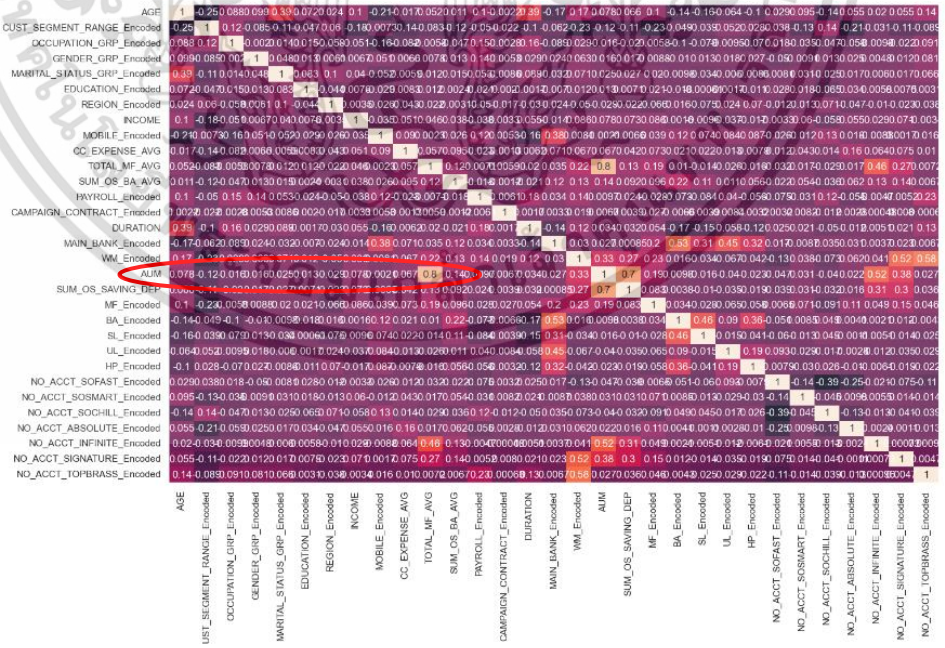
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เรียนรู้และพยากรณ์ได้อย่างแม่นยำ ดังนั้นต้องมีการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำเข้าแบบจำลอง การเตรียมข้อมูลอย่างถูกต้องและครบถ้วนช่วยให้ได้ผลลัพธ์ที่ถูกต้องและน่าเชื่อถือและสามารถใช้ในการพยากรณ์หรือการตัดสินใจในการดำเนินงานต่อไปได้อย่างมีประสิทธิภาพ

3.5.5 การเลือกคุณลักษณะที่สำคัญ (Feature Selection)

เป็นกระบวนการในการเลือกเฉพาะตัวแปรหรือคุณลักษณะที่มีความสำคัญและมีประสิทธิภาพสูงสุดในการวิเคราะห์ข้อมูล เพื่อนำไปใช้ในการสร้างแบบจำลองการเรียนรู้ของเครื่องที่มีประสิทธิภาพและมีความแม่นยำในการพยากรณ์หรือจัดกลุ่มข้อมูล ดังนั้นงานวิจัยนี้ได้ทำการเลือกคุณลักษณะด้วยเทคนิคทางสถิติ 2 วิธี คือ การทำตารางสัมประสิทธิ์สหสัมพันธ์ (Correlation Matrix) และวิเคราะห์ความแปรปรวน (ANOVA)

1) การเลือกคุณลักษณะที่สำคัญด้วยค่าสัมประสิทธิ์สหสัมพันธ์ เป็นกระบวนการที่ใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรและตัวแปร เพื่อเลือกคุณสมบัติที่มีความสำคัญสำหรับการสร้างแบบจำลองใช้ค่า ‘r’ เพื่อวัดความสัมพันธ์ระหว่างคู่ของตัวแปร โดยการเลือกคุณลักษณะที่สำคัญด้วยค่าสัมประสิทธิ์สหสัมพันธ์เหมาะสำหรับค่าต่อเนื่อง (Continuous Data) หรือข้อมูลประเภทไบนารี (Binary Data) เช่น ข้อมูลที่เป็นตัวเลข หรือ ข้อมูลที่มีค่าเป็น 0 หรือ 1 เพราะค่าสัมประสิทธิ์สหสัมพันธ์สามารถวัดความสัมพันธ์ระหว่างคู่ของตัวแปรได้



รูปที่ 3.12 ตารางแสดงสัมประสิทธิ์สหสัมพันธ์แต่ละตัวแปร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.12 ตารางแสดงสัมประสิทธิ์สหสัมพันธ์แต่ละตัวแปรเพื่อหาความสัมพันธ์ระหว่างตัวแปรและทำการตัดตัวแปรที่มีความสัมพันธ์กันมากออกเพื่อลดปัญหาภาวะร่วมเส้นตรงหลายตัวแปร การทำตารางสัมประสิทธิ์สหสัมพันธ์ ผู้จัดทำได้ใช้คำสั่ง .Corr() สร้างตารางสัมประสิทธิ์สหสัมพันธ์และผู้จัดทำได้กำหนดเกณฑ์การตัดคุณลักษณะไว้ที่ Correlation>0.7 ซึ่งเป็นเกณฑ์ที่แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันในระดับสูง (สุจิตรา, 2563) หากตัวแปรใดที่มีค่าเกิน 0.7 จะต้องตัดตัวแปรนั้นออกจากแบบจำลองเพื่อลดความซับซ้อนของแบบจำลอง จึงทำให้ตัวแปร AUM ที่มีค่า Correlation=0.8 ที่ จะถูกตัดออกจากแบบจำลองในที่สุด

2) การเลือกคุณลักษณะวิเคราะห์ความแปรปรวน (ANOVA) งานวิจัยนี้ได้ใช้คำสั่งในของ SelectKBest ใช้วิธี f_classif เพื่อเลือกคุณลักษณะที่มีความสำคัญและประสิทธิภาพสูงที่สุด k ตัว โดยหลักการของ SelectKBest โดยวิธี f_classif คือการคำนวณค่าสถิติเอฟ (F-value) ในการวิเคราะห์ความแปรปรวน (One Way ANOVA test) สำหรับแต่ละตัวแปรและคำนวณค่าความน่าจะเป็น (p-value) ที่สามารถใช้ในการประเมินความสำคัญของคุณลักษณะด้วยกัน โดยถ้าค่าความน่าจะเป็น สูงกว่าระดับนัยสำคัญที่ 0.05 หมายถึงค่าเฉลี่ยของคุณลักษณะนั้นเท่ากันและมีความสำคัญน้อย ในขณะที่ค่าความน่าจะเป็น ต่ำกว่าระดับนัยสำคัญ 0.05 หมายถึงค่าเฉลี่ยของคุณลักษณะนั้นแตกต่างกันอย่างน้อย 1 กลุ่ม และมีความสำคัญสูง

ตั้งสมมติฐาน

H_0: ค่าเฉลี่ยของตัวแปรตามเท่ากันในทุกกลุ่ม

H_1: ค่าเฉลี่ยของตัวแปรตามไม่เท่ากันอย่างน้อย 1 กลุ่ม

ตารางที่ 3.12 วิเคราะห์ความแปรปรวนสถิติทดสอบเอฟและค่าความน่าจะเป็นแต่ละตัวแปร

คุณลักษณะ	สถิติทดสอบเอฟ	P-value
AGE	15.261	9.363×10^{-5}
CUST_SEGMENT_RANGE	107.189	4.093×10^{-25}
OCCUPATION_GRP	23.534	1.227×10^{-6}
GENDER_GRP	26.519	2.610×10^{-7}
MARITAL_STATUS_GRP	0.267	0.604
EDUCATION	0.547	0.459
REGION	54.956	1.236×10^{-13}
INCOME	38.754	4.814×10^{-10}

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.12 (ต่อ) วิเคราะห์ความแปรปรวนสถิติทดสอบเอฟและค่าความน่าจะเป็นแต่ละตัวแปร

คุณลักษณะ	สถิติทดสอบเอฟ	P-value
MOBILE	1.822	0.177
CC_EXPENSE_AVG	33.394	7.531×10^{-9}
TOTAL_MF_AVG	600.907	1.521×10^{-132}
SUM_OS_BA_AVG	155.653	1.032×10^{-35}
PAYROLL	123.636	1.027×10^{-28}
CAMPAIGN_CONTACT	2.442	0.118
DURATION	63.257	1.821×10^{-15}
MAIN_BANK	74.304	6.734×10^{-18}
WM	1253.981	5.375×10^{-274}
SUM_OS_SAVING_DEP	1306.424	2.462×10^{-285}
MF	343.188	1.450×10^{-76}
BA	54.717	1.396×10^{-13}
SL	33.427	7.405×10^{-9}
UL	26.016	3.387×10^{-7}
HP	35.404	2.682×10^{-25}
NO_ACCT_ TYPE_A	4.857	0.027
NO_ACCT_ TYPE_B	56.924	4.543×10^{-14}
NO_ACCT_ TYPE_C	26.898	2.146×10^{-7}
NO_ACCT_ TYPE_D	27.941	1.251×10^{-7}
NO_ACCT_ TYPE_E	309.616	2.912×10^{-69}
NO_ACCT_ TYPE_F	599.947	2.458×10^{-132}
NO_ACCT_ TYPE_G	8.787	0.003

จากตารางที่ 3.12 แสดงวิเคราะห์ความแปรปรวนสถิติทดสอบเอฟและค่าความน่าจะเป็นของแต่ละตัวแปรหลังจากการใช้คำสั่ง f_classif เพื่อคำนวณ One way ANOVA Test แล้วจะพบว่าคุณลักษณะ MARITAL_STATUS_GRP, EDUCATION, MOBILE และCAMPAIGN_CONTACT มีค่าความน่าจะเป็น (P-value) สูงกว่าระดับนัยสำคัญที่ผู้วิจัยตั้งไว้ 0.05 หมายความว่า ยอมรับสมมติฐานหลักที่ว่า ค่าเฉลี่ยระหว่างกลุ่มของคุณลักษณะดังกล่าวนั้นไม่แตกต่างกัน ทำให้ไม่มีหลักฐานที่เพียงพอในการสรุปว่าค่าเฉลี่ยของตัวแปรตามขึ้นอยู่กับค่าของคุณลักษณะทั้ง 4 ตัว จึง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเอกสารนี้ทุกครั้งที่มีการนำไปใช้

สรุปว่า MARITAL_STATUS_GRP, EDUCATION, MOBILE และ CAMPAIGN_CONTACT ไม่มีผลต่อตัวแปรเป้าหมายหรือตัวแปรตาม

3.5.6 จัดการความไม่สมดุลของข้อมูล (Data Imbalance Handling)

การจัดการความไม่สมดุลของข้อมูล (Data Imbalance Handling) เป็นกระบวนการหรือเทคนิคที่ใช้ในการปรับปรุงปัญหาที่เกี่ยวข้องกับความไม่สมดุลในจำนวนตัวอย่างของคลาสต่าง ๆ ในชุดข้อมูล โดยทั่วไปแล้วเรามักพบว่า คลาสหนึ่งมีจำนวนตัวอย่างที่มากกว่าอีกคลาสหนึ่ง ซึ่งอาจส่งผลให้แบบจำลองเรียนรู้ไม่แม่นยำ เนื่องจากมองข้ามคลาสที่มีจำนวนน้อยกว่าส่งผลให้แบบจำลองพยากรณ์คลาสที่มีจำนวนมากกว่าเป็นหลัก นำไปสู่ผลลัพธ์ที่ไม่เสถียรและไม่น่าเชื่อถือ โดยปกติแล้วเราสามารถเพิ่มข้อมูลได้จากการค้นคว้าหาข้อมูลเพิ่มเติมจากแหล่งต่าง ๆ แต่การค้นคว้าหาข้อมูลเพิ่มมักจะต้องใช้งบประมาณค่าใช้จ่าย ในการค้นคว้าเพื่อให้ได้ซึ่งข้อมูลมาทำให้มีข้อจำกัดในการเพิ่มข้อมูล แต่การใช้เทคนิคทางคณิตศาสตร์ที่เพิ่มข้อมูลจากข้อมูลที่มีอยู่ เป็นเทคนิคที่หลายงานวิจัยใช้อย่างแพร่หลายและมีหลายเทคนิคให้เลือกใช้โดยไม่มีข้อจำกัดทางด้านค่าใช้จ่าย ดังนั้นในงานวิจัยนี้จึงใช้วิธี "Combination of Oversampling and Undersampling" หรือการผสมผสานเทคนิคสุ่มเพิ่มจำนวนตัวอย่างในคลาสน้อย (Oversampling) และการสุ่มลดจำนวนตัวอย่างในคลาสมาก (Undersampling) พร้อมกันเพื่อจัดการกับปัญหาความไม่สมดุลของคลาสในข้อมูล โดยมีเหตุผลดังนี้

1) การสุ่มเพิ่มจำนวนตัวอย่างในคลาสน้อย จะช่วยเพิ่มจำนวนของตัวอย่างที่อยู่ในคลาสน้อยให้มีจำนวนมากขึ้น อาจช่วยให้แบบจำลองสามารถเรียนรู้ และจำแนกคลาสน้อยได้ดีขึ้น เพราะมีข้อมูลในการเรียนรู้เพิ่มมากขึ้น

2) การสุ่มลดจำนวนตัวอย่างในคลาสมากพร้อมกัน อาจจะช่วยลดจำนวนของตัวอย่างที่มีจำนวนมากเกินไปเพื่อทำให้มีน้ำหนักที่เท่าเทียมกันมากขึ้น ดังนั้นการผสมผสานเทคนิคการสุ่มเพิ่มจำนวนตัวอย่างในคลาสน้อยกับการลดจำนวนตัวอย่างในคลาสมากพร้อมกัน อาจช่วยลดการกระจายของข้อมูลและช่วยให้แบบจำลองสามารถเรียนรู้ได้ดียิ่งขึ้น และให้น้ำหนักในการจำแนกแต่ละคลาสเท่ากันมากขึ้น

หลังจากทำขั้นตอนทำความสะอาดข้อมูลพบว่า จำนวนทั้งสองตัวแปรเป้าหมายหรือคลาสของข้อมูลชุดฝึกสอนมีจำนวนต่างกันเป็นอย่างมากโดยคลาสที่เป็น Positive Class หรือ คลาสบวกคือ ‘DEPOSIT’ มีเพียง 3,907 ข้อมูล และ Negative Class หรือ คลาสลบ ‘NON_DEPOSIT’ มี 248,023 ข้อมูล สัดส่วนคลาสบวกประมาณ 1.5 เปอร์เซ็นต์ของชุดข้อมูลทั้งหมด และสัดส่วนของคลาสลบประมาณ 98.5 เปอร์เซ็นต์ของชุดข้อมูล หากนำข้อมูลเหล่านี้ไปฝึกสอนแบบจำลองจะทำให้แบบจำลองเกิดความเอนเอียงในการพยากรณ์ผล ยกตัวอย่าง เช่น แบบจำลองจะจดจำข้อมูลคลาสลบที่เป็นกลุ่มคนที่ไม่เปิดบัญชีมากกว่า เพราะมีข้อมูลที่ใช้ฝึกสอนมากกว่า ทำให้การนำแบบจำลอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ไปใช้งานจริง จะพยากรณ์ข้อมูลส่วนใหญ่เป็นกลุ่มคนที่ไม่เปิดบัญชีมากกว่าคนที่เปิดบัญชี ดังนั้นเพื่อเพิ่มขนาดของข้อมูลและปรับสมดุลของข้อมูลเพื่อให้แบบจำลองสามารถจำแนกข้อมูลที่ได้ขึ้นจึงจัดการความไม่สมดุลของข้อมูลด้วยวิธี SMOTE (Synthetic Minority Oversampling Technique) และ ENN (Edited Nearest Neighbors) โดยเป็นเทคนิคที่ใช้ในการปรับปรุงการสมดุลของคลาสในชุดข้อมูล โดยมีวิธีการดังต่อไปนี้

SMOTE จะสร้างตัวอย่างของคลาสน้อย (Minority Class) ใหม่ในที่นี้คือคลาส ‘DEPOSIT’ โดยใช้ข้อมูลของคลาสน้อยที่มีอยู่แล้วและสร้างตัวอย่างเสมือน โดยการสร้างข้อมูลเพิ่มขึ้นอย่างสุ่มในบริเวณของข้อมูลน้อยซึ่งทำให้คลาสน้อยมีจำนวนมากขึ้น

ENN จะลบตัวอย่างของคลาสมาก (Majority class) หรือคลาส ‘NON_DEPOSIT’ ที่อยู่ใกล้เคียงกับข้อมูลของคลาสน้อย โดยตรวจสอบว่าตัวอย่างของคลาสมากที่อยู่ใกล้เคียงกับคลาสน้อยนั้นอาจทำให้เกิดความสับสนในในการสร้างแบบจำลองและจะถูกลบทิ้งออกไป

ตารางที่ 3.13 เปรียบเทียบตัวแปรเป้าหมายหลังจากจัดการความไม่สมดุลโดยวิธี SMOTE และ ENN

ตัวแปรเป้าหมาย	ข้อมูลดั้งเดิม	ข้อมูลใหม่
DEPOSIT	3,907	17,076
NON_DEPOSIT	248,023	228,854

จากตารางที่ 3.13 แสดงการเปรียบเทียบตัวแปรเป้าหมายล่าสุดหลังกำจัดค่านอกเกณฑ์กับตัวแปรเป้าหมายใหม่หลังจากทำขั้นตอนการผสมผสานการสุ่มเกินกับการสุ่มลด โดยวิธี SMOTE (Synthetic Minority Oversampling Technique) และ ENN (Edited Nearest Neighbors) จากการใช้คำสั่ง SMOTEEN () ในภาษา Python ที่ปรับค่าพารามิเตอร์ Sampling_strategy = 0.1/1 หมายถึงการเพิ่มจำนวนตัวอย่างในคลาสน้อยขึ้นมาในอัตราส่วน 0.1 เท่าของข้อมูลทั้งหมดและลดจำนวนตัวอย่างในคลาสมากอัตราส่วน 1 เท่าของข้อมูลชนกลุ่มน้อยทั้งหมด Kubalik and Sramek (2019) กล่าวไว้ว่าการเพิ่มตัวอย่างที่มากเกินไปหรือลดตัวอย่างที่มากเกินไป ทำให้แบบจำลองพยากรณ์ข้อมูลอย่างเอนเอียง และพยากรณ์ข้อมูลได้เพียงบางลักษณะเท่านั้น โดยการเพิ่มตัวอย่างและลดตัวอย่างสำหรับชุดข้อมูลนี้เป็นเสมือนการค้นคว้าหาข้อมูล จากแหล่งข้อมูลเพิ่มเติมเพื่อให้ได้ข้อมูลที่มีความสมดุลกันมากขึ้น เพียงแต่ใช้เทคนิคการสร้างข้อมูลเสมือนขึ้นมาจากข้อมูลต้นฉบับ ด้วย SMOTE โดยเทคนิคนี้จะสร้างข้อมูลขึ้นมาระหว่างจุดที่สนใจกับข้อมูลที่อยู่ใกล้เคียงที่สุด เพื่อให้ได้ข้อมูลที่เป็นข้อมูลใหม่ไม่เหมือนกับข้อมูลต้นฉบับด้วยค่าทางสถิติต่าง ๆ ทำให้แบบจำลอง มีข้อมูลเพิ่มมากขึ้น เปรียบเสมือนการค้นคว้าข้อมูลเพิ่มจากแหล่งข้อมูลอื่น ๆ เพียงแต่สร้างข้อมูลจำลองขึ้นมาด้วยเทคนิคการสุ่มเพิ่ม โดยเทคนิคนี้ใช้รวมกับการสุ่มลดข้อมูลที่มีค่าใกล้เคียงกันระหว่างชนกลุ่มน้อย (Minority) กับชนกลุ่มมาก (Majority) ด้วย ENN ซึ่งเป็นการปรับสมดุล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลไม่ให้มีข้อมูลที่ซ้ำซ้อนมากเกินไป และเพื่อป้องกันปัญหาการรั่วไหลของข้อมูล (Data Leakage) เพราะเกิดจากมีข้อมูลที่เพิ่มขึ้นมากเกินไปจนทำให้ข้อมูลรั่วไหลไปยังข้อมูลชุดทดสอบ งานวิจัยนี้จึงปรับพารามิเตอร์การสร้างตัวอย่างใหม่เพียง 0.1/1 หมายถึงจะสร้างตัวอย่างใหม่ให้กลุ่มข้อมูลน้อยมีจำนวนเพียง 10 เปอร์เซ็นต์ ของจำนวนตัวอย่างในกลุ่มข้อมูลใหญ่เพื่อให้ได้ชุดข้อมูลที่สมดุลกันมากขึ้น

จากผลที่เกิดขึ้นหลังจากการทำการผสมผสานการสุ่มเกินกับการสุ่มลดโดยวิธี SMOTE + ENN ทำให้ข้อมูลใหม่ของตัวแปรเป้าหมาย DEPOSIT มีจำนวนเพิ่มขึ้นจาก 3,907 เป็น 17,076 คิดเป็นสัดส่วนคือ 0.069 หรือประมาณ 6.9 เปอร์เซ็นต์ ของข้อมูลทั้งหมดและข้อมูลใหม่ของตัวแปรเป้าหมาย NON_DEPOSIT มีจำนวนลดลงจาก 248,023 เป็น 228,854 ซึ่งคิดเป็นสัดส่วนเท่ากับ 0.930 หรือประมาณ 93.0 เปอร์เซ็นต์ของข้อมูลทั้งหมดทำให้คลาสน้อยและคลาสมากมีการสมดุลกันมากขึ้นนำไปสู่การสร้างโมเดลที่มีความสามารถในการพยากรณ์ทั้งคลาสบวกและคลาสลบได้ดีมากขึ้น

3.6 การสร้างแบบจำลอง (Modeling)

ขั้นตอนที่ 4 จะเป็นการสร้างแบบจำลองเพื่อนำมาใช้ในการพยากรณ์กลุ่มลูกค้าที่มีโอกาสเปิดบัญชีเงินฝากประเภท D กับธนาคาร จะใช้ไลบรารี Scikit-learn สำหรับการเรียนรู้ของเครื่องแบบจัดหมวดหมู่ 3 วิธีคือ วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) วิธีนาอีฟเบย์ (Naïve Bayes) และวิธีป่าไม้สุ่ม (Random Forest) โดยการเลือกใช้แบบจำลองทั้ง 3 เป็นที่นิยมในการทำงานด้านการเรียนรู้ของเครื่องและวิทยาการข้อมูลเนื่องจากคุณสมบัติและประสิทธิภาพที่แตกต่างกันตามลักษณะของข้อมูล ยกตัวอย่างเช่น

1) วิธีเพื่อนบ้านใกล้สุด k ตัว มีความสามารถในการจัดกลุ่มหรือพยากรณ์ผลลัพธ์ที่มีการกระจายซับซ้อนได้ดีในบางกรณี โดยเฉพาะเมื่อข้อมูลมีคุณสมบัติที่คล้ายคลึงกันมาก

2) วิธีนาอีฟเบย์ มีความสามารถในการทำงานรวดเร็วและใช้งานง่าย ใช้ในงานที่ต้องการการจำแนกแบบจำลองที่เรียนรู้รวดเร็ว

3) วิธีป่าไม้สุ่ม มีความสามารถในการจัดกลุ่มหรือพยากรณ์ผลลัพธ์ที่มีความซับซ้อนและคุณลักษณะหลายตัวแปรได้ดีและมักจะมีประสิทธิภาพดีในการจัดกลุ่มข้อมูล

ก่อนการสร้างแบบจำลองได้มีการแบ่งข้อมูลออกเป็น 2 ส่วนคือ ชุดข้อมูลฝึกสอน 80 เปอร์เซ็นต์เพื่อนำฝึกสอนแบบจำลอง และชุดข้อมูลทดสอบ 20 เปอร์เซ็นต์เพื่อวัดประสิทธิภาพแบบจำลอง โดยแต่ละแบบจำลองจะปรับพารามิเตอร์จากการลองผิดลองถูกเพื่อหาพารามิเตอร์ที่ดีที่สุดสำหรับชุดข้อมูลนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6.1 วิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors)

วิธีเพื่อนบ้านใกล้สุดเป็นวิธีที่ง่ายต่อการทำความเข้าใจแบบจำลอง และง่ายต่อการพยากรณ์คลาสของข้อมูลใหม่ เนื่องจากใช้ข้อมูลตัวอย่างที่ใกล้ที่สุด k ตัวอย่างเพื่อพยากรณ์คลาสของข้อมูลใหม่ สามารถใช้กับข้อมูลที่มีคุณสมบัติหลายตัวแปร โดยไม่ต้องคำนวณค่าพารามิเตอร์หรือปรับแต่งโมเดลเพิ่มเติมมากมายนัก

ตารางที่ 3.14 ไฮเปอร์พารามิเตอร์ของวิธีเพื่อนบ้านใกล้สุด k ตัว

ไฮเปอร์พารามิเตอร์	กำหนดค่า
n_neighbors	125
weights	uniform
metric	euclidean

จากตารางที่ 3.14 แสดงการกำหนดไฮเปอร์พารามิเตอร์ในการสร้างแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว (K-Nearest Neighbors) นั้น ผู้จัดทำได้กำหนดพารามิเตอร์ จำนวนของเพื่อนบ้าน (Neighbors) เท่ากับ 125 ซึ่งสอดคล้องกับจำนวนข้อมูลที่มีจำนวนมาก ถ้าปรับจำนวนของเพื่อนบ้านน้อยเกินไปจะทำให้แบบจำลองเรียนรู้ข้อมูลได้น้อยและไม่เหมาะสมกับข้อมูลขนาดใหญ่ การเลือกค่า k ที่เหมาะสมตามลักษณะของข้อมูล ค่า k ควรเป็นจำนวนเต็มบวกที่ไม่เกินจำนวนตัวอย่างทั้งหมดในชุดข้อมูล ส่วนใหญ่ใช้ค่า k ที่เป็นเลขคี่เพื่อหลีกเลี่ยงกรณีที่มีผลลัพธ์ที่ไม่ชัดเจนในกรณีที่มีเพื่อนบ้านเสมอกัน

น้ำหนักระยะห่างระหว่างจุด (Weights) การปรับ น้ำหนักระยะห่างระหว่างจุด (weights) เป็น ยูนิฟอร์ม (Uniform) คือให้น้ำหนักเท่ากันสำหรับทุก ๆ ตัวอย่างที่อยู่ใกล้เคียงกับจุดที่เราต้องการพยากรณ์ กล่าวคือจะไม่พยายามให้ให้ความสำคัญกับตัวอย่างที่อยู่ใกล้เคียงกับจุดที่เราต้องการพยากรณ์มากกว่าตัวอย่างที่อยู่ไกล ซึ่งอาจจะเหมาะสำหรับข้อมูลที่มีลักษณะใกล้เคียงกันและมีการกระจายที่ใกล้กัน

คำนวณระยะห่างระหว่างจุด (Metric) การปรับ คำนวณระยะห่างระหว่างจุด (metric) เป็น ยูคลิเดียน (Euclidean) เนื่องจากเป็นวิธีการคำนวณระยะห่างที่นิยมใช้งานมากในการแบ่งกลุ่มข้อมูล โดยเฉพาะงานที่มีความคล้ายคลึงระหว่างข้อมูลมาก เช่น ลักษณะของกลุ่มคนที่เปิดบัญชีและไม่เปิดบัญชีมีความคล้ายคลึงกันอย่างมาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6.2 วิธีนาอีฟเบย์ (Naïve Bayes)

เทคนิคการจำแนกด้วยนาอีฟเบย์เป็นการใช้หลักการของทฤษฎีการคำนวณความน่าจะเป็น (Probability) ในการพยากรณ์คลาสของข้อมูล โดยพิจารณาความสัมพันธ์ระหว่างคลาสและตัวแปรคุณลักษณะ (Feature) ของข้อมูล วิธีนาอีฟเบย์มีความเร็วในการทำงานสูง เนื่องจากไม่ต้องคำนวณค่าพารามิเตอร์หรือปรับแต่งแบบจำลองเพิ่มเติมมากมาย ทำให้สามารถนำไปใช้กับข้อมูลขนาดใหญ่ได้โดยไม่เสียเวลาในกระบวนการสร้างแบบจำลอง

ตารางที่ 3.15 ไฮเปอร์พารามิเตอร์ของวิธีนาอีฟเบย์

ไฮเปอร์พารามิเตอร์	กำหนดค่า
var_smoothing	1e-9
priors	[0.2, 0.8]

จากตารางที่ 3.15 ผู้จัดทำได้ทำการปรับไฮเปอร์พารามิเตอร์ การปรับค่าสมดุลจากข้อมูลเดิม (var_smoothing) เป็นพารามิเตอร์ที่ใช้ในการปรับค่าความสามารถในการประมาณค่าความน่าจะเป็น (likelihood) สำหรับแต่ละคลาสเท่ากับ 1e-9 หรือ 0.000000001 ซึ่งคือค่าเริ่มต้นของพารามิเตอร์ หากปรับ var_smoothing มากขึ้นจะทำให้โมเดลของเรามีความสมบูรณ์แบบมากขึ้น แต่อาจทำให้โมเดลเรียนรู้ข้อมูลฝึกสอนเกินไป ซึ่งอาจทำให้โมเดลไม่สามารถพยากรณ์ข้อมูลที่มีความซับซ้อนได้ดีพอ กลับกันเมื่อค่า var_smoothing น้อยลง โมเดลจะสามารถเรียนรู้ข้อมูลฝึกสอนได้ดีกว่า แต่อาจเกิดการ Overfitting ได้ง่ายขึ้น ดังนั้นจึงปรับเป็นค่าเริ่มต้นของแบบจำลอง

ค่าล่วงหน้าของความน่าจะเป็น (Priors) เป็นพารามิเตอร์ที่ใช้กำหนดค่าความน่าจะเป็น (Prior) ของแต่ละคลาส โดยค่าเริ่มต้นของพารามิเตอร์นี้จะเป็น None โดยทั่วไปแล้วการปรับ “Priors” แต่ละงานวิจัยไม่มีกำหนดตายตัวขึ้นอยู่กับข้อมูลแต่ละงานวิจัย หากเราปรับ “Priors” ให้มีค่าไม่เท่ากันจะทำให้แบบจำลองนาอีฟเบย์ให้น้ำหนักกับคลาสที่มี Prior Probability มากกว่าคลาสอีกฝ่าย ซึ่งผู้วิจัยได้มองว่าข้อมูลชุดนี้ไม่สมดุลกันจึงปรับเป็น 0.2 ต่อ 0.8 จะเป็นการให้โมเดลให้น้ำหนักกับคลาส ‘DEPOSIT’ มากกว่าคลาส ‘NON_DEPOSIT’ ทำให้โมเดลจะมีแนวโน้มที่จะจำแนกตัวอย่างให้เป็นคลาส ‘DEPOSIT’ มากกว่า

3.6.3 วิธีป่าไม้สุ่ม (Random Forest)

เทคนิคป่าไม้สุ่มเป็นวิธีในการจำแนกหมวดหมู่ที่มีความแม่นยำสูง โดยใช้แนวคิดของการสร้างต้นไม้ (Decision Tree) หลาย ๆ ต้นแล้วนำมาประกอบกันเป็นป่า (Forest) รวมผลลัพธ์ของทุกต้นแบบเพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำ ซึ่งสามารถให้ความแม่นยำในการพยากรณ์สูง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.16 ไฮเปอร์พารามิเตอร์ของวิธีป่าไม้สุ่ม

ไฮเปอร์พารามิเตอร์	กำหนดค่า
n_estimators	1200
max_depth	30
min_samples_split	5
min_samples_leaf	3
class_weight	Balanced

จากตารางที่ 3.16 แสดงการกำหนดไฮเปอร์พารามิเตอร์ของวิธีป่าไม้สุ่ม ซึ่งผู้จัดทำได้กำหนด จำนวนของต้นไม้ (n_estimators) เท่ากับ 1200 หมายความว่าโมเดลจะใช้ 1200 ต้นไม้ในการสร้างและพยากรณ์ข้อมูล ซึ่งสามารถช่วยเพิ่มความแม่นยำของโมเดลได้แต่อาจทำให้การสร้างโมเดลใช้เวลานานขึ้น ซึ่งการปรับจำนวนของต้นไม้มากๆเหมาะสำหรับข้อมูลที่มีขนาดใหญ่และข้อมูลที่มีความซับซ้อนสูง ๆ

ความลึกของต้นไม้ (max_depth) ในงานนี้ผู้วิจัยได้กำหนดความลึกของต้นไม้เท่ากับ 30 ซึ่งค่าเริ่มต้นของ “max_depth” คือ None ซึ่งหมายความว่าโมเดลจะไม่มีการจำกัดความลึกของต้นไม้ ถ้าไม่ระบุ “max_depth” จะทำให้โมเดลเกิดความซับซ้อนและเสี่ยงต่อการ Overfitting ดังนั้นงานวิจัยต่าง ๆ จึงไม่ระบุพารามิเตอร์ที่ดีที่สุด แต่ควรมีการปรับพารามิเตอร์ให้สอดคล้องกับความซับซ้อนของข้อมูลและคุณลักษณะ (Feature) ที่ฝึกสอนแบบจำลอง

จำนวนตัวอย่างขั้นต่ำสำหรับการแยกสองโหนด (min_samples_split) พารามิเตอร์ที่ใช้กำหนดจำนวนตัวอย่างขั้นต่ำที่จะต้องมีในการแบ่งกิจกรรมสร้างต้นไม้ (Decision Tree) ซึ่งจะแบ่งกิจกรรมต่อไปเป็นโหนดย่อย ๆ โดยจะหยุดแบ่งกิจกรรมเมื่อจำนวนตัวอย่างที่เหลือน้อยกว่าค่าที่กำหนด ซึ่งค่าเริ่มต้นจะเท่ากับ 2 การปรับค่า “min_samples_split” สูงจะทำให้โมเดลมีความเข้ากันได้น้อยขึ้นและมีโอกาสเกิด Overfitting ในข้อมูลชุดฝึกสอนในขณะที่การปรับค่าต่ำจะทำให้โมเดลมีความยืดหยุ่นมากขึ้นแต่อาจทำให้เกิด Underfitting ในข้อมูลชุดฝึกสอน ดังนั้นผู้จัดทำจึงกำหนดเท่ากับ 5 ซึ่งมีค่าที่ไม่มากเกินไปและไม่น้อยเกินไป

จำนวนตัวอย่างขั้นต่ำในใบไม้ (min_samples_leaf) เป็นพารามิเตอร์ที่ใช้กำหนดจำนวนตัวอย่างขั้นต่ำที่ต้องมีในใบไม้ทุกใบ หากจำนวนตัวอย่างในใบน้อยกว่า “min_samples_leaf” จะไม่ทำการแยกกิ่งต่อ ซึ่งค่าเริ่มต้นอยู่ที่ 1 แต่ผู้วิจัยได้กำหนดค่าเป็น 3 หมายความว่าจะต้องมีจำนวนตัวอย่างในแต่ละใบ (Leaf) ไม่น้อยกว่า 3 ตัวอย่าง ถ้ามีตัวอย่างน้อยกว่า 3 ตัวอย่างในใบนั้น จะไม่ทำการแยกกิ่งต่อไป

น้ำหนักของคลาส (class_weight) คือพารามิเตอร์ที่ใช้กำหนดน้ำหนักให้กับแต่ละคลาสในการสร้างแบบจำลอง ซึ่งเป็นวิธีหนึ่งที่ใช้สำหรับการจัดการกับปัญหาการไม่สมดุลของคลาสใน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การจัดแบ่งชุดข้อมูล เช่น ในกรณีที่มีจำนวนตัวอย่างของคลาสหนึ่งมากกว่าอีกคลาส การไม่ปรับการจัดการกับคลาสที่ไม่สมดุลอาจทำให้โมเดลที่สร้างขึ้นมีความแม่นยำลดลง และเพิ่มโอกาสในการพยากรณ์ผิดกับคลาสน้อย ดังนั้นผู้จัดทำได้ปรับพารามิเตอร์เท่ากับ ‘balanced’ หมายถึงการกำหนดน้ำหนักให้กับแต่ละคลาสให้สมดุลกัน ซึ่งช่วยเพิ่มประสิทธิภาพในการจำแนกคลาสที่มีจำนวนตัวอย่างไม่สมดุลกัน

3.7 การวัดประสิทธิภาพ (Evaluation)

ขั้นตอนที่ 5 การวัดประสิทธิภาพของแบบจำลองมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง โดยเลือกแบบจำลองที่เหมาะสมสำหรับข้อมูลชุดนี้มากที่สุด และสามารถพยากรณ์กลุ่มลูกค้าที่มีแนวโน้มที่จะเปิดบัญชีได้แม่นยำ ซึ่งขั้นตอนนี้จะแยกกลุ่มข้อมูลฝึกสอน 80 เปอร์เซ็นต์ และกลุ่มข้อมูลทดสอบ 20 เปอร์เซ็นต์ เพื่อนำชุดข้อมูลทดสอบมาทำการประเมินผลและวัดประสิทธิภาพจากการพิจารณาจาก 3 เทคนิคดังนี้

1) เมทริกซ์ความสับสน (Confusion Matrix) พิจารณาจาก 1. ค่าความแม่น (Accuracy) ใช้วัดประสิทธิภาพทั้งหมดของโมเดลในการพยากรณ์ที่ถูกต้องเทียบกับจำนวนทั้งหมดของตัวอย่าง 2. ค่าความเที่ยง (Precision) ใช้วัดประสิทธิภาพเน้นไปที่ความแม่นยำในการพยากรณ์ให้ถูกต้องในคลาสที่เป็นบวก (Positive Class) หรือ กลุ่มคนที่เปิดบัญชี ‘DEPOSIT’ ว่ามีความสามารถในการจำแนกคลาสบวกได้แม่นยำ 3. ค่าระลึก (Recall) ใช้วัดประสิทธิภาพที่เน้นไปที่ความสามารถในการตรวจจับตัวอย่างในคลาสที่เป็นบวก (Positive Class) ว่ามีความสามารถในการจำแนกตัวอย่างที่เป็นบวกทั้งหมดในข้อมูล และ 4. ค่าประสิทธิภาพโดยรวม (F-Measure) ใช้เพื่อการประเมินความสมดุลระหว่างความแม่นยำและความเรียกคืน โดยคำนวณค่าเฉลี่ยเฉพาะของค่าความแม่นและค่าระลึกเพื่อให้ได้ค่าที่นำมาวัดประสิทธิภาพของโมเดลที่มีการพยากรณ์ความสามารถทั้งสองด้านได้อย่างครอบคลุม

2) วิธีการตรวจสอบไขว้แบบแบ่งชั้น (Stratified K-fold Cross Validation) ช่วยในการประเมินประสิทธิภาพของแบบจำลองที่มีปัญหาคลาสไม่สมดุลได้อย่างถูกต้องโดยจะทำการแบ่งกลุ่มทดสอบ หรือ ‘Validation Data’ ให้มีคลาสหรือตัวแปรเป้าหมายที่สมดุลกันซึ่งเหมาะสำหรับกลุ่มข้อมูลที่มีคลาสไม่สมดุลกัน

3) กราฟ ROC Curve (Receiver Operating Characteristic Curve) เป็นกราฟที่ใช้ในการประเมินประสิทธิภาพของแบบจำลองในการจำแนกข้อมูล สามารถใช้ ROC Curve เพื่อเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองที่ต่างกันได้ กราฟ ROC Curve ที่ดี คือกราฟที่มีพื้นที่ใต้เส้นโค้ง (Area under the Curve, AUC) มากที่สุดเท่าที่เป็นไปได้ ซึ่งค่า AUC มีค่าตั้งแต่ 0 ถึง 1 ขึ้นอยู่กับเกณฑ์ที่กำหนด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.7.1 การวัดประสิทธิภาพวิธีเพื่อนบ้านใกล้สุด k ตัว

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว เสร็จสิ้น นำข้อมูลชุดทดสอบมาประเมินประสิทธิภาพ ค่าความแม่น ค่าความเที่ยง ค่าระลึก และค่าประสิทธิภาพโดยรวม การวิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบแบ่งชั้น และกราฟเส้นโค้ง ROC Curve

ตารางที่ 3.17 ประสิทธิภาพการพยากรณ์ของแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว

การเรียนรู้ของเครื่อง	ค่าความแม่น (Accuracy)	ค่าความเที่ยง (Precision)	ค่าระลึก (Recall)	ค่าประสิทธิภาพโดยรวม (F-Measure)
วิธีเพื่อนบ้านใกล้สุด k ตัว	94.68%	75.33%	33.27%	46.16%

จากตารางที่ 3.17 แสดงเปอร์เซ็นต์ประสิทธิภาพการพยากรณ์ของแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัว พบว่าประสิทธิภาพการพยากรณ์ของแบบจำลองมีค่าความแม่น (Accuracy) เท่ากับ 94.68% ค่าความเที่ยง (Precision) เท่ากับ 75.33% ค่าระลึก (Recall) เท่ากับ 33.27% และค่าประสิทธิภาพโดยรวม (F-Measure) เท่ากับ 46.16% ซึ่งตีความหมายได้ว่า แบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัวมีการพยากรณ์คลาสบวก (Positive Class) หรือในที่นี้คือคลาส ‘DEPOSIT’ ค่อนข้างน้อยครั้งจึงทำให้มีค่าความเที่ยง (Precision) สูงในขณะที่ ค่าระลึก (Recall) น้อยหมายถึงแบบจำลองพยากรณ์ คลาสบวกน้อย (Positive Class) ถูกน้อยเมื่อเทียบจากคลาสบวกทั้งหมดของข้อมูล

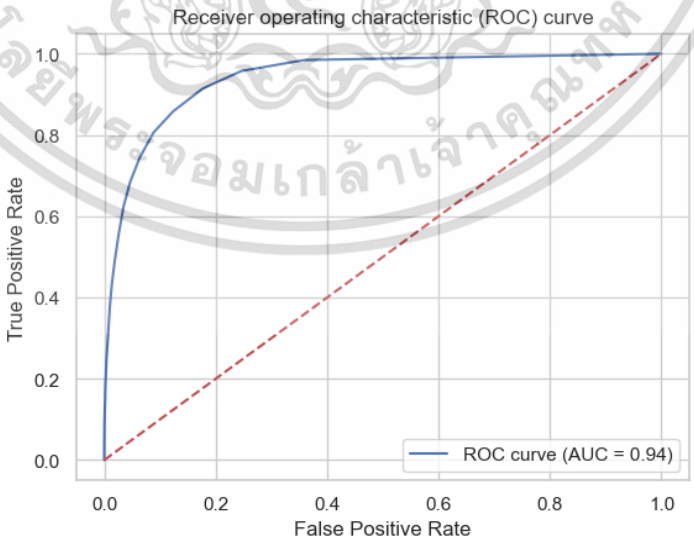
จากนั้นทำการตรวจสอบความแม่นตรงของแบบจำลองพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบแบ่งชั้นของวิธีเพื่อนบ้านใกล้สุด k ตัว เพื่อตรวจสอบว่าแบบจำลองพยากรณ์ผลได้อย่างแม่นยำและสามารถพยากรณ์ข้อมูลที่มีความหลากหลายได้ดีหรือไม่ ดังผลลัพธ์แสดงดังตารางที่ 3.18

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.18 การตรวจสอบไขว้แบบแบ่งชั้นของวิธีเพื่อนบ้านใกล้สุด k ตัว

Fold	**ค่าความแม่น (Accuracy)**
1	0.9453
2	0.9465
3	0.9441
4	0.9454
5	0.9464
6	0.9439
7	0.9438
8	0.9439
9	0.9443
10	0.9454

จากตารางที่ 3.18 แสดงตารางการตรวจสอบไขว้แบบแบ่งชั้นของวิธีเพื่อนบ้านใกล้สุด k ตัว ทั้งหมด 10 Fold จะพบว่า แต่ละ Fold อยู่ในช่วง 0.9438 ถึง 0.9454 ซึ่งเป็นค่าที่สูงและใกล้เคียงกัน บ่งบอกว่าแบบจำลองวิธีเพื่อนบ้านใกล้สุด k ตัวมีความสามารถในการพยากรณ์ข้อมูลอย่างมีประสิทธิภาพและมีความเสถียรในการทำงาน โดยมีค่าค่าเฉลี่ยอยู่ที่ 0.9449 หรือประมาณ 94.49 เปอร์เซ็นต์



รูปที่ 3.13 กราฟเส้นโค้ง ROC Curve ของวิธีเพื่อนบ้านใกล้สุด k ตัว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.13 แสดงกราฟเส้นโค้ง ROC Curve ของวิธีเพื่อนบ้านใกล้สุด k ตัว โดยจะเห็นว่า กราฟมีลักษณะโค้งไล่เรียงจากซ้ายล่างไปขวาบน โดยจุดที่อยู่บนสุดของกราฟคือจุดที่มี ความไวในการเจาะจง (Sensitivity) และความสามารถในการระบุ (Specificity) สูง และจากตารางเกณฑ์การตัดสินใจทั่วไปของค่าประมาณพื้นที่ใต้โค้ง ROC หรือ AUC เท่ากับ 0.94 ซึ่งมีเกณฑ์ที่ดีมากในการจำแนก

3.7.2 การวัดประสิทธิภาพวิธีนาอีฟเบย์

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีนาอีฟเบย์เสร็จสิ้น นำข้อมูลชุดทดสอบมา ทำการประเมินประสิทธิภาพ ค่าความแม่น ค่าความเที่ยง ค่าระลึก และค่าประสิทธิภาพโดยรวม การวิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบแบ่งชั้น และกราฟเส้นโค้ง ROC Curve

ตารางที่ 3.19 ประสิทธิภาพการพยากรณ์ของแบบจำลองวิธีนาอีฟเบย์

การเรียนรู้ของเครื่อง	ค่าความแม่น (Accuracy)	ค่าความเที่ยง (Precision)	ค่าระลึก (Recall)	ค่าประสิทธิภาพโดยรวม (F-Measure)
วิธีนาอีฟเบย์	91.13%	28.34%	19.26%	22.93%

จากตารางที่ 3.19 แสดงเปอร์เซ็นต์ประสิทธิภาพการพยากรณ์ของแบบจำลองวิธีนาอีฟเบย์ พบว่าประสิทธิภาพการพยากรณ์ของแบบจำลองมีค่าความแม่น เท่ากับ 91.13 เปอร์เซ็นต์ ค่าความเที่ยง เท่ากับ 28.34 เปอร์เซ็นต์ ค่าระลึก เท่ากับ 19.26 เปอร์เซ็นต์ และค่าประสิทธิภาพโดยรวม (F-Measure) เท่ากับ 22.93% พบว่า ค่าความเที่ยง และค่าระลึก มีค่าน้อยมากหมายความว่า แบบจำลองของนาอีฟเบย์นั้นมีความสามารถพยากรณ์ข้อมูลที่เป็นคลาสบวกค่อนข้างต่ำ แต่จากค่าความแม่นมีค่าที่สูงอาจจะเป็นเพราะว่า แบบจำลองนาอีฟเบย์พยากรณ์ข้อมูลที่เป็นคลาสลบได้ดี อันเนื่องมากจากมีความไม่สมดุลข้อมูลที่มีจำนวนคลาสลบเยอะกว่าคลาสบวก ทำให้แบบจำลองมีการเรียนรู้คลาสลบ หรือ คลาส ‘NON_DEPOSIT’ ได้ดี

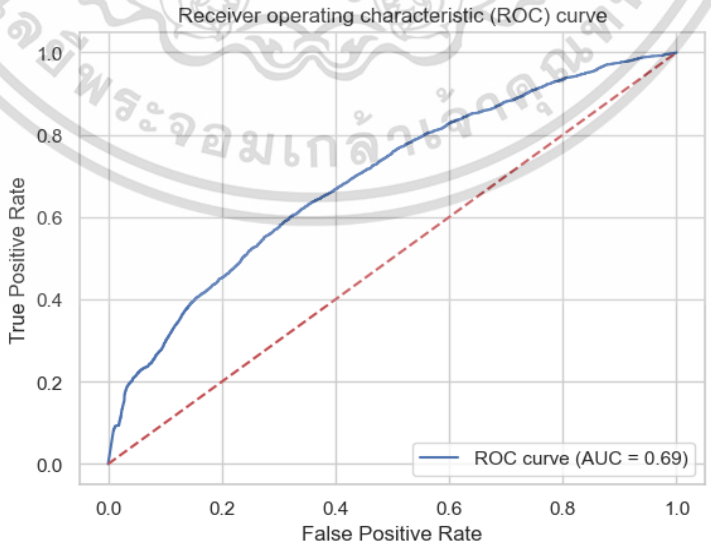
จากนั้นทำการตรวจสอบความแม่นตรงของแบบจำลองพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบแบ่งชั้นของวิธีนาอีฟเบย์ เพื่อตรวจสอบว่าแบบจำลองพยากรณ์ผลได้อย่างแม่นยำและสามารถพยากรณ์ข้อมูลที่มีความหลากหลายได้ดีหรือไม่ ดังผลลัพธ์แสดงดังตารางที่ 3.20

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.20 การตรวจสอบไขว้แบบแบ่งชั้นของวิธีนาอีฟเบย์

Fold	ค่าความแม่น (Accuracy)
1	0.9113
2	0.9123
3	0.9102
4	0.9119
5	0.9097
6	0.9143
7	0.9112
8	0.9120
9	0.9112
10	0.9104

จากตารางที่ 3.20 แสดงตารางการตรวจสอบไขว้แบบแบ่งชั้นของวิธีนาอีฟเบย์ ทั้ง 10 Fold จะพบว่า แต่ละ Fold อยู่ในช่วง 0.9097 ถึง 0.9143 ซึ่งเป็นค่าที่สูงและใกล้เคียงกัน ซึ่งบ่งบอกว่าแบบจำลองวิธีนาอีฟเบย์ ตัวโมเดลที่ใช้มีความสามารถในการพยากรณ์ข้อมูลอย่างมีประสิทธิภาพและมีความเสถียรในการทำงาน แต่ยังมีค่าความแม่นน้อยกว่าวิธีของเพื่อนบ้านใกล้สุด k โดยค่าเฉลี่ยอยู่ที่ 0.9114 หรือประมาณ 91.14 เปอร์เซ็นต์



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 3.14 กราฟเส้นโค้ง ROC Curve ของวิธีนาอีฟเบย์

จากรูปที่ 3.14 แสดงกราฟเส้นโค้ง Receiver Operating Characteristic Curve (ROC Curve) ของวิธีนาอีฟเบย์ โดยจะเห็นว่า กราฟมีลักษณะเป็นเส้นตรงหรือเป็นเส้นทแยงมุมใกล้เส้นการแยกระหว่าง True Positive Rate (TPR) และ False Positive Rate (FPR) หมายความว่าทำให้ไม่ค่อยมีความแม่นยำในการพยากรณ์ค่า True Positive Rate (TPR) และ False Positive Rate (FPR) ในระดับที่สูงขึ้น และจากตารางเกณฑ์การตัดสินใจทั่วไปของค่าประมาณพื้นที่ใต้โค้ง ROC หรือ AUC เท่ากับ 0.69 ซึ่งมีเกณฑ์ที่ค่อนข้างต่ำ

3.7.3 การวัดประสิทธิภาพวิธีป่าไม้สุ่ม

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีป่าไม้สุ่มเสร็จสิ้น นำข้อมูลชุดทดสอบมา ทำการประเมินประสิทธิภาพ ค่าความแม่น ค่าความเที่ยง ค่าระลึก และค่าประสิทธิภาพโดยรวม การวิเคราะห์ความแม่นตรงของแบบจำลองพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบแบ่งชั้นและกราฟเส้นโค้ง ROC Curve

ตารางที่ 3.21 ประสิทธิภาพการพยากรณ์ของแบบจำลองวิธีป่าไม้สุ่ม

การเรียนรู้ของเครื่อง	ค่าความแม่น (Accuracy)	ค่าความเที่ยง (Precision)	ค่าระลึก (Recall)	ค่าประสิทธิภาพโดยรวม (F-Measure)
วิธีป่าไม้สุ่ม	97.40%	87.31%	72.72%	79.35%

จากตารางที่ 3.21 แสดงเปอร์เซ็นต์ประสิทธิภาพการพยากรณ์ของแบบจำลองวิธีป่าไม้สุ่ม พบว่า ประสิทธิภาพการพยากรณ์ของแบบจำลองมีค่าความแม่น เท่ากับ 97.40% ค่าความเที่ยง เท่ากับ 87.31% ค่าระลึก เท่ากับ 72.72% และค่าประสิทธิภาพโดยรวม เท่ากับ 79.35% ซึ่งจะเห็นว่า ค่าความเที่ยง กับค่าระลึก ค่อนข้างสูงและมีค่าใกล้เคียงกันนั้นหมายความว่า แบบจำลองวิธีป่าไม้สุ่มนั้นมีประสิทธิภาพในการพยากรณ์คลาสบวกหรือ คลาส ‘DEPOSIT’ ค่อนข้างสูงทำให้ค่าประสิทธิภาพโดยรวมสูงตามไปด้วย เนื่องจากมีการปรับไฮเปอร์พารามิเตอร์ที่สามารถทำให้การให้ความสำคัญของแต่ละคลาสมีน้ำหนักเท่ากัน และวิธีป่าไม้สุ่มเป็นแบบจำลองที่เป็นการเรียนรู้แบบรวมกลุ่ม เป็นการเรียนรู้หลาย ๆ แบบจำลองเพื่อโหวตแบบจำลองที่ดีที่สุดในการพยากรณ์ ทำให้ลดความผิดพลาดในการพยากรณ์จึงทำให้มีประสิทธิภาพในการพยากรณ์สูง

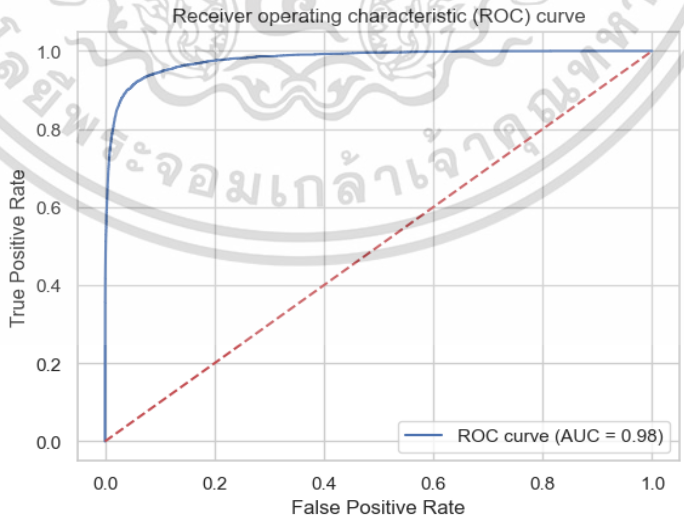
จากนั้นทำการตรวจสอบความแม่นตรงของแบบจำลองพยากรณ์ด้วยวิธีการตรวจสอบไขว้แบบแบ่งชั้นของวิธีป่าไม้สุ่ม เพื่อตรวจสอบว่าแบบจำลองพยากรณ์ผลได้อย่างแม่นยำและสามารถพยากรณ์ข้อมูลที่มีความหลากหลายได้ดีหรือไม่ ดังผลลัพธ์แสดงดังตารางที่ 3.22

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.22 การตรวจสอบไขว้แบบแบ่งชั้นของวิธีป่าไม้สุ่ม

Fold	**ค่าความแม่น (Accuracy)**
1	0.9718
2	0.9727
3	0.9719
4	0.9734
5	0.9726
6	0.9728
7	0.9729
8	0.9743
9	0.9726
10	0.9715

จากตารางที่ 3.22 แสดงตารางการตรวจสอบไขว้แบบแบ่งชั้นของวิธีป่าไม้สุ่ม ทั้ง 10 Fold จะพบว่า แต่ละ Fold อยู่ในช่วง 0.9715 ถึง 0.9743 ซึ่งเป็นค่าที่สูงและใกล้เคียงกัน ซึ่งบ่งบอกว่าแบบจำลองวิธีป่าไม้สุ่ม มีความสามารถในการพยากรณ์ข้อมูลอย่างมีประสิทธิภาพและมีความเสถียรในการทำงาน ค่าเฉลี่ยอยู่ที่ 0.972700571 หรือประมาณ 97.27 เปอร์เซ็นต์



รูปที่ 3.15 กราฟเส้นโค้ง ROC Curve ของวิธีป่าไม้สุ่ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.15 กราฟเส้นโค้ง ROC Curve ของวิธีป่าไม้สุ่ม พบว่า กราฟเส้นโค้ง ROC มีลักษณะเส้นการวาดที่ใกล้เคียงกับเส้นทแยงมุม ซึ่งเส้นทแยงมุมแสดงถึงการพยากรณ์แบบสุ่ม ซึ่งเส้น ROC อยู่ใกล้เคียงกับส่วนบนซ้ายของกราฟ แสดงถึงความแม่นยำที่สูงและสามารถพยากรณ์คลาสบวกได้ดี และจะเห็นได้ว่าพื้นที่ใต้เส้น ROC ที่กว้าง ซึ่งแสดงถึงความแม่นยำและความแม่นยำที่สูงทั้งในการตรวจจับคลาสบวกและคลาสลบ ซึ่งมีค่า AUC (Area Under the Curve) เท่ากับ 0.98 อ้างอิงจากตารางเกณฑ์การตัดสินใจทั่วไปของค่าประมาณพื้นที่ใต้โค้ง ROC หรือ AUC ซึ่งมีเกณฑ์ที่ค่อนข้างสูงมาก

3.7.4 ผลการเปรียบเทียบประสิทธิภาพแบบจำลอง

ผลการวัดประสิทธิภาพการเรียนรู้ของเครื่องทั้ง 3 วิธี วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และ วิธีป่าไม้สุ่ม พบว่ากราฟเส้นโค้ง ROC Curve ของแต่ละวิธีมีเกณฑ์ที่อยู่ในระดับสามารถจำแนกข้อมูลได้ และ การตรวจสอบไขว้แบบแบ่งชั้นแต่ละวิธี ที่สามารถบ่งบอกความแม่นตรงของตัวแบบจำลองได้ในชุดข้อมูลใหม่ ดังนั้นการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองจึงขึ้นอยู่กับค่าในเมทริกซ์ความสับสน ได้แก่

ค่าความแม่น (Accuracy) เป็นการวัดความแม่นยำของแบบจำลองโดยรวม กล่าวคือแบบจำลองพยากรณ์ถูกกี่ครั้งจากจำนวนการพยากรณ์ทั้งหมด

ค่าความเที่ยง (Precision) เป็นค่าที่แบบจำลองพยากรณ์เป็นคลาสที่กำลังพิจารณาและถูกต้องต่อค่าที่แบบจำลองพยากรณ์ว่าเป็นคลาสที่กำลังพิจารณาทั้งถูกและผิด

ค่าระลึก (Recall) เป็นค่าที่แบบจำลองพยากรณ์เป็นคลาสที่กำลังพิจารณาและถูกต้องต่อคลาสที่สนใจทั้งหมด

ค่าประสิทธิภาพโดยรวม (F-Measure) เป็นการวัดความเที่ยงและค่าระลึกของแบบจำลองไปพร้อม ๆ กัน

งานวิจัยนี้ได้จัดทำเพื่อ ต้องการพยากรณ์ลูกค้าที่มีโอกาสเปิดบัญชีให้แม่นยำมากที่สุด เพื่อให้เกิด "Maximum Customer Acquisition" คือ การเน้นการเพิ่มจำนวนลูกค้าที่มีโอกาสเปิดบัญชีมากที่สุด โดยพิจารณาจาก ค่าระลึก (Recall) เพราะเป็นค่าที่วัดสัดส่วนของตัวอย่างที่ถูกต้องที่พยากรณ์เป็นคลาสบวกที่งานวิจัยนี้สนใจ หรือ ลูกค้าที่เปิดบัญชี จากจำนวนทั้งหมดในชุดข้อมูล ดังนั้น ค่าระลึก (Recall) จึงเป็นตัวชี้วัดที่สำคัญ เมื่อต้องการให้การจำแนกคลาสบวก หรือ ลูกค้าที่เปิดบัญชี มีความครอบคลุม และไม่พลาดไปมากเกิดไป ซึ่งจะช่วยให้ธนาคารลูกค้าที่มีโอกาสสูงที่สุดในการเปิดบัญชี เพื่อสร้างรายได้และความมีเสถียรภาพของธนาคารมากที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.23 เปรียบเทียบประสิทธิภาพการพยากรณ์ของแบบจำลองทั้ง 3 วิธี

การเรียนรู้ของเครื่อง	ค่าความแม่น (Accuracy)	ค่าความเที่ยง (Precision)	ค่าระลึก (Recall)	ค่าประสิทธิภาพโดยรวม (F-Measure)
วิธีเพื่อนบ้านใกล้สุด k ตัว	94.68%	75.33%	33.27%	46.16%
วิธีนาอีฟเบย์	91.13%	28.34%	19.26%	22.93%
วิธีป่าไม้สุ่ม	97.40%	87.31%	72.72%	79.35%

จากตารางที่ 3.23 แสดงการเปรียบเทียบเปอร์เซ็นต์ประสิทธิภาพการพยากรณ์ของแบบจำลองทั้ง 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และวิธีป่าไม้สุ่ม พบว่า วิธีป่าไม้สุ่มมีประสิทธิภาพการพยากรณ์สูงที่สุดทุกค่า ได้แก่ มีค่าความแม่น ค่าความเที่ยง ค่าระลึก และค่าประสิทธิภาพโดยรวม มีค่าเท่ากับ 94.68% 75.33% 33.27% และ 46.16% ตามลำดับ

เนื่องจากวิธีเพื่อนบ้านใกล้สุด k ตัว ข้อมูลต้องไม่มีความสัมพันธ์ที่ซับซ้อน ถ้าข้อมูลมีความสัมพันธ์ที่ซับซ้อนหรือไม่มีรูปแบบที่ชัดเจนอาจทำให้แบบจำลองให้ผลลัพธ์ที่ไม่แม่นยำ (สุเมธ และสมพร, 2564) ซึ่งข้อมูลชุดนี้มีความสัมพันธ์ที่ซับซ้อนแต่ละตัวแปรและมีรูปแบบที่ไม่ชัดเจนจึงทำให้ไม่เหมาะกับข้อมูลชุดนี้

เนื่องจากวิธีนาอีฟเบส์มักจะไม่เหมาะสำหรับชุดข้อมูลที่มีความซับซ้อนหรือความสัมพันธ์ที่ซับซ้อนกันมาก เนื่องจากวิธีนาอีฟเบส์สมมติว่าตัวแปรอิสระในข้อมูลเป็นอิสระต่อกัน ดังนั้นหากข้อมูลมีความซับซ้อนมากอาจทำให้สมมติฐานนี้ไม่เป็นไปตามความเป็นจริง ซึ่งอาจทำให้ผลลัพธ์ของแบบจำลองไม่แม่นยำและไม่เสถียรต่อการพยากรณ์ และข้อมูลที่ไม่สมดุลทำให้มีตัวอย่างจำนวนน้อยบางคลาสเป็นอีกเหตุผลที่ทำให้แบบจำลองวิธีนาอีฟเบย์ไม่มีประสิทธิภาพในการพยากรณ์ข้อมูลชุดนี้ (สุรวัชร และสายชล, 2560)

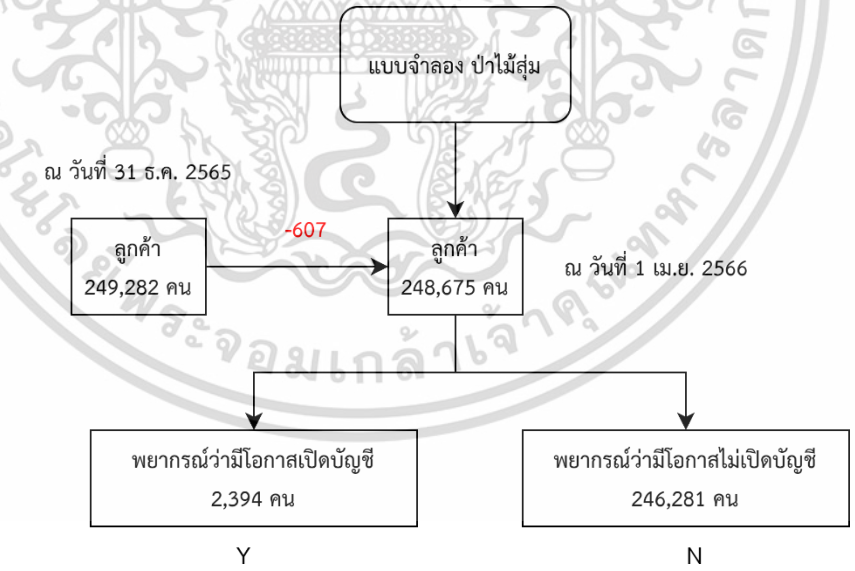
เนื่องจากวิธีป่าไม้สุ่มพบว่า การวิเคราะห์ความแม่นตรงตัวแบบพยากรณ์ ด้วยการตรวจสอบไขว้ที่บ่งบอกความสามารถในการพยากรณ์ข้อมูลใหม่อย่างมีประสิทธิภาพ กราฟเส้นโค้ง ROC Curve ที่มีค่า AUC เท่ากับ 0.98 มีเกณฑ์ที่ค่อนข้างสูงมาก และมีค่าความแม่น ค่าความเที่ยง ค่าระลึก และค่าประสิทธิภาพโดยรวม สูงที่สุด ดังนั้นวิธีป่าไม้สุ่มสามารถเป็นแบบจำลองที่เหมาะสมต่อการพยากรณ์ลูกค้าในการเปิดบัญชีเงินฝากประเภท D

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.8 การนำแบบจำลองที่เหมาะสมไปใช้งาน (Deployment)

หลังจากสร้างแบบจำลองและวัดประสิทธิภาพของแบบจำลองทั้ง 3 วิธีพบว่า วิธีป่าไม้สุ่ม เป็นแบบจำลองที่เหมาะสมที่สุด ในพยากรณ์กลุ่มลูกค้าเปิดบัญชีเงินฝากประเภท D โดยการนำไปใช้งานจะสร้างแคมเปญที่เป็นลักษณะแบ่งกลุ่มลูกค้าตามข้อความเฉพาะบุคคล ซึ่งจะพยากรณ์กับกลุ่มลูกค้าที่ยังไม่เปิดบัญชี ณ วันที่ 1 เม.ย 2566 คือกลุ่มที่เคยถูกนำเข้าแบบจำลองเพื่อฝึกสอนมาก่อน ซึ่งเป็นกลุ่มลูกค้าที่ภายใน 1 ปี ไม่มีการเปิดบัญชี ณ วันที่ 31 ธ.ค 2565 ซึ่งทางผู้วิจัยได้เห็นว่าระยะเวลาที่เปลี่ยนไป 4 เดือนทำให้ข้อมูลกลุ่มลูกค้าเหล่านั้นเปลี่ยนไป เช่น ลักษณะส่วนบุคคลต่าง ๆ ที่เปลี่ยนไป พฤติกรรมการใช้เงินที่เปลี่ยนไป การถือครองผลิตภัณฑ์ของธนาคารที่เปลี่ยนไป การเปลี่ยนแปลงของข้อมูลดังกล่าวอาจจะทำให้ลูกค้าบางคนที่อยู่ในกลุ่ม 249,282 ที่เคยนำเข้าแบบจำลองซึ่งเป็นข้อมูลตัวอย่างของกลุ่มที่ไม่มีเปิดบัญชี อาจจะมีแนวโน้มในการเปิดบัญชีเงินฝากประเภท D ในช่วงเวลาปัจจุบัน

จากข้อมูลเมื่อวันที่ 31 ธ.ค 2565 กลุ่มที่ไม่เปิดบัญชีเท่ากับ 249,282 คน ระยะเวลาผ่านไปประมาณ 4 เดือนจนถึง 1 เม.ย. 2566 พบว่ามีลูกค้าที่เปิดบัญชีไปแล้วจำนวน 607 คน ทำให้จำนวนลูกค้าที่นำแบบจำลองไปพยากรณ์มีจำนวนทั้งสิ้น 248,675 คน โดยกลุ่มนี้ก็ยังมีลักษณะคือ เป็นกลุ่มที่มีบัญชีออมทรัพย์แบบธรรมดา เป็นกลุ่มที่มีเครดิตอย่างน้อย 1 ประเภท และยังไม่มีบัญชีเงินฝากประเภท D ณ ปัจจุบัน



รูปที่ 3.16 กลุ่มที่ได้หลังจากแบบจำลองพยากรณ์

จากรูปที่ 3.16 แสดงถึงกลุ่มหลังจากที่นำข้อมูลของลูกค้าทั้งสิ้น 248,675 คน เข้าแบบจำลอง จากนั้นแบบจำลองจะพยากรณ์แยกกลุ่มลูกค้าออกเป็น 2 กลุ่มได้แก่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

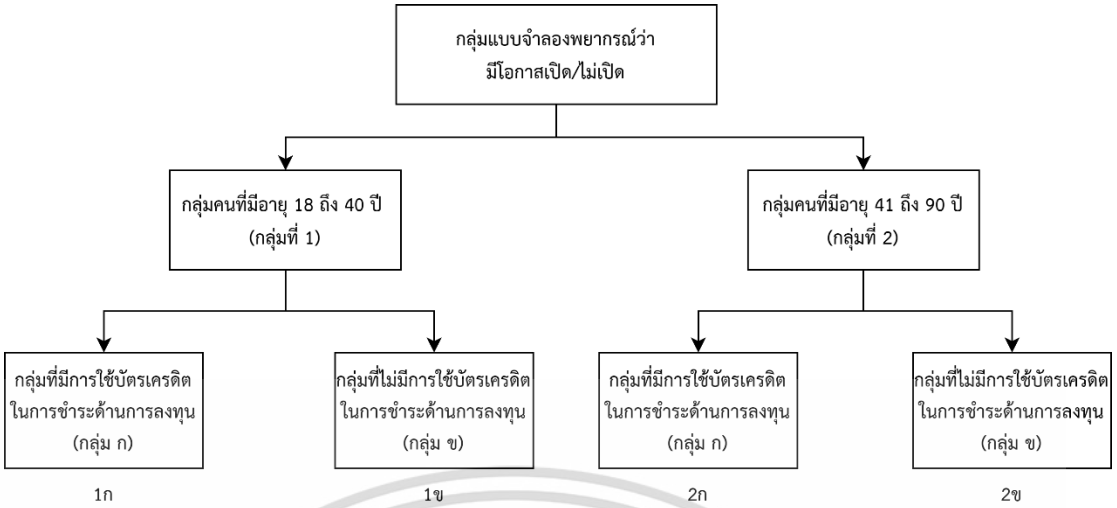
1. กลุ่มที่แบบจำลองพยากรณ์ว่า กลุ่มคนเหล่านี้มีโอกาสเปิดบัญชีเงินฝากประเภท D หรือ กลุ่ม ‘Y’ เท่ากับ 2,394 คน

2. กลุ่มที่แบบจำลองพยากรณ์ว่า กลุ่มคนเหล่านี้มีโอกาสที่จะไม่เปิดบัญชีเงินฝากประเภท D หรือกลุ่ม ‘N’ เท่ากับ 246,281 คน

เนื่องจากการวิจัยนี้จะตีความผลลัพธ์การส่งแคมเปญออกเป็น 2 มุมมองได้แก่ มุมผลการพยากรณ์ของแบบจำลอง และมุมมองผลการแยกกลุ่มตามคุณลักษณะของลูกค้า ทางผู้วิจัยจึงส่งแคมเปญให้แก่ลูกค้าทั้ง 2 กลุ่ม คือ 1. กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มคนเหล่านี้มีโอกาสเปิดบัญชี (Y) 2. กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มคนเหล่านี้มีโอกาสไม่เปิดบัญชี (N) โดยส่งแคมเปญในช่วงระยะเวลาเดียวกัน ข้อความเดียวกัน และเก็บผลพร้อมกัน การส่งแคมเปญให้กลุ่มลูกค้าดังกล่าว จะมีการสร้างข้อความ (Message) เพื่อสื่อสารกับลูกค้าผ่านทางช่องทางแอปพลิเคชันธนาคาร (Application) โดยปกติแล้วจะมีการสร้างข้อความตามในแคมเปญให้เป็นข้อความที่กำหนดเฉพาะบุคคล (Personalized Message) หมายถึงข้อความที่ถูกปรับแต่งให้เข้ากับผู้รับข้อความแต่ละคนหรือ กลุ่มลูกค้าเป้าหมายที่ต้องการสื่อสาร โดยคำว่า "เฉพาะบุคคล" ในที่นี้หมายถึงการปรับแต่งหรือกำหนดเนื้อหาในข้อความให้เป็นพิเศษสำหรับผู้รับ ใช้ข้อมูลส่วนบุคคล เช่น อายุ พฤติกรรม เพื่อให้ข้อความในแคมเปญ มีความเป็นเอกลักษณ์และเชื่อมโยงกับผู้รับข้อความได้อย่างมีประสิทธิภาพ และสร้างความสัมพันธ์ระหว่างลูกค้าและธนาคารได้ดียิ่งขึ้น โดยจากงานวิจัย Fridh and Dahl (2019) ได้พบว่า ผลของการทำการตลาดแบบส่วนบุคคล (Personalized Marketing) มีผลต่อกระบวนการตัดสินใจของผู้บริโภคในทางที่แตกต่างกัน หมายความว่า การทำการตลาดแบบส่วนบุคคลอาจจะทำให้ผู้บริโภคสนใจแคมเปญมากขึ้น และการประยุกต์แบบจำลองก็อาจเพิ่มโอกาสที่ลูกค้าจะเข้ามาใช้ผลิตภัณฑ์มากขึ้น

โดยการสร้างข้อความที่กำหนดเฉพาะส่วนบุคคล ทางผู้วิจัยจึงได้แบ่งกลุ่มลูกค้าออกเป็นแต่ละกลุ่มตามคุณลักษณะที่สนใจ เพื่อสร้างข้อความแคมเปญข้อเสนอที่เกี่ยวกับบัญชีเงินฝากประเภท D ให้แก่กลุ่มลูกค้าที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y) และ มีโอกาสไม่เปิดบัญชี (N) โดยข้อความจะปรับแต่งตามลักษณะที่แบ่งกลุ่มเอาไว้เพื่อให้เหมาะสมกับความต้องการ และพฤติกรรมของกลุ่มเป้าหมายของผู้ใช้งาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.17 กลุ่มที่แยกตามคุณลักษณะเพื่อส่งแคมเปญ

รูปที่ 3.17 แสดงกลุ่มที่แยกตามคุณลักษณะเพื่อส่งแคมเปญขาย การส่งแคมเปญจะถูกส่งให้แก่กลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y) และพยากรณ์ว่ามีโอกาสไม่เปิดบัญชี (N) พร้อมกันด้วยข้อความที่เหมือนกัน ซึ่งทางผู้วิจัยมีความต้องการที่จะสร้างข้อความกำหนดเฉพาะบุคคลเพื่อเพิ่มประสิทธิภาพการตอบกลับแคมเปญ

จากการอ้างอิงจากงานวิจัยของ อภิชญาฎา และบุษรา (2563) ที่ได้ศึกษาพฤติกรรมการวางแผนทางการเงินเพื่อการอยู่อาศัยในวัยเกษียณฯ แต่ละช่วงอายุ พบว่าช่วงอายุที่เริ่มวางแผนทางการเงินมากที่สุดอยู่ในช่วง 40 ปีขึ้นไป และจากการแบ่งช่วงอายุของธนาคารที่แบ่งออกเป็น 8 ช่วง ดังนี้ 1. นักเรียน (อายุ 7 ถึง 21 ปี) 2. หนุ่มสาว (อายุ 22 ถึง 26 ปี) 3. วัยสร้างครอบครัว (อายุ 27 ถึง 40 ปี) 4.วัยมีครอบครัว (อายุ 41 ถึง 60 ปี) 5. วัยเกษียณฯ1 (อายุ 61 ถึง 65 ปี) 6. วัยเกษียณฯ2 (อายุ 66 ถึง 70 ปี) 7. วัยเกษียณฯ3 (อายุ 71 ถึง 90 ปี) 8. อื่น ๆ (อายุน้อยกว่า 7 และอายุมากกว่า 90 ปี) ดังนั้นงานวิจัยนี้จึงได้แบ่งกลุ่มลูกค้าที่จะสร้างข้อความส่วนบุคคลให้แก่ลูกค้าออกเป็น 2 กลุ่ม ตามอายุได้แก่

กลุ่มที่ 1) มีอายุ 18 ถึง 40 ปี เป็นกลุ่มที่รวม วัยนักเรียน วัยหนุ่มสาว วัยสร้างครอบครัว

กลุ่มที่ 2) มีอายุ 41 ถึง 90 ปี เป็นกลุ่มที่รวม วัยมีครอบครัว วัยเกษียณฯ1 วัยเกษียณฯ2 วัยเกษียณฯ3

จากงานวิจัยของ Sanou et al. (2018) กล่าวไว้ว่า บุคคลที่มีความอดทนยอมรับความเสี่ยงได้น้อย มักจะลงทุนกับการเปิด บัญชีฝากเงิน มากกว่าการลงทุนที่มีความเสี่ยงสูง เช่น หุ้น หรือกองทุน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ดังนั้น จากงานวิจัยที่กล่าวบ่งบอกถึงกลุ่มที่อาจจะสนใจในบัญชีแตกต่างกันไป และจากการสอบถามผู้เชี่ยวชาญทำให้ จึงแบ่งลูกค้าออกเป็น 2 กลุ่ม ตามพฤติกรรมในการใช้บัตรเครดิต เพื่อประเมินการยอมรับความเสี่ยงของแต่ละบุคคล งานวิจัยนี้จึงได้แบ่งกลุ่มออกอีก 2 กลุ่ม ดังนี้

กลุ่ม ก) กลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุน เป็นกลุ่มที่มีการใช้บัตรเครดิตซื้อผลิตภัณฑ์หรือทรัพย์สินที่เกี่ยวกับการลงทุน เช่น ซื้อกองทุน ซื้อหุ้น หรือลงทุนในทรัพย์สินต่าง ๆ อย่างน้อยหนึ่งครั้ง

กลุ่ม ข) กลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุน ซึ่งเป็นการชำระบัตรเครดิตเพื่อใช้จ่ายทั่วไป เช่น ใช้จ่ายซื้อของออนไลน์ ใช้จ่ายร้านอาหาร จะไม่มีการใช้จ่ายที่เกี่ยวกับการลงทุน

ดังนั้นการสร้างข้อความที่กำหนดเฉพาะบุคคลหรือข้อความที่ถูกปรับแต่งให้เหมาะสมและเป็นพิเศษตามลักษณะส่วนบุคคลของผู้รับข้อความ เพื่อให้ลูกค้าเกิดความเชื่อใจและสนิทสนมกับธนาคาร ทำให้มีโอกาสตอบรับการเปิดบัญชีมากขึ้น โดยทำการรวมกลุ่มที่ 1 และ2 เข้ากับกลุ่ม ก และข เพื่อสร้างข้อความที่ไม่เหมือนกันอยู่ 4 ประเภท ให้กับกลุ่มลูกค้า 4 กลุ่ม

ข้อความประเภท 1ก) เป็นข้อความที่ถูกส่งให้แก่กลุ่มลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

ข้อความประเภท 1ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

ข้อความประเภท 2ก) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

ข้อความประเภท 2ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

งานวิจัยนี้จะมีการส่งแคมเปญในช่วงวันที่ 20 เมษายน 2566 และเก็บผลวันที่ 11 พฤษภาคม 2566 โดยจะส่งแคมเปญให้แก่ทั้ง 2 กลุ่มคือ

1. กลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y)
2. กลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสไม่เปิดบัญชี (N)

โดย 2 กลุ่มนี้จะได้รับข้อความที่เหมือนกัน 4 ข้อความ ตามคุณลักษณะที่แบ่งกลุ่มไว้ ได้แก่ ประเภทข้อความ 1ก, 1ข, 2ก และ 2ข แต่ละประเภทจะมีเนื้อหาข้อความที่แตกต่างกัน ดังตารางที่ 3.24

ตารางที่ 3.24 จำนวนกลุ่มลูกค้าแต่ละกลุ่มที่ได้รับข้อความที่มีเนื้อหาแตกต่างกัน

ประเภทข้อความ	กลุ่ม	ข้อความ	Y	N
1ก	อายุ 18 ถึง 40 ปี และใช้บัตรเครดิตชำระด้านการลงทุน	“อีกทางเลือกเพิ่มที่ช่วยให้เงินของคุณงอกเงยได้และ ได้ผลตอบแทนที่แน่นอน แนะนำบัญชีประเภท D บัญชีเงินออม ที่ให้ดอกเบี้ยรวมโบนัสสูงสุด 2% ตั้งแต่บาทแรก ช่วยเพิ่มมูลค่าเงินออมของคุณให้งอกเงยได้เร็วขึ้น 😳”	152	12,860
ข1	อายุ 18 ถึง 40 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน	“บัญชีประเภท D บัญชีเงินฝาก ดิจิทัลตัวช่วยในการเก็บเงิน ที่ผลตอบแทนสูง 😳 และแน่นอนตั้งแต่บาทแรก รับดอกเบี้ยรวมโบนัสสูงสุด 2% เพียงมีฝากเงินเพิ่มขึ้นทุกเดือน ไม่มีขั้นต่ำในการฝาก”	490	69,072
2ก	อายุ 41 ถึง 90 ปี และใช้บัตรเครดิตชำระด้านการลงทุน	“บัญชีเงินออม ให้ผลตอบแทนสูง บัญชีประเภท D บัญชีเงินฝาก รับดอกเบี้ยรวมโบนัสสูงสุด 2% ตั้งแต่บาทแรก ให้เงินช่วยทำงานเพื่อรับผลตอบแทนที่แน่นอน 😳”	365	25,416
2ข	อายุ 41 ถึง 90 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน	“ฝากเงินที่บัญชี บัญชีประเภท D บัญชีเงินฝาก รับดอกเบี้ยรับดอกเบี้ยรวมโบนัสสูงสุด 2% ตั้งแต่บาทแรก ถึง 1 แสนบาท 😳 ผลตอบแทนสูงง่าย ๆ ได้ทุกวัน”	1,387	138,933

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 3.24 แสดงจำนวนกลุ่มลูกค้าแต่ละกลุ่มที่ได้รับข้อความที่มีเนื้อหาแตกต่างกัน โดยเนื้อหาของข้อความแต่ละกลุ่มได้สร้างขึ้นจากความรู้ความเข้าใจทางผู้จัดทำและปรึกษาคณะผู้เชี่ยวชาญโดยมีเหตุผลดังต่อไปนี้

ประเภทข้อความ 1ก: กลุ่มอายุ 18 ถึง 40 ปี และใช้บัตรเครดิตชำระด้านการลงทุน ข้อความที่ถูกสร้างขึ้นโดยการวิเคราะห์ตีความหมายว่า เนื่องจากกลุ่มดังกล่าวเป็นกลุ่มวัยกลางคนรวมถึงวัยที่อายุยังน้อย อาจมีความต้องการในการสร้างทรัพย์สินและเตรียมความพร้อมทางการเงินเพื่ออนาคต และการใช้จ่ายบัตรเครดิตลงทุนอาจแสดงถึงความสนใจในการลงทุนและการเพิ่มมูลค่าทางการเงินในระยะยาว ดังนั้นจึงขอแนะนำบัญชีเงินฝากดิจิทัลให้เป็นทางเลือกเพื่อให้เป็น “ตัวช่วยในการทำให้เงินงอกเงย” เพราะกลุ่มนี้อาจจะมีการลงทุนอยู่แล้วจึงเสนอเป็นตัวช่วยมากกว่าการเสนอขายธรรมดา และข้อความมีการใส่อิโมจิเพื่อเพิ่มอารมณ์ความรู้สึกในการอ่าน

ประเภทข้อความ 1ข: กลุ่มอายุ 18 ถึง 40 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน ข้อความที่ถูกสร้างขึ้นโดยการวิเคราะห์ตีความหมายว่า เนื่องจากกลุ่มดังกล่าวอาจมีความต้องการในการสร้างฐานะการเงินที่แข็งแรงในอนาคต การเปิดบัญชีออมเงินเป็นวิธีที่เหมาะสมในการเริ่มต้นการออมเงิน แต่ไม่มีการใช้จ่ายบัตรเครดิตที่เกี่ยวกับการลงทุน อาจแสดงถึงการระมัดระวังในการใช้เงินและความสนใจในการเพิ่มมูลค่าทางการเงินในระยะยาว โดยอาจมีความต้องการในการสร้างฐานะการเงินที่เสถียรและการสร้างกองทุนสำรองเพื่อมีความมั่นคงในการเงินในอนาคต จึงเสนอให้บัญชีเงินฝากดิจิทัลนี้เป็นตัวช่วยในการเก็บเงินและย้ำว่าได้รับผลตอบแทนที่แน่นอนเพื่อให้เกิดความสบายใจต่อการฝากเงิน และข้อความมีการใส่อิโมจิเพื่อเพิ่มอารมณ์ความรู้สึกในการอ่าน

ประเภทข้อความ 2ก: กลุ่มอายุ 41 ถึง 90 ปี และใช้บัตรเครดิตชำระด้านการลงทุน ข้อความที่ถูกสร้างขึ้นโดยการวิเคราะห์ตีความหมายว่า เนื่องจากกลุ่มดังกล่าวเป็นกลุ่มที่มีสถานะทางการเงินที่เสถียรและมั่นคงกว่า มีรายได้ที่มั่นคงและมีสมรรถภาพในการลงทุน กลุ่มนี้มักมีเป้าหมายในการเตรียมความพร้อมในช่วงชีวิตต่อไป เช่น ภาระหน้าที่ที่ต้องดูแลคนในครอบครัว หรือการเกษียณอย่างมั่นคงหรือการลงทุนในรายได้ที่มากขึ้น ซึ่งเป็นกลุ่มที่ไม่ต้องการความเสี่ยงที่เพิ่มขึ้นไปจากเดิมมากนักเพื่ออนาคตหลังเกษียณที่มั่นคง จึงเสนอบัญชีเงินฝากดิจิทัลนี้ด้วยข้อความ "ให้เงินช่วยทำงาน" เพราะกลุ่มนี้อาจจะมีการลงทุนอยู่แล้วจึงเสนอเป็นตัวช่วยมากกว่าการเสนอขายธรรมดา และ "รับผลตอบแทนที่แน่นอน" หมายถึงการได้รับผลตอบแทนจากการลงทุนที่มีความเสี่ยงต่ำหรือความเสี่ยงที่รู้สึกถึงขั้นต่ำ ทำให้ลูกค้าอาจจะเกิดความสมในในบัญชีนี้ และข้อความมีการใส่อิโมจิเพื่อเพิ่มอารมณ์ความรู้สึกในการอ่าน

ประเภทข้อความ 2ข: กลุ่มอายุ 41 ถึง 90 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน ข้อความที่ถูกสร้างขึ้นโดยการวิเคราะห์ตีความหมายว่า เนื่องจากกลุ่มนี้เป็นกลุ่มที่มีอายุเยอะและไม่ใช้จ่ายบัตรเครดิตที่เกี่ยวกับการลงทุนเลย อาจไม่มีประสบการณ์และความคุ้นเคยกับการลงทุนแบบต่าง ๆ และอาจรู้สึกสบายใจกับวิธีการอื่นในการออมเงินที่มีความเสี่ยงต่ำ เช่น เปิดบัญชีออมทรัพย์ที่มีการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

คาดการณ์ผลตอบแทนที่แน่นอนมากขึ้น จึงเสนอขายบัญชีเงินฝากดิจิทัลที่บอกถึงดอกเบี้ยที่ชัดเจนที่ผลตอบแทนที่สูงและความง่ายของการได้ดอกเบี้ยทุกวัน และข้อความมีการใส่อิโมจิเพื่อเพิ่มอารมณ์ความรู้สึกในการอ่าน

โดยกลุ่มที่แบบจำลองพยากรณ์ว่า กลุ่มลูกค้ากลุ่มนี้มีโอกาสเปิดบัญชี (Y) มีจำนวนทั้งสิ้น 2,394 คน แบ่งเป็นกลุ่มข้อความประเภท 1ก เท่ากับ 152 กลุ่มข้อความประเภท 1ข, 2ก, 2ข เท่ากับ 490, 365 และ 1,387 คนตามลำดับ และกลุ่มที่แบบจำลองพยากรณ์ว่า กลุ่มลูกค้ากลุ่มนี้มีโอกาสไม่เปิดบัญชี (N) มีจำนวนทั้งสิ้น 246,281 คน แบ่งเป็นกลุ่มข้อความประเภท 1ก เท่ากับ 12,860 กลุ่มข้อความประเภท 1ข, 2ก, 2ข เท่ากับ 69,072, 25,416 และ 138,933 ตามลำดับ โดยการส่งแคมเปญทั้ง 8 กลุ่มนี้พร้อมกัน เพื่อเปรียบเทียบผลการเข้ามาใช้ผลิตภัณฑ์ ระหว่างกลุ่มที่พยากรณ์ด้วยแบบจำลอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิจัยและอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและพัฒนาแบบจำลองการเรียนรู้ของเครื่อง สำหรับการส่งแคมเปญเสนอขายผลิตภัณฑ์บัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท (บัญชีเงินฝากประเภท D) โดยการนำข้อมูลสร้างแบบจำลองมาจากคลังข้อมูลของธนาคารพาณิชย์แห่งหนึ่ง เพื่อสร้างแบบจำลองทั้งสิ้น 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และ วิธีป่าไม้สุ่ม การทดสอบประสิทธิภาพ พบว่า วิธีป่าไม้สุ่ม เป็นแบบจำลองที่ให้ประสิทธิภาพดีที่สุด จากนั้นนำแบบจำลองไปพยากรณ์ลูกค้าที่ต้องการส่งแคมเปญเพื่อเสนอขายบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท (บัญชีเงินฝากประเภท D) โดยการส่งแคมเปญจะจัดส่งให้แก่ลูกค้าที่มีข้อความแตกต่างกัน เป็นข้อความเฉพาะบุคคล เพื่อกระตุ้นประสิทธิภาพการเข้ามาใช้ผลิตภัณฑ์ของแคมเปญ โดยบทนี้ไปจะแสดงผลลัพธ์ของการประยุกต์ใช้การเรียนรู้ของเครื่องสำหรับการส่งแคมเปญและผลตอบกลับของแคมเปญดังนี้

4.1 ผลลัพธ์การเข้ามาใช้ผลิตภัณฑ์

จากการได้แบบจำลองที่เหมาะสมที่สุดสำหรับการนำไปใช้งาน และจากการวางแผนการสร้างข้อความเฉพาะกลุ่มให้แก่ลูกค้า 4 กลุ่ม ซึ่งได้มีการส่งให้แก่ลูกค้า ที่มีบัญชีออมทรัพย์ธรรมดา มีบัตรเครดิต และไม่มีบัญชีเงินฝากประเภท D จำนวน 246,281 คน ที่ได้ข้อความวันที่ 20 เมษายน พ.ศ. 2566 เพียงวันเดียว และเก็บผลการเปิดบัญชีในวันที่ 11 พฤษภาคม พ.ศ. 2566 ระยะเวลา 21 วัน เก็บผล ซึ่งมีจำนวนลูกค้าที่เปิดบัญชีเงินฝากประเภท D ภายใต้การถูกนำเสนอด้วยแคมเปญนี้ทั้งสิ้น 74 คน จากทั้งหมด โดยหัวข้อนี้จะอธิบายผลจากการประยุกต์การเรียนรู้ของเครื่องและผลตอบรับแคมเปญทั้งหมด

4.1.1 ผลลัพธ์การเข้ามาใช้ผลิตภัณฑ์จำแนกตามการพยากรณ์ด้วยแบบจำลอง

จากการนำแบบจำลองที่เหมาะสมไปใช้งานแคมเปญได้มีการส่งให้แก่ลูกค้าจำนวน 246,281 คน โดยแบบจำลองจำแนกเป็น 2 กลุ่ม คือ 1.กลุ่มที่แบบจำลองพยากรณ์ ‘Y’ หรือกลุ่มลูกค้ามีโอกาสเปิดบัญชี และ 2. กลุ่มที่แบบจำลองพยากรณ์ว่า ‘N’ หรือกลุ่มลูกค้าที่ไม่มีโอกาสเปิดบัญชี โดยจะตารางเปรียบเทียบการตอบกลับระหว่างกลุ่มที่แบบจำลองพยากรณ์ดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.1 ผลการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี

แบบจำลองพยากรณ์	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
มีโอกาสเปิดบัญชี (Y)	2,394	2	0.0835

จากตารางที่ 4.1 แสดงผลการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี หลังจากการเก็บผลระหว่างวันที่ 20 เมษายน พ.ศ. 2566 ถึงวันที่ 11 พฤษภาคม พ.ศ. 2566 ผลลัพธ์จากตารางที่รวมจำนวนคนที่เปิดบัญชีเงินฝากประเภท D พบว่า กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มลูกค้ากลุ่มนี้มีโอกาสเปิดบัญชี (Y) และได้มีการส่งแคมเปญให้แก่ลูกค้าทั้งหมด 2,394 มีจำนวนลูกค้าเข้ามาใช้ผลิตภัณฑ์เท่ากับ 2 คน คิดเป็นการเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0835

ตารางที่ 4.2 ผลการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสที่จะไม่เปิดบัญชี

แบบจำลองพยากรณ์	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
มีโอกาสที่จะไม่เปิดบัญชี (N)	246,281	72	0.0292

จากตารางที่ 4.2 แสดงผลการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสไม่เปิดบัญชี หลังจากการเก็บผลระหว่างวันที่ 20 เมษายน พ.ศ. 2566 ถึงวันที่ 11 พฤษภาคม พ.ศ. 2566 ผลลัพธ์จากตารางที่รวมจำนวนคนที่เปิดบัญชีเงินฝากประเภท D พบว่า กลุ่มที่แบบจำลองพยากรณ์ว่ากลุ่มลูกค้ากลุ่มนี้ไม่มีโอกาสเปิดบัญชี (N) และได้มีการส่งแคมเปญให้แก่ลูกค้าทั้งหมด 246,281 มีจำนวนลูกค้าเข้ามาใช้ผลิตภัณฑ์เท่ากับ 72 คน คิดเป็นการเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0292

4.1.2 ผลลัพธ์การเข้ามาใช้ผลิตภัณฑ์จำแนกตามประเภทข้อความ

ผลลัพธ์ของการเข้ามาใช้ผลิตภัณฑ์บัญชีเงินฝากประเภท D ภายในวันที่ 20 เมษายน พ.ศ. 2566 ถึง วันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นเวลา 21 วันเก็บผล ที่จำแนกตามประเภทข้อความ โดยแต่ละข้อความจะจำแนกตามคุณลักษณะของลูกค้าดังนี้

ข้อความประเภท 1ก : เป็นข้อความที่ถูกส่งให้แก่กลุ่มลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อความประเภท 1ข : เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

ข้อความประเภท 2ก : เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

ข้อความประเภท 2ข : เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

กลุ่มข้อความประเภท 1ก ที่ได้รับข้อความในแคมเปญคือ “อีกทางเลือกเพิ่มที่ช่วยให้เงินของคุณงอกเงยได้และ ได้ผลตอบแทนที่แน่นอน แนะนำบัญชีประเภท D บัญชีเงินออม ที่ให้ดอกเบี้ยรวมโบนัสสูงสุด 2% ตั้งแต่บาทแรก ช่วยเพิ่มมูลค่าเงินออมของคุณให้งอกเงยได้เร็วขึ้น 😲” จะแสดงการเข้ามาใช้ผลิตภัณฑ์ดังตารางที่ 4.3

ตารางที่ 4.3 การเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 1ก

ประเภทข้อความ	กลุ่ม	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
1ก	อายุ 18 ถึง 40 ปี และใช้บัตรเครดิตชำระด้านการลงทุน	13,012	5	0.0384

จากตารางที่ 4.3 แสดงการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 1ก หรือ กลุ่มอายุ 18 ถึง 40 ปี และใช้บัตรเครดิตชำระด้านการลงทุน ที่ได้รับข้อเสนอแคมเปญในวันที่ 20 เมษายน พ.ศ. 2566 และรอรวบรวมผล วันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นเวลารอผล 21 วัน พบว่า จากการส่งแคมเปญไป 13,012 คน มีการเข้ามาใช้ผลิตภัณฑ์เท่ากับ 5 คน คิดเป็นการเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0384

กลุ่มข้อความประเภท 1ข ที่ได้รับข้อความในแคมเปญคือ “บัญชีประเภท D บัญชีเงินฝาก ดิจิทัลตัวช่วยในการเก็บเงิน ที่ผลตอบแทนสูง 😲 และแน่นอนตั้งแต่บาทแรก รับดอกเบี้ยรวมโบนัสสูงสุด 2% เพียงมีฝากเงินเพิ่มขึ้นทุกเดือน ไม่มีขั้นต่ำในการฝาก” จะแสดงการเข้ามาใช้ผลิตภัณฑ์ดังตารางที่ 4.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.4 การเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 1ข

ประเภทข้อความ	กลุ่ม	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
1ข	อายุ 18 ถึง 40 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน	69,562	30	0.0431

จากตารางที่ 4.4 แสดงการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 1ข หรือ กลุ่มอายุ 18 ถึง 40 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน ที่ได้รับข้อเสนอแคมเปญในวันที่ 20 เมษายน พ.ศ. 2566 และรอรวบรวมผล วันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นเวลารอผล 21 วัน พบว่า จากการส่งแคมเปญไป 69,562 คน มีการเข้ามาใช้ผลิตภัณฑ์เท่ากับ 30 คน คิดเป็นการเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0431

กลุ่มข้อความประเภท 2ก ที่ได้รับข้อความในแคมเปญคือ “บัญชีเงินออม ให้ผลตอบแทนสูง บัญชีประเภท D บัญชีเงินฝาก รับดอกเบี้ยรวมโบนัสสูงสุด 2%ตั้งแต่บาทแรก ให้เงินช่วยทำงานเพื่อรับผลตอบแทนที่แน่นอน 😱” จะแสดงการเข้ามาใช้ผลิตภัณฑ์ดังตารางที่ 4.5

ตารางที่ 4.5 การเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 2ก

ประเภทข้อความ	กลุ่ม	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
2ก	อายุ 41 ถึง 90 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน	140,320	29	0.0207

จากตารางที่ 4.5 แสดงการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 2ก หรือ กลุ่มอายุ 41 ถึง 90 ปี และใช้บัตรเครดิตชำระด้านการลงทุน ที่ได้รับข้อเสนอแคมเปญในวันที่ 20 เมษายน พ.ศ. 2566 และรอรวบรวมผล วันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นเวลารอผล 21 วัน พบว่า จากการส่งแคมเปญไป 25,781 คน มีการเข้ามาใช้ผลิตภัณฑ์เท่ากับ 10 คน คิดเป็นการเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0388

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กลุ่มข้อความประเภท 2ข ที่ได้รับข้อความในแคมเปญคือ “ฝากเงินที่บัญชี บัญชีประเภท D บัญชีเงินฝาก รับดอกเบี้ยรับดอกเบี้ยรวมโบนัสสูงสุด 2% ตั้งแต่บาทแรก ถึง 1 แสนบาท 😳 ผลตอบแทนสูงง่าย ๆ ได้ทุกวัน” จะแสดงการเข้ามาใช้ผลิตภัณฑ์ดังตารางที่ 4.6

ตารางที่ 4.6 การเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 2ข

ประเภทข้อความ	กลุ่ม	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
2ข	อายุ 41 ถึง 90 ปี และใช้บัตรเครดิตชำระด้านการลงทุน	25,781	10	0.0388

จากตารางที่ 4.6 แสดงการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มลูกค้าที่ได้รับข้อความประเภท 2ข หรือ กลุ่มอายุ 41 ถึง 90 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน ที่ได้รับข้อเสนอแคมเปญในวันที่ 20 เมษายน พ.ศ. 2566 และรอรวบรวมผล วันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นเวลารอผล 21 วัน พบว่า จากการส่งแคมเปญไป 140,320 คน มีการเข้ามาใช้ผลิตภัณฑ์เท่ากับ 29 คน คิดเป็นการเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0207

4.1.3 ผลลัพธ์การเข้ามาใช้ผลิตภัณฑ์จำแนกตามการพยากรณ์ด้วยแบบจำลองและประเภทข้อความ

ผลลัพธ์ของการเข้ามาใช้ผลิตภัณฑ์บัญชีเงินฝากประเภท D ภายในวันที่ 20 เมษายน พ.ศ. 2566 ถึง วันที่ 11 พฤษภาคม พ.ศ. 2566 โดยจำแนกตามการพยากรณ์ด้วยแบบจำลองและประเภทข้อความ ผลลัพธ์ในรูปแบบตารางเมทริกซ์ความสับสน ที่บ่งบอกถึงการพยากรณ์ด้วยแบบจำลอง (Pred) กับผลลัพธ์ที่ลูกค้าเข้ามาเปิดบัญชีจริง (Real) โดนในตารางจะถูกจำแนกเป็นลูกค้าที่เปิดบัญชีหรือการพยากรณ์ว่ามีโอกาสเปิดบัญชี แทนด้วย ‘Y’ และ ลูกค้าไม่เปิดบัญชีหรือการพยากรณ์ว่ามีโอกาสไม่เปิดบัญชี แทนด้วย “N” ซึ่งประเภทข้อความจะถูกจำแนกตามคุณลักษณะของลูกค้าดังนี้

ข้อความประเภท 1ก) เป็นข้อความที่ถูกส่งให้แก่กลุ่มลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อความประเภท 1ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

ข้อความประเภท 2ก) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

ข้อความประเภท 2ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

ตารางที่ 4.7 ผลการเข้ามาใช้ผลิตภัณฑ์ที่จำแนกตามการพยากรณ์ด้วยแบบจำลองและประเภทข้อความ ในรูปแบบตารางเมทริกซ์ความสับสน

	อายุ 18 ถึง 40 ปี				**อายุ 41 ถึง 90 ปี**			
			Y (Pred)	**N (Pred)**			**Y (Pred)**	**N (Pred)**
ใช้บัตรเครดิตชำระด้านการลงทุน	**1ก**	**Y (Real)**	0	5	**2ข**	**Y (Real)**	0	10
		N (Real)	152	12,855		**N (Real)**	365	25,406
ไม่ใช้บัตรเครดิตชำระด้านการลงทุน	**1ข**	**Y (Real)**	2	28	**2ข**	**Y (Real)**	0	29
		N (Real)	478	69,044		**N (Real)**	1,387	138,904

ตารางที่ 4.7 แสดงผลการเข้ามาใช้ผลิตภัณฑ์ที่จำแนกตามแบบจำลองพยากรณ์และกลุ่มข้อความ เป็นตารางที่แสดงการพยากรณ์ของแบบจำลอง (Pred) กับการเปิดบัญชีจริง (Real)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กลุ่มประเภทข้อความ 1ก หรือ กลุ่มที่อายุ 18 ถึง 40 ปี ใช้บัตรเครดิตด้านการลงทุน แบบจำลองพยากรณ์ว่า มีโอกาสเปิดบัญชี (Y) ทั้งหมด 152 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 0 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 152 คน

แบบจำลองพยากรณ์ว่า มีโอกาสไม่เปิดบัญชี (N) ทั้งหมด 12,860 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 5 คน ดังนั้นไม่มีลูกค้ามาเปิดบัญชี 12,855 คน

กลุ่มประเภทข้อความ 1ข หรือ กลุ่มที่อายุ 18 ถึง 40 ปี ไม่ใช้บัตรเครดิตด้านการลงทุน แบบจำลองพยากรณ์ว่า มีโอกาสเปิดบัญชี (Y) ทั้งหมด 480 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 2 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 478 คน

และแบบจำลองพยากรณ์ว่า มีโอกาสไม่เปิดบัญชี (N) ทั้งหมด 69,072 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 28 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 69,044 คน

กลุ่มประเภทข้อความ 2ก หรือ กลุ่มที่อายุ 41 **ถึง** 90 **ปี ใช้บัตรเครดิตด้านการลงทุน** แบบจำลองพยากรณ์ว่า มีโอกาสเปิดบัญชี (Y) ทั้งหมด 365 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 0 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 365 คน

แบบจำลองพยากรณ์ว่า มีโอกาสไม่เปิดบัญชี (N) ทั้งหมด 25,416 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 10 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 25,406 คน

กลุ่มประเภทข้อความ 2ข หรือ กลุ่มที่อายุ 41 **ถึง** 90 **ปี ไม่ใช้บัตรเครดิตด้านการลงทุน** แบบจำลองพยากรณ์ว่า มีโอกาสเปิดบัญชี (Y) ทั้งหมด 1,387 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 0 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 1,387 คน

และแบบจำลองพยากรณ์ว่า มีโอกาสไม่เปิดบัญชี (N) ทั้งหมด 138,933 คน มีลูกค้ามาเปิดบัญชีจริง (Real) 29 คน ดังนั้น ไม่มีลูกค้ามาเปิดบัญชี 138,904 คน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้ทำการประยุกต์ใช้การเรียนรู้ของเครื่องในการส่งแคมเปญ บัญชีเงินฝากดิจิทัล โดยเป็นกรณีศึกษาธนาคารพาณิชย์แห่งหนึ่ง โดยในบทนี้จะนำเสนอสรุปผลการวิจัยตามแต่ละวัตถุประสงค์ของงานวิจัย รวมถึงข้อจำกัดและข้อเสนอแนะของงานวิจัย ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 เพื่อตอบวัตถุประสงค์ข้อที่ 1 คือ เพื่อศึกษาและพัฒนาแบบจำลองการเรียนรู้ของเครื่องสำหรับการส่งแคมเปญเสนอขายผลิตภัณฑ์บัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ซึ่งจากงานวิจัยครั้งนี้ได้ทำการนำข้อมูลของลูกค้า ย้อนหลัง 1 ปี จากคลังข้อมูลของธนาคารพาณิชย์แห่งหนึ่ง โดยนำข้อมูลลูกค้ามาเป็นต้นแบบสำหรับการฝึกสอนแบบจำลอง เป็นกลุ่มลูกค้าที่ตรงตาม 3 เงื่อนไขของงานวิจัยนี้ ได้แก่ เป็นลูกค้าที่มีบัญชีออมทรัพย์แบบธรรมดา เป็นลูกค้าที่มีบัตรเครดิตอย่างน้อย 1 ประเภท และเป็นลูกค้าที่ไม่มีบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท (บัญชีเงินฝากประเภท D) จำนวนทั้งสิ้น 253,594 รายการ ซึ่งแบ่งออกเป็น 2 กลุ่ม ได้แก่ กลุ่มที่มีการเปิดบัญชีในระหว่างวันที่ 1 มกราคม พ.ศ. 2565 ถึงวันที่ 31 ธันวาคม พ.ศ. 2565 เท่ากับ 4,312 รายการ และ กลุ่มที่ไม่เปิดบัญชีระหว่างวันที่ 1 มกราคม พ.ศ. 2565 ถึงวันที่ 31 ธันวาคม พ.ศ. 2565 เท่ากับ 249,282 รายการ จากนั้นได้มีการจัดเตรียมข้อมูลก่อนสร้างแบบจำลอง เช่น การตัดค่าสูญหาย การกำจัดค่านอกเกณฑ์ การรวมข้อมูล การแปลงข้อมูล การคัดเลือกคุณลักษณะ และการจัดการข้อมูลที่ไม่สมดุลกัน เพื่อนำข้อมูลที่ถูกจัดเตรียมเสร็จสิ้นเข้าสู่แบบจำลอง โดยแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอน 80 เปอร์เซ็นต์ของข้อมูลทั้งหมด และเป็นชุดข้อมูลทดสอบ 20 เปอร์เซ็นต์ของข้อมูลทั้งหมดเพื่อนำไปวัดประสิทธิภาพแบบจำลอง โดยการสร้างแบบจำลองด้วยข้อมูลชุดฝึกสอนมีทั้งหมด 3 วิธี ได้แก่

วิธีเพื่อนบ้านใกล้สุด k ตัว เป็นวิธีที่นิยมใช้ เพราะไม่ต้องการสมการที่ซับซ้อนหรือการปรับพารามิเตอร์ที่ซับซ้อน เพียงแค่คำนวณระยะห่างระหว่างตัวอย่างและหาตัวอย่างที่ใกล้ที่สุดเพื่อทำนายข้อมูลใหม่ เป็นวิธีที่เข้าใจง่ายและเรียนรู้ได้รวดเร็ว

วิธีนาอีฟเบย์ เป็นวิธีที่นิยมใช้ เพราะการประมวลผลที่รวดเร็วเนื่องจากไม่ต้องประมวลผลซับซ้อนหรือทำการค้นหาพารามิเตอร์

วิธีป่าไม้สุ่ม เป็นวิธีที่นิยมใช้ เพราะเป็นวิธีที่มีความแม่นยำสูงในการจำแนกข้อมูล โดยสร้างชุดต้นไม้เพื่อประมวลผลและตัดสินใจ มีความเสถียรในการจำแนกข้อมูลที่มีลักษณะที่ซับซ้อน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.1.2 เพื่อตอบวัตถุประสงค์ข้อที่ 2 คือ เพื่อเปรียบเทียบวิธีการเรียนรู้ของเครื่องที่เหมาะสมสำหรับการนำไปใช้พยากรณ์ กลุ่มลูกค้าที่มีโอกาสจะเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ซึ่งงานวิจัยนี้ได้ทำการเตรียมข้อมูลเพื่อสร้างแบบจำลองทั้ง 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และ วิธีป่าไม้สุ่ม มีการปรับไฮเปอร์พารามิเตอร์เพื่อให้แบบจำลองมีความแม่นยำในการทำนายข้อมูลชุดนี้มากที่สุด แล้วนำผลลัพธ์การเรียนรู้ของเครื่องทั้ง 3 วิธีมาเปรียบเทียบกัน พบว่า แบบจำลองที่ดีที่สุด คือ วิธีป่าไม้สุ่มที่มี ค่าความแม่น 97.40% ค่าความเที่ยง 87.31% ค่าระลึก 72.72% และค่าประสิทธิภาพโดยรวม 79.35% สูงที่สุดเมื่อเทียบกับอีก 2 วิธี สอดคล้องกับงานวิจัยของ Waritpon et al. (2022) ซึ่งทำการศึกษาเปรียบเทียบกันระหว่างแบบจำลองการเรียนรู้ของเครื่องเพื่อจำแนกว่าลูกค้าจะเปิดบัญชีเงินฝาก พบว่าวิธีป่าไม้สุ่ม มีค่าความแม่น ค่าความเที่ยง ค่าประสิทธิภาพโดยรวม สูงที่สุด เนื่องจาก วิธีป่าไม้สุ่ม เป็นวิธีที่สุ่มผสมผสานคุณสมบัติของการแบ่งส่วน และการตัดสินใจของต้นไม้หลาย ๆ ต้นไม้เข้าด้วยกัน ทำให้มีความยืดหยุ่นในการจัดการกับข้อมูลที่ซับซ้อน และเกิดการเรียนรู้แบบรวมกลุ่ม (Ensemble Learning) ที่มีประสิทธิภาพสูง สามารถทำงานกับข้อมูลที่มีคุณลักษณะที่หลากหลาย และไม่ว่าจะเป็นข้อมูลที่มีจำนวนตัวอย่างน้อยหรือจำนวนตัวอย่างมาก ซึ่งช่วยให้ได้ความแม่นยำและประสิทธิภาพที่ดีในการทำนาย

ซึ่งงานวิจัยนี้ได้วิเคราะห์ความแม่นตรงตัวแบบพยากรณ์ ด้วยการตรวจสอบไขว้แบ่งชั้น 10 Fold ซึ่ง แต่ละ Fold อยู่ในช่วง 0.9715 ถึง 0.9743 ซึ่งเป็นค่าที่สูงและใกล้เคียงกัน ซึ่งบ่งบอกว่าแบบจำลองวิธีป่าไม้สุ่ม มีความสามารถในการพยากรณ์ข้อมูลอย่างมีประสิทธิภาพ และมีความเสถียรในการทำงาน ค่าเฉลี่ยอยู่ที่ 0.9727 หรือประมาณ 97.27 เปอร์เซ็นต์ และการวิเคราะห์กราฟเส้นโค้ง ROC Curve มีลักษณะเส้นการวาดที่ใกล้เคียงกับเส้นทแยงมุม ซึ่งเส้นทแยงมุมแสดงถึงการพยากรณ์แบบสุ่ม เส้น ROC อยู่ใกล้เคียงกับส่วนบนซ้ายของกราฟ จึงแสดงถึงความแม่นยำที่สูงและสามารถพยากรณ์คลาสบวกได้ดี มีพื้นที่ใต้เส้น ROC ที่กว้าง แสดงถึงความแม่นยำสูง ในการตรวจจับคลาสบวกและคลาสลบ มีค่า AUC (Area Under the Curve) เท่ากับ 0.98 อ้างอิงจากตารางเกณฑ์การตัดสินใจทั่วไปของค่าประมาณพื้นที่ใต้โค้ง ROC หรือ AUC ซึ่งมีเกณฑ์ที่ค่อนข้างสูงมาก ดังนั้น วิธีป่าไม้สุ่ม จึงเป็นแบบจำลองที่เหมาะสมสำหรับการพยากรณ์ ลูกค้าที่มีโอกาสเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท (บัญชีเงินฝากประเภท D)

5.1.3 เพื่อตอบวัตถุประสงค์ข้อที่ 3 คือ เพื่อทดสอบการนำแบบจำลองไปใช้งาน ในการส่งแคมเปญเสนอขายบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท ซึ่งงานวิจัยนี้ได้ทำการเลือกแบบจำลอง วิธีป่าไม้สุ่ม ที่เป็นแบบจำลองที่เหมาะสมที่สุดนำมาใช้พยากรณ์กลุ่มลูกค้า ที่มีบัญชีออมทรัพย์แบบธรรมดา มีบัตรเครดิตอย่างน้อย 1 ประเภท และไม่มีบัญชีเงินฝากประเภทดิจิทัลดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท (บัญชีเงินฝากประเภท D) จำนวนทั้งสิ้น 248,675 คน พบว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แบบจำลองได้พยากรณ์กลุ่มที่มีโอกาสเปิดบัญชี (Y) จำนวน 2,394 คน และกลุ่มที่มีโอกาสไม่เปิดบัญชีจำนวน (N) 246,281 คน โดยการส่งแคมเปญจะมีการสร้างข้อความเฉพาะบุคคลให้แก่ลูกค้าแต่ละกลุ่มที่แตกต่างกัน 4 ประเภท

ข้อความประเภท 1ก) เป็นข้อความที่ถูกส่งให้แก่กลุ่มลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

ข้อความประเภท 1ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 18 ถึง 40 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

ข้อความประเภท 2ก) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่การใช้บัตรเครดิตในการชำระด้านการลงทุนอย่างน้อย 1 ครั้ง

ข้อความประเภท 2ข) เป็นข้อความที่ถูกส่งให้แก่ลูกค้าที่มีอายุ 41 ถึง 90 ปี และเป็นกลุ่มที่ไม่มีการใช้บัตรเครดิตในการชำระด้านการลงทุนแม้แต่ครั้งเดียว

จากการส่งแคมเปญให้แก่กลุ่มลูกค้าที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชีและไม่มีโอกาสเปิดบัญชีพร้อมกันโดยแยกข้อความเฉพาะบุคคลให้แก่ลูกค้าทั้ง 4 กลุ่ม แต่ละกลุ่มมีการตอบกลับเข้ามาใช้ผลิตภัณฑ์ของธนาคารที่แตกต่างกัน โดยจะเปรียบเทียบแต่ละกลุ่มดังตารางต่อไปนี้

ตารางที่ 5.1 เปรียบเทียบการเข้ามาใช้ผลิตภัณฑ์จำแนกตามประเภทข้อความ 1ก, 1ข, 2ก และ 2ข

ประเภทข้อความ	กลุ่ม	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
1ก	อายุ 18 ถึง 41 ปี และใช้บัตรเครดิตชำระด้านการลงทุน	13,012	5	0.0384
1ข	อายุ 18 ถึง 41 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน	69,562	30	0.0431
2ก	อายุ 41 ถึง 90 ปี และใช้บัตรเครดิตชำระด้านการลงทุน	25,781	10	0.0388
2ข	อายุ 41 ถึง 90 ปี และไม่ใช้บัตรเครดิตชำระด้านการลงทุน	140,320	29	0.0207

จากตารางที่ 5.1 แสดงเปรียบเทียบการเข้ามาใช้ผลิตภัณฑ์จำแนกตามประเภทข้อความ 1ก, 1ข, 2ก และ 2ข พบว่า ข้อความที่มีลูกค้าเข้ามาใช้ผลิตภัณฑ์มากที่สุดคือ ประเภทข้อความ 1ข เข้ามาใช้ผลิตภัณฑ์ 30 คน รองลงมาคือประเภทข้อความ 2ข, 2ก และ 1ก ที่มีการเข้ามาใช้ผลิตภัณฑ์เท่ากับ 29 10 และ 5 ตามลำดับ หากวัดจากเปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์พบว่า ประเภทข้อความที่มี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์สูงที่สุดคือ ประเภทข้อความ 1ข ร้อยละเท่ากับ 0.0431 รองลงมาคือประเภทข้อความ 2ก, 1ก และ 2ข เปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0388 0.0384 และ 0.0207 ตามลำดับ

จากการประยุกต์การเรียนรู้ของเครื่องสำหรับการส่งแคมเปญเสนอขายผลิตภัณฑ์ของธนาคารแห่งหนึ่ง ได้พยากรณ์กลุ่มลูกค้าที่มีบัญชีออมทรัพย์แบบธรรมดา มีบัตรเครดิต แต่ไม่มีบัญชีเงินฝากประเภท D ทั้งสิ้น 248,675 คน พบว่า 2,394 คน เป็นกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y) จึงเป็นกลุ่มเป้าหมายของการส่งแคมเปญครั้งนี้ โดยหลังจากสร้างข้อความเฉพาะบุคคลให้แก่ลูกค้า 2,394 คน ได้ส่งข้อความเสนอขายผลิตภัณฑ์บัญชีเงินฝากประเภท D วันที่ 20 เมษายน พ.ศ. 2566 และรอเก็บผลวันที่ 11 พฤษภาคม พ.ศ. 2566 เป็นเวลา 21 วัน พบว่า มีลูกค้าเข้ามาใช้ผลิตภัณฑ์ เท่ากับ 2 คน คิดเป็นเปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์ร้อยละ 0.0835 หากเปรียบเทียบกับแคมเปญในอดีตเมื่อปี พ.ศ. 2565 ที่ได้ส่งแคมเปญเกี่ยวกับบัญชีเงินฝากประเภท D จะสามารถเปรียบเทียบได้ดังตารางต่อไปนี้

ตารางที่ 5.2 เปรียบเทียบแคมเปญที่ประยุกต์การเรียนรู้ของเครื่องกับแคมเปญในอดีต

ชื่อแคมเปญ	จำนวนที่ส่งแคมเปญทั้งหมด (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)	ระยะเวลาเก็บผล (วัน)
X sell deposit special rate type D	2,746	0	0	31
Apply Machine Learning Campaign	2,394	2	0.0835	21

จากตารางที่ 5.2 แสดงการเปรียบเทียบแคมเปญที่ประยุกต์การเรียนรู้ของเครื่องกับแคมเปญจะพบว่า แคมเปญที่ไม่ได้ประยุกต์ใช้การเรียนรู้ของเครื่องชื่อแคมเปญว่า X sell with deposit special rate type D คือแคมเปญในอดีตที่ถูกส่งให้แก่ลูกค้า 2,746 คน ระยะเวลาเก็บผล 31 วัน มีการเข้ามาใช้ผลิตภัณฑ์เท่ากับ 0 คน คิดเป็นร้อยละ 0 แต่แคมเปญที่ประยุกต์ใช้การเรียนรู้ของเครื่องชื่อแคมเปญว่า Apply Machine Learning Campaign คือแคมเปญที่ส่งให้แก่ลูกค้า 2,394 คน ระยะเวลาเก็บผล 21 วัน มีลูกค้าเข้ามาใช้ผลิตภัณฑ์หรือเปิดบัญชี 2 คน คิดเป็นร้อยละ 0.0835

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 5.3 เปรียบเทียบการเข้ามาใช้ผลิตภัณฑ์จำแนกตามการพยากรณ์ด้วยแบบจำลอง

แบบจำลองพยากรณ์	**จำนวนที่ส่งแคมเปญทั้งหมด (คน)**	**การเข้ามาใช้ผลิตภัณฑ์ (คน)**	**การเข้ามาใช้ผลิตภัณฑ์ (%)**
มีโอกาสเปิดบัญชี (Y)	2,394	2	0.0835
มีโอกาสไม่เปิดบัญชี (N)	246,281	72	0.0292

ตารางที่ 5.3 แสดงการเปรียบเทียบการเข้ามาใช้ผลิตภัณฑ์ของแคมเปญที่ส่งไปจำแนกตามแบบจำลอง การเข้ามาใช้ผลิตภัณฑ์ของกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y) เท่ากับ 2 คน คิดเป็นร้อยละเท่ากับ 0.0835 ซึ่งการเข้ามาใช้ผลิตภัณฑ์ของกลุ่มที่แบบจำลองพยากรณ์ว่าไม่มีโอกาสเปิดบัญชี (N) เท่ากับ 72 คน คิดเป็นร้อยละ 0.0292 หากเปรียบเทียบเปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์จากสัดส่วนที่ส่งแคมเปญไปพบว่า กลุ่มลูกค้าที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y) มีเปอร์เซ็นต์การเข้ามาใช้บริการสูงกว่า กลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสไม่เปิดบัญชี (N)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 5.4 เปรียบเทียบการเข้ามาใช้ผลิตภัณฑ์จำแนกตามการพยากรณ์ด้วยแบบจำลองและประเภทข้อความ

		ประเภทข้อความ	แบบจำลองพยากรณ์	จำนวนที่ส่งแคมเปญ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (คน)	การเข้ามาใช้ผลิตภัณฑ์ (%)
อายุ 18 ถึง 40 ปี	ใช้บัตรเครดิต	1ก	Y	152	0	0
			N	12,860	5	0.0388
	ไม่ใช้บัตรเครดิต	1ข	Y	490	2	0.4081
			N	69,072	28	0.0405
อายุ 41 ถึง 90 ปี	ใช้บัตรเครดิต	2ก	Y	365	0	0
			N	25,416	10	0.0393
	ไม่ใช้บัตรเครดิต	2ข	Y	1,387	0	0
			N	138,933	29	0.0202

ตารางที่ 5.4 ตารางแสดงการเปรียบเทียบการเข้ามาใช้ผลิตภัณฑ์จำแนกตามกลุ่มข้อความและแบบจำลองพยากรณ์ พบว่ากลุ่มที่มีเปอร์เซ็นต์การเข้ามาใช้ผลิตภัณฑ์มากที่สุดคือ กลุ่มประเภทข้อความ B และเป็นกลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี (Y) ร้อยละ 0.4081 แต่อย่างไรก็ตาม หากพิจารณากลุ่มที่แบบจำลองพยากรณ์ว่ามีโอกาสเปิดบัญชี ที่จำแนกตามประเภทข้อความ 1ก, 2ก และ 2ข พบว่า ไม่มีลูกค้าเข้ามาใช้ผลิตภัณฑ์แม้แต่คนเดียว

ดังนั้นแบบจำลองที่สร้างขึ้น ยังคงมีประสิทธิภาพไม่เพียงพอสำหรับการพยากรณ์ลูกค้าที่มีโอกาสเปิดบัญชีเงินฝากประเภทดิจิทัล ดอกเบี้ยสูง ฝากไม่เกิน 1 แสนบาท (บัญชีเงินฝากประเภท D) หรือ อาจจะมีปัจจัยอื่นที่ส่งผลกระทบต่อการเข้ามาใช้ผลิตภัณฑ์ของลูกค้าในการสร้างแคมเปญแบบแบ่งข้อความเฉพาะบุคคล เป็นต้น

5.2 ข้อจำกัดและข้อเสนอแนะ

5.2.1 ข้อจำกัด

1) เนื่องจากแบบจำลองการวิธีการเรียนรู้เครื่องมีหลากหลายให้เลือกใช้ งานวิจัยนี้ได้เลือกใช้แบบจำลองเพียง 3 วิธี ได้แก่ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีนาอีฟเบย์ และ วิธีป่าไม้สุ่ม ซึ่งเป็นแบบจำลองที่ได้รับความนิยม เข้าใจง่าย ดังนั้นจึงอาจจะมีแบบจำลองอื่น ๆ ที่สามารถทำให้ค่าประสิทธิภาพการพยากรณ์ได้ดีกว่าหรือเหมาะสมมากกว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2) การรอผลการเข้ามาใช้ผลิตภัณฑ์มีเวลาจำกัดเพียง 21 วัน ซึ่งหากมีเวลาเก็บผลที่มากขึ้น คาดว่าผลการเข้ามาใช้ผลิตภัณฑ์อาจจะเปลี่ยนไป

3) ฮาร์ดแวร์ที่ใช้ในงานวิจัยนี้มีข้อจำกัด เช่น ความเร็วของเครื่องคอมพิวเตอร์พกพา ที่เกิดจากการเสื่อมสภาพของอุปกรณ์ ส่งผลให้การพัฒนาแบบจำลองใช้เวลานาน

4) ข้อความแคมเปญที่ส่งไปหาลูกค้าทั้ง 4 กลุ่ม มีความแตกต่างกัน จึงอาจจะเป็นปัจจัยที่ทำให้ผลการเข้ามาใช้ผลิตภัณฑ์ที่แตกต่างกัน

5.2.1 ข้อเสนอแนะ

1) นอกจากการเรียนรู้ของเครื่อง (Machine Learning) แล้ว ยังมีการเรียนรู้ในรูปแบบอื่น ๆ หากผู้วิจัยท่านถัดไปจะพัฒนาแบบจำลองเพื่อต่อยอด งานวิจัยนี้จึงเสนอหลักการแบบจำลองการเรียนรู้เชิงลึก (Deep Learning) เพิ่มเติม เช่น Recurrent Neural Networks (RNN), Multilayer Perceptrons (MLPs), Radial Basis Function Networks (RBFNs) ซึ่งอาจจะให้ประสิทธิภาพในการพยากรณ์ดียิ่งขึ้น

2) งานวิจัยนี้ใช้เทคนิค SMOTE และ ENN เข้ามาแก้ไขความไม่สมดุลของข้อมูล ซึ่งเทคนิคการแก้ไขความไม่สมดุลมีหลากหลาย ควรมีการทดสอบการจัดการปัญหาความไม่สมดุลของข้อมูลด้วยวิธีอื่น ๆ เพิ่มเติม เพื่อเปรียบเทียบประสิทธิภาพของแต่ละวิธี ที่ทำให้แบบจำลองมีประสิทธิภาพดีขึ้น

3) การสร้างแบบจำลองพยากรณ์ข้อมูลที่ดี ควรมีการปรับปรุงแบบจำลองอยู่บ่อยครั้ง เพื่อให้แบบจำลอง สามารถเรียนรู้กับข้อมูลชุดใหม่ได้อย่างหลากหลาย จึงต้องนำข้อมูลเข้ามาฝึกสอนแบบจำลองอย่างสม่ำเสมอ

4) เนื่องจากมีผลิตภัณฑ์ที่หลากหลาย การพิจารณาการสร้างแบบจำลองควรมีการคำนึงถึงตัวแปรที่เกี่ยวข้องทั้งหมด ซึ่งอาจจะต้องใช้ผู้เชี่ยวชาญเฉพาะทางมาให้คำแนะนำในการคัดเลือกตัวแปร

5) ควรจะพิจารณาพฤติกรรมการใช้เงินของลูกค้า เช่น ช่วงเงินโบนัสออก ช่วงเงินเดือนออก ฯลฯ เนื่องจากพฤติกรรมดังกล่าว อาจจะเป็นปัจจัยที่ทำให้ลูกค้าเข้ามาใช้ผลิตภัณฑ์ของธนาคารมากขึ้น ดังนั้นควรพิจารณาการส่งแคมเปญในช่วงระยะเวลาที่เหมาะสม หรือ ระยะเวลาในการรอผลการเข้ามาใช้ผลิตภัณฑ์ที่อาจจะทำให้ผลลัพธ์ของการส่งแคมเปญเปลี่ยนไป

6) ข้อความที่ส่งแคมเปญควรมีการวางแผนการสร้างให้คอบคลุมมากขึ้น มีข้อความทดลอง (Experiment) และ ข้อความควบคุม (Control) เพื่อให้การวัดผลเป็นไปได้อย่างแม่นยำมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

คณะกรรมการกิจการกระจายเสียง กิจการโทรทัศน์ และกิจการโทรคมนาคมแห่งชาติ. 2563. **ข้อมูลสถิติโทรคมนาคม.** [Online]. เข้าได้ถึงจาก https://www.nbtc.go.th.

จิตพนธ์ ชุมเกต. 2560. **การพัฒนาผลิตภัณฑ์จากภูมิปัญญาท้องถิ่นเพื่อเพิ่มประสิทธิภาพทางการจัดการชุมชนอย่างยั่งยืนของชุมชนไทยมุสลิม อำเภอชะอำ จังหวัดเพรชบุรี.** ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการจัดการ. มหาวิทยาศิลปากร.

จิรภา โคมเดือน. 2560. **การเปรียบเทียบประสิทธิภาพของสถิติทดสอบค่าเฉลี่ยของสามประชากร.** วิทยาศาสตรมหาบัณฑิต สาขาสถิติประยุกต์. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

จิรวัฒน์ จันทองพูน, พัณณิตา ไล่สาม, สัณห์ฤทัย แซ่หว่อง, และ พรนรายณ์ บุญราศรี. **การศึกษาการขยายตัวของเมืองด้วยเทคนิควิธีป่าไม้สุ่ม กรณีศึกษา อำเภอเมืองสงขลา จังหวัดสงขลา.** 24-26 ในการประชุมวิชาการวิศวกรรมโยธาแห่งชาติ ครั้งที่ 27. สงขลา.

ณัฏฐณิชา ยงยิ่ง. 2562. **การประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกในการจำแนกข้อมูลถนนจากภาพถ่าย Drone เพื่อการสำรวจถนนในเขตชนบท.** วิทยาศาสตรบัณฑิต สาขาวิชาภูมิศาสตร์. มหาวิทยาลัยนเรศวร.

ธนาภัทร ภัทรวินิจ. 2563. **การทำนายความผิดพลาดระยะต้นของเครื่องวิเคราะห์อินทรีย์คาร์บอนโดยการเรียนรู้เชิงลึก.** วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์. จุฬาลงกรณ์มหาวิทยาลัย.

ปิยวรรณ นิลถนอม, ธนพร มาลัย และสายชล สินสมบูรณ์ทอง. 2564. **การเปรียบเทียบประสิทธิภาพการทานายผลการแปลงข้อมูลในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล.** วารสารวิทยาศาสตร์และเทคโนโลยีไทย. 1: 15-24

พงศกร พลธีระเสถียร. 2565. **การศึกษาเปรียบเทียบประสิทธิผลการโฆษณาที่ทำให้เกิดการซื้อสินค้า ระหว่างแพลตฟอร์มกูเกิล (Google Ads) และเฟซบุ๊ก(Facebook Ads) กรณีศึกษา สินค้าแบรนด์ “พริกแกงน้ำใจ” ภายใต้บริษัท เคอร์รี่ แอนด์ สไปร์จำกัด.** นิเทศศาสตรมหาบัณฑิต สาขาวิชาการสื่อสารการตลาดดิจิทัล. มหาวิทยาลัยกรุงเทพ.

พัชนา สุวรรณแสน. 2562. **การจัดการข้อมูลสูญหาย วิธีเคเนียร์เรสเนเบอร์.** วารสารวิจัยวิทยาศาสตร์และเทคโนโลยี. 1: 1-9.

เพ็ญศรี บางบอน, พอตา วุฒิพรชัย และรุจิภา สินสมบูรณ์ทอง. 2560. **ระบบปัญญาประดิษฐ์กับการปฏิวัติอุตสาหกรรม.** วารสารวิชาการสถาบันวิทยาการจัดการแห่งแปซิฟิค. 4: 213-218.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง (ต่อ)

ภาสพิชญ์ ชูใจ. 2557. **การเรียนรู้ร่วมกันสำหรับปัญหาการจำแนกข้อมูลไม่สมดุล.** วิศวกรรมศาสตรดุษฎีบัณฑิต. สาขาวิชาวิศวกรรมคอมพิวเตอร์. มหาวิทยาลัยเทคโนโลยีสุรนารี

มาณวิกา กิตติพร. 2562. **ระบบแนะนำทางการศึกษาและเทคนิคการเรียนรู้ของเครื่องจักร.** วารสารสารสนเทศศาสตร์. 4: 98-99.

รัชพล กลัดชื่น และจรัญ แสนราช. 2561. **การเปรียบเทียบประสิทธิภาพอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสม เพื่อการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษาระดับอาชีวศึกษา.** วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี. 17: 1-10.

วีระพันธ์ พานิชย์. 2564. **การประยุกต์ใช้ Machine Learning พยากรณ์ผลการเรียนวิชา Web Database.** วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม. มหาวิทยาลัยธุรกิจบัณฑิตย์.

สุจิตรา สุคนธมัต. 2564. **เอกสารประกอบการเรียนการสอนโปรแกรมสำเร็จรูปทางสถิติ.** ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

สุภาภรณ์ พัฒนวงศ์ปราการ. 2563. **การวิเคราะห์เทคนิคการจำแนกประเภทข้อมูล กรณีศึกษาการพยากรณ์ระดับชั้นผู้รับเหมาก่อสร้างสำหรับโครงการก่อสร้างของภาครัฐ.** วิทยาศาสตรมหาบัณฑิต สาขาวิชาระบบสารสนเทศเพื่อการจัดการ. มหาวิทยาลัยธรรมมศาสตร์

สุเมธ จุฑาจันทร์ และสมพร ปั่นโภชา. 2564. **เปรียบเทียบผลลัพธ์ของการอนุมัติสินเชื่อด้วย 3 แบบจำลองของระเบียบวิธีการเรียนรู้ของเครื่องโดยใช้โปรแกรมอาร์.** 189-199. ในการประชุมนำเสนอผลงานวิจัยบัณฑิตศึกษาระดับชาติ ครั้งที่ 15. ปทุมธานี

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง. 2560. **การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย.** วารสารวิทยาศาสตร์และเทคโนโลยี. 5: 840-853.

อภิชญาฎา ภู่ระหงษ์ และบุษรา โพวาทอง. 2563. **การวางแผนทางการเงินเพื่อการอยู่อาศัยในวัยเกษียณสาหรับกลุ่มคนอายุ 40 ปีขึ้นไป ในเขตกรุงเทพมหานคร.** วารสารวิชาการสาระศาสตร์. 4: 867-878.

เอกพันธ์ บุญเสริม. 2563. **การประยุกต์ใช้การเรียนรู้ของเครื่องในการพยากรณ์ความรุนแรงของผู้บาดเจ็บจากอุบัติเหตุทางถนนในช่วงเทศกาลปีใหม่จากข้อมูลเปิดภาครัฐของประเทศไทย.** วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ. มหาวิทยาลัยศรีนครินทรวิโรฒ.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง (ต่อ)

Blessie, E and Karthikeyan, E. 2012. **A feature selection algorithm using correlation Based Method.** Journal of Algorithms & Computational Technology 3: 387-388

Brown, R. 2019. **What are features in Machine Learning.** [Online]. https://cogitotech.medium.com

Dietrich, D., Heller, B. and Yang, B. 2015. **Data science & big data analytics discovering, analyzing, visualizing and presenting data. Indiana.** john wiley & sons,Inc.

Donghai, G. Weiwei, Y. Young-Koo, L. and Sungyoung, L 2009. **Nearest neighbor editing aided by unlabeled data.** Information Sciences. 179: 2273–2282.

Feng, H. and Hang, L. 2013. **A novel boundary oversampling algorithm based on neighborhood rough set model: nrsboundary-smote.** hindawi Publishing Corporation. Mathematical Problems in Engineering. 1: 1-10.

Fridh, D. and Dahl, T. 2019. **A consumer perspective of personalized marketing an exploratory study on consumer perception of personalized marketing and how it affects the purchase decision making.** bachelor of science. högskolan kristianstad.

Guiso, L. and Japelli, T. 2009. **Financial literacy and portfolio diversification.** bachelor of economics. university of naples.

Hawkins, D. 2003. **The problem of overfitting. school of statistics.** university of minnesota.

Ibrahim, O. and Osman, A. 2014. **A novel feature selection based on one-way anova f-test for e-mail spam classification.** journal of applied sciences engineering and technology. 7: 625-638.

Kubalik, J. and Sramek, M. 2019. **Addressing the curse of imbalanced training sets one-sided selection. international journal of artificial intelligence and applications.** 10: 35-47.

Kumar, S. 2023. **An outliers detection and elimination framework in classification task of data mining.** Decision Analytics Journal. 6: 1-8.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง (ต่อ)

Lopez, A. Fernandez, J. Moreno-Torres, G. and Herrera, F. 2012. **Analysis of preprocessing vs cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics.** Expert Systems with Applications, 7: 145-146.

Prusty, S. Patnaik, S. and Dash, S. 2022. **Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. frontiers in nanotechnology.** 4:972421.

Sanou, A., Liverpool, L. and Shupp, R. 2018. **Eliciting risk attitudes in the field: surveys or experimental methods an empirical comparison in rural niger.** The Journal of Development Studies. 8: 1450-1470

Shearer, C. 2000. **The crisp-dm model the new blueprint for data mining.** Journal of Data Warehousing. 5: 13–22.

Wang, K. Tian, J. Zheng, C. Yang, H. Ren, J. Li, C. Han, Q. Zhang, Y 2021. **Improving risk Identification of adverse outcomes in chronic heart failure using smote+enn and machine learning.** Risk Management and Healthcare Policy. 14: 2453–2463.

Waritpon, S. Anamai, N. and Jaratsri, R. 2022. **A comparison of machine learning techniques for classification in bank marketing data.** thai journal of mathematics. 17: 157-168.

Yu, L. Liu, H. 2003. **Feature Selection for high-dimensional data a fast correlation-based filter solution. bachelor of science (computer science).** arizona state university.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ก

คำสั่งไพทอนที่ใช้ในการสร้างแบบจำลอง

ตารางที่ ก.1 ชุดคำสั่งไพทอนในการนำเข้าไลบรารี่ทั้งหมด

```
import warnings
warnings.simplefilter(action='ignore',category=FutureWarning)
import pandas as pd
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.graph_objects as go
import plotly.express as px
from pandas.api.types import CategoricalDtype
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import KFold,cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.2 ชุดคำสั่งไพทอนในการนำข้อมูลเข้า

```
# ระบุรายชื่อไฟล์ CSV ที่ต้องการอ่าน
files = ['ตำแหน่งไฟล์ที่ต้องการอ่าน']
# สร้าง DataFrame เพื่อรวมข้อมูลทั้งหมด
df = pd.DataFrame()
#สร้าง for loop เพื่ออ่านหลายไฟล์
for file in files:
    data = pd.read_csv(file) # อ่านไฟล์ CSV
# รวมข้อมูลเข้าด้วยกันใน DataFrame
    df= df.append(data,ignore_index=True
# ตรวจสอบข้อมูลที่รวมเสร็จแล้ว
print(df.head()) # สามารถแสดงตัวอย่างข้อมูลเบื้องต้นได้
```

ตารางที่ ก.2 ชุดคำสั่งไพทอนลบข้อมูลที่สูญหาย

```
#ลบค่าสูญหายออกจากข้อมูล
df.dropna(inplace=True)
print(df .isnull().sum().sum())
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.3 ชุดคำสั่งไพทอนกำจัดค่านอกเกณฑ์ด้วย **IQR**

```
#สร้างตัวแปรเก็บค่าควอร์ไทล์ของ age
Q1=np.percentile(df["AGE"],25)
Q3=np.percentile(df["AGE"],75)
#คำนวณค่า IQR
IQR= Q3-Q1
lower_limit = Q1-1.5*IQR
upper_limit = Q3+1.5*IQR
#ลบค่าที่อยู่นอกช่วง IQR
mask = (df['AGE']>=lower_limit)&(df['AGE']<=upper_limit)
df=df.drop(df[(df['AGE']<lower_limit)|(df['AGE']>upper_limit)].index)
```

ตารางที่ ก.3 ชุดคำสั่งไพทอนกำจัดค่านอกเกณฑ์ด้วยแผนภาพการกระจาย

```
#สร้างแผนภาพการกระจาย
fig, ax = plt.subplots(figsize = (10,10))
ax.scatter(df[‘ตัวแปร X’],df[‘ตัวแปร Y’], c ="blue")
# ค่าในแกน x
ax.set_xlabel(‘ตัวแปร X’’)
#  ค่าในแกน y
ax.set_ylabel(‘ตัวแปร Y’)
plt.show()
#ลบค่านอกเกณฑ์ที่กำหนด
df=df.drop(df[df[‘แกน X และ Y ที่ต้องการลบ’]>’กำหนดค่าต้องการลบ’].index)
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.4 ชุดคำสั่งไพทอนรวมคุณลักษณะเข้าด้วยกัน

```
#รวมคุณลักษณะ cc_spending

df['CC_EXPENSE_AVG'] = np.where((df['V1'] + df['V2'] + df['V3']) == 0, 0, (df['V1'] +
df['V2'] + df['V3'])/(3))

#รวมคุณลักษณะ total_mf_os

df['TOTAL_MF_AVG'] = np.where((df['TOTAL_MF_OS_01'] + df['TOTAL_MF_OS_02']
+ df['TOTAL_MF_OS_03']) == 0, 0, (df['TOTAL_MF_OS_01'] + df['TOTAL_MF_OS_02']
+ df['TOTAL_MF_OS_03'])/3)

#รวมคุณลักษณะ sum_os_ba

df['SUM_OS_BA_AVG'] = np.where((df['SUM_OS_BA01'] + df['SUM_OS_BA02'] +
df['SUM_OS_BA03']) == 0, 0, (df['SUM_OS_BA01'] + df['SUM_OS_BA02'] +
df['SUM_OS_BA03'])/3)
```

ตารางที่ ก.5 ชุดคำสั่งไพทอนแปลงข้อมูลด้วยการเข้ารหัสป้าย

```
#เรียกฟังก์ชัน

le=LabelEncoder()

#แปลงตัวแปร

df [‘ตัวชื่อตัวแปรใหม่’]=le.fit_transform(df[ชื่อตัวแปรเดิม])
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.6 ชุดคำสั่งไพทอนแปลงข้อมูลด้วยค่าต่ำสุดและสูงสุด

```
# สร้าง MinMaxScaler object

scaler = MinMaxScaler()

# ทำการสร้าง DataFrame ใหม่ที่เป็น copy ของ DataFrame เดิม

x_scaled = x.copy()

# ทำการ fit และ transform ข้อมูลใน DataFrame ใหม่

x_scaled[['ตัวแปรต้องการเปลี่ยน']] = scaler.fit_transform(x_scaled[['ตัวแปร']])
```

ตารางที่ ก.7 ชุดคำสั่งไพทอนสร้างตารางสัมประสิทธิ์สหสัมพันธ์เพื่อคัดเลือกคุณลักษณะ

```
#สร้างตารางสัมประสิทธิ์สหสัมพันธ์เพื่อคัดเลือกคุณลักษณะ

corr=x_scaled.corr()

plt.figure(figsize=(20,10))

sns.heatmap(corr,annot=True)
```

ตารางที่ ก.8 ชุดคำสั่งไพทอนวิเคราะห์ความแปรปรวนเพื่อคัดเลือกคุณลักษณะ

```
#สร้างตารางความแปรปรวน

f_scores,p_values=f_classif('ตัวแปร X'))

for i in range(len('ตัวแปร X')):
```

ตารางที่ ก.9 ชุดคำสั่งไพทอนการทำ smote และ enn

```
# กำหนดตัวเลือกการทำ SMOTEENN และปรับค่า sampling_strategy

smote_enn = SMOTEENN(sampling_strategy=0.1/1,random_state=42)

# oversample และ undersample

x_resampled, y_resampled = smote_enn.fit_resample('ตัวแปร X','ตัวแปร Y')
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.10 ชุดคำสั่งไพทอนสร้างแบบจำลอง 3 วิธี

```
#แบ่งชุดข้อมูลฝึกสอนและข้อมูลทดสอบ 80:20

x_train, x_test, y_train, y_test = train_test_split('ตัวแปร X','ตัวแปร Y', test_size=0.2,
random_state=42)

# สร้างโมเดลเพื่อนบ้านใกล้สุด k ตัวและกำหนดพารามิเตอร์

knn = KNeighborsClassifier(n_neighbors=125,
                           weights='uniform',
                           metric='euclidean')

# สร้างโมเดลนาอีฟเบย์และกำหนดพารามิเตอร์

nb = GaussianNB(var_smoothing=1e-9,priors=[0.2, 0.8])

# สร้างโมเดลนาป่าไม้สุ่มและกำหนดพารามิเตอร์

rfc = RandomForestClassifier(n_estimators=1200,
                           max_depth=30,
                           min_samples_split=5,
                           min_samples_leaf=3,
                          class_weight='balanced')

#ฝึกสอนโมเดล

'ชื่อฟังก์ชันที่เก็บค่าพารามิเตอร์ไว้'.fit(x_train,y_train.values.ravel())
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.11 ชุดคำสั่งไพทอนสร้างเส้นโค้ง **ROC**

```
# พยากรณ์ผลลัพธ์ของชุดข้อมูลทดสอบ
y_prob = ‘ตัวแปรเก็บแบบจำลองที่ฝึกสอน’.predict_proba(x_test)
# สกัดความน่าจะเป็นของคลาสบวก
y_prob_pos = y_prob[:, 1]
# แสดงผล ROC curve
fpr, tpr, thresholds = roc_curve(y_test, y_prob_pos, pos_label='DEPOSIT')
auc_score = auc(fpr, tpr)
plt.plot(fpr, tpr, 'b', label='ROC curve (AUC = %0.2f)' % auc_score)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([-0.05, 1.05])
plt.ylim([-0.05, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic (ROC) curve')
plt.legend(loc="lower right")
plt.show()
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ ก.12 ชุดคำสั่งไพทอนวิเคราะห์ความแม่นตรง

```
# สร้าง StratifiedKFold object และกำหนดให้มีจำนวน fold เท่ากับ 10
stratified_kfold = StratifiedKFold(n_splits=10)
# ใช้ cross_val_score กับ StratifiedKFold และกำหนด scoring เป็น 'accuracy'
cv_scores = cross_val_score(‘ตัวแปรเก็บแบบจำลองที่ฝึกสอน’, x_train, y_train.ravel(), cv=stratified_kfold, scoring='accuracy')
# แสดงผลค่า accuracy ที่ได้จาก cross-validation
print(f'Stratified Cross-validation scores: {cv_scores}')
print(f'Mean accuracy score: {np.mean(cv_scores):.2f}')
```

ตารางที่ ก.13 ชุดคำสั่งไพทอนรายงานผลการทดสอบเมทริกซ์ความสับสน

```
#สร้างฟังก์ชันพยากรณ์ข้อมูล
‘ชื่อตัวแปรที่เก็บข้อมูลการพยากรณ์’ = ‘ชื่อแบบจำลอง’.predict(‘ข้อมูลทดสอบ’)
#รายงานผลการทดสอบ KNN
print("รายงานผลการทดสอบ KNN")
print(classification_report(y_test, y_pred_knn, target_names=[‘ชื่อคลาส’,‘ชื่อคลาส’]))
#รายงานผลการทดสอบ NB
print("รายงานผลการทดสอบ NB")
print(classification_report(y_test, y_pred_nb, target_names=[‘ชื่อคลาส’,‘ชื่อคลาส’]))
#รายงานผลการทดสอบ RFC
print("รายงานผลการทดสอบ RFC")
print(classification_report(y_test, y_pred_rfc, target_names=[‘ชื่อคลาส’,‘ชื่อคลาส’]))
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้