การตรวจจับการฉ้อโกงบัตรเครดิตโดยเทคนิคการเรียนรู้ของเครื่อง

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES

ณัฏฐา ดวงศรี

NATTA DUANGSRI

การศึกษาค้นคว้าอิสระเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาตรมหาบัณฑิต สาขาวิชาสถิติและการวิเคราะห์ธุรกิจ

ภาควิชาสถิติ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2565

KMITL-2022-SC-M-050-119

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES

ณัฏฐา ดวงศรี

NATTA DUANGSRI

An INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS AND BUSINESS ANALYTICS

DEPARTMENT OF STATISTICS SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2022

KMITL-2022-SC-M-050-119

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2022

SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อการค้นคว้าอิสระ การตรวจจับการฉ้อโกงบัตรเครดิตโดยเทคนิคการเรียนรู้ของเครื่อง
ชื่อนักศึกษา นางสาวณัฏฐา ดวงศรี
รหัสประจำตัว 63605123
ปริญญา วิทยาศาสตรมหาบัณฑิต (สถิติและการวิเคราะห์ธุรกิจ)
ภาควิชา สถิติ
พ.ศ. 2565
อาจารย์ที่การค้นคว้าอิสระ ผู้ช่วยศาสตราจารย์ ดร.พรพิมล ชัยวุฒิศักดิ์

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาเทคนิคการจำแนกกลุ่มสำหรับการตรวจจับการฉ้อโกงบัตรเครดิต 5 วิธี ได้แก่ วิธีการถดถอยลอจิสติกทวิภาค วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท พร้อมทั้งเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่มที่ไม่อาศัยและอาศัยเทคนิคการแก้ปัญหาข้อมูลที่ไม่สมดุลกันโดยพิจารณาจาก ค่าความไว ความถูกต้องและพื้นที่ใต้โค้ง ผลลัพธ์จากการศึกษาพบว่าค่าความไวของโมเดลจำแนกกลุ่มที่ใช้เทคนิค SMOTE สำหรับจัดการข้อมูลที่ไม่สมดุล สูงกว่าค่าความไวของโมเดลจำแนกกลุ่มที่ไม่ใช้เทคนิค SMOTE เนื่องจากชุดข้อมูลมีปัญหาความไม่สมดุลสูงมาก โดยร้อยละของรายการข้อมูลทุจริตคือ 0.67 นอกจากนี้วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง เป็นวิธีที่มีความเหมาะสมมากที่สุด สำหรับการทำนายการจำแนกกลุ่ม โดยมีค่าความไวคิดเป็นร้อยละ 88 ค่าความถูกต้องคิดเป็นร้อยละ 87 และ ค่าพื้นที่ใต้โค้งคิดเป็นร้อยละ 87.37

คำสำคัญ : วิธีการถดถอยลอจิสติกทวิภาค, วิธีต้นไม้การตัดสินใจ, วิธีการสุ่มป่าไม้, วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง, วิธีเอดาบูท

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Independent Study Title	Credit Card Fraud Detection Using Machine Learning Techniques
Student Name	Miss Natta Duangsri
Student ID	63605123
Degree	Master of Science (Statistics and Business Analytics)
Department	Statistics
Year	2022
Independent Study Advisor	Asst. Prof. Dr. Pornpimol Chaiwuttisak

Abstract

The purpose of this research was to study five classification methods for credit card fraud detection such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Adaboost (ADA). Along with comparing the efficiency of imbalanced classes by using SMOTE based on recall, consideration into account. recall, accuracy, and area under the curve. The results found that the recall by using SMOTE to solve the imbalanced classes has high performance because of the dataset's highly imbalanced problem. The percentage of fraudulent transactions was 0.67. In addition, Extreme Gradient Boosting is an effective method for classification. The recall was 88%, the accuracy was 87%, and the area under the curve was 87.37% respectively.

Keywords: Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, Adaboost

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

การค้นคว้าอิสระฉบับนี้สำเร็จลุล่วงได้ ด้วยความกรุณาจากผู้ช่วยศาสตราจารย์ ดร.พรพิมล ชัยวุฒิศักดิ์ อาจารย์ที่ปรึกษาการค้นคว้าอิสระ ที่ได้ให้คำแนะนำ คำปรึกษา ช่วยเหลือและตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ตลอดจนติดตามผลงานทุกขั้นตอนของการดำเนินงานในการทำการค้นคว้าอิสระนี้ให้สำเร็จลุล่วงสมบูรณ์ ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณกรรมการสอบการค้นคว้าอิสระ ทั้งสองท่าน ได้แก่ ผู้ช่วยศาสตราจารย์ ดร.สิทธิชัย เจริญเศรษฐศิลป์ และรองศาสตราจารย์ ดร.วลัยลักษณ์ อัตธีรวงศ์ กรรมการที่กรุณาให้คำแนะนำ และข้อเสนอแนะที่เป็นประโยชน์ สำหรับปรับปรุงแก้ไขทำให้การค้นคว้าอิสระเล่มนี้ให้สมบูรณ์ยิ่งขึ้น ตลอดจนกรุณาสละเวลามาเป็นกรรมการสอบการค้นคว้าอิสระ

ขอขอบพระคุณธนาคารพาณิชย์แห่งหนึ่ง ที่เอื้อเฟื้อเผยแพร่ข้อมูล ที่ให้ความอนุเคราะห์ข้อมูลเพื่อใช้สำหรับการทำการค้นคว้าอิสระในครั้งนี้

สุดท้ายนี้ขอกราบขอบพระคุณครอบครัวของข้าพเจ้า ที่คอยเป็นกำลังใจ และให้การสนับสนุนในทุกเรื่องด้วยดีเสมอมา และขอขอบคุณเพื่อน ๆ ที่ให้คำปรึกษาและช่วยเหลือให้การทำงานมาโดยตลอดจนการค้นคว้าอิสระ นี้สำเร็จไปได้ด้วยดี

นางสาวณัฏฐา ดวงศรี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ก
บทคัดย่อภาษาอังกฤษ ..... ข
กิตติกรรมประกาศ ..... ค
สารบัญ ..... ง
สารบัญตาราง ..... ช
สารบัญรูปภาพ ..... ฌ
บทที่ 1 บทนำ 1
1.1 ความสำคัญและที่มาของปัญหา 1
1.2 วัตถุประสงค์ของงานวิจัย 3
1.3 ขอบเขตของงานวิจัย 3
1.4 ประโยชน์ที่คาดว่าจะได้รับ 4
1.5 นิยามศัพท์ 4
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 5
2.1 แนวคิดเกี่ยวกับบัตรเครดิต 5
2.1.1 ประเภทของบัตรเครดิต 7
2.1.2 ความสำคัญของบัตรเครดิต 7
2.2 แนวคิดเกี่ยวกับการฉ้อโกงบัตรเครดิต 8
2.2.1 ประเภทของเทคนิคการตรวจจับการฉ้อโกง 9
2.2.2 ลักษณะการฉ้อโกงบัตรเครดิต 10
2.3 วิธีการวิเคราะห์ถดถอยลอจิสติกทวิภาค 11
2.3.1 ประเภทของการวิเคราะห์ความถดถอยลอจิสติก 12
2.3.2 วัตถุประสงค์การวิเคราะห์ความถดถอยลอจิสติก 12
2.3.3 เงื่อนไขการวิเคราะห์ความถดถอยลอจิสติก 12
2.3.4 ตัวแบบการวิเคราะห์ความถดถอยลอจิสติก 13
2.3.5 การตรวจสอบความเหมาะสมของสมการถดถอยลอจิสติก 14
2.3.6 การทดสอบสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัว 14
2.3.7 การวัดระดับความสัมพันธ์ 15

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

2.4 วิธีต้นไม้การตัดสินใจ....16
2.5 วิธีการสุ่มป่าไม้....18
2.6 วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง....19
2.7 วิธีเอดาบูท....20
2.8 การแทนค่าของตัวแปรประเภทกลุ่ม....21
2.9 เทคนิคการปรับเพิ่มและลดข้อมูล....22
2.9.1 วิธีการสุ่มเพิ่ม....22
2.9.2 วิธีการสุ่มเกินโดยเทคนิค SMOTE....23
2.9.3 วิธีการสุ่มลด....24
2.10 การคัดเลือกคุณลักษณะ....24
2.11 การวัดประสิทธิภาพและการประเมินตัวแบบ....25
2.12 งานวิจัยที่เกี่ยวข้อง....27
บทที่ 3 วิธีการดำเนินงานวิจัย 32
3.1 ขั้นตอนการดำเนินงานวิจัย....32
3.2 ขั้นตอนการปฏิบัติงานของหน่วยงาน Fraud....34
3.3 ตัวแปรที่ใช้การวิจัย....35
3.4 เครื่องมือที่ใช้ในการวิจัย....36
3.5 การเก็บรวบรวมข้อมูล....36
3.6 การเตรียมข้อมูล....37
3.7 การวิเคราะห์การจำแนกกลุ่มของชุดข้อมูล....40
3.8 การวัดประเมินประสิทธิภาพของตัวแบบ....42
บทที่ 4 ผลการวิจัยและอภิปรายผล.... 44
4.1 ผลการสำรวจข้อมูลเบื้องต้น....44
4.2 ผลการวิเคราะห์ความสำคัญของตัวแปร....49
4.3 การตรวจสอบความเหมาะสมของตัวแบบการถดถอยลอจิสติกทวิภาค....50
4.4 ผลการจำแนกรายการธุรกรรมฉ้อโกง....52
4.4.1 วิธีการถดถอยลอจิสติกทวิภาค....52
4.4.2 วิธีต้นไม้การตัดสินใจ....54

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

4.4.3 วิธีการสุ่มป่าไม้....56
4.4.4 วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง....57
4.4.5 วิธีเอดาบูท....59
4.5 ผลการจำแนกรายการธุรกรรมฉ้อโกงโดยอาศัยการปรับความไม่สมดุลด้วยเทคนิค SMOTE....60
4.5.1 วิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE....60
4.5.2 วิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE....62
4.5.3 วิธีการสุ่มป่าไม้ร่วมกับเทคนิคSMOTE....63
4.5.4 วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิคSMOTE....65
4.5.5 วิธีเอดาบูทร่วมกับเทคนิค SMOTE....66
4.6 ผลการเปรียบเทียบการจำแนกรายการธุรกรรมฉ้อโกงระหว่างเทคนิคการจำแนกกลุ่มที่ใช้การปรับความไม่สมดุล และไม่ใช้การปรับความไม่สมดุล....68
4.7 อภิปรายผล....70
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ.... 71
5.1 สรุปผลการวิจัย....71
5.1.1 สรุปผลและตรวจสอบประสิทธิภาพของตัวแบบไม่ใช้การปรับความไม่สมดุล....71
5.1.2 สรุปผลและตรวจสอบประสิทธิภาพของตัวแบบใช้การปรับความไม่สมดุล....71
5.2 ข้อเสนอแนะ....72
เอกสารอ้างอิง.... 73
ภาคผนวก.... 75
ภาคผนวก ก.... 76
ภาคผนวก ข.... 84
ประวัติผู้เขียน....86

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่ **หน้า**

2.1 สรุปการทบทวนวรรณกรรมที่เกี่ยวข้อง 29
3.1 ตัวแปรและระดับค่าของข้อมูลที่ใช้ในการวิจัย 35
3.2 การแปลงข้อมูลตัวแปรที่ใช้ในการวิจัย 37
4.1 ค่าสถิติเชิงพรรณาของตัวแปรเชิงปริมาณ 45
4.2 ตารางแจกแจงความถี่ของตัวแปรเชิงคุณภาพ 45
4.3 การวิเคราะห์ความสำคัญของตัวแปร 49
4.4 ผลลัพธ์จากการพยากรณ์ด้วยเทคนิคการถดถอยลอจิสติกทวิภาค วิธี Forward: LR (Likelihood Ratio) 49
4.5 ค่าความเป็นไปได้ (Likelihood Value) และค่าสัมประสิทธิ์การอธิบาย (Pseudo R Square) 52
4.6 ผลการจำแนกกลุ่มโดยวิธีการถดถอยลอจิสติกทวิภาค 53
4.7 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการถดถอยลอจิสติกทวิภาค 53
4.8 ผลการจำแนกกลุ่มโดยวิธีต้นไม้การตัดสินใจ 54
4.9 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีต้นไม้การตัดสินใจ 55
4.10 ผลการจำแนกกลุ่มโดยวิธีการสุ่มป่าไม้ 56
4.11 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการสุ่มป่าไม้ 56
4.12 ผลการจำแนกกลุ่มโดยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง 57
4.13 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง 58
4.14 ผลการจำแนกกลุ่มโดยวิธีเอดาบูท 59
4.15 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอดาบูท 59
4.16 ผลการจำแนกกลุ่มโดยวิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE 60
4.17 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE 61
4.18 ผลการจำแนกกลุ่มโดยวิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE 62
4.19 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE 62

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง (ต่อ)

4.20 ผลการจำแนกกลุ่มโดยวิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE 63
4.21 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE.....64
4.22 ผลการจำแนกกลุ่มโดยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE..... 65
4.23 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE.....65
4.24 ผลการจำแนกกลุ่มโดยวิธีเอดาบูทร่วมกับเทคนิค SMOTE..... 66
4.25 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอดาบูทร่วมกับเทคนิค SMOTE.....67
4.26 การจำแนกรายการธุรกรรมฉ้อโกงระหว่างเทคนิคการจำแนกกลุ่มที่ไม่ใช้การปรับความไม่สมดุล.....68
4.27 การจำแนกรายการธุรกรรมฉ้อโกงระหว่างเทคนิคการจำแนกกลุ่มที่ใช้การปรับความไม่สมดุล.....68

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่ **หน้า**

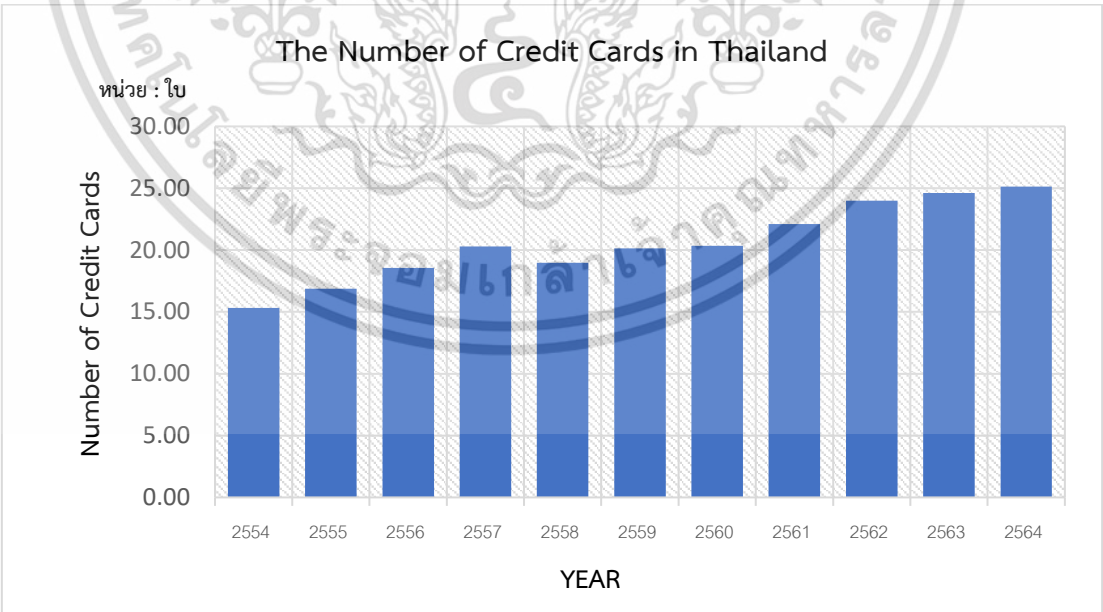
1.1 จำนวนบัตรเครดิตตั้งแต่ปี 2554 - 2564 ในประเทศไทย 1
1.2 มูลค่าการใช้บัตรเครดิตตั้งแต่ปี 2554 - 2564 2
2.1 วิธีการทำงานของการตรวจจับการฉ้อโกง 8
2.2 โครงสร้างต้นไม้ตัดสินใจ 16
2.3 การรวมกลุ่มกันของโครงสร้างต้นไม้ตัดสินใจ 18
2.4 ความแตกต่างระหว่าง Bagging และ Boosting 20
2.5 ตัวอย่างการสุ่มเพิ่ม 22
2.6 ตัวอย่างวิธีการสังเคราะห์ข้อมูลเพิ่ม 23
2.7 ตัวอย่างการสุ่มลด 24
3.1 ขั้นตอนการดำเนินงานวิจัย 33
3.2 ขั้นตอนการปฏิบัติงานของหน่วยงาน Fraud 34
4.1 สัดส่วนการฉ้อโกงของรายการธุรกรรมการใช้บัตรเครดิต 44
4.2 กราฟ Boxplot ของตัวแปร Credit limit และ Amount 48
4.3 กราฟของตัวแปรเชิงคุณภาพ 48
4.4 ค่า AUC ของวิธีการถดถอยลอจิสติกทวิภาค 54
4.5 ค่า AUC ของวิธีต้นไม้การตัดสินใจ 55
4.6 ค่า AUC ของวิธีการสุ่มป่าไม้ 57
4.7 ค่า AUC ของวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง 58
4.8 ค่า AUC ของวิธีเอดาบูท 60
4.9 ค่า AUC ของวิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE 61
4.10 ค่า AUC ของวิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE 63
4.11 ค่า AUC ของวิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE 64
4.12 ค่า AUC ของวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE 66
4.13 ค่า AUC ของวิธีเอดาบูทร่วมกับเทคนิค SMOTE 67
4.14 เปรียบเทียบเทคนิคการจำแนกกลุ่มที่ใช้การปรับความไม่สมดุลโดย SMOTE และไม่ใช้การปรับความไม่สมดุล 69

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

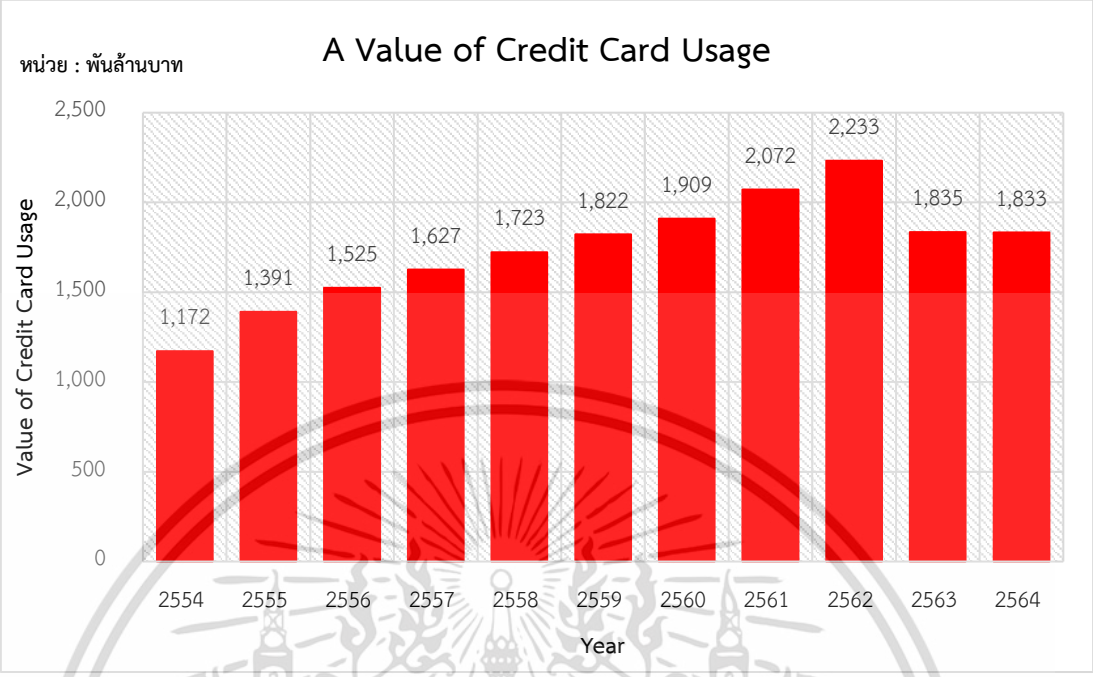
ปัจจุบันเป็นสังคมแห่งเทคโนโลยีที่เน้นการเข้าถึงข้อมูลได้รวดเร็วและมีประสิทธิภาพ ในปัจจุบันมนุษย์มีความต้องการที่เพิ่มมากขึ้นอย่างไม่มีที่สิ้นสุด เพื่อก้าวให้ทันกับสภาพแวดล้อมทางสังคมที่มีการเปลี่ยนแปลงอย่างรวดเร็ว มีการซื้อขายแลกเปลี่ยนสินค้าโดยใช้เงินสดเป็นสื่อกลาง แต่ในปัจจุบันบัตรเครดิตได้เข้ามามีบทบาทและได้รับความนิยมเข้ามากมีการใช้เป็นสื่อกลางแทนเงินสดเพิ่มมากขึ้น เนื่องจากบัตรเครดิตถือเป็นตัวช่วยในการจ่ายชำระค่าบริการหรือการซื้อสินค้าทำให้สังคมปัจจุบันนั้นได้ก้าวเข้าสู่สังคมไร้เงินสดและยังนิยมมาใช้ในการซื้อขายผ่านอินเทอร์เน็ตอีกด้วย เมื่อมีการซื้อขายสินค้าผ่านบัตรเครดิต ผู้ใช้บัตรเครดิตจะต้องแสดงความยินยอมว่าการซื้อขายนั้นได้เกิดขึ้นจริง ด้วยการเซ็นชื่อในใบเสร็จ หากเป็นการซื้อขายทางอินเทอร์เน็ต ผู้ใช้อาจจะกรอกหมายเลขบัตรเครดิต และรหัสลับหลังบัตร เพื่อเป็นการแสดงความจำนงในการซื้อขายออนไลน์เติบโตอย่างรวดเร็ว ทำให้การใช้งานบัตรเครดิตซื้อสินค้า จากข้อมูลของธนาคารแห่งประเทศไทย พบว่าตั้งแต่ปี พ.ศ. 2554 ต่อเนื่องจนถึงปี พ.ศ. 2564 ที่ผ่านมามีจำนวนบัตรเครดิตรวม 226,349,836 ล้านใบ และมีมูลค่าการใช้บัตรเครดิต 19,142 พันล้านบาท (ธนาคารแห่งประเทศไทย, 2564)



รูปที่ 1.1 จำนวนบัตรเครดิตตั้งแต่ปี 2554 - 2564 ในประเทศไทย

ที่มา : ธนาคารแห่งประเทศไทย, 2564

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 1.2 มูลค่าการใช้บัตรเครดิตตั้งแต่ปี 2554 - 2564
ที่มา : ธนาคารแห่งประเทศไทย (2564)

การทำธุรกรรมแบบไร้เงินสดกว้างขวางมากขึ้น ทำให้มีปัญหาอย่างหนึ่งของธนาคารผู้ออกบัตร ก็คือ การตรวจสอบการทำธุรกรรมออนไลน์ว่าเป็นการกระทำที่ผิดปกติหรือไม่ (Fraud Detection) ซึ่งการฉ้อโกงเหล่านี้ได้สร้างความเสียหายให้กับธุรกิจธนาคารเป็นอย่างมาก โดยเฉพาะอย่างยิ่งการฉ้อโกงที่เกิดขึ้นจากการซื้อสินค้าผ่านช่องทางพาณิชย์อิเล็กทรอนิกส์ การฉ้อโกงบัตรเครดิตเป็นรูปแบบหนึ่งของการโจรกรรมข้อมูลประจำตัวที่เกี่ยวข้องกับการนำข้อมูลบัตรเครดิตของผู้อื่นโดยไม่ได้รับอนุญาต ปัญหาการฉ้อโกงมาพร้อมกับความก้าวหน้าทางเทคโนโลยีที่ทันสมัยได้ถูกพัฒนาขึ้นเพื่อป้องกันและคาดการณ์เหตุการณ์ต่าง ๆ และปิดช่องว่างของการกระทำผิดในรูปแบบที่แพร่หลายตลอดจนตรวจหาและจัดการกับรูปแบบการฉ้อโกงใหม่ๆที่เปลี่ยนแปลงไปตามกาลเวลาและกระบวนการทางอาชญากรที่มีความซับซ้อนเพิ่มมากขึ้น จากข้อมูลของธนาคารแห่งประเทศไทยกรณีการตัดเงินที่ผิดปกติผ่านบัตรเครดิตและบัตรเดบิตของลูกค้า โดยเกิดจากมิจฉาชีพสุ่มข้อมูลบัตรและนำไปสวมรอยทำธุรกรรมโดยพบว่าระหว่างวันที่ 1 - 17 ตุลาคม 2564 มีบัตรของธนาคารต่าง ๆ จำนวน 10,700 ใบ ถูกนำไปทำธุรกรรมลักษณะดังกล่าว มูลค่ารวมประมาณ 130 ล้านบาท โดยแบ่งเป็นบัตรเดบิตจำนวน 4,800 ใบ เป็นจำนวนเงิน 30 ล้านบาท และบัตรเครดิตจำนวน 5,900 ใบ เป็นจำนวนเงิน 100 ล้านบาท (ธนาคารแห่งประเทศไทย, 2564)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากความสำคัญและปัญหาดังกล่าว ผู้วิจัยจึงเห็นความสำคัญในการศึกษาแบบจำลองการป้องกันการฉ้อโกงเพื่อเป็นแนวทางในการเพิ่มขีดความสามารถขององค์กรและลดความเสียหายที่จะเกิดขึ้นจากการฉ้อโกง โดยมุ่งเน้นพัฒนาแบบจำลองการจำแนกกลุ่ม โดยใช้การเรียนรู้ของเครื่องที่ประสิทธิภาพความแม่นยำในการจำแนกฉ้อโกง งานวิจัยนี้ สามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการวิเคราะห์พฤติกรรมของลูกค้าเพื่อป้องกันการฉ้อโกงได้มีประสิทธิภาพมากยิ่งขึ้นและการป้องกันการฉ้อโกงบัตรเครดิตที่จะทำงานได้อย่างมีประสิทธิภาพนั้น จำเป็นต้องใช้ความสามารถมากกว่าการวิเคราะห์ทั่วไป และอาศัยการวิเคราะห์ทั้งเชิงคาดการณ์และการปรับตัวที่รวดเร็ว รวมทั้งศักยภาพของด้านปัญญาประดิษฐ์ เช่น การเรียนรู้ของเครื่อง เป็นต้น จึงนำมาสู่การประยุกต์ใช้แหล่งข้อมูลขนาดใหญ่ร่วมกับการตรวจตราแบบเรียลไทม์ที่ทันท่วงที ซึ่งสิ่งเหล่านี้ได้สร้างความได้เปรียบกลับมาสู่การป้องกันการฉ้อโกงบัตรเครดิตและการประพฤติผิดในรูปแบบต่าง ๆ ในปัจจุบัน การฉ้อโกงในการทำธุรกรรมบัตรเครดิตเป็นเรื่องปกติในปัจจุบัน ดังนั้นจึงจำเป็นต้องมีวิธีการจำแนกกลุ่มที่มีประสิทธิภาพเพื่อลดการสูญเสีย

จากการศึกษาครั้งนี้จึงได้นำเอาความก้าวหน้าของเทคโนโลยีด้านปัญญาประดิษฐ์มาใช้ในการตรวจจับการฉ้อโกงได้อย่างมีประสิทธิภาพโดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อเป็นกุญแจสำคัญในการลดความสูญเสีย ช่วยสร้างความมั่นใจให้กับลูกค้าและองค์กรต่าง ๆ ธุรกิจที่เกี่ยวข้องกับบัตรเครดิต

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อศึกษาการจำแนกกลุ่มสำหรับการตรวจจับการฉ้อโกงบัตรเครดิตโดยวิธีการถดถอยลอจิสติกทวิภาค วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท
2) เพื่อเปรียบเทียบประสิทธิภาพของการจำแนกกลุ่มก่อนและหลังการแก้ปัญหาข้อมูลที่ไม่สมดุลกันโดยพิจารณาจาก ค่าความแม่นยำ (Precision) ค่าความไว (Recall)

1.3 ขอบเขตของงานวิจัย

ในการศึกษาครั้งนี้ ข้อมูลที่ใช้ในการศึกษาประกอบด้วย รายการธุรกรรมการใช้จ่ายบัตรเครดิตและร้านค้า รวมทั้งผลการตรวจสอบการฉ้อโกงบัตรเครดิต เพื่อใช้ในการสร้างตัวแบบด้วยเทคนิคการเรียนรู้ของเครื่องสำหรับการตรวจจับการฉ้อโกงบัตรเครดิตและเปรียบเทียบประสิทธิภาพของตัวแบบต่าง ๆ ผู้วิจัยได้อาศัยข้อมูลทุติยภูมิของกรณีศึกษาธนาคารพาณิชย์ไทยในประเทศไทยแห่งหนึ่ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ช่วยเพิ่มประสิทธิภาพในการตรวจสอบธุรกรรมที่คาดว่าฉ้อโกงบัตรเครดิตเบื้องต้นได้
2) ลดมูลค่าความสูญเสียของการฉ้อโกงบัตรเครดิตและสร้างความมั่นใจให้กับลูกค้าได้

1.5 นิยามศัพท์

บัตรเครดิต หมายถึง เป็นบัตรที่ธนาคารพาณิชย์หรือผู้ประกอบการธุรกิจบัตรเครดิตที่ไม่ใช่ธนาคารพาณิชย์ (Non-bank) ออกให้แก่ผู้ถือบัตรเพื่อใช้ชำระสินค้า ค่าบริการหรือค่าอื่นใดแทนการชำระเงินด้วยเงินสดหรือเพื่อใช้เบิกถอนเงินสด โดยลูกค้าต้องชำระค่าธรรมเนียม ค่าบริการและดอกเบี้ย

การฉ้อโกง หมายถึง การกระทำใด ๆ ไม่ว่าจะเป็นสิ่งที่พูดหรือทำในลักษณะที่ไม่ซื่อสัตย์ ที่อาศัยการหลอกหลวงผู้อื่นให้ได้มาเพื่อผลประโยชน์ต่อตนเองและผู้อื่นได้รับความเสียหาย

การเรียนรู้ของเครื่อง หมายถึง ระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเองโดยปราศจากการป้อนคำสั่งของโปรแกรมเมอร์ ความก้าวหน้าในครั้งนี้มาพร้อมกับความคิดที่ว่าเครื่องคอมพิวเตอร์สามารถเรียนรู้เพียงแค่จากข้อมูลอย่างเดียวเพื่อที่จะผลิตผลลัพธ์ที่แม่นยำออกมาได้ (สมาคมโปรแกรมเมอร์ไทย, 2561)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาเกี่ยวกับการตรวจจับการฉ้อโกงบัตรเครดิตโดยเทคนิคการเรียนรู้ของเครื่องของธนาคารพาณิชย์ไทยในประเทศไทยแห่งหนึ่ง เพื่อตรวจจับการฉ้อโกงบัตรเครดิตและลดมูลค่าความสูญเสียของการฉ้อโกงบัตรเครดิตต่อองค์กรให้ได้มากที่สุด ผู้วิจัยได้ศึกษาแนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

2.1 แนวคิดเกี่ยวกับบัตรเครดิต

ธุรกิจบัตรเครดิตมีต้นกำเนิดและได้พัฒนาขึ้นในประเทศสหรัฐอเมริกาประมาณ 88 ปี ในปีค.ศ. 1914 บริษัทน้ำมันในแคลิฟอร์เนีย ปัจจุบันคือชื่อ โมบิล ออยส์ ผลิตบัตรเครดิตออกมาให้ลูกค้าและพนักงาน เพื่อไว้ซื้อผลิตภัณฑ์ของบริษัท ซึ่งจัดเป็นบัตรเครดิตสำหรับการซื้อสินค้าใบแรกของโลก ต่อมาได้ถูกนิยมนำไปใช้ในกิจการต่าง ๆ เช่น โรงแรมและร้านค้าต่างๆ โดยร้านค้าต่างๆ และโรงแรมได้มีการทำเงินโลหะขึ้นมา ให้แก่ลูกค้าของตนเองซึ่งลูกค้าเหล่านั้นจะต้องเปิดบัญชีเป็นรายเดือน หมายเลขบัญชีของลูกค้าและชื่อบริษัทผู้ออกเงินนั้นจะปรากฏอยู่บนเงินเครดิต ลูกค้าจะใช้เงินเครดิตนั้นเพื่อซื้อสินค้าหรือใช้บริการของบริษัทผู้ออกเงินนั้นโดยไม่ต้องจ่ายค่าธรรมเนียมหรือค่าใช้จ่ายอื่นๆ เพิ่มเลย

ในปีค.ศ. 1950 นายแฟรงค์ นักธุรกิจชาวนิวยอร์คร่วมกับเพื่อนได้ทำการก่อตั้งบริษัท ดินเนอร์ คลับ ขึ้นมาเพื่อออกบัตรชื่อ ดินเนอร์ คลับ ให้แก่ลูกค้า เพื่อจะให้ลูกค้าสามารถนำบัตรเครดิตของตนเองไปใช้ในการชำระค่าบริการร้านภัคตตาคารและบริการโรงแรมในนิวยอร์คได้ โดยการชำระเงินให้ชำระตอนสิ้นเดือนและสามารถใช้วงเงินได้ไม่จำกัด ลูกค้าที่เป็นสมาชิกจะได้รับรายงานรายชื่อโรงแรมร้านค้าและภัตตาคารต่าง ๆ ที่สามารถใช้บัตรเครดิตได้โดยไม่ต้องพกพาเงินสดไปด้วย บริษัท ดินเนอร์ คลับ จะเป็นตัวแทนการชำระเงินแทนลูกค้า ในขณะผู้ใช้บริการจะเรียกเก็บเงินจากผู้ใช้บริการในภายหลังและจัดทำลำดับความน่าเชื่อถือของผู้ใช้บริการไว้ทั้งหมดตามพฤติกรรมการชำระเงินตามที่ใช้จ่ายบัตรเครดิต

ธุรกิจของบริษัท ดินเนอร์ คลับ ได้รับความนิยมแพร่หลายในประเทศสหรัฐอเมริกา ต่อมาได้ทำการขยายไปเปิดสาขาในต่างประเทศอีกหลายประเทศ โดยกิจการร้านค้าของผู้รับบัตรมีทุกประเภทตั้งแต่โรงแรม ภัตตาคาร บริการน้ำมัน ธุรกิจสายการบิน ธุรกิจท่องเที่ยวและร้านค้า นับเป็นจุดเริ่มต้นของบัตรเครดิตชนิดเพื่อการเดินทางและการรับรอง ซึ่งจะแตกต่างจากบัตรเครดิตที่ออกจากบริษัท

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

น้ำมัน เพราะ ดินเนอร์ คลับ ไม่ได้เป็นผู้จัดจำหน่ายสินค้าเองแต่มีหน้าที่เป็นตัวกลางในการให้บริการและอำนวยความสะดวกแก่ผู้ถือบัตรและร้านค้า

ในปีค.ศ. 1959 ธนาคารแห่งอเมริกา ในรัฐแคลิฟอร์เนีย ได้ออกบัตรของชื่อ “แบงค์อเมริการ์ด” ได้รับความนิยมจากประชาชนมากถึง หนึ่งล้านคนภายใน 2 ปี และมากขึ้นเกือบ สามล้านคน

ในปีค.ศ. 1967 ต่อมาธนาคารได้จัดตั้ง Bank of America Service Corporation ในการทำหน้าที่ออกบัตรให้ความสะดวกแก่ผู้ถือบัตรและดูแลเก็บค่าบริการต่าง ๆ

ในปีค.ศ. 1970 มีสถาบันการเงินกว่า 3,000 แห่ง บัตรเครดิตได้รับความนิยมมากขึ้นทั้งในประเทศและนอกประเทศสหรัฐอเมริกาจนต้องตั้งองค์กรอิสระระหว่างประเทศ เพื่อทำหน้าที่ดูแลและขยายบัตรไปทั่วโลกและมีการเปลี่ยนเป็น วีซ่าอินเตอร์เนชั่นแนล มาออกบัตรเครดิตวีซ่า ซึ่งเป็นที่นิยมทั่วไปในปัจจุบัน

บัตรเครดิตเป็นบัตรที่ธนาคารพาณิชย์หรือสถาบันการเงินอื่น ๆ ออกให้แก่ลูกค้าเพื่อใช้ชำระค่าสินค้าและบริการไปก่อนแทนการใช้เงินสดแล้วค่อยผ่อนชำระหรือผ่อนชำระคืนภายหลัง ซึ่งบัตรเอทีเอ็มและบัตรเดบิตไม่สามารถให้บริการในด้านสินเชื่อได้ บัตรเครดิตจัดเป็นการให้สินเชื่อประเภทหนึ่ง ซึ่งจะกำหนดวงเงินเพื่อใช้จ่ายสำหรับการเบิกเงินสดล่วงหน้าไว้ส่วนหนึ่งและกำหนดวงเงินที่ใช้ซื้อสินค้าไว้ส่วนหนึ่ง ในปัจจุบันธนาคารพาณิชย์ต่าง ๆ นิยมออกบัตรเครดิตประเภทหนึ่งที่เรียกว่าบัตรร่วม ซึ่งเป็นความร่วมมือระหว่างธนาคารพาณิชย์และธุรกิจต่าง ๆ ในการออกบัตรเครดิตร่วมกัน ลูกค้าจะได้ประโยชน์จากการให้บริการของธุรกิจนั้น ๆ และยังสามารถชำระค่าสินค้าและบริการจากธนาคารผู้ออกบัตรได้ เช่นเดียวกับบัตรเครดิตธรรมดาสำหรับประเทศไทย เริ่มมีการใช้บัตรเครดิตครั้งแรกในปี ค.ศ. 1969 ผู้ออกบัตรเครดิต คือ บริษัทไดเนอร์ส คลับ (ประเทศไทย) จำกัด (ธนาคารกรุงไทยจำกัด (มหาชน), 2547)

ปี พ.ศ 2521 ธนาคารกสิกรไทย จำกัด ได้ออกบัตรเครดิตของตนเองที่มีชื่อว่า “บัตรเครดิตธนาคารกสิกรไทย” และขณะเดียวกัน ธนาคารศรีนคร จำกัด ก็ได้ออกบัตรเครดิตที่มีชื่อว่า“บัตรเอนกประสงค์ธนาคารศรีนคร” และได้สิทธิให้เป็นตัวแทน ในการออกบัตรมาสเตอรการ์ด (Master Card) ซึ่งเป็นบัตรเครดิตต่างประเทศที่เผยแพร่ในประเทศไทยเป็นครั้งแรก

ปี ค.ศ 1978 ธนาคารกสิกรไทย จำกัด ได้เข้าเป็นสมาชิกของวีซ่าอินเตอร์เนชั่นแนล และเป็นผู้ได้รับสิทธิในการเป็นผู้แทนออกบัตรเครดิตวีซ่าในประเทศไทย

ปี พ.ศ. 1981 บริษัท ซี ทัวร์ จำกัด ได้เข้าร่วมกับบริษัท อเมริกัน เอ็กซ์เพรส จำกัด โดยมีธนาคารกรุงเทพ จำกัด (มหาชน) ร่วมทุนกันในการจัดตั้งบริษัท อเมริกันเอ็กซ์เพรส (ประเทศไทย) จำกัดออกบัตรเครดิตอเมริกันเอ็กซ์เพรส

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.1 ประเภทของบัตรเครดิต

1) Charge Card ผู้ถือบัตรเครดิตประเภทนี้ ส่วนใหญ่เป็นนักธุรกิจ นักบริหาร หรือเป็นบุคคลที่มีฐานะทางการเงินดี เนื่องจากจุดประสงค์หลักของบัตรเครดิตประเภทนี้ คือ ใช้เพื่อชำระค่าบริการการเดินทางและท่องเที่ยวพักผ่อนเป็นหลัก ทำให้วงเงินบัตรเครดิตประเภทนี้ไม่มีการจำกัดวงเงินค่าใช้จ่ายล่วงหน้า แต่ต้องชำระยอดหนี้บัตรเครดิตให้เสร็จสิ้นไปภายในระยะเวลาที่กำหนดไว้โดยไม่เสียดอกเบี้ย

2) Credit Card หรือ Bank Card เป็นบัตรเครดิตที่ออกโดยสถาบันการเงิน หรือทำการออกร่วมกับสถาบันบัตรเครดิตจากต่างประเทศ เพื่ออำนวยความสะดวกในการชำระค่าสินค้าและบริการใช้ดำรงในชีวิตประจำวันของผู้ถือบัตร โดยบัตร Credit Card หรือ Bank Card ที่พบเห็นบ่อย ๆ อาทิ บัตร VISA กับ Master Card ของแต่ละสถาบันการเงิน

2.1.2 ความสำคัญของบัตรเครดิต

ในปัจจุบันบัตรเครดิตมีการเพิ่มปริมาณการใช้มากขึ้นทุกวัน ธุรกิจบัตรเครดิตได้สร้างผลกำไรให้กับธนาคารและสถาบันที่ออกบัตรเครดิต ผู้ใช้บัตรเครดิตมีจำนวนไม่น้อยที่มีบัตรเครดิตหลายใบ และมียอดการใช้จ่ายหลายบัตรในเวลาเดียวกัน ซึ่งเป็นสิ่งที่มองเห็นอย่างชัดเจนว่า บัตรเครดิตมีความสำคัญมาก ซึ่งสามารถแบ่งได้ดังนี้ (ธนาคารกรุงไทยจำกัด (มหาชน), 2547)

1. ด้านความสะดวกในการซื้อสินค้า (Convenience) ในการซื้อสินค้านั้นผู้ถือบัตรเครดิตไม่จำเป็นต้องเตรียมเงินสดมามากกว่าหรือเท่ากับราคาสินค้า ถึงแม้การซื้อสินค้าในบางครั้งผู้ซื้อไม่ได้วางแผนมาก่อน แต่มีความอยากได้หรือจำเป็นต้องใช้ ก็สามารถซื้อสินค้าได้ถ้าเป็นสมัยก่อนที่จะมีบัตรเครดิตเข้ามา ผู้ซื้อสินค้าจำเป็นต้องพกเงินสดติดตัวไปด้วยจำนวนมากหรือต้องใช้อย่างอื่นแทน แต่ในบางครั้งผู้ขายจะไม่ยอมรับอย่างอื่นที่ไม่ใช่เงินสดจากผู้ซื้อ เพราะไม่แน่ใจว่าจะมีเงินหรือไม่ โดยเฉพาะคนที่ไม่รู้จักก็จะไม่รับสิ่งที่ไม่ใช่เงินสดเลย นั่นทำให้บัตรเครดิตทำให้ผู้ขายมั่นใจได้ว่าเมื่อขายสินค้าไปแล้ว จะสามารถเรียกเก็บเงินจากธนาคารหรือผู้ออกบัตรได้ทันที ขณะที่ผู้ซื้อสินค้าสามารถซื้อสินค้าได้โดยยังที่ไม่ต้องชำระเงินสดก่อน

2. ด้านการเดินทาง (Traveling) ในปัจจุบันการเดินทางติดต่อธุรกิจนั้นสะดวกมากขึ้น ไม่ต้องเดินทางไปต่างถิ่นเป็นเวลานานๆ ที่จำเป็นต้องใช้เงินมาก เช่น ค่าอาหารค่าโรงแรมและอื่นๆ ที่จะต้องใช้จ่าย เช่น เดินทางไป 1 เดือน จะต้องเสียค่าโรงแรม ค่าอาหาร และค่าใช้จ่ายอื่นๆ เป็นเงินจำนวนมากจำเป็นต้องพกพาเงินสดจำนวนมาก ทำให้เกิดความไม่สะดวกและเสี่ยงต่อการสูญหาย หรือถูกโจรกรรมแต่ถ้ามีบัตรพลาสติกหรือบัตรเครดิต ซึ่งสามารถใช้แทนเงินสดที่ต้องใช้จ่ายได้

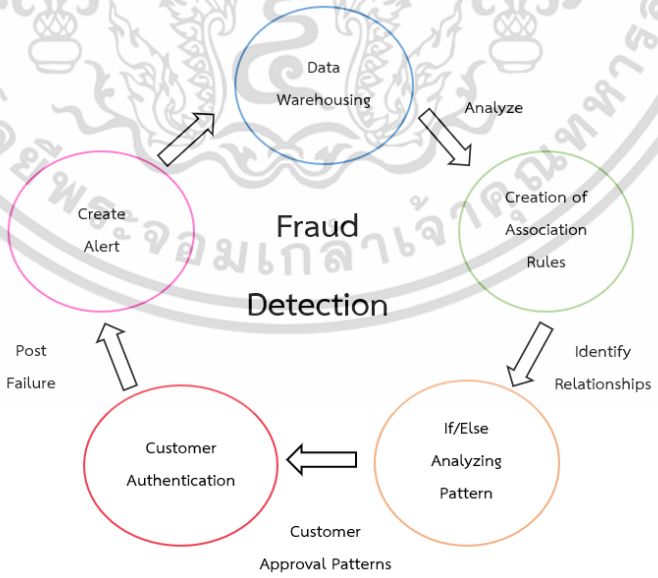
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ระหว่างเดินทาง ปัญหาการถูกโจรกรรมหรือการสูญหายที่อาจเกิดขึ้นก็จะไม่ก่อให้เกิดการสูญเสียเหมือนเงินสดหรือทรัพย์สินสูญหายเพราะกรณีบัตรเครดิตสูญหาย เจ้าของบัตรสามารถแจ้งบัตรสูญหายหรืออายัดการใช้บัตรได้ทันที ผู้ที่เจอบัตรหรือเก็บได้จึงไม่สามารถนำไปใช้ได้

3. ด้านความน่าเชื่อถือทางด้านการเงิน (Financial Stability) ธนาคารหรือผู้ออกบัตรในแต่ละใบผู้ออกบัตรจะมีการตรวจสอบสถานะทางการเงินของผู้ทำเรื่องขอยื่นมีบัตร ซึ่งคำขอมีบัตรเครดิตของธนาคารหรือบริษัทผู้ออกบัตรจะเป็นผู้กำหนดวงเงินให้ใช้จ่ายผ่านบัตรเครดิต สามารถแบ่งออกได้เป็น บัตรเงิน และบัตรทอง ซึ่งบัตรทั้งสองประเภทนี้สามารถเป็นตัวแทนในการแสดงให้กับร้านค้าต่าง ๆ หรือผู้ถือบัตรสามารถทราบถึงสถานะของผู้ถือบัตรว่ามีความน่าเชื่อถืออยู่ในระดับใด (สุรเชษฐ, 2541)

2.2 แนวคิดเกี่ยวกับการฉ้อโกงบัตรเครดิต

การตรวจจับการฉ้อโกงเป็นกระบวนการที่ตรวจจับและป้องกันมิให้ผู้หลอกลวงได้รับเงินหรือทรัพย์สินด้วยวิธีการอันเป็นเท็จ เป็นกิจกรรมที่ดำเนินการเพื่อตรวจจับและป้องกันความพยายามของผู้ฉ้อโกงไม่ให้ได้รับเงินหรือทรัพย์สินโดยฉ้อโกง การตรวจจับการฉ้อโกงเป็นที่แพร่หลายในด้านการธนาคาร การประกันภัย การแพทย์ ภาครัฐ และภาครัฐ ตลอดจนในหน่วยงานบังคับใช้กฎหมาย



รูปที่ 2.1 วิธีการทำงานของการตรวจจับการฉ้อโกง

ที่มา : Kanade (2021)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิจกรรมที่เป็นการฉ้อโกงรวมถึงการฟอกเงิน การโจมตีทางอินเทอร์เน็ต การเรียกร้องค่าสินไหมทดแทนจากธนาคาร การปลอมแปลงเช็คธนาคาร การขโมยข้อมูลประจำตัว และการกระทำที่ผิดกฎหมายหลายอย่าง ด้วยเหตุนี้ องค์กรต่าง ๆ จึงใช้เทคโนโลยีการตรวจจับและป้องกันการฉ้อโกงที่ทันสมัย และกลยุทธ์การจัดการความเสี่ยง เพื่อต่อสู้กับธุรกรรมฉ้อโกงที่เพิ่มขึ้นในแพลตฟอร์มที่หลากหลาย เทคนิคเหล่านี้ใช้การวิเคราะห์แบบคาดการณ์ได้ (เช่น การเรียนรู้ของเครื่อง) เพื่อสร้างคะแนนความเสี่ยงในการฉ้อโกง ควบคู่ไปกับการตรวจสอบเหตุการณ์ฉ้อโกงแบบเรียลไทม์ ซึ่งช่วยให้สามารถตรวจสอบธุรกรรมและอาชญากรรมได้อย่างต่อเนื่องในแบบเรียลไทม์ นอกจากนี้ยังช่วยถอดรหัสมาตรการป้องกันใหม่และซับซ้อนผ่านระบบอัตโนมัติ

2.2.1 ประเภทของเทคนิคการตรวจจับการฉ้อโกง (Kanade, 2021)

การตรวจจับการฉ้อโกงโดยทั่วไปเกี่ยวข้องกับเทคนิคการวิเคราะห์ข้อมูล เทคนิคเหล่านี้จัดอยู่ในประเภทกว้างๆ เป็นเทคนิคการวิเคราะห์ข้อมูลทางสถิติและปัญญาประดิษฐ์

เทคนิคการวิเคราะห์ข้อมูลทางสถิติ การวิเคราะห์ข้อมูลทางสถิติสำหรับการตรวจจับการฉ้อโกงดำเนินการทางสถิติต่าง ๆ เช่น การรวบรวมข้อมูลการฉ้อโกง การตรวจจับการฉ้อโกง และการตรวจสอบการฉ้อโกงโดยดำเนินการตรวจสอบอย่างละเอียด เทคนิคเหล่านี้แบ่งออกเป็นประเภทดังนี้

1. การคำนวณค่าพารามิเตอร์ทางสถิติ คือ การคำนวณพารามิเตอร์ทางสถิติต่าง ๆ เช่น ค่าเฉลี่ย ควอไทล์ และการแจกแจงความน่าจะเป็นสำหรับข้อมูลที่เกี่ยวข้องกับการฉ้อโกงที่รวบรวมระหว่างกระบวนการรวบรวมข้อมูล

2. การวิเคราะห์การถดถอย ช่วยตรวจสอบความสัมพันธ์ระหว่างตัวแปรที่สนใจตั้งแต่สองตัวขึ้นไป นอกจากนี้ยังประเมินความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม ซึ่งช่วยให้เข้าใจและระบุความสัมพันธ์ระหว่างตัวแปรการฉ้อโกงต่าง ๆ ซึ่งช่วยในการคาดการณ์กิจกรรมการฉ้อโกงในอนาคต

3. การแจกแจงความน่าจะเป็นและตัวแบบ เทคนิคนี้ตัวแบบและการแจกแจงความน่าจะเป็นการฉ้อโกงทางธุรกิจต่าง ๆ ทั้งในแง่ของพารามิเตอร์ต่าง ๆ หรือการแจกแจงความน่าจะเป็น

4. การจับคู่ข้อมูล การจับคู่ข้อมูลใช้เพื่อเปรียบเทียบข้อมูลที่รวบรวมได้สองชุด เช่น ข้อมูลการฉ้อโกง กระบวนการนี้สามารถดำเนินการได้โดยใช้อัลกอริทึม นอกจากนี้การจับคู่ข้อมูลยังใช้เพื่อลบเรกคอร์ดที่ซ้ำกันและระบุลิงก์ระหว่างชุดข้อมูลสองชุด เพื่อการตลาดและความปลอดภัย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เทคนิคปัญญาประดิษฐ์ การนำมาใช้ในการป้องกันการฉ้อโกงช่วยให้บริษัทต่าง ๆ ปรับปรุงการรักษาความปลอดภัยภายในและทำให้กระบวนการทางธุรกิจคล่องตัวขึ้น ด้วยการเพิ่มประสิทธิภาพ AI ได้กลายเป็นเทคโนโลยีที่จำเป็นในการป้องกันการฉ้อโกงที่สถาบันการเงิน เทคนิคการตรวจจับการฉ้อโกงแบบเทคนิคปัญญาประดิษฐ์ ประกอบด้วยวิธีการดังนี้:

1. การทำเหมืองข้อมูล คือ การตรวจจับและป้องกันการฉ้อโกงจะจำแนก จัดกลุ่ม และแบ่งกลุ่มข้อมูล และค้นหาการเชื่อมโยงและกฎเกณฑ์ในข้อมูลที่อาจบ่งบอกถึงรูปแบบที่น่าสนใจโดยอัตโนมัติ รวมถึงสิ่งที่เกี่ยวข้องกับการฉ้อโกง

2. โครงข่ายประสาทเทียม คือ การตรวจจับการฉ้อโกงจะทำการจัดประเภท การจัดกลุ่ม และการคาดการณ์ข้อมูลที่เกี่ยวข้องกับการฉ้อโกง ซึ่งสามารถเปรียบเทียบกับข้อสรุปที่หยิบยกขึ้นมาในการตรวจสอบภายในหรือเอกสารทางการเงินที่เป็นทางการ

3. การเรียนรู้ของเครื่อง การเรียนรู้ของเครื่องจากรูปแบบการฉ้อโกงในอดีตและจดจำรูปแบบดังกล่าวในการทำธุรกรรมในอนาคต การเรียนรู้ของเครื่องใช้วิธีการเรียนรู้แบบมีผู้สอนหรือแบบไม่มีผู้สอน ในการเรียนรู้แบบมีผู้สอน

4. การจดจำรูปแบบ อัลกอริทึมการรู้จำรูปแบบจะตรวจจับคลาส คลัสเตอร์ หรือรูปแบบของพฤติกรรมที่น่าสงสัยโดยประมาณ ทั้งแบบอัตโนมัติ หรือด้วยตนเอง เทคนิคอื่น ๆ เช่น ทฤษฎีเบย์เซียน ทฤษฎีการตัดสินใจ และการจับคู่ลำดับ ยังใช้เพื่อวัตถุประสงค์ในการตรวจจับการฉ้อโกง

2.2.2 ลักษณะการฉ้อโกงด้านบัตรเครดิต

การทุจริตด้านบัตรเครดิตได้มีการพัฒนารูปแบบให้มีความสลับซับซ้อนขึ้นมากกว่าเดิม ทำให้การตรวจสอบป้องกันทำได้ยากขึ้น ทั้งนี้ในกระบวนการป้องกันจะต้องได้รับความร่วมมือทั้งจากผู้ถือบัตร ธนาคาร หรือสถาบันผู้ออกบัตร สถานธุรกิจประกอบการที่รับบัตรเครดิตชำระค่าสินค้าและบริการซึ่งลักษณะการฉ้อโกงด้านบัตรเครดิต (Asha and Suresh , 2021) สามารถแบ่งได้ดังนี้

1. การฉ้อโกงแอปพลิเคชัน (Application Fraud) เมื่อผู้ฉ้อโกงเข้าควบคุมแอปพลิเคชัน ขโมยข้อมูลประจำตัวของลูกค้า และสร้างบัญชีปลอม จากนั้นการทำธุรกรรมจะเกิดขึ้น

2. ตราประทับบัตรอิเล็กทรอนิกส์ (Electronic or Manual Card Imprints) ผู้ฉ้อโกงจะอ่านข้อมูลจากแถบแม่เหล็กที่ปรากฏบนบัตร จากนั้นจึงใช้ข้อมูลประจำตัวและดำเนินการธุรกรรมฉ้อโกง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. ไม่มีบัตร (Card Not Present) ประเภทของบัตรเครดิตที่ไม่มีบัตรจริงในระหว่างการทำธุรกรรม

4. การฉ้อโกงบัตรปลอม (Counterfeit Card Fraud) ผู้ฉ้อโกงจะคัดลอกข้อมูลทั้งหมดจากแถบแม่เหล็กและบัตรจริงดูเหมือนบัตรเดิมใช้งานได้เหมือนบัตรเดิมเท่านั้น บัตรนี้ใช้ในการฉ้อโกง

5. บัตรหาย/ถูกขโมย (Lost/Stolen Card) การฉ้อโกงประเภทนี้เกิดจากการทำบัตรหายโดยผู้ถือบัตรหรือจากการขโมยบัตรจากผู้ถือบัตร

6. การขโมยรหัสบัตร (Card Id Theft) ประเภทของการฉ้อโกงที่ ID ของผู้ถือบัตรถูกขโมยและเกิดการฉ้อโกง

7. การฉ้อโกงทางบัตรที่ไม่ได้รับทางไปรษณีย์ (Mail Non-Received Card Fraud) ในขณะที่ออกบัตรเครดิต จะมีขั้นตอนการส่งอีเมลไปยังผู้รับ การฉ้อโกงอาจเกิดขึ้นได้โดยการฉ้อโกงอีเมลหรือฟิชชิ่ง

8. การเข้าครอบครองบัญชี (Account Takeover) ที่นี่ผู้ฉ้อโกงจะเข้าควบคุมเจ้าของบัญชีอย่างสมบูรณ์และทำการฉ้อโกง

9. การฉ้อโกงปลอมบนเว็บไซต์ (Fake Fraud In Website) ผู้ฉ้อโกงจะแนะนำรหัสที่เป็นอันตรายซึ่งทำงานบนเว็บไซต์

10. การชนกันของผู้ค้า (Merchant Collision) ในการฉ้อโกงประเภทนี้ รายละเอียดผู้ถือบัตรจะถูกเปิดเผยโดยบุคคลที่สามหรือผู้ฉ้อโกงโดยผู้ค้าโดยไม่ได้รับอนุญาตจากผู้ถือบัตร

2.3 วิธีการวิเคราะห์ถดถอยลอจิสติกทวิภาค (Logistic Regression Analysis)

การวิเคราะห์ถดถอยลอจิสติกทวิภาค (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพ ใช้สำหรับหาความสัมพันธ์ระหว่างตัวแปรอิสระ (มากกว่าหรือเท่ากับ 1 ตัว) กับตัวแปรตาม และพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ เมื่อแบ่งตามระดับการวัดของตัวแปรสามารถแบ่งออกเป็น 2 ประเภท คือ การวิเคราะห์ความถดถอยลอจิสติกทวิภาค (Binary logistic regression) และการวิเคราะห์ลอจิสติกแบบพหุ (Multinomial logistic regression) ซึ่งการวิเคราะห์ความถดถอยลอจิสติกทวิภาคดังกล่าวแตกต่างกันในด้านของตัวแปรตาม โดยการวิเคราะห์ความถดถอยลอจิสติกทวิภาคใช้กับตัวแปรตามที่มีค่าได้เพียง 2 ค่า คือ 0 กับ 1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3.1 ประเภทของการวิเคราะห์ความถดถอยลอจิสติก (สายชล สินสมบูรณ์ทอง, 2559)

- <u>แบ่งตามจำนวนตัวแปรอิสระ</u>

1) ตัวแปรอิสระ (x) มีเพียง 1 ตัว เป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพ เรียกว่า การวิเคราะห์การถดถอยลอจิสติกอย่างง่าย (Simple Logistic Regression Analysis)

2) ตัวแปรอิสระ (x) มี 2 ตัวขึ้นไป เป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพหรือมีทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ เรียกว่า การวิเคราะห์การถดถอยลอจิสติกพหุคูณ (Multiple Logistic Regression Analysis)

- <u>แบ่งตามจำนวนกลุ่มของตัวแปรตาม</u>

1) ตัวแปรตาม (y) เป็นตัวแปรเชิงกลุ่ม (categorical variable) ที่มีค่าได้เพียง 2 ค่า (dichotomous or binary variable) หรือ แบ่งหน่วยข้อมูลออกเป็น 2 กลุ่ม เรียกว่า การวิเคราะห์การถดถอยลอจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis)

2) ตัวแปรตาม (y) เป็นตัวแปรเชิงกลุ่ม (Categorical Variable) ที่มีค่ามากกว่า 2 ค่า(Multinomial Variable) หรือ แบ่งหน่วยข้อมูลออกได้มากกว่า 2 กลุ่ม เรียกว่า การวิเคราะห์การถดถอยลอจิสติกแบบหลายกลุ่ม (Multinomial Logistic Regression Analysis)

2.3.2 วัตถุประสงค์การวิเคราะห์การถดถอยลอจิสติก (กัลยา, 2548)

1) เพื่อศึกษาความสัมพันธ์ของตัวแปรอิสระที่มีโอกาสที่จะเกิดเหตุการณ์ตัวแปรตาม พร้อมทั้งศึกษาระดับความสัมพันธ์ของตัวแปรอิสระแต่ละตัว

2) เพื่อพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ จากสมการลอจิสติกที่เหมาะสม โดยเลือกตัวแปรที่เหมาะสมเพื่อทำให้เปอร์เซ็นต์ของความถูกต้องในการพยากรณ์มีค่าสูงสุด

2.3.3 เงื่อนไขของการวิเคราะห์การถดถอยลอจิสติก (กัลยา, 2548)

การวิเคราะห์การถดถอยลอจิสติกจะมีเงื่อนไขน้อยกว่าการวิเคราะห์การถดถอยแบบปกติ แต่อย่างไรก็ตามการวิเคราะห์การถดถอยลอจิสติกก็ยังมีเงื่อนไขหลายข้อดังนี้

1) ตัวแปรอิสระ (x) อาจจะเป็นตัวแปรเชิงกลุ่มที่มีค่าได้ 2 ค่า หรือ เป็นมาตรวัดแบบอันตรภาค (Interval Scale) และมาตรวัดแบบอัตราส่วน (Ratio Scale) ก็ได้

2) ค่าความคลาดเคลื่อนเป็นศูนย์หรือ $\mathbf{E}(\varepsilon) = 0$

3) ε_i และ ε_j เป็นอิสระกัน

4) ε_i และ x_i เป็นอิสระกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5) ตัวแปรอิสระไม่ควรมีความสัมพันธ์กันหรือไม่ควรเกิดปัญหาพหุสัมพันธ์เชิงเส้น (Multicollinearity)

สำหรับเงื่อนไขของการวิเคราะห์การถดถอยแบบปกตินอกจากจะมีเงื่อนไขทั้ง 5 ข้อข้างต้นจะต้องเพิ่มเงื่อนไขอีก 2 ข้อคือ

1) ค่าคลาดเคลื่อนมีการแจกแจงแบบปกติ

2) ค่าแปรปรวนของค่าคลาดเคลื่อนคงที่

หมายเหตุ การวิเคราะห์การถดถอยลอจิสติกจะต้องใช้ขนาดตัวอย่าง (n) มากกว่าการวิเคราะห์การถดถอยแบบปกติโดยทั่วไป $n \geq 30k$ โดยที่ k เป็นจำนวนตัวแปรอิสระ

2.3.4 ตัวแบบการวิเคราะห์การถดถอยลอจิสติก (กัลยา, 2548)

การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression Analysis)

1) กรณีตัวแปรอิสระ 1 ตัว

$$Y = \beta_0 + \beta_1 X + e \qquad (2.1)$$

หรือ $E(Y) = \beta_0 + \beta_1 X$

สำหรับการวิเคราะห์การถดถอยลอจิสติกนั้น เมื่อ Y มีค่าได้เพียง 2 ค่า จะพบว่าความสัมพันธ์ระหว่าง X และ Y ไม่ได้อยู่ในรูปเชิงเส้น แต่จะอยู่ในรูป

$$E(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \qquad (2.2)$$

เรียกสมการที่ 2.2 ว่า Logistic Response Function Function

ดังนั้น P(เกิดเหตุการณ์ที่สนใจ) = $E(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$ (2.3)

P(ไม่เกิดเหตุการณ์ที่สนใจ) = $E(Y) = \frac{1}{1+e^{\beta_0+\beta_1 X}}$ (2.4)

2) กรณีตัวแปรอิสระมากกว่า 1 ตัว

เมื่อมีตัวแปรอิสระมากกว่า 1 ตัว หรือมีตัวแปรอิสระ p ตัว (p ≥ 2) Logistic Response Function หรือสมการที่ 2.3 จะกลายเป็น

ดังนั้น P(เกิดเหตุการณ์ที่สนใจ) = $\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots+\beta_p X_p)}}$ (2.5)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

P(ไม่เกิดเหตุการณ์ที่สนใจ) = 1- P(เกิดเหตุการณ์ที่สนใจ) (2.6)

การปรับรูปแบบความสัมพันธ์ที่ไม่ได้อยู่ในรูปเชิงเส้นให้อยู่ในรูปเชิงเส้น

$$\text{Odd Ratio} = \frac{\mathbf{P}(\text{เกิดเหตุการณ์ที่สนใจ})}{\mathbf{P}(\text{ไม่เกิดเหตุการณ์ที่สนใจ})} \quad (2.7)$$

Odd Ratio (OR) เป็นอัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจกับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ จากสมการที่ 2.7 หาค่า $\log_e(OR)$

$$\log_e(OR) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \quad (2.8)$$

สมการที่ 2.8 จะอยู่ในรูปเชิงเส้น และเรียกว่า Logit Response Function จากสูตรของ Odd Ratio จะพบว่า ถ้า Odd Ratio ในสมการที่ 2.7 มีค่ามากกว่า 1 แสดงว่าเหตุการณ์นั้นมีโอกาสเกิดมากกว่าที่จะไม่เกิด

สำหรับการประมาณค่า Y เป็นการประมาณ P(เกิดเหตุการณ์ที่สนใจ) จะใช้สมการ 2.5 สำหรับการประมาณค่าสัมประสิทธิ์ $\beta_0, \beta_1, ..., \beta_p$ ในสมการที่ 2.5 จะใช้วิธี Maximum Likelihood ในขณะที่ในสมการความถดถอยทั่วไปจะใช้วิธี Least Square ในการประมาณค่า $\beta_0, \beta_1, ..., \beta_p$ จากสมการ $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + e$

2.3.5 การตรวจสอบความเหมาะสมของสมการถดถอยลอจิสติก (กัลยา, 2548)

พิจารณาค่าความเป็นไปได้ เพื่อวัดค่าความเหมาะสมของสมการลอจิสติก จะศึกษาจากค่า -2LL (-2 log likelihood) ซึ่งเป็นคำมาจาก loglikelihood ที่คูณด้วย -2 เพื่อต้องการให้ค่าที่ได้มีการแจกแจงมีลักษณะเป็นการแจกแจง χ^2 สำหรับการทดสอบนัยสำคัญทางสถิติการพิจารณาค่า -2LL ถ้ามีค่าต่ำ สมการลอจิสติกมีความเหมาะสมที่สุดในการทดสอบนัยสำคัญความเหมาะสมของสมการลอจิสติก ใช้สถิติ $\chi^2 - test$

2.3.6 การทดสอบสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัว (สายชล สินสมบูรณ์ทอง ,2559)

<u>สถิติทดสอบของวอลด์ (Wald statistic)</u>

วัลด์ (Wald statistic) เป็นการทดสอบสมมติฐานที่กำหนด ดังนี้

$H_0 : \beta_i = 0$; I = 1,2,...,p

$H_1 : \beta_i = 0$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สถิติทดสอบของ Wald คือ

$$\text{Wald}=\left[\frac{b_i}{SE(b_i)}\right]^2 \qquad (2.9)$$

จะปฏิเสธ H_0 ถ้า $Wald > \chi^2_{\alpha;1}$ โดยที่ $\chi^2_{\alpha;1}$ เปิดจากตารางไคสแควร์

ในการทดสอบ ถ้าไม่สามารถปฏิเสธ แสดงว่า ตัวแปรอิสระที่ i ไม่มีผลต่อกาเปลี่ยนแปลง Odds Ratio ดังนั้นจึงไม่มีผลต่อความน่าจะเป็นของความน่าจะเป็นของการเกิดเหตุการณ์นั้น และถ้าทดสอบพบว่ามีนัยสำคัญทางสถิติหรือปฏิเสธและค่าสัมประสิทธิ์เป็นบวก (+) แสดงว่า ตัวแปรอิสระที่ i นั้นมีผลต่อการเพิ่มความน่าจะเป็นของการเกิดเหตุการณ์ และถ้าค่าสัมประสิทธิ์เป็นลบ (-) แสดงว่าตัวแปรอิสระที่ i นั้นลดความน่าจะเป็นของการเกิดเหตุการณ์ (ยุทธ, 2555)

2.3.7 การวัดระดับความสัมพันธ์

ค่าสัมประสิทธิ์การพยากรณ์ (R Square) ในสมการถดถอยลอจิสติกทวิภาค คือค่าความแปรปรวนของตัวแปรตามที่สามารถอธิบายได้โดยตัวแปรอิสระ ซึ่งมีสถิติที่ทดสอบระดับความสัมพันธ์หลายค่า ได้แก่ Cox & Snell R Square

<u>สถิติทดสอบ Cox & Snell R Square</u>

$$R^2_{CS} = 1-\left[\frac{L(0)}{L(B)}\right]^{\frac{2}{n}} \qquad (2.10)$$

โดยที่ L(0) คือ ฟังก์ชันความควรจะเป็นของฟังก์ชันที่มีค่าคงที่เท่านั้น

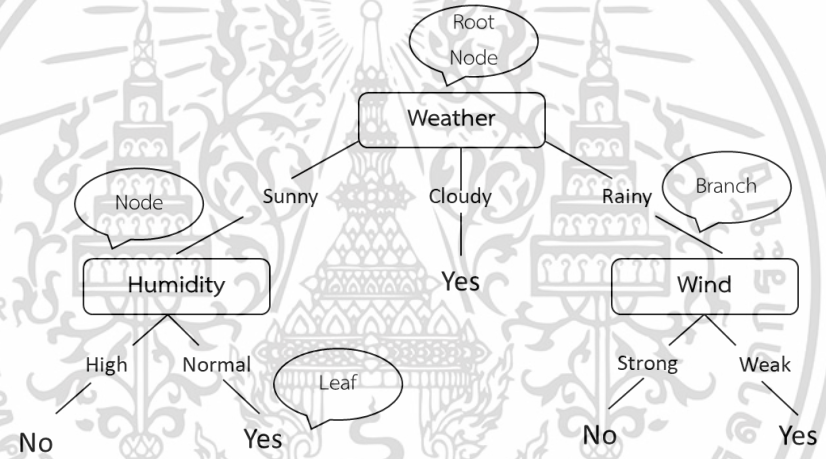
L(B) คือ ฟังก์ชันความควรจะเป็นของฟังก์ชันที่มีตัวแปรอิสระที่กำหนด

สถิติทดสอบ Cox & Snell R Square เป็นการตรวจสอบความสอดคล้องของกับตัวแบบ หรือเปอร์เซ็นต์ที่สามารถอธิบายความแปรปรวนในการวิเคราะห์ความถดถอยลอจิสติกทวิภาค ซึ่ง Cox & Snell R Square มีค่าน้อยกว่า 1 เสมอ สถิติค่านี้จะคล้ายกับค่า R Square ในการวิเคราะห์ถดถอยพหุ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.4 วิธีต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจเป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ ซึ่งเมื่อมีข้อมูลที่ต้องการจำแนกก็จะนำคุณลักษณะต่าง ๆ ของข้อมูลนั้นไปเทียบกับเส้นทางในต้นไม้ จนกระทั่งกลุ่ม (Class) ปลายทางซึ่งก็คือกลุ่มของข้อมูลที่เหมือนกัน ภายในต้นไม้จะประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดจะมีคุณลักษณะเป็นตัวทดสอบกิ่งของต้นไม้ (Branch) ซึ่งแสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (Leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่ม (Class) ก็คือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก (Root Node) โครงสร้างของต้นไม้ตัดสินใจแสดง ดังรูปที่ 2.2 (ภรันยา, 2559)



รูปที่ 2.2 โครงสร้างต้นไม้ตัดสินใจ

ที่มา : ภรัณยา ปาลวิสุทธิ์ (2016)

การสร้างต้นไม้การตัดสินใจ

ขั้นตอนที่สำคัญในเทคนิคการเรียนรู้แบบต้นไม้การตัดสินใจก็ คือการสร้างแบบจำลองของต้นไม้การตัดสินใจ โดยจะต้องเริ่มพิจารณาที่โหนดราก เป็นอันดับแรกก่อนจะดำเนินการพิจารณาใบและกิ่งก้าน การหาโหนดรากจะใช้วิธีการคำนวณหาค่า Information Gain ที่มากที่สุด

การสร้างต้นไม้การตัดสินใจเพื่อใช้ในการจัดหมวดหมู่ (Classification) มีทฤษฎีที่นิยมใช้ในการสร้างต้นไม้การตัดสินใจคือ ทฤษฎี Information Gain เป็นทฤษฎีที่นำมาใช้ในการเลือก Attribute ในแต่ละโหนดของต้นไม้ ซึ่ง Attribute ตัวใดที่มีค่า Information Gain สูงสุด หรือว่ามีค่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Entropy น้อย ก็จะถูกเลือกเป็นแอทริบิวต์ของโหนดนั้น โดยมีสมการการคำนวณค่า Entropy แสดงดังสมการที่ 2.11

$$Entropy(s) = 1 - \sum_{i=1}^{n} \left[P(V_i) \log_2 V_i \right] \tag{2.11}$$

โดยที่

Entropy (s) หรือ E(S) คือ ค่า Entropy ของเซตข้อมูล (S)

S คือ เซตข้อมูลความน่าจะเป็น $P(V_1), P(V_2), ..., P(V_n)$

$P(V_i)$ คือ ความน่าจะเป็นของข้อมูล i ที่สนใจ

เมื่อมีการคำนวณหาค่า Entropy ของระบบแล้ว ต่อไปคือการพิจาณาแอทริบิวต์ ว่าแอทริบิวต์ ไหนที่มีผลกระทบต่อผลลัพธ์มากที่สุด วิธีที่นิยมใช้พิจารณาหาแอทริบิวต์ ที่มีผลกระทบต่อผลลัพธ์มากที่สุดคือ การคำนวณหาค่า Information Gain ซึ่งสามารถคำนวณได้จากสมการต่อไปนี้

$$Gain(S, A) = Entropy(s) - \sum_{v \in Value(A)} \left(\frac{S_v}{S} Entropy(S_v) \right) \tag{2.12}$$

โดยที่

Gain(S,A) คือ ค่า Gain ของแอททริบิวต์ที่สนใจ

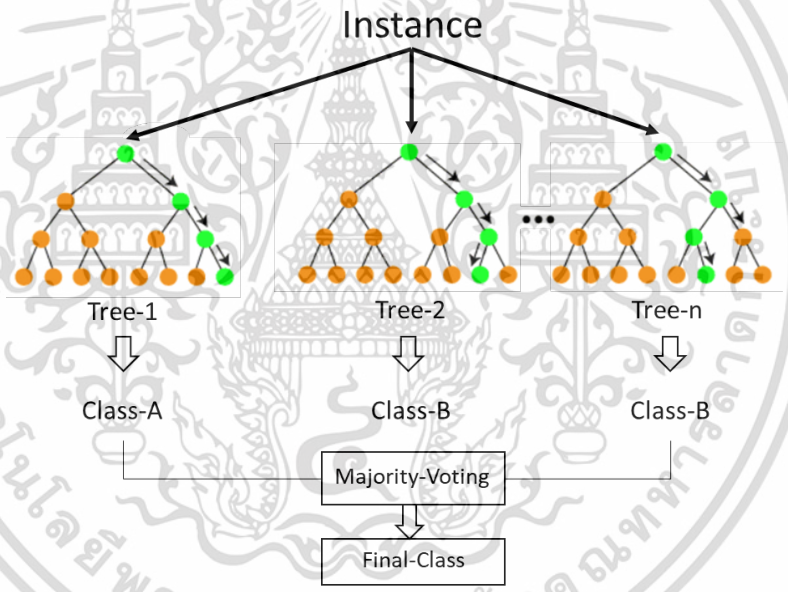
$E(S_y)$ คือ ค่า Entropy ของเซตข้อมูลที่สนใจ

S_y คือ เซตข้อมูลที่สนใจ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.5 วิธีการสุ่มป่าไม้ (Random Forest)

การสุ่มป่าไม้เกิดจากการรวมกลุ่มกันของโครงสร้างต้นไม้ (Breiman , 2001) ซึ่งมีค่าความคลาดเคลื่อนโดยรวมของป่าไม้ถูกเปลี่ยนให้เป็นค่าลิมิต ทำให้จำนวนของต้นไม้ในป่าเพิ่มขึ้น ค่าความคลาดเคลื่อนโดยรวมจะขึ้นกับความมั่นคง (Strength) ของต้นไม้แต่ละต้น รวมถึงความสัมพันธ์กันระหว่างต้นไม้เหล่านั้นโดยจะใช้วิธีการสุ่มเลือกคุณสมบัติเพื่อการแบ่งแยกโหนดทำให้ค่าความผิดพลาดลดลง ขั้นตอนวิธีนี้จะมีประสิทธิภาพมากเมื่อนำไปใช้วิเคราะห์เกี่ยวกับเสียงและการประมาณการขนาดใหญ่ เราสามารถสร้างแบบจำลองที่ใช้ต้นไม้หลาย ๆ ต้นในการตัดสินใจเพื่อนำมาประมวลผล ซึ่งมีความแม่นยำสูง สามารถจัดการข้อมูลได้มากและเหมาะสำหรับข้อมูลที่มีความสำคัญ ดังรูปที่ 2.3



รูปที่ 2.3 การรวมกลุ่มกันของโครงสร้างต้นไม้ตัดสินใจ

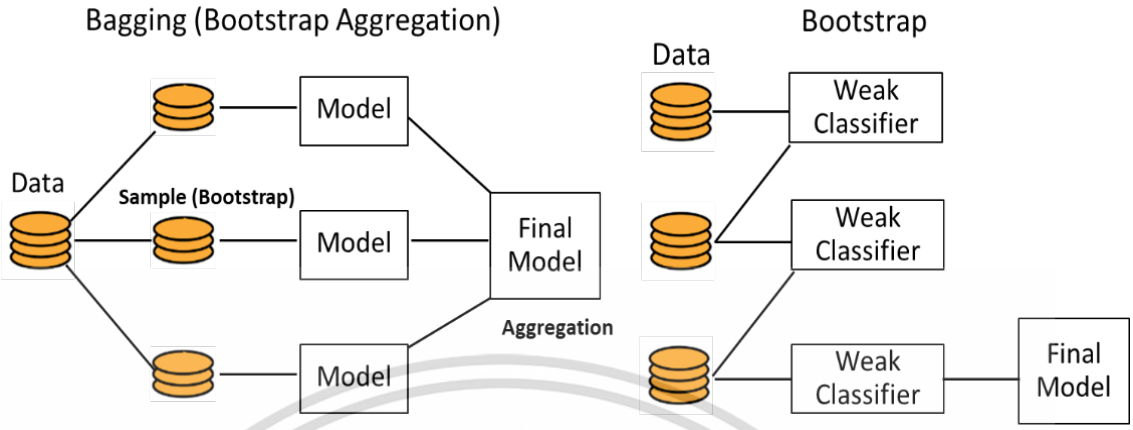
ที่มา : Breiman (2001)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.6 วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting)

บูทติ้ง (Boosting) เป็นเทคนิคเพื่อลดความแปรปรวนและเพิ่มความแม่นยำในการทำนายของตัวจำแนกประเภท โดยวิธีลดความอคติ และมีแนวคิดที่ให้ตัวเรียนรู้ที่อ่อนแอชุดหนึ่ง ทำงานร่วมกันจนสามารถพัฒนามาเป็นตัวเรียนรู้ที่เข้มแข็งได้ การสร้างตัวเรียนรู้ที่อ่อนแอ แต่ละตัวสามารถทำได้โดยการปรับเพิ่มน้ำหนักของการทำนายที่ผิดพลาดให้มากขึ้นในแต่ละรอบ แล้วทำการเรียนรู้ใหม่ ซึ่งจะทำให้โมเดลของตัวจำแนกกลุ่ม เปลี่ยนไปโดยให้ความสำคัญกับความผิดพลาดในรอบที่แล้วมากขึ้น เมื่อได้จำนวนตัวเรียนรู้ที่อ่อนแอมากพอแล้ว จึงนำมารวมกันสร้างเป็นตัวเรียนรู้ที่เข้มแข็งต่อไป ในขั้นตอนการทำงานบูทติ้ง บูทติ้ง คือ การนำตัวจำแนกกลุ่มที่อ่อนแอที่มีความแม่นยำต่ำมาทำนายข้อมูลที่มี หลังจากนั้นเราจะนำตัวใหม่มาแก้ไขข้อผิดพลาดที่มี โดยผลรวมของตัวจำแนกประเภทจะเกิดเป็นตัวจำแนกประเภทใหม่ขึ้นมา เราจะทำแบบนี้ไปเรื่อยๆ จนได้โมเดลที่ดีที่สุดจากผลรวมของการจำแนก ถ้าให้มองภาพรวมการทำงานของบูทติ้ง (Boosting) ก็เหมือนการทำงานเป็นทีมนั่นเอง โดยการเอาตัวจำแนกที่ไม่ได้ดีมากมารวมกันจนทำนายข้อมูลที่ซับซ้อนมากๆได้ แต่ข้อเสียของการใช้บูทติ้ง (Boosting) ก็คือต้องรันหลายครั้งและเป็นลำดับ กว่าจะได้โมเดลที่ต้องการ ต่างจากแบ็กกิ้ง (Bagging) ที่สามารถสุ่มข้อมูลได้แล้วเทรนโมเดลได้พร้อมกันเลย แต่ความฉลาดของเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) นั้น จะสามารถเลือกชนิดของตัวเรียนรู้ที่อ่อนแอ หรือ ตัวจำแนกที่อ่อนแอได้ด้วย ไม่ว่าจะเป็นในรูปแบบต้นไม้ หรือแบบเส้นตรง และในหลาย ๆ ครั้งนั้นเทคนิคบูทติ้ง สามารถทำนายข้อมูลที่มีความซับซ้อนมาก ๆ ได้ มากกว่าการใช้แบ็กกิ้ง เอ็กซ์ทรีมกาเดียนบูทติ้งมีพื้นฐานอยู่บนหลักการของ Gradient Boosting ซึ่งจัดเป็นการใช้โมเดลตัวจำแนกประเภทหลายๆโมเดล มาช่วยในการหาคำตอบ (Ensemble Classifier) ในกลุ่มบูทติ้งประเภทหนึ่ง แต่ได้รับการปรับปรุงให้สามารถทำงานได้อย่างรวดเร็วและมีประสิทธิภาพมากขึ้นกว่าเดิม สามารถใช้ประโยชน์จาก Multithread ได้อย่างเต็มที่ และเพิ่มตัวแปร Regularization เข้าไปเพื่อลดการเกิดการที่โมเดลจดจำรูปแบบของข้อมูล (Overfitting) ลงได้ (Robert et al., 2014)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.4 ความแตกต่างระหว่าง Bagging และ Boosting

ที่มา : Wu and Weng, (2021)

เอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Boosting) ใช้หลักการสร้างต้นไม้แต่ละต้นจะเป็นแบบเรียงลำดับ โดยข้อมูลนำเข้า (Input) แต่ละของต้นไม้แต่ละต้น จะเป็นเอาท์พุต (Output) จากต้นไม้ก่อนหน้า โดยหลักการคือเอ็กซ์ทรีมกาเดียนบูทติ้ง จำทำการสร้างต้นไม้แต่ละต้น เพื่อลดค่าความผิดพลาด ที่เกิดจากต้นไม้ก่อนหน้า โดยใช้วิธี Gradient Descend แล้วนำผลลัพธ์ที่ได้มารวมกันก็จะทำให้ได้ค่าใกล้เคียงกับค่าที่จะทำนาย ซึ่งข้อดีของเอ็กซ์ทรีมกาเดียนบูทติ้ง คือ ความอคติและความแปรปรว0น ลดลงเนื่องจากความผิดพลาดก่อนหน้าถูกแก้ไข ความลึกของชั้นต้นไม้ แค่หนึ่งชั้นก็เพียงพอที่จะได้ค่าประสิทธิภาพที่ดีขึ้นมาก เมื่อเทียบกับ (Bagging Tree) และแรนดอมฟอเรส ที่ต้องเพิ่มชั้นความลึกมากขึ้น เพื่อให้ได้ประสิทธิภาพที่ใกล้เคียง

2.7 วิธีเอดาบูท (AdaBoost)

เอดาบูท เป็นการเรียนรู้ต่อกันเป็นลูกโซ่ (Sequential Ensemble Method) ที่รวม Weak Learners หลาย ๆ ตัวแล้วสร้างเป็น Strong Learning ทำให้ประสิทธิภาพของแบบจำลองจำแนกประเภทมีประสิทธิภาพเพิ่มขึ้น โดยการเรียนรู้ในแต่ละรอบจะมีการกำหนดค่าน้ำหนักของข้อมูลแต่ละรายการ (Classifier Instance) จะเรียนรู้จากค่าน้ำหนักเหล่านั้น โดยถ้าจำแนกประเภทผิด ค่าน้ำหนักของรายการนั้นจะมีค่าน้ำหนักมากขึ้น การจำแนกประเภทของ เอดาบูท คือการถ่วงคะแนนเข้ากับคำตอบของแต่ละ Instance แล้วเลือกคำตอบจากค่าเป้าหมายที่ได้รวมแล้วค่าน้ำหนักมากที่สุด สังเกตว่า เอดาบูท ให้ค่ากับ Instance ที่ "เก่ง" มากกว่าที่ "ไม่เก่ง" ในขณะที่ Ensemble Learning แบบ Random Forest .ให้ค่าน้ำหนัก Instance เท่ากัน เอดาบูท ไม่ได้มีเป้าหมายที่ทำให้ Classifier Instance ตัวถัด ๆ ไปมีความแม่นยำขึ้น แต่ทำงานโดยการเก็บคะแนนความแม่นยำของแต่ละ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Instance เอาไว้เพื่อเอามาถ่วงน้ำหนักการตัดสินใจเมื่อต้องการพยากรณ์ในภายหลัง (สุรพงค์, 2562) ดังสมการที่ 2.13

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{2.13}$$

โดยที่ H คือ Strong Learning

t คือ จำนวนรอบการเรียนรู้ t = 1,2,3,...,T

α คือ Voting Power

h_t คือ Weak Learner

2.8 การแทนค่าของตัวแปรประเภทกลุ่ม (Categorical Encoding)

การเตรียมข้อมูลสำหรับสร้างตัวแบบการเรียนรู้ของเครื่อง (Machine Learning Model) โดยข้อมูลนำเข้าจะต้องเป็นตัวเลข แต่ในการทำงานกับข้อมูลจริง พบว่าข้อมูลที่ได้เป็นข้อความ เช่น (ใช่, ไม่ใช่) หรือ (เพศชาย, เพศหญิง, เพศทางเลือก) เป็นต้น ดังนั้นจำเป็นต้องแปลงข้อมูลที่เป็นข้อความหรือหมวดหมู่ ให้เป็นตัวเลขก่อน โดยวิธีการแทนค่าของตัวแปรประเภทกลุ่มในภาษาไพธอนมีหลายวิธี ได้แก่

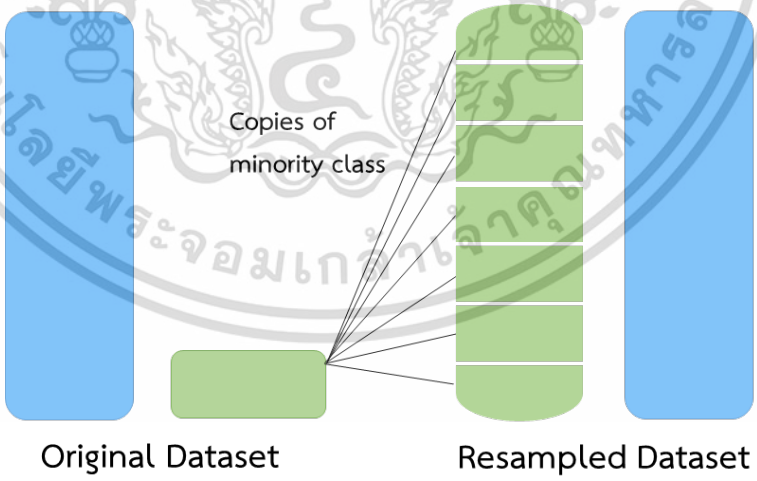
1. **One-Hot Encoding** คือการทำข้อมูลที่ถูกเก็บในลักษณะ Categorical ทั้งในลักษณะที่มีลำดับ (Ordinal Number) และไม่มีลำดับ (Nominal Number) เปลี่ยนให้อยู่ในรูปแบบของ Binary Values ที่มีค่า 0 หรือ 1 เท่านั้น
2. **Dummy Variable Encoding** ซึ่งมีลักษณะเหมือนกับ One-Hot Encoding ทั้งหมด แต่แตกต่างกันที่จำนวนคอลัมน์ น้อยกว่า 1 คอลัมน์
3. **Label Encoding** เป็นวิธีอย่างง่ายโดยแปลงแต่ละค่าในคอลัมน์/ตัวแปรให้เป็นตัวเลขเริ่มเรียงลำดับจากศูนย์ เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.9 เทคนิคการปรับเพิ่มและลดข้อมูล

การจัดการความไม่สมดุลของข้อมูล (Imbalance) เกิดขึ้นเมื่อจำนวนข้อมูลของแต่ละคลาสแตกต่างกันมากทำให้ผลลัพธ์จากการจำแนกข้อมูลมีความโน้มเอียงไปทางข้อมูลที่มีจำนวนมาก ส่งผลให้ตัวแบบเรียนรู้ข้อมูลจริงผิดพลาดไป เป็นการประยุกต์เอาวิธีสุ่มตัวอย่างซึ่งเป็นวิธีการทางสถิติ เพื่อสร้างข้อมูลสำหรับการสอน โดยมีจุดประสงค์เพื่อให้จำนวนสมาชิกในข้อมูลทั้งสองกลุ่มมีความสมดุลกัน ซึ่งสามารถแบ่งออกได้เป็น 3 ประเภท ดังนี้

2.9.1 วิธีการสุ่มเพิ่ม (Over Sampling) เป็นเทคนิคการสุ่มตัวอย่างแบบเพิ่ม คือ วิธีการเพิ่มขึ้นของข้อมูล การเพิ่มจำนวนของข้อมูลที่เป็นกลุ่มตัวอย่างขนาดเล็กจนกระทั่งจำนวนข้อมูลเท่ากับจำนวนกลุ่มตัวอย่างขนาดใหญ่ หลักการพื้นฐานของการสุ่มตัวอย่างแบบเพิ่มคือการสุ่มสุ่มตัวอย่างแบบเพิ่ม ซึ่งเป็นการเพิ่มข้อมูลของตัวอย่างโดยสุ่มเลือกข้อมูลของกลุ่มตัวอย่างขนาดเล็ก และทำการเพิ่มจำนวนข้อมูลของกลุ่มตัวอย่างขนาดเล็ก ทำเช่นนี้ไปจนกระทั่งจำนวนข้อมูลมีจำนวนเท่ากับจำนวนกลุ่มตัวอย่างขนาดใหญ่ ยกตัวอย่างเช่น ข้อมูลมีอัตราส่วนอยู่ที่ ข้อมูลส่วนมากอยู่ที่ 7 ส่วน ต่อข้อมูลส่วนน้อย 1 ส่วน จากวิธีการสุ่มเพิ่มแบบสุ่ม จะทำการนำข้อมูลส่วนน้อยมาทำซ้ำเพิ่มขึ้นจนมี 7 ส่วนเท่ากันกับข้อมูลส่วนมาก ดังรูปที่ 2.5



รูปที่ 2.5 ตัวอย่างการสุ่มเพิ่ม

ที่มา : Lahera (2019)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.9.2 วิธีการสุ่มเกินโดยเทคนิค SMOTE (Synthetic Minority Over-Sampling Technique) เป็นเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์สำหรับแก้ไขปัญหาความไม่สมดุลระหว่างกลุ่ม โดยจะเป็นกระบวนการจัดการข้อมูลก่อนการสร้างตัวแบบสำหรับการแก้ไขปัญหาความไม่สมดุลระหว่างกลุ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมเหมือนกับการวิธีการสุ่มเพิ่ม แต่จะทำการสังเคราะห์ข้อมูลขึ้นมาใหม่จากข้อมูลเดิมที่มีอยู่ตามหลักการเพื่อนบ้านที่อยู่ใกล้ที่สุดในการขยายขอบเขตการตัดสินใจของตัวแบบ ซึ่งในขั้นตอนการสังเคราะห์ข้อมูลใหม่มีขั้นตอนดังนี้ คือระบุเพื่อนบ้านและทำการสังเคราะห์ข้อมูลขึ้นมาระหว่างข้อมูลสองจุด

ตัวอย่างเช่นการสร้างจุด $m_1(a,b,...,z)$ ระหว่าง $M_1(x_a,x_b,...,x_z)$ และ $M_2(y_a,y_b,...,y_z)$ เมื่อ ดังสมการที่ 2.14

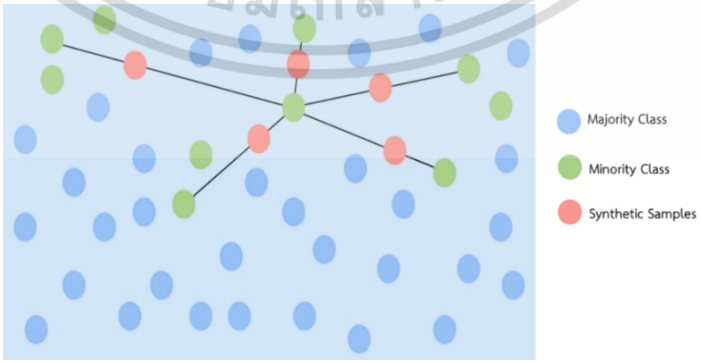
$$\begin{aligned} a &= x_a + (y_a - x_a) \times \text{rand(UNIFORM)} \\ a &= x_b + (y_b - x_b) \times \text{rand(UNIFORM)} \\ &\vdots \\ a &= x_z + (y_z - x_z) \times \text{rand(UNIFORM)} \end{aligned} \tag{2.14}$$

โดยที่

m_1 คือ จุดที่สังเคราะห์ขึ้นมาใหม่ระหว่าง $M_1(x_a,x_b,...,x_z)$ และ $M_2(y_a,y_b,...,y_z)$

$x_a,x_b,...,x_z$ คือ ข้อมูลในค่าสังเกตที่จุด M_1

$y_a,y_b,...,y_z$ คือ ข้อมูลในค่าสังเกตที่จุด M_2

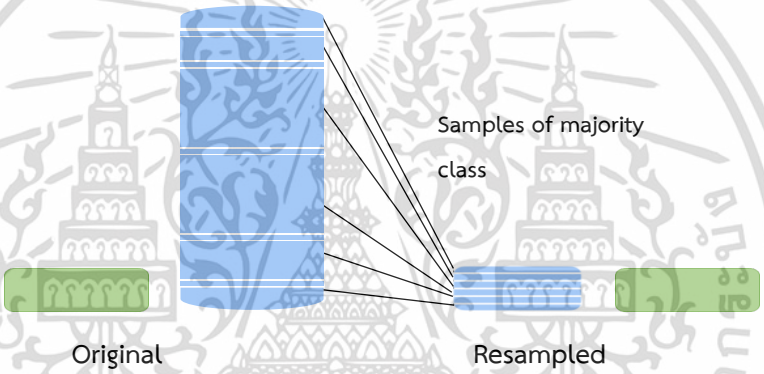


รูปที่ 2.6 ตัวอย่างวิธีการสังเคราะห์ข้อมูลเพิ่ม

ที่มา : Lahera (2019)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.9.3 วิธีการสุ่มลด (Under Sampling) เป็นเทคนิคการสุ่มตัวอย่างแบบลด คือ การลดจำนวนของข้อมูลที่เป็นกลุ่มตัวอย่างขนาดใหญ่จนกระทั่งจำนวนข้อมูลเท่ากับจำนวนกลุ่มตัวอย่างขนาดเล็ก หลักการพื้นฐานของการสุ่มตัวอย่างแบบลดคือการสุ่มสุ่มตัวอย่างแบบลด ซึ่งเป็นการลบข้อมูลของตัวอย่างโดยสุ่มเลือกข้อมูลของกลุ่มตัวอย่างขนาดใหญ่ และทำการลดจำนวนข้อมูลของกลุ่มตัวอย่างขนาดใหญ่ ทำเช่นนี้ไป จนกระทั่งจำนวนข้อมูลมีจานวนเท่ากับจำนวนกลุ่มตัวอย่างขนาดเล็ก ยกตัวอย่างเช่น ข้อมูลมีอัตราส่วนของข้อมูลส่วนมากอยู่ที่ 7 ส่วน ต่อ ข้อมูลส่วนน้อย 1 ส่วน จากวิธีการสุ่มลดแบบสุ่ม จะทำการสุ่มหาชุดข้อมูลที่เก็บไว้ในจำนวนที่เท่ากันกับข้อมูลส่วนน้อย จากนั้นจะลบข้อมูลที่เหลือออก ข้อมูลทั้งสองกลุ่มจะมีจำนวนเท่ากัน ดังรูปที่ 2.7



รูปที่ 2.7 ตัวอย่างการสุ่มลด

ที่มา : Lahera (2019)

2.10 การคัดเลือกคุณลักษณะ (Feature Selection)

การคัดเลือกคุณลักษณะ (Feature Selection) คือขั้นตอนสำหรับคัดเลือกคุณลักษณะที่มีความสำคัญกับการสร้างแบบจำลองให้มีคุณภาพ นอกจากนี้ยังมีประโยชน์ในด้านการลดจำนวนคุณลักษณะของชุดข้อมูลที่มีจำนวนมากเกินความจำเป็น เพื่อลดระยะเวลาในการสร้างแบบจำลอง โดยจะคัดเลือกคุณลักษณะที่มีความสัมพันธ์ (Correlation) สูงกับข้อมูลคำตอบ ซึ่งจะส่งผลให้ประสิทธิภาพในการทำงานของแบบจำลองดียิ่งขึ้น ในขณะเดียวกันคุณลักษณะนั้นต้องมีความสัมพันธ์ระหว่างคุณลักษณะด้วยกันต่ำ เพื่อให้ข้อมูลมีความเป็นอิสระต่อกันมากที่สุด สำหรับเทคนิค Feature Selection สามารถแบ่งได้ 2 วิธีหลัก ๆ คือ วิธีฟิลเตอร์ (Filter Methods), วิธีแรปเปอร์ (Wrapper Methods)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1. **Filter Methods** เป็นวิธีที่พยายามคัดเลือกคุณลักษณะที่สำคัญโดยการคำนวณค่าน้ำหนักหรือค่าความสัมพันธ์ของแต่ละคุณลักษณะ โดยเลือกเฉพาะคุณลักษณะที่มีความสำคัญเก็บเอาไว้ เช่น เทคนิค Correlation Based Feature Selection (CFS)
2. **Wrapper Methods** เป็นวิธีคัดเลือกคุณลักษณะที่สำคัญ โดยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูลมาสร้างเซตของคุณลักษณะใหม่โดยการเพิ่มหรือลดจำนวนคุณลักษณะจากเซตเดิม เช่น เทคนิค Forward Selection เทคนิค Backward Elimination เทคนิค Evolutionary Selection

2.11 การวัดประสิทธิภาพและการประเมินตัวแบบ (Evaluation Phase)

เป็นขั้นตอนในการประเมินผลจากการเรียนรู้ด้วยเครื่อง โดยตัวแบบจะถูกประเมินด้วยวิธีดังนี้

1) ค่าความถูกต้อง (Accuracy) เป็นค่าที่ใช้วัดประสิทธิภาพหรือเป็นตัววัดขนาดของความผิดพลาด หากค่าความถูกต้องมีค่ามากจะมีความผิดพลาดน้อย ดังสมการที่ 2.15

$$\text{ค่าความถูกต้อง} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.15}$$

2) ค่าความแม่นยำ (Precision) เป็นอัตราส่วนของการทำนายข้อมูลในคลาสได้ถูกต้องจากจำนวนข้อมูลทั้งหมดในคลาสนั้น ดังสมการที่ 2.16

$$\text{ค่าความแม่นยำ} = \frac{TP}{TP+FP} \tag{2.16}$$

3) ค่าความไว (Recall) เป็นอัตราส่วนของการค้นพบคลาสที่ใกล้เคียงกับคำขอและมีการค้นคืนให้กับผู้ใช้กับเอกสารที่ตรงตามคำขอทั้งหมดแม่นยำ ดังสมการที่ 2.17

$$\text{ค่าความไว} = \frac{TP}{TP+FN} \tag{2.17}$$

4) ค่าความถ่วงดุล (F-Measure) เป็นค่าที่แสดงความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความไว เพื่อหาค่าความถ่วงดุล โดยค่าที่ได้จากการคำนวณจะมีค่าอยู่ระหว่าง 0 ถึง 1 ถ้าค่าที่คำนวณ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ได้เข้าใกล้ 1 หมายความว่า การให้ผลในการจำแนกมีประสิทธิภาพสูง และถ้าค่าคำนวณได้เข้าใกล้ 0 หมายความว่า การให้ผลในการจำแนกมีประสิทธิภาพต่ำ ดังสมการที่ 2.18

$$\text{ค่าความถ่วงดุล} = \frac{2(\text{Precision})(\text{Recall})}{(\text{Precision})+(\text{Recall})} \quad (2.18)$$

5) ค่าอัตราความถูกต้องเชิงลบ (TN Rate) เป็นอัตราส่วนของจำนวนข้อมูลที่ทำนายได้ถูกต้องในคลาส Negative จากจำนวนข้อมูลที่แท้จริงในคลาส Negative ดังสมการที่ 2.19

$$\text{อัตราความถูกต้องเชิงลบ} = \frac{\text{TN}}{\text{FP}+\text{TN}} \quad (2.19)$$

6) ค่าอัตราความผิดพลาดเชิงบวก (FP Rate) เป็นอัตราส่วนของจำนวนข้อมูลที่ทำนายได้ผิดว่าคำตอบอยู่ในคลาส Positive จากจำนวนข้อมูลที่แท้จริงในคลาส Positive ดังสมการที่ 2.20

$$\text{อัตราความผิดพลาดเชิงบวก} = \frac{\text{FP}}{\text{FP}+\text{TN}} \quad (2.20)$$

7) ค่าอัตราความผิดพลาดเชิงลบ (FN Rate) เป็นอัตราส่วนของจำนวนข้อมูลที่ทำนายได้ผิดว่าคำตอบอยู่ในคลาส Negative จากจำนวนข้อมูลที่แท้จริงในคลาส Negative ดังสมการที่ 2.21

$$\text{อัตราความผิดพลาดเชิงลบ} = \frac{\text{FN}}{\text{TP}+\text{FN}} \quad (2.21)$$

8) ค่าพื้นที่ใต้เส้นโค้ง (AUC) เป็นการวัดประสิทธิภาพ โดยที่เส้นในแกนนอนจะเป็ True Positive Rate ส่วนเส้นในแกนตั้งจะเป็น False Positive Rate ค่าการวัดโดยใช้ AUC นี้ จะเริ่มที่ 0 ถึง 1 โดยที่ 0 หมายถึง ตัวแบบนั้นมีประสิทธิภาพต่ำ ส่วน 1 หมายถึง แบบจำลองนั้นมีประสิทธิภาพสูงที่สุด

โดยที่ True Positive (TP) คือ จำนวนข้อมูลที่ทายถูกว่าเป็นคลาสที่กำลังสนใจอยู่
True Negative (TN) คือ จำนวนข้อมูลที่ทายถูกว่าเป็นคลาสที่ไม่ได้กำลังสนใจอยู่
False Positive (FP) คือ จำนวนข้อมูลที่ทายผิดมาเป็นคลาสที่กำลังสนใจอยู่
False Negative (FN) คือ จำนวนข้อมูลที่ทายผิดมาเป็นคลาสที่ไม่ได้กำลังสนใอยู่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.12 งานวิจัยที่เกี่ยวข้อง

Khare et al. (2018) ได้อธิบายการทำงานเรื่อง Decision Tree, Random Forest, Support Vector Machine Learning และ Logistic Regression โดยใช้กับชุดข้อมูลที่มีเบ้สูงและทำงานกับชุดข้อมูลประเภทดังกล่าว การประเมินประสิทธิภาพขึ้นอยู่กับความถูกต้อง ความไว และความแม่นยำ ผลปรากฏว่าความถูกต้องในการถดถอยลอจิสติกคือ 97.7% สำหรับ Decision Tree คือ 95.5% สำหรับ Random Forest คือ 98.6% สำหรับ Support Vector Machine Learning คือ 97.5% ซึ่งสามารถได้ข้อสรุปดังนี้ อัลกอริทึม Random Forest มีความถูกต้องสูงสุดเมื่อเทียบกับอัลกอริทึมอื่น ๆ และถือเป็นอัลกอริทึมที่ดีที่สุดในการตรวจจับการฉ้อโกงสำหรับชุดข้อมูลที่มีปัญหาความไม่สมดุล

Rajora et al. (2018) ได้ทำการวิจัยเชิงเปรียบเทียบการตรวจสอบการฉ้อโกงบัตรเครดิตโดยใช้วิธี ML กับชุดข้อมูลผู้ถือบัตรยุโรป ได้แก่ Random Forest (RF) และ K-Nearest Neighbors (KNN) ซึ่งพิจารณาความถูกต้องและพื้นที่ด้านล่างเส้นโค้งเป็นตัวชี้วัดประสิทธิภาพ ผลลัพธ์ความถูกต้องของ RF คิดเป็นร้อยละ 94.9 และ AUC คิดเป็น 0.94 ในทางตรงกันข้ามความถูกต้องของ KNN คิดเป็นร้อยละ 93.2 และ AUC คิดเป็น 0.93 ทั้ง ๆ ที่ผลที่ได้คือหวังว่าการศึกษานี้ไม่ได้ตรวจสอบชุดข้อมูลที่ใช้มีปัญหาระดับความไม่สมดุล ผู้เขียนพิจารณาความถูกต้องและพื้นที่ใต้เส้นโค้ง (AUC) เป็นตัวชี้วัดประสิทธิภาพหลัก ผลการวิจัยพบว่า Random Forest มีความแม่นยำคิดเป็นร้อยละ 94.9 และ AUC คิดเป็น 0.94 ในทางตรงกันข้าม K-Nearest Neighbors ให้ความแม่นยำคิดเป็นร้อยละ 93.2 และ AUC คิดเป็น 0.93 แม้ว่าผลลัพธ์เหล่านี้จะมีแนวโน้มที่ดี แต่งานวิจัยนี้ไม่ได้ตรวจสอบปัญหาความไม่สมดุลของคลาสที่มีอยู่ในชุดข้อมูลที่ใช้

Varmedja et al. (2019) ได้เสนอวิธีการตรวจจับการฉ้อโกงบัตรเครดิตโดยใช้ Machine Learning ซึ่งใช้ชุดข้อมูลการฉ้อโกงบัตรเครดิตที่มาจาก Kaggle ชุดข้อมูลนี้มีธุรกรรมที่ทำขึ้นภายใน 2 วันโดยผู้ถือบัตรเครดิตในยุโรป เพื่อจัดการกับปัญหาความไม่สมดุลของคลาสที่มีอยู่ในชุดข้อมูล ผู้วิจัยได้ใช้เทคนิค SMOTE และวิธีการ Machine Learning ได้แก่ RF, NB และ MLP ผลการทดลองแสดงให้เห็นว่าอัลกอริทึม RF ทำงานได้ดีที่สุดโดยมีความถูกต้องในการตรวจจับการฉ้อโกงคิดเป็นร้อยละ 99.96 วิธี NB และ MLP ได้คะแนนความถูกต้องคิดเป็นร้อยละ 99.23 และ 99.93 ตามลำดับ ผู้วิจัยยอมรับว่าควรมีการวิจัยเพิ่มเติมเพื่อใช้วิธีการเลือกคุณลักษณะที่สามารถปรับปรุงความถูกต้องของวิธี ML อื่น ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Tran and Dang (2021) เสนอให้ใช้วิธีการสุ่มตัวอย่างโดยวิธีของเทคนิค Synthetic Minority Oversampling Technique (SMOTE) และ Adaptive Synthetic (ADASYN) เพื่อทำงานกับชุดข้อมูลที่ไม่สมดุลเพื่อให้ได้ชุดข้อมูลที่สมดุล อัลกอริทึม ML ต่อไปนี้ใช้กับข้อมูลที่สมดุลได้แก่ Random Forest, K-Nearest Neighbors, Logistic Regression และ Decision Tree ในการประเมินประสิทธิภาพของแบบจำลองเหล่านี้ จะใช้การวัดการจำแนกที่ครอบคลุมรวมทั้งการวัดพื้นฐานและการวัดกราฟิก ซึ่งสังเกตเห็นว่าขั้นตอนวิธี ML แสดงเป็นบวกผลของการฉ้อโกงหลังจากการสุ่มตัวอย่างชุดข้อมูลใหม่

Awoyemi et al. (2017) ได้ทำการวิเคราะห์ข้อมูล K-Nearest Neighbors, Logistic Regression และ Naive Bayes เมื่อนำไปใช้กับข้อมูลธุรกรรมบัตรเครดิต ซึ่งสุ่มตัวอย่างเพิ่มเติมโดยใช้เทคนิค Synthetic Minority Over-sampling Technique (SMOTE) ผลปรากฏว่า KNN ทำได้ดีกว่าอีก 2 วิธี ประสิทธิภาพถูกเปรียบเทียบโดยค่าความไว ค่าความแม่นยำและค่าสัมประสิทธิ์สหสัมพันธ์

Trivedi et al. (2020) เสนอเครื่องมือตรวจจับการฉ้อโกงบัตรเครดิตที่มีประสิทธิภาพโดยใช้วิธี ML ในงานวิจัยนี้ผู้เขียนได้พิจารณาเทคนิค ML ที่มีการเรียนรู้แบบมีผู้สอน ซึ่งรวมถึง Gradient Boosting (GB) และ Random Forest ผู้เขียนประเมินวิธีการเหล่านี้โดยใช้ชุดข้อมูลผู้ถือบัตรของยุโรป ตัวชี้วัดประสิทธิภาพที่ใช้ในการประเมินประสิทธิผลของแนวทางที่เสนอรวมถึงความถูกต้องและความแม่นยำ ผลการทดลองแสดงให้เห็นว่า GB ให้ความถูกต้องคิดเป็นร้อยละ 94.01 และความแม่นยำคิดเป็นร้อยละ 93.99 ในทางกลับกัน RF มีความถูกต้องคิดเป็นร้อยละ 94.00 และความแม่นยำคิดเป็นร้อยละ 95.98

Tanouz et al. (2020) นำเสนอกรอบการตรวจจับการฉ้อโกงบัตรเครดิตโดยใช้อัลกอริทึม ML ในงานวิจัยนี้ใช้ชุดข้อมูลผู้ถือบัตรของยุโรปเพื่อประเมินประสิทธิภาพของวิธีการที่เสนอ นอกจากนี้ผู้วิจัยยังใช้เทคนิคการสุ่มตัวอย่างลดเพื่อแก้ปัญหาความไม่สมดุลของคลาสที่มีอยู่ในชุดข้อมูลที่ใช้ วิธีการ ML ที่พิจารณาในงานนี้ ได้แก่ Random Forest และ Logistic Regression ทั้งนี้ใช้ความแม่นยำเป็นตัวชี้วัดประสิทธิภาพหลัก ผลการวิจัยพบว่าความถูกต้องในการตรวจจับการฉ้อโกงโดยวิธี RF คิดเป็นร้อยละ 91.24 ในทางตรงกันข้าม วิธี LR ให้ความถูกต้องคิดเป็นร้อยละ 95.16 นอกจากนี้ ได้นำเสนอเมทริกซ์ความสับสนเพื่อยืนยันว่าวิธีการที่เสนอเหล่านี้ดำเนินการอย่างเหมาะสมสำหรับกลุ่มบวกและลบ ผลการศึกษาพบว่าปัญหาความไม่สมดุลของคลาสที่มีอยู่ในชุดข้อมูลผู้ถือบัตรเครดิตของยุโรปจำเป็นต้องมีการตรวจสอบเพิ่มเติม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการทบทวนวรรณกรรมและศึกษางานวิจัยที่เกี่ยวข้องผู้วิจัยจึงได้รวบรวมตัวแปรและวิธีการเบื้องต้นสำหรับการตรวจจับการฉ้อโกงบัตรเครดิตได้ดังตารางที่ 2.1

ตารางที่ 2.1 สรุปการทบทวนวรรณกรรมที่เกี่ยวข้อง

ผู้วิจัย	วัตถุประสงค์	ตัวแปรที่ใช้	คำอธิบาย	วิธีการ
Y. Lucas. et al. (2020)	เปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่องต่าง ๆ เพื่อทำนายการฉ้อโกงบัตรเครดิตให้มีประสิทธิภาพที่ดีที่สุด	Step	ชั่วโมงของเวลา	Random Forest Logistic Regression Decision Tree AdaBoost
		type	ประเภทการชำระเงิน	
		amount	จำนวนเงิน	
		nameOrig	ชื่อลูกค้า	
		oldbalanceOrg	ยอดเงินเริ่มต้นก่อนการทำธุรกรรม	
		newbalanceOrig	ยอดคงเหลือใหม่หลังการทำธุรกรรม	
		nameDest	ชื่อร้านค้า	
		oldbalanceDest	ยอดเงินร้านค้าเริ่มต้นก่อนการทำธุรกรรม	
		newbalanceDest	ยอดร้านค้าคงเหลือใหม่หลังการทำธุรกรรม	
		isFraud	ผลการตรวจสอบการฉ้อโกง	
		isFlaggedFraud	ผลการตรวจสอบการโอนเงินจำนวนมาก	

ผู้วิจัย	วัตถุประสงค์	ตัวแปรที่ใช้	คำอธิบาย	วิธีการ
Fang. et al. (2019)	ตรวจจับการฉ้อโกงบัตรเครดิตโดยอาศัยการเรียนรู้ของเครื่องและแก้ไขชุดข้อมูลให้สมดุลกันเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึม	Age zip Amount Category of consumption Job Telephone number	อายุ รหัสไปรษณีย์ จำนวนเงิน กลุ่มร้านค้า อาชีพ เบอร์โทรศัพท์	Random Forest Gradient Boosting
Gupta. et al. (2019)	การปรับสมดุลของชุดข้อมูลและเปรียบเทียบอัลกอริทึม	V1-V28 Time Amount Class	ชื่อตัวแปรที่ใช้แทน เวลาที่ทำธุรกรรม จำนวนเงิน ผลการตรวจสอบการฉ้อโกง	Random Forest Logistic Regression Decision Tree
Ayorinde (2021)	เพื่อตรวจสอบแนวทางที่เป็นไปได้ในการฉ้อโกงบัตรเครดิตที่ส่งผลกระทบในทางลบต่อสถาบันการเงินและเปรียบเทียบว่าอัลกอริทึมใดสามารถทำนายธุรกรรมที่เป็นการฉ้อโกงได้ดีที่สุด	Transaction Date and Time Merchant Category Amount First name	วันที่และเวลาที่ทำธุรกรรม ร้านค้า กลุ่มร้านค้า จำนวนเงิน ชื่อ	Random Forest Logistic Regression Decision Tree Xgboost

ผู้วิจัย	วัตถุประสงค์	ตัวแปรที่ใช้	คำอธิบาย	วิธีการ
		Last name	นามสกุล	
		Gender	เพศ	
		Street	ที่อยู่	
		isFraud	ผลการตรวจสอบการฉ้อโกง	

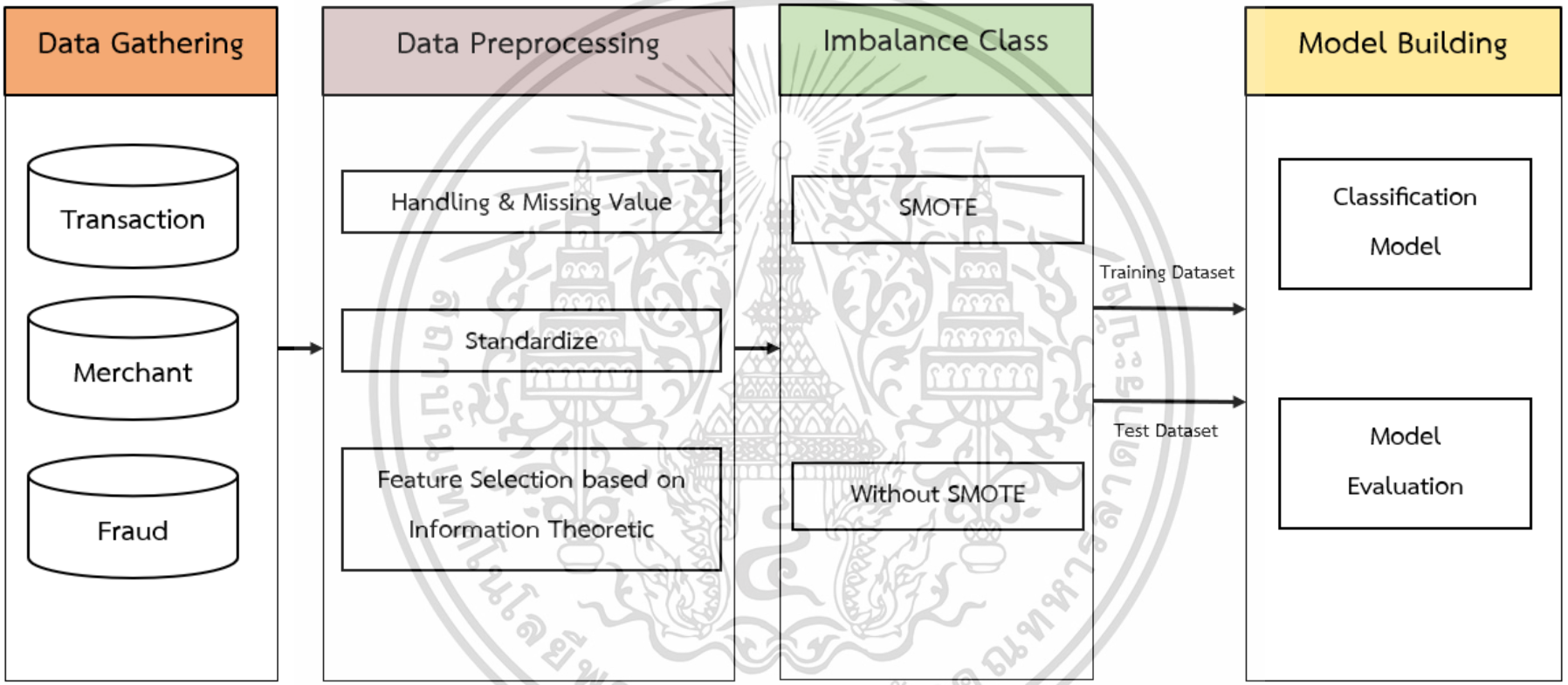
บทที่ 3
วิธีการดำเนินงานวิจัย

การวิจัยครั้งนี้จะทำการตรวจจับฉ้อโกงบัตรเครดิตโดยเทคนิคการเรียนรู้ของเครื่อง ซึ่งจะนำชุดข้อมูลที่เกี่ยวข้องกับบัตรเครดิตและการฉ้อโกงบัตรเครดิตมาทำการพิจารณา โดยใช้การวิเคราะห์วิธีการวิเคราะห์การถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท และเปรียบเทียบประสิทธิภาพ ในการศึกษาแต่ละวิธีจะใช้โปรแกรมไพธอนเข้ามาช่วยในการวิเคราะห์ข้อมูลและดำเนินการตามวัตถุประสงค์ของงานวิจัยต่อไป บทนี้จะกล่าวถึงการวางแผนการวิจัยและวิธีการดำเนินการวิจัยดังนี้

3.1 ขั้นตอนการดำเนินงานวิจัย

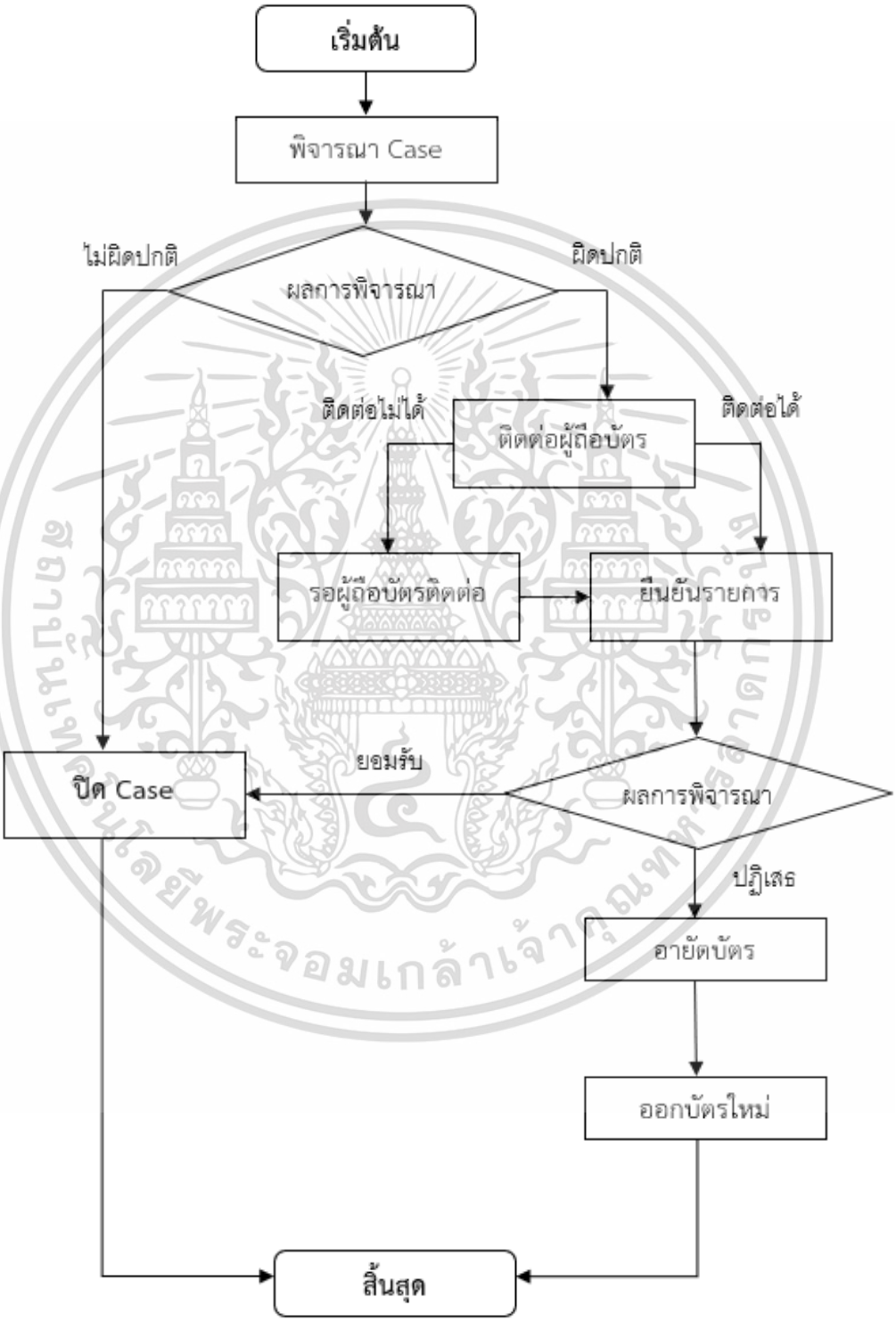
1. ทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง
2. เก็บรวบรวมข้อมูลเกี่ยวกับรายการธุรกรรมการใช้จ่ายบัตรเครดิตและร้านค้า รวมทั้งผลการตรวจสอบการฉ้อโกงบัตรเครดิต
3. จัดเตรียมข้อมูลสำหรับใช้ในการวิเคราะห์ ได้แก่ การจัดการ Missing Value, ความซ้ำซ้อนของข้อมูล การแปลงข้อมูลให้อยู่ในรูปแบบมาตรฐานเดียวกัน
4. พิจารณาความไม่สมดุลของชุดข้อมูล (Imbalance Class) โดยใช้เทคนิค SMOTE
5. ทำการจัดแบ่งชุดข้อมูลเป็น Training dataset และ Test dataset
6. สร้างตัวแบบพยากรณ์ด้วยวิธีการจำแนกกลุ่มทั้งหมด 5 วิธี ได้แก่ วิธีการวิเคราะห์การถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท ประเมินประสิทธิภาพของตัวแบบการจำแนกกลุ่มก่อนและหลังการแก้ปัญหาข้อมูลที่ไม่สมดุลกันโดยพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ ค่าความไว
7. สรุปและข้อเสนอแนะ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.1 ขั้นตอนการดำเนินงานวิจัย

3.2 ขั้นตอนการปฏิบัติงานของหน่วยงาน Fraud



รูปที่ 3.2 ขั้นตอนการปฏิบัติงานของหน่วยงาน Fraud

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 3.2 ขั้นตอนการปฏิบัติงานของหน่วยงาน Fraud มีดังนี้

1. พนักงานเลือกพิจารณา Case เพื่อตรวจสอบรายการผิดปกติ
2. ผลพิจารณา เป็นรายการผิดปกติหรือไม่
   - ถ้าผิดปกติ พนักงานดำเนินการติดต่อผู้ถือบัตรทางโทรศัพท์เพื่อยืนยันรายการใช้บัตร
   - ถ้าไม่ผิดปกติ พนักงานเลือกปิด Case
3. ผลการพิจารณา สามารถติดต่อผู้ถือบัตรได้หรือไม่
   - ถ้าติดต่อไม่ได้ รอผู้ถือบัตรติดต่อกลับและยืนยันรายการใช้บัตรกับผู้ถือบัตร
   - ถ้าติดต่อได้ ยืนยันรายการใช้บัตรกับผู้ถือบัตร
4. ผลการยืนยันรายการ ผู้ถือบัตรปฏิเสธรายการใช่หรือไม่
   - ถ้าไม่ปฏิเสธ ปิด Case ลูกค้าสามารถใช้รายการใช้จ่ายได้ปกติ
   - ถ้าปฏิเสธ อายัดและออกบัตรใหม่เพื่อให้ลูกค้าสามารถใช้รายการใช้จ่ายได้ปกติ

3.3 ตัวแปรที่ใช้ในการวิจัย

จากตารางที่ 3.1 ตัวแปรที่ใช้ในการวิเคราะห์แบ่งออกเป็น

- **ตัวแปรอิสระ ได้แก่** Amount, Online, ECOM, Group Merchant, Credit limit, Country, Card, Date, Time, State
- **ตัวแปรตาม** คือ Fraud

ตารางที่ 3.1 ตัวแปรและระดับค่าของข้อมูลที่ใช้ในการวิจัย

ลำดับที่	ชื่อตัวแปร	สเกลข้อมูล	คำอธิบาย
1	Amount	Ratio Scale	จำนวนเงินที่ทำรายการ
2	Online	Nominal Scale	การทำธุรกรรมออนไลน์
3	ECOM	Nominal Scale	ระบบการยืนยันการชำระเงิน
4	Group Merchant	Nominal Scale	ประเภทกลุ่มของร้านค้า
5	Credit limit	Ratio Scale	จำนวนวงเงินบัตรเครดิต
6	Group Country	Nominal Scale	โซนกลุ่มประเทศที่ลูกค้าทำธุรกรรม
7	Card	Nominal Scale	ประเภทการชำระเงินบัตรเครดิต
8	Date	Nominal Scale	ช่วงวันที่ที่ลูกค้าทำธุรกรรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับที่	ชื่อตัวแปร	สเกลข้อมูล	คำอธิบาย
9	Time	Nominal Scale	ช่วงเวลาที่ลูกค้าทำธุรกรรม
10	State	Nominal Scale	กลุ่มทวีปที่ลูกค้าทำธุรกรรม
11	Fraud	Nominal Scale	ผลการตรวจสอบการฉ้อโกง

3.4 เครื่องมือที่ใช้ในการวิจัย

ผู้วิจัยได้เลือกซอฟต์แวร์ที่นำมาใช้ในการวิจัย โปรแกรม Jupyter notebook

1. Scikit Learn หรือ sklearn เป็นไลบรารีในภาษา Python สำหรับการพัฒนาโปรแกรมโดยใช้การเรียนรู้ของเครื่อง
2. Imbalanced-Learn หรือ imblearn เป็นโมดูลในภาษา Python สำหรับการจัดการกับกลุ่มข้อมูลที่ไม่มีความสมดุลกัน
3. Matplotlib เป็นโมดูลในภาษา Python สำหรับการแสดงผลในรูปแบบการสร้างภาพแผนผัง หรือ ภาพเคลื่อนไหว
4. NumPy เป็นโมดูลส่วนเสริมของ Python ที่มีฟังก์ชันเกี่ยวกับคณิตศาสตร์และการคำนวณ
5. Pandas เป็นไลบรารีในภาษา Python ที่ทำให้สามารถจัดการข้อมูลต่างๆได้ง่ายขึ้น
6. Seaborn เป็นโมดูลในภาษา Python สำหรับการแสดงผลในรูปแบบการสร้างภาพแผนผังที่มีความหลากหลายโดยมีพื้นฐานมาจากโมดูล Matplotlib

3.5 การเก็บรวบรวมข้อมูล

ข้อมูลที่นำมาใช้ในการดำเนินงานครั้งนี้ ได้ทำการค้นหาและศึกษาจากข้อมูลทุติยภูมิ (Secondary Data) ของกรณีศึกษาธนาคารพาณิชย์ โดยจะมีข้อมูลที่หาได้ คือ รายการธุรกรรมการใช้จ่ายบัตรเครดิตและร้านค้า รวมทั้งผลการตรวจสอบการฉ้อโกงบัตรเครดิต จำนวนข้อมูลทั้งหมด 617,703 ระเบียน พบว่ารายการใช้จ่ายปกติมีจำนวน 613,572 ระเบียน คิดเป็นร้อยละ 99.33 และการฉ้อโกงมีจำนวน 4,131 ระเบียน คิดเป็นร้อยละ 0.67

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6 การเตรียมข้อมูล

1. การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนการคัดกรองข้อมูลที่ไม่เกี่ยวข้อง ข้อมูลที่สูญหาย ข้อมูลที่มีความผิดปกติออกไป

2. การแปลงรูปข้อมูล (Data Transform) เป็นขั้นตอนในการแปลงข้อมูลให้เหมาะสมกับการใช้งาน ซึ่งในการแปลงข้อมูลนั้นจะใช้วิธีแทนค่าของตัวแปรกลุ่มในภาษาไพธอน คือ One Hot Encoding ก่อนการวิเคราะห์ข้อมูล ซึ่งข้อมูลบางส่วนจะถูกแปลงให้อยู่ในรูปข้อมูลเชิงปริมาณ ได้แก่ Online, ECOM, Group Merchant, Group Country, Card, Date, Time, State โดยข้อมูลที่ถูกแปลงสามารถแบ่งได้ดังตารางที่ 3.2

ตารางที่ 3.2 การแปลงข้อมูลตัวแปรที่ใช้ในการวิจัย

ลำดับ	ตัวแปร	รายละเอียด	แปลงค่า
1	ONLINE	Online	0
2	ECOM	3DS	0
		NON 3DS	1
		Other	2
3	Group Merchant	101-Department Store	0
		102-Supermarket	1
		103-Restaurant	2
		105-Hotel & Resort	3
		106-Hospital & Healthcare	4
		107-Gold & Jewelry	5
		108-Clothing & Apparel	6
		109-Household Supplies	7
		110-Automobile Related	8
		111-Gas Station	9
		112-Telecomm. & Comp	10
		113-Travelling	11

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับ	ตัวแปร	รายละเอียด	แปลงค่า
		114-Souvenir	12
		116-Bookstore And Stationery	13
		117-Building And Renovation	14
		118-Business Services	15
		119-Direct Marketing	16
		120-Education	17
		121-Financial Services	18
		122-Bill Payment	19
		123-Insurance	20
		124-Medical	21
		125-Personal Care	22
		126-Recreation	23
		127-Retail Cosmetic	24
		128-Transportation	25
		129-Other Merchants	26
4	Group Country	Domestic	0
		International	1
5	Card	Card Not Present	0
6	Date	Early Month	0
		Mid Month	1
		Late Month	2
7	Time	Midnight	0
		Early Morning	1
		Morning	2
		Early Afternoon	3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับ	ตัวแปร	รายละเอียด	แปลงค่า
		Afternoon	4
		Evening	5
		Night	6
8	State	Africa	0
		Antarctic	1
		Asia	2
		Australia	3
		Europe	4
		North America	5
		South America	6
9	Fraud	Not Fraud	0
		Fraud	1

จากตารางที่ 3.2 มีการแบ่งช่วงของตัวแปร Date และ Time ดังนี้

Date	Early Month	ระหว่างวันที่ 1 - 10
	Mid Month	ระหว่างวันที่ 11 - 20
	Late Month	ระหว่างวันที่ 21 - 31
Time	Midnight	ระหว่างช่วงเวลา 00.00 – 03.59
	Early Morning	ระหว่างช่วงเวลา 04.00 – 06.59
	Morning	ระหว่างช่วงเวลา 07.00 – 09.59
	Early Afternoon	ระหว่างช่วงเวลา 10.00 – 12.59
	Afternoon	ระหว่างช่วงเวลา 13.00 – 16.59
	Evening	ระหว่างช่วงเวลา 17.00 – 20.59
	Night	ระหว่างช่วงเวลา 21.00 – 23.59

และพบว่าตัวแปร 2 ตัว คือ Online และ Card มีค่าเหมือนกันทุกรายการ จึงไม่สามารถนำไปจำแนกกลุ่มได้ ดังนั้น จึงตัดตัวแปร 2 ตัวนี้ออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3. แบ่งข้อมูลเป็นชุดข้อมูลฝึกหัด (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) ในการศึกษาครั้งนี้จะทำการแบ่งข้อมูลทั้งหมดออกเป็น ชุดข้อมูลฝึกหัด คิดเป็นร้อยละ 70 และชุดข้อมูลทดสอบ คิดเป็นร้อยละ 30 สำหรับชุดข้อมูลฝึกหัดแบ่งออกเป็น 432,392 โดยรายการธุรกรรมปกติแบ่งออกเป็น 429,500 รายการ และการฉ้อโกงแบ่งออกเป็น 2,892 รายการ สำหรับชุดข้อมูลทดสอบแบ่งออกเป็น 185,311 โดยรายการธุรกรรมปกติแบ่งออกเป็น 184,072 รายการ และการฉ้อโกงแบ่งออกเป็น 1,239 รายการ

3.7 การวิเคราะห์การจำแนกกลุ่มของชุดข้อมูล

ในการศึกษาครั้งนี้ ได้มีการนำวิธีการจำแนกกลุ่มของชุดข้อมูลมาเปรียบเทียบกัน ประกอบไปด้วยขั้นตอนต่างๆ ดังนี้

1. **วิธีการถดถอยลอจิสติก (Logistic Regression)**

ชุดคำสั่ง

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression (solver = 'saga', penalty = 'l2',
class_weight = 'balanced')
```

- Solver: อัลกอริทึมที่จะใช้ในปัญหาการปรับให้เหมาะสม ค่าเริ่มต้นคือ Saga สำหรับปัญหาหลายคลาส
- Penalty: กำหนดให้เป็น l2 สามารถลดค่าสัมประสิทธิ์บางส่วนให้เป็นศูนย์ ทำให้เป็นมาตรฐานเพื่อป้องกันไม่ให้เกิดการ Overfitting โดยการลดขนาดสัมประสิทธิ์ โดย L2 มันจะมีน้ำหนักน้อย มีค่ามากกว่าศูนย์
- Class_Weight: น้ำหนักที่เกี่ยวข้องกับคลาสโดยกำหนดให้เป็น balanced ทุกคลาสควรจะมีน้ำหนักที่สมดุลกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. วิธีต้นไม้ตัดสินใจ (Decision Tree)

ชุดคำสั่ง

```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier (criterion = 'entropy', max_depth = 7,
min_samples_split = 3, splitter = 'best')
```

- Criterion: ฟังก์ชันวัดคุณภาพของ Split โดยกำหนดให้เป็น entropy
- Max_depth: ความลึกสูงสุดของต้นไม้ โดยกำหนดให้เท่ากับ 7
- Min_samples_split: จำนวนตัวอย่างขั้นต่ำที่ต้องการ Split ภายในโหนด โดยกำหนดให้เท่ากับ 3
- Splitter: กลยุทธ์ใช้ในการ Split แต่ละโหนด โดยกำหนดให้เป็น best ในการเลือกการแบ่งที่ดีที่สุด

3. วิธีการสุ่มป่าไม้ (Random Forest)

ชุดคำสั่ง

```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(criterion = 'entropy', max_depth = 8,
min_samples_leaf = 3)
```

- Criterion: ฟังก์ชันวัดคุณภาพของ Split โดยกำหนดให้เป็น entropy
- Max_depth: ความลึกสูงสุดของต้นไม้ โดยกำหนดให้เท่ากับ 8
- Min_samples_split: จำนวนตัวอย่างขั้นต่ำที่ต้องการ Split ภายในโหนด โดยกำหนดให้เท่ากับ 3
- Min_samples_leaf: จำนวนตัวอย่างขั้นต่ำที่ต้องอยู่ที่โหนดปลายสุด โดยกำหนดให้เท่ากับ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง

ชุดคำสั่ง

```
from xgboost import XGBClassifier
model = XGBClassifier(, max_depth = 5)
```

Max_depth: ความลึกสูงสุดของต้นไม้ โดยกำหนดให้เท่ากับ 5

5. วิธีเอดาบูท

ชุดคำสั่ง

```
form sklearn.ensemble.AdaBoostClassifier
model = AdaBoostClassifier (base_estimator = (criterion = 'entropy',
max_depth = 7, min_samples_split = 3, splitter = 'best'))
```

- Criterion: ฟังก์ชันวัดคุณภาพของ Split โดยกำหนดให้เป็น entropy
- Max_depth: ความลึกสูงสุดของต้นไม้ โดยกำหนดให้เท่ากับ 7
- Min_samples_split: จำนวนตัวอย่างขั้นต่ำที่ต้องการ Split ภายในโหนด โดยกำหนดให้เท่ากับ 3
- Splitter: กลยุทธ์ใช้ในการ Split แต่ละโหนด โดยกำหนดให้เป็น best ในการเลือกการแบ่งที่ดีที่สุด

3.8 การวัดประเมินประสิทธิภาพของตัวแบบ

การวัดประสิทธิภาพของตัวแบบที่เหมาะสม โดยนำตัวแบบจำลองที่สร้างขึ้นด้วยวิธีการวิเคราะห์การถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท นำมาเปรียบเทียบประสิทธิภาพ โดยพิจารณาจากค่า Recall, Precision, F1-score, ROC Curve, AUC ซึ่งได้จาก Confusion Matrix

- Recall แสดงถึงประสิทธิภาพของตัวแบบ เมื่อเราคิดว่าผลการทำนาย False Negative มีผลเป็นอย่างมาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- Precision แสดงถึงประสิทธิภาพของตัวแบบ เมื่อเราคิดว่าผลการทำนาย False Positive เช่น กรณีการทำนายว่าเป็นการฉ้อโกงบัตรเครดิตหนือไม่ฉ้อโกงบัตรเครดิต ซึ่งเมื่อเกิดผลการทำนายรายการธุรกรรมปกติเป็นการฉ้อโกง จะทำให้อาจสูญเสียลูกค้าปกติไปได้

- F1-score มีการพัฒนาตัวแบบ แบบไม่ต้องคำนึงถึง False Positive และ False Negative เป็นหลัก

- ROC Curve เป็นกราฟที่แต่ละจุด (x, y) เป็นค่า False Positive Rate และ True Positive Rate (Recall) ซึ่งเกิดจากการขยับค่า Threshold ที่ต่างกัน โดยยิ่งมีการซ้อนทับของการกระจายตัวของค่าความเชื่อมั่นของทั้งสอง Class มากเท่าไหร่ เราก็จะเห็น ROC Curve มีความชันน้อยลงเท่านั้น

- AUC คือ พื้นที่ใต้กราฟ สามารถบอกได้ว่าตัวแบบของเรามีความสามารถในการแยกการฉ้อโกงออกจากการไม่ฉ้อโกงได้ดีแค่ไหน

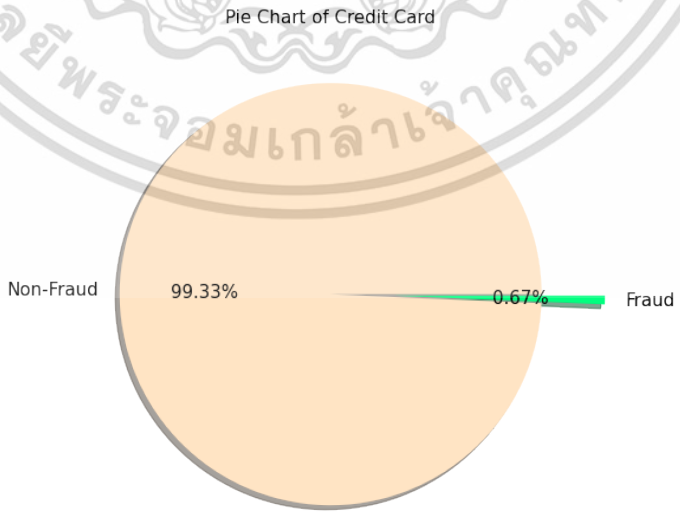
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในบทนี้ได้นำเสนอผลการสำรวจข้อมูลรายการธุรกรรมการใช้บัตรเครดิตและผลการจำแนกรายการฉ้อโกงบัตรเครดิตโดยอาศัยเทคนิคการจำแนกกลุ่มข้อมูล ได้แก่ วิธีการถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) วิธีต้นไม้การตัดสินใจ (Decision Tree) วิธีการสุ่มป่าไม้ (Random Forest) วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง (Extreme Gradient Descent) และวิธีเอดาบูท (Ada) นอกจากนี้ แสดงผลการจำแนกกลุ่มข้อมูลโดยใช้เทคนิค SMOTE เพื่อจัดการความไม่สมดุลของข้อมูล พร้อมทั้งเปรียบเทียบประสิทธิภาพของตัวแบบการจำแนกกลุ่มด้วยวิธีการต่างๆ โดยรายละเอียดจะแสดงตามลำดับดังนี้

4.1 ผลการสำรวจข้อมูลเบื้องต้น

รายการธุรกรรมการใช้บัตรเครดิตที่ใช้ในการศึกษาจำนวนทั้งสิ้น 617,703 ระเบียน พบว่ารายการใช้จ่ายปกติ (Non-Fraud) จำนวน 613,572 ระเบียน คิดเป็นร้อยละ 99.33 และ รายการที่มีการฉ้อโกง (Fraud) จำนวน 4,131 ระเบียน คิดเป็นร้อยละ 0.67 แสดงดังรูปที่ 4.1 ซึ่งจะเห็นได้ว่าสัดส่วนของรายการธุรกรรมที่มีการฉ้อโกงและรายการธุรกรรมปกติไม่สมดุล กล่าวคือข้อมูลคำตอบของแต่ละคลาสมีจำนวนไม่เท่ากัน เรียกว่า “Imbalanced Data”



รูปที่ 4.1 สัดส่วนการฉ้อโกงของรายการธุรกรรมการใช้บัตรเครดิต

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการสำรวจข้อมูลเบื้องต้นไม่พบว่ามีข้อมูลสูญหาย (Missing Value) และ ไม่พบค่าผิดปกติ (Outlier) นอกจากนี้ได้นำเสนอตารางค่าสถิติเชิงพรรณนาของตัวแปรเชิงปริมาณและตารางแจกแจงความถี่ของตัวแปรเชิงคุณภาพ แสดงดังตารางที่ 4.1 และ 4.2

ตารางที่ 4.1 ค่าสถิติเชิงพรรณนาของตัวแปรเชิงปริมาณ

ตัวแปร / ค่าสถิติ	Amount	Credit Limit
ค่าต่ำสุด	0	1000
ค่าสูงสุด	18,104,644.66	15,000,000.00
ค่าเฉลี่ยเลขคณิต	4,116.72	237,792.17
ส่วนเบี่ยงเบนมาตรฐาน	52,162.68	454,810.27
ค่าความโด่ง	30120.55	57.78
ค่าความเบ้	121.72	5.65
ค่ามัธยฐาน	1,000	100,000

ตารางที่ 4.2 ตารางแจกแจงความถี่ของตัวแปรเชิงคุณภาพ

ค่าของตัวแปร	ความถี่	ร้อยละ
ระบบการยืนยันการชำระเงิน		
3DS	46,132	7.5
NON 3DS	151,505	24.5
Other	420,066	68.0
การทำธุรกรรมออนไลน์		
Online	617,703	100.0
ประเภทกลุ่มของร้านค้า		
101-DEPARTMENT STORE	68,529	11.1
102-SUPERMARKET	88,326	14.3
103-RESTAURANT	52,943	8.6
104-ENTERTAINMENT	478	0.1
105-HOTEL & RESORT	6,498	1.1
106-HOSPITAL & HEALTHCARE	16,202	2.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

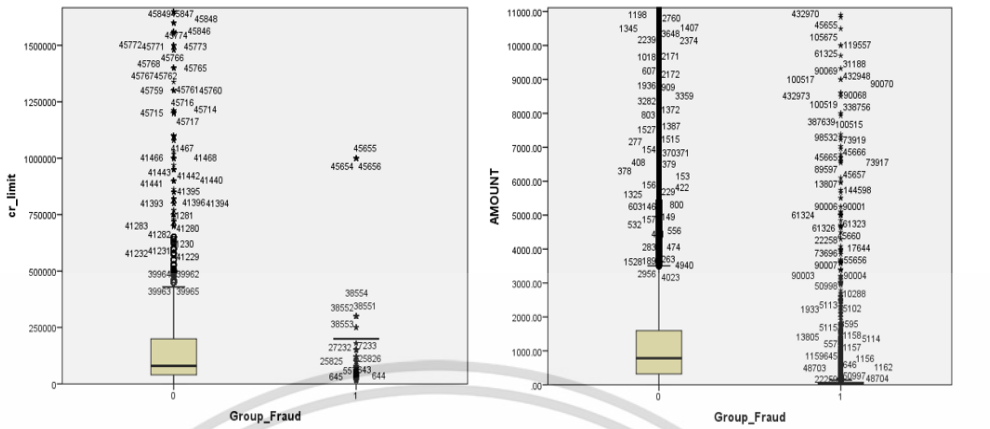
ค่าของตัวแปร	ความถี่	ร้อยละ
107-GOLD & JEWELRY	675	0.1
108-CLOTHING & APPAREL	10,528	1.7
109-HOUSEHOLD SUPPLIES	15,577	2.5
110-AUTOMOBILE RELATED	5,198	0.8
111-GAS STATION	134,302	21.7
112-TELECOMM. & COMP	58,866	9.5
113-TRAVELLING	2,795	0.5
114-SOUVENIR	331	0.1
116-BOOKSTORE AND STATIONERY	2,539	0.4
117-BUILDING AND RENOVATION	719	0.1
118-BUSINESS SERVICES	15,611	2.5
119-DIRECT MARKETING	5,531	0.9
120-EDUCATION	984	0.2
121-FINANCIAL SERVICES	597	0.1
122-BILL PAYMENT	7,029	1.1
123-INSURANCE	30,635	5.0
124-MEDICAL	2,689	0.4
125-PERSONAL CARE	885	0.1
126-RECREATION	11,991	1.9
127-RETAIL COSMETIC	2,388	0.4
128-TRANSPORTATION	31,039	5.0
129-OTHER MERCHANTS	43,818	7.1
ประเภทการชำระเงินบัตรเครดิต		
Card Not Present	617,703	100.0
การแบ่งกลุ่มทวีป		
Africa	150	0.0
Antarctic	12	0.0
Asia	555,859	90.0
Australia	469	0.1
Europe	38,043	6.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

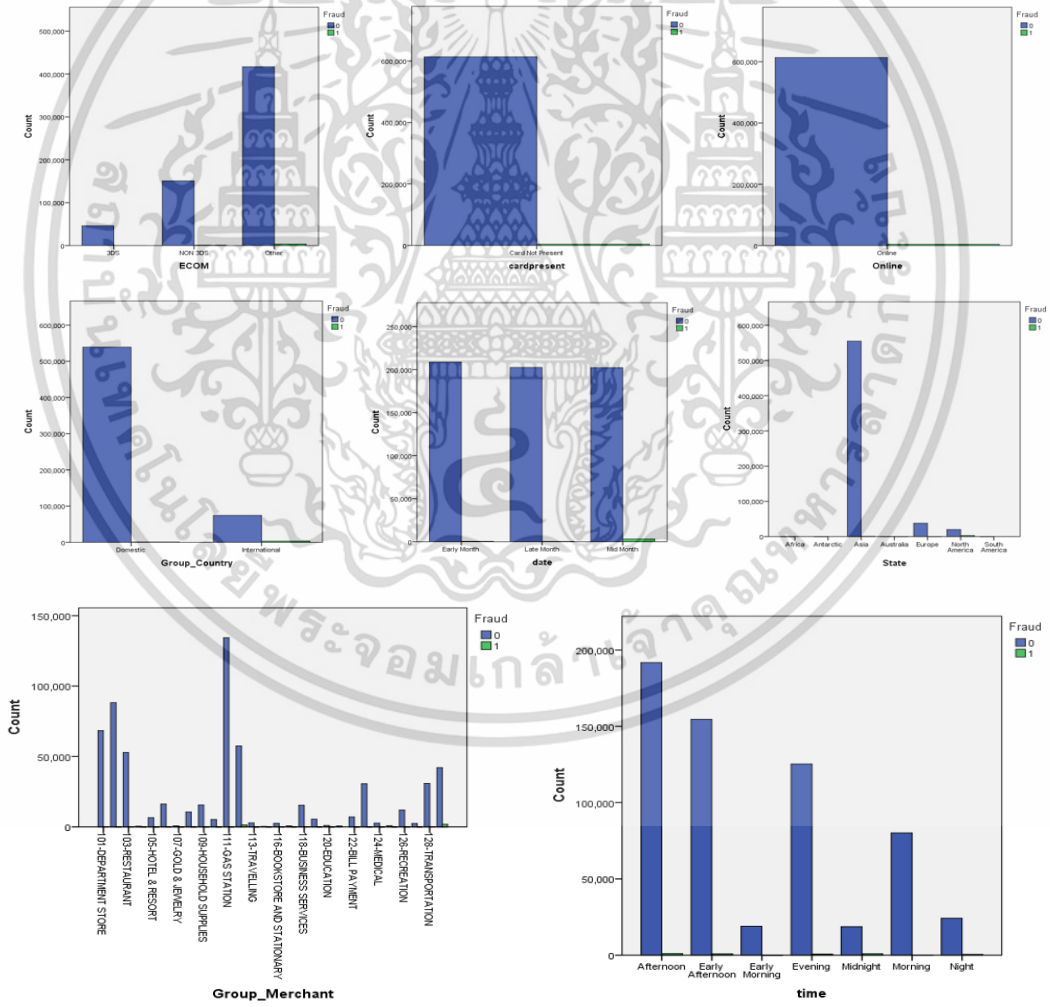
ค่าของตัวแปร	ความถี่	ร้อยละ
North America	23,116	3.7
South America	54	0.0
การแบ่งช่วงวันที่		
Early Month	209,288	33.9
Mid Month	205,458	33.3
Late Month	202,957	32.9
การแบ่งช่วงเวลา		
Afternoon	192,774	31.2
Early Afternoon	155,436	25.2
Early Morning	18,971	3.1
Evening	125,997	20.4
Midnight	19,566	3.2
Morning	80,240	13.0
Night	24,719	4.0
การแบ่งกลุ่มประเทศ		
Domestic	539,904	87.4
International	77,799	12.6

นำตัวแปรเชิงปริมาณ 2 ตัวแปร คือ Credit limit และ Amount มาพล็อตกราฟ Boxplot ตัวแปร Credit limit พบว่ารายการธุรกรรมปกติมีค่าเฉลี่ยน้อยกว่าการฉ้อโกงและการฉ้อโกงมีความผันแปรที่ต่ำกว่า ส่วนตัวแปร Amount พบว่ารายการธุรกรรมปกติมีค่าเฉลี่ยสูงกว่าการฉ้อโกงแต่การฉ้อโกงมีความผันแปรที่ต่ำกว่ากับรายการธุรกรรมปกติและลักษณะข้อมูลมีการเบ้ขวา ดังรูปที่ 4.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.2 กราฟ Boxplot ของตัวแปร Credit limit และ Amount



รูปที่ 4.3 กราฟของตัวแปรเชิงคุณภาพ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 4.3 จะพบว่าตัวแปร Card แบ่งออกเป็น Card Present อย่างเดียว และ Online แบ่งออกเป็น Online อย่างเดียว ซึ่งไม่ส่งผลต่อการจำแนกกลุ่ม จึงทำการตัดตัวแปร 2 ตัวนี้ออกไป

4.2 ผลการวิเคราะห์ความสำคัญของตัวแปร (Feature Importance)

จากการวิเคราะห์ความสำคัญของตัวแปร โดยวิธีการถดถอยลอจิสติก วิธีการสุ่มป่าไม้และวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง จะได้ค่าความสำคัญของตัวแปรดังตารางที่ 4.3

ตารางที่ 4.3 การวิเคราะห์ความสำคัญของตัวแปร

Feature	คะแนนความสำคัญ		
	Logistic Regression	Random Forest	xgboost
Amount	-5.13419	0.57944	0.35968
ECOM	0.00000	0.00000	0.00000
Group Merchant	-0.35000	0.06236	0.03469
Credit limit	-0.13642	0.18521	0.13841
Group Country	0.43069	0.01853	0.05231
Date	-0.02085	0.04154	0.03142
Time	0.14945	0.02651	0.01725
State	0.60291	0.08641	0.35633

จากตารางที่ 4.3 ค่าความสำคัญของตัวแปรของวิธีการถดถอยลอจิสติก วิธีการสุ่มป่าไม้และวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งคุณลักษณะสูงที่สุด คือ Amount มีค่าเท่ากับ -5.13419, 0.57944 และ, 0.35968 ตามลำดับ และตัวแปร ECOM มีค่าเท่ากับ 0 ซึ่งไม่มีความสำคัญในการวิเคราะห์การจำแนกกลุ่ม จึงทำการตัดตัวแปร ECOM ออกไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3 การตรวจสอบความเหมาะสมของตัวแบบการถดถอยลอจิสติกทวิภาค

ผลลัพธ์จากการพยากรณ์ด้วยเทคนิคการถดถอยลอจิสติกทวิภาค วิธี Forward: LR (Likelihood Ratio) ดังตารางที่ 4.4

ตารางที่ 4.4 ผลลัพธ์จากการพยากรณ์ด้วยเทคนิคการถดถอยลอจิสติกทวิภาค วิธี Forward: LR (Likelihood Ratio)

ตัวแปรอิสระ		B	S.E.	Wald	Sig.	Exp(B)
Step 1	State	-1.4530	0.0048	92102.61	0.0000	0.2339
Step 2	Time	-0.9417	0.0092	10548.61	0.0000	0.3899
	State	-0.3662	0.0082	1981.00	0.0000	0.6934
Step 3	Ecom	-1.1662	0.0213	2997.04	0.0000	0.3116
	Time	-0.7041	0.0087	6496.18	0.0000	0.4945
	State	0.1447	0.0113	162.55	0.0000	1.1557
Step 4	Ecom	-0.9682	0.0217	1983.44	0.0000	0.3798
	Date	-0.9324	0.0228	1669.08	0.0000	0.3936
	Time	-0.6048	0.0090	4495.69	0.0000	0.5462
	State	0.3768	0.0119	1005.70	0.0000	1.4576
Step 5	Ecom	-0.9488	0.0212	2009.54	0.0000	0.3872
	Country	-3.0356	0.1483	419.04	0.0000	0.0480
	Date	-0.9177	0.0230	1592.49	0.0000	0.3994
	Time	-0.6545	0.0093	4902.78	0.0000	0.5197
	State	1.4674	0.0537	745.78	0.0000	4.3379
Step 6	Ecom	-0.9667	0.0212	2071.49	0.0000	0.3803
	Country	-3.2851	0.1530	461.26	0.0000	0.0374
	Date	-0.9081	0.0229	1566.06	0.0000	0.4033
	Time	-0.6607	0.0094	4906.03	0.0000	0.5165

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวแปรอิสระ		B	S.E.	Wald	Sig.	Exp(B)
	State	1.4943	0.0548	742.64	0.0000	4.4562
	mcc	0.0200	0.0018	117.61	0.0000	1.0202
Step 7	cr_limit	0.0000	0.0000	76.29	0.0000	1.0000
	Ecom	-0.9897	0.0212	2171.09	0.0000	0.3717
	Country	-3.1757	0.1529	431.47	0.0000	0.0418
	Date	-0.8981	0.0229	1536.92	0.0000	0.4073
	Time	-0.6545	0.0095	4789.32	0.0000	0.5197
	State	1.4869	0.0547	739.92	0.0000	4.4235
	mcc	0.0192	0.0019	107.19	0.0000	1.0194
Step 8	AMOUNT	-0.0001	0.0000	154.33	0.0000	0.9999
	cr_limit	0.0000	0.0000	31.32	0.0000	1.0000
	Ecom	-0.9759	0.0217	2029.71	0.0000	0.3768
	Country	-3.0930	0.1521	413.69	0.0000	0.0454
	Date	-0.8837	0.0229	1490.30	0.0000	0.4132
	Time	-0.6447	0.0095	4645.52	0.0000	0.5248
	State	1.4778	0.0543	740.86	0.0000	4.3833
	mcc	0.0174	0.0019	86.73	0.0000	1.0175
Hosmer and Lemeshow test χ^2 = 854.936, Sig. = 0.000						

จากตารางที่ 4.4 ได้ค่า χ^2 = 854.936 และค่า Sig. เท่ากับ 0.000 < 0.05 จึงปฏิเสธสมมติฐานหลัก (H_0) แสดงว่า ตัวแบบไม่มีความเหมาะสม เนื่องจากตัวอย่างมีขนาดใหญ่ Hosmer and Lemeshow test จึงมีค่ามาก ทำให้สรุปผลได้ว่าปฏิเสธสมมติฐานหลัก (H_0) ทั้งที่ ตัวแบบมีความเหมาะสม (กัลยา, 2548) ดังตารางที่ 4.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การทดสอบสัมประสิทธิ์การอธิบาย (Pseudo R Square) ของตัวแบบการถดถอยลอจิสติกทวิภาค

ตารางที่ 4.5 ค่าความเป็นไปได้ (Likelihood Value) และ ค่าสัมประสิทธิ์การอธิบาย (Pseudo R Square)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	33462.813	.026	.334

จากตารางที่ 4.5 พิจารณาจากค่าความเป็นไปได้ (Likelihood Value) โดยวัดจากค่า -2LL (-2 Log Likelihood) ซึ่งทดสอบความมีนัยสำคัญทางสถิติได้โดยใช้ Model Chi–Square และพบว่า Model Chi-Square Test เท่ากับ 33,462.813 และค่า Sig เท่ากับ 0.000 < 0.05 จึงปฏิเสธสมมติฐานหลัก (H_0) แสดงว่า ตัวแบบขึ้นอยู่กับตัวแปรอิสระ อย่างน้อย 1 ตัว และ ค่า Cox & Snell R Square เท่ากับ 0.026 หมายความว่า ตัวแปรอิสระ 7 ตัว สามารถอธิบายโอกาสที่เกิดเหตุการณ์ที่สนใจได้ร้อยละ 2.6 และค่า Nagelkerke R Square เท่ากับ 0.334 หมายความว่า ตัวแปรอิสระ 7 ตัว สามารถอธิบายโอกาสที่เกิดเหตุการณ์ที่สนใจได้ร้อยละ 33.4

4.4 ผลการจำแนกรายการธุรกรรมฉ้อโกง

ผลลัพธ์การจำแนกรายการธุรกรรมฉ้อโกง ด้วยเทคนิคการจำแนกกลุ่ม 5 วิธี คือ 1) วิธีการถดถอยลอจิสติกทวิภาค 2) วิธีต้นไม้การตัดสินใจ 3) วิธีการสุ่มป่าไม้ 4) วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง 5) วิธีเอดาบูท รายละเอียดผลลัพธ์ที่ได้ในแต่ละวิธี แสดงรายละเอียดดังต่อไปนี้

4.4.1 วิธีการถดถอยลอจิสติกทวิภาค

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริงของข้อมูลชุดทดสอบ (Test Dataset) แสดงในรูปเมทริกซ์ความสับสน (Confusion Matrix) ดังตารางที่ 4.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.6 ผลการจำแนกกลุ่มโดยวิธีการถดถอยลอจิสติกทวิภาค

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	167,993	16,079
Fraud	251	988

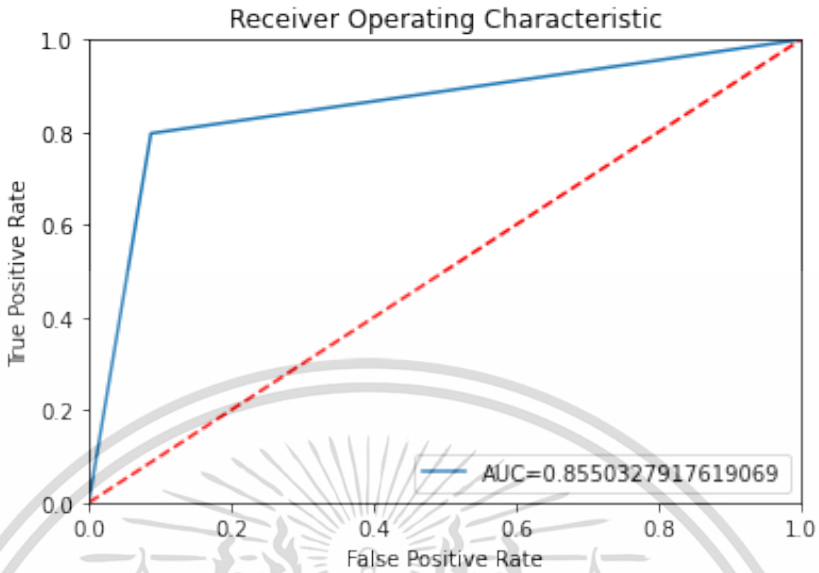
จากตารางที่ 4.6 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.7

ตารางที่ 4.7 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการถดถอยลอจิสติกทวิภาค

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	91%	95%	91%	85.50%
1	6%	81%	11%		
Weighted-Avg.	99%	91%	95%		

จากตารางที่ 4.7 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 91 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 91 ค่าความถ่วงดุลร้อยละ 95 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 85.50 และธุรกรรมฉ้อโกงเท่ากับร้อยละ 9 มีค่าความแม่นยำร้อยละ 6 ค่าความไวร้อยละ 81 ค่าความถ่วงดุลร้อยละ 11

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.4 ค่า AUC ของวิธีการถดถอยลอจิสติกทวิภาค

4.4.2 วิธีต้นไม้การตัดสินใจ

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริงของข้อมูลชุดทดสอบ (Test Dataset) แสดงในรูปเมทริกซ์ความสับสน (Confusion Matrix) ดังตารางที่ 4.8

ตารางที่ 4.8 ผลการจำแนกกลุ่มโดยวิธีต้นไม้การตัดสินใจ

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	184,049	23
Fraud	374	865

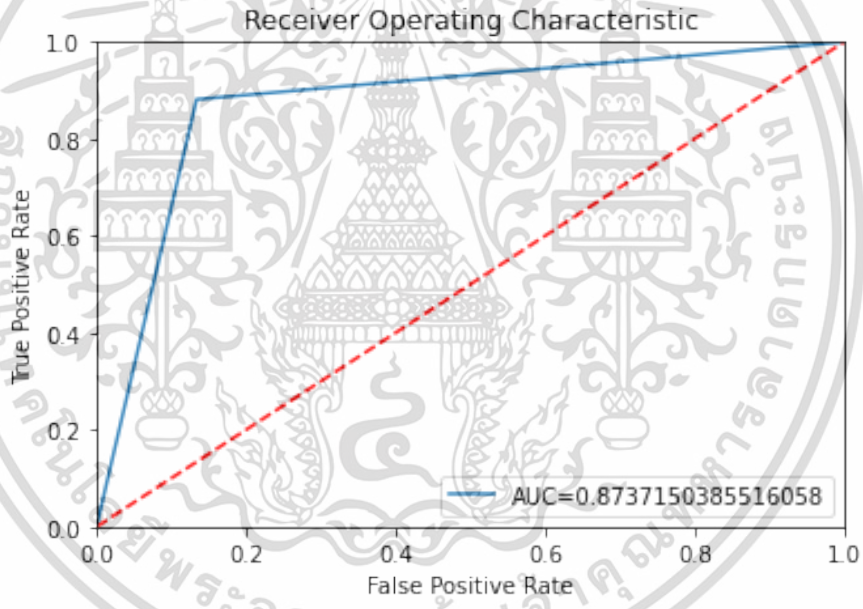
จากตารางที่ 4.8 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.9 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีต้นไม้การตัดสินใจ

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	100%	100%	100%	87.37%
1	97%	70%	81%		
Weighted-Avg.	100%	100%	100%		

จากตารางที่ 4.9 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 100 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 100 ค่าความถ่วงดุลร้อยละ 100 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 87.37 และธุรกรรมฉ้อโกงมีค่าความแม่นยำร้อยละ 97 ค่าค่าความไวร้อยละ 70 ค่าความถ่วงดุลร้อยละ 81



รูปที่ 4.5 ค่า AUC ของวิธีต้นไม้การตัดสินใจ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.3 วิธีการสุ่มป่าไม้

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริง และสร้างตารางเมทริกซ์ความสับสน แสดงดังตารางที่ 4.10

ตารางที่ 4.10 ผลการจำแนกกลุ่มโดยวิธีการสุ่มป่าไม้

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	184,066	6
Fraud	380	859

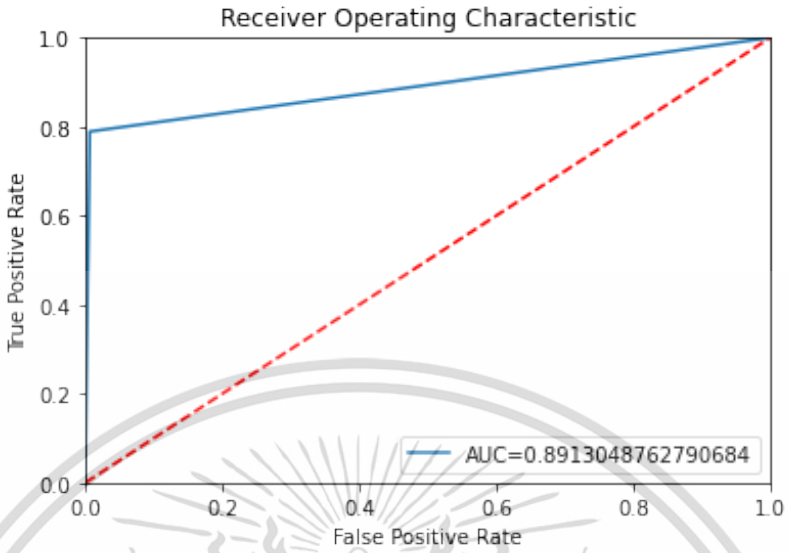
จากตารางที่ 4.10 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.11

ตารางที่ 4.11 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการสุ่มป่าไม้

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	100%	100%	100%	89.13%
1	99%	69%	82%		
Weighted-Avg.	100%	100%	100%		

จากตารางที่ 4.11 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 100 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 100 ค่าความถ่วงดุลร้อยละ 100 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 89.13 และธุรกรรมฉ้อโกงมีค่าความแม่นยำร้อยละ 99 ค่าความไวร้อยละ 69 ค่าความถ่วงดุลร้อยละ 82

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.6 ค่า AUC ของวิธีการสุ่มป่าไม้

4.4.4 วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริง และสร้างตารางเมทริกซ์ความสับสน แสดงดังตารางที่ 4.12

ตารางที่ 4.12 ผลการจำแนกกลุ่มโดยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	184,043	29
Fraud	382	857

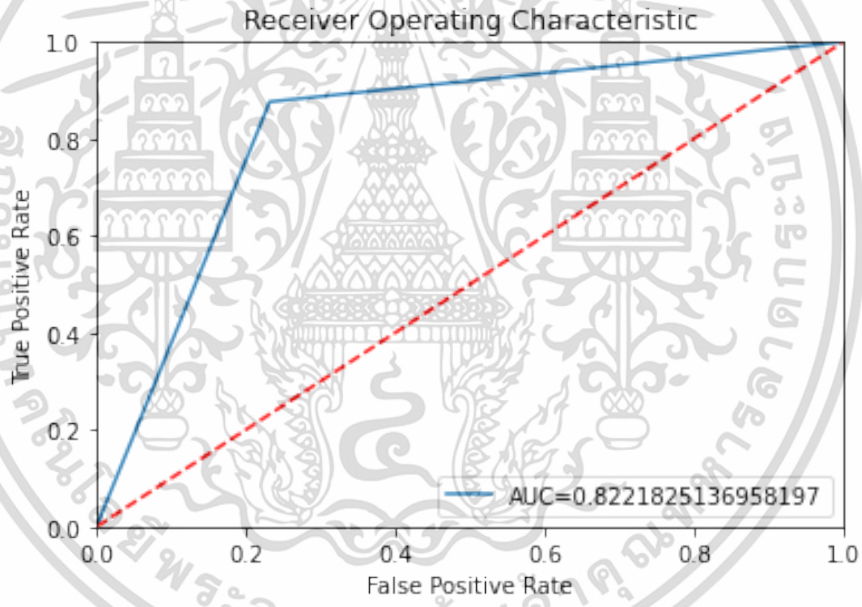
จากตารางที่ 4.12 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.13

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.13 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	100%	100%	100%	82.22%
1	97%	69%	81%		
Weighted-Avg.	100%	100%	100%		

จากตารางที่ 4.13 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 100 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 100 ค่าความถ่วงดุลร้อยละ 100 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 82.22 และธุรกรรมฉ้อโกงมีค่าความแม่นยำร้อยละ 97 ค่าความไวร้อยละ 69 ค่าความถ่วงดุลร้อยละ 81



รูปที่ 4.7 ค่า AUC ของวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4.5 วิธีเอดาบูท

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริง และสร้างตารางเมทริกซ์ความสับสน แสดงดังตารางที่ 4.14

ตารางที่ 4.14 ผลการจำแนกกลุ่มโดยวิธีเอดาบูท

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	183,9673	105
Fraud	303	936

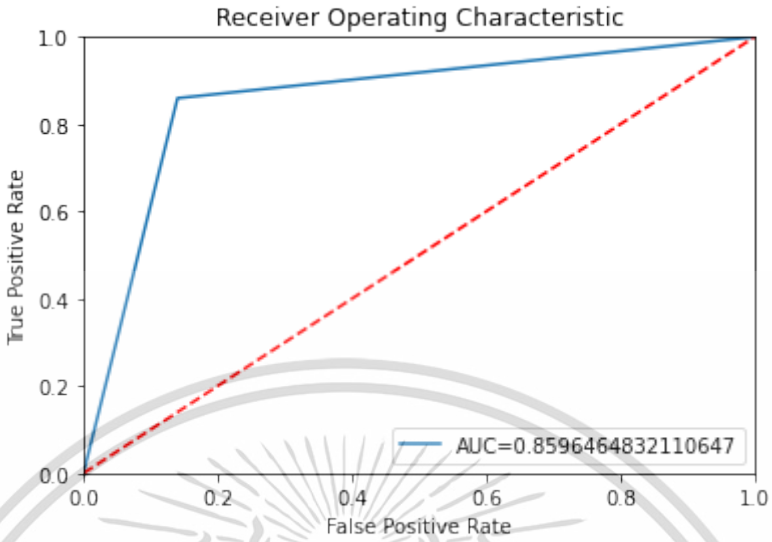
จากตารางที่ 4.14 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.15

ตารางที่ 4.15 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอดาบูท

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	100%	100%	100%	85.96%
1	90%	76%	82%		
Weighted-Avg.	100%	100%	100%		

จากตารางที่ 4.15 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 100 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 100 ค่าความถ่วงดุลร้อยละ 100 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 85.96 และธุรกรรมฉ้อโกงมีค่าความแม่นยำร้อยละ 90 ค่าความไวร้อยละ 76 ค่าความถ่วงดุลร้อยละ 82

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.8 ค่า AUC ของวิธีเอดาบูท

4.5 ผลการจำแนกรายการธุรกรรมฉ้อโกงโดยอาศัยการปรับความไม่สมดุลด้วยเทคนิค SMOTE

ผลลัพธ์การจำแนกรายการธุรกรรมฉ้อโกงโดยอาศัยการปรับความไม่สมดุลด้วยเทคนิค SMOTE ร่วมกับเทคนิคการจำแนกกลุ่ม 5 วิธี คือ 1) วิธีการถดถอยลอจิสติกทวิภาค 2) วิธีต้นไม้การตัดสินใจ 3) วิธีการสุ่มป่าไม้ 4) วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง 5) วิธีเอดาบูท รายละเอียดผลลัพธ์ที่ได้ในแต่ละวิธี แสดงรายละเอียดดังต่อไปนี้

4.5.1. วิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริงของข้อมูลชุดทดสอบ (Test Dataset) แสดงในรูปเมทริกซ์ความสับสน (Confusion Matrix) ดังตารางที่ 4.16

ตารางที่ 4.16 ผลการจำแนกกลุ่มโดยวิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	141,340	42,732
Fraud	153	1,086

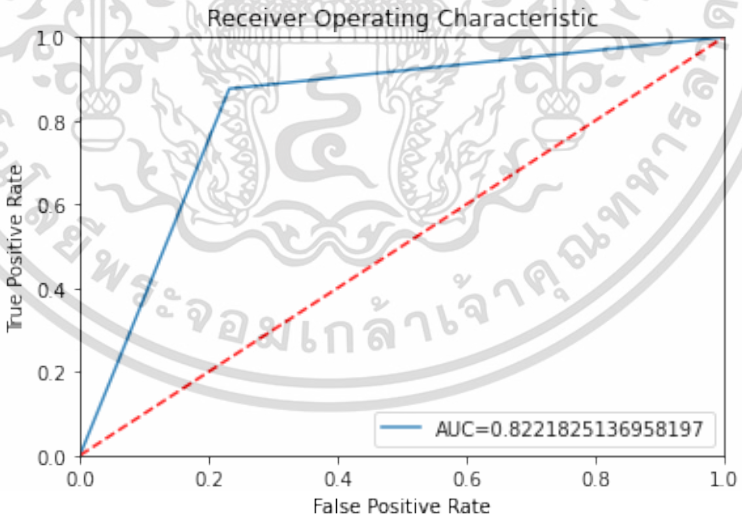
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตารางที่ 4.16 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.17

ตารางที่ 4.17 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	77%	87%	77%	82.22%
1	2%	88%	5%		
Weighted-Avg.	99%	77%	86%		

จากตารางที่ 4.17 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 77 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 77 ค่าความถ่วงดุลร้อยละ 87 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 82.22 และธุรกรรมฉ้อโกงเท่ากับร้อยละ 23 มีค่าความแม่นยำร้อยละ 2 ค่าความไวร้อยละ 88 ค่าความถ่วงดุลร้อยละ 5



รูปที่ 4.9 ค่า AUC ของวิธีการถดถอยลอจิสติกทวิภาคร่วมกับเทคนิค SMOTE

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.2. วิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริงของข้อมูลชุดทดสอบ (Test Dataset) แสดงในรูปเมทริกซ์ความสับสน (Confusion Matrix) ดังตารางที่ 4.18

ตารางที่ 4.18 ผลการจำแนกกลุ่มโดยวิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	158,252	25,820
Fraud	174	1,065

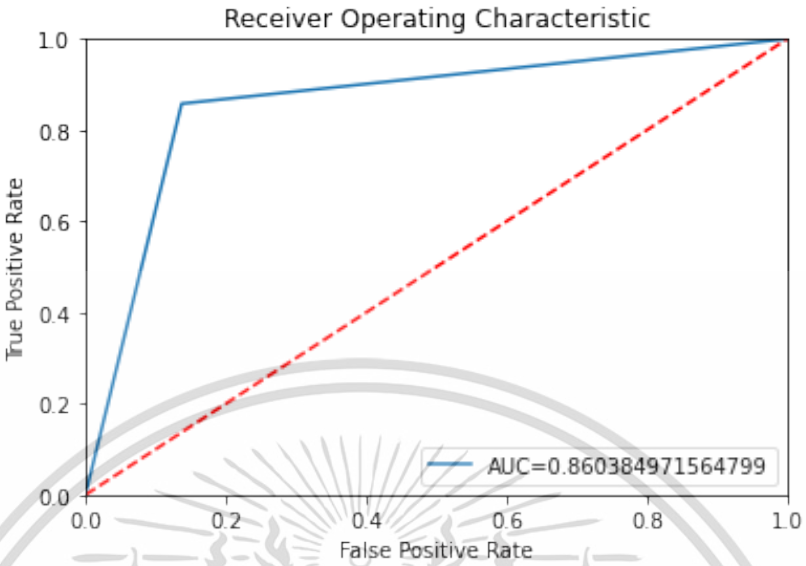
จากตารางที่ 4.18 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.19

ตารางที่ 4.19 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	86%	92%	86%	86.04%
1	4%	86%	8%		
Weighted-Avg.	99%	86%	92%		

จากตารางที่ 4.19 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 86 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 86 ค่าความถ่วงดุลร้อยละ 92 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 86.04 และธุรกรรมฉ้อโกงเท่ากับร้อยละ 14 มีค่าความแม่นยำร้อยละ 4 ค่าความไวร้อยละ 86 ค่าความถ่วงดุลร้อยละ 8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.10 ค่า AUC ของวิธีต้นไม้การตัดสินใจร่วมกับเทคนิค SMOTE

4.5.3. วิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริง และสร้างตารางเมทริกซ์ความสับสน แสดงดังตารางที่ 4.20

ตารางที่ 4.20 ผลการจำแนกกลุ่มโดยวิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	166,679	17,393
Fraud	192	1,047

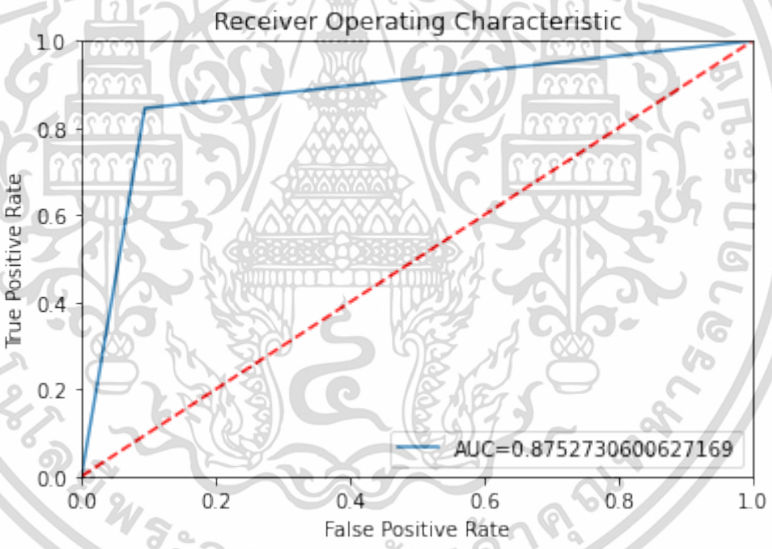
จากตารางที่ 4.20 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.21

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.21 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	91%	95%	91%	87.53%
1	6%	85%	11%		
Weighted-Avg.	99%	91%	94%		

จากตารางที่ 4.21 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 91 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 91 ค่าความถ่วงดุลร้อยละ 95 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 87.53 และธุรกรรมฉ้อโกงเท่ากับร้อยละ 9 มีค่าความแม่นยำร้อยละ 6 ค่าความไวร้อยละ 85 ค่าความถ่วงดุลร้อยละ 11



รูปที่ 4.11 ค่า AUC ของวิธีการสุ่มป่าไม้ร่วมกับเทคนิค SMOTE

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5.4. วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริง และสร้างตารางเมทริกซ์ความสับสน แสดงดังตารางที่ 4.22

ตารางที่ 4.22 ผลการจำแนกกลุ่มโดยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	183,865	196
Fraud	307	1,092

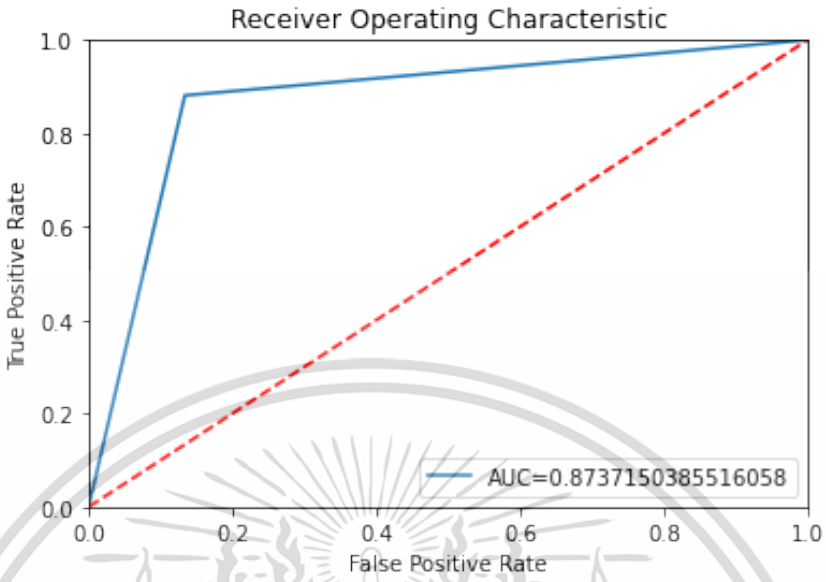
จากตารางที่ 4.22 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.23

ตารางที่ 4.23 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	87%	93%	87%	87.37%
1	4%	88%	8%		
Weighted-Avg.	99%	87%	92%		

จากตารางที่ 4.23 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 87 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 87 ค่าความถ่วงดุลร้อยละ 93 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 87.37 และธุรกรรมฉ้อโกงเท่ากับร้อยละ 13 มีค่าความแม่นยำร้อยละ 4 ค่าความไวร้อยละ 88 ค่าความถ่วงดุลร้อยละ 8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.12 ค่า AUC ของวิธีเอ็กซ์ทรีมกาเดียนบูทติ้งร่วมกับเทคนิค SMOTE

4.5.5. วิธีเอดาบูทร่วมกับเทคนิค SMOTE

ผลลัพธ์จากการพยากรณ์ที่ได้เปรียบเทียบกับค่าจริง และสร้างตารางเมทริกซ์ความสับสน แสดงดังตารางที่ 4.24

ตารางที่ 4.24 ผลการจำแนกกลุ่มโดยวิธีเอดาบูทร่วมกับเทคนิค SMOTE

ค่าจริง	ค่าพยากรณ์	
	Not Fraud	Fraud
Not Fraud	182,832	1240
Fraud	261	978

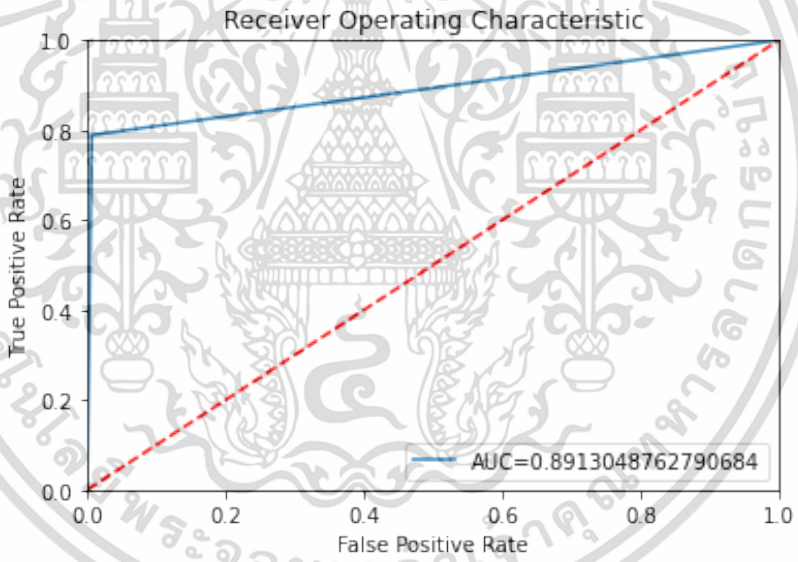
จากตารางที่ 4.24 คำนวณหาตัวชี้วัดประสิทธิภาพของตัวแบบการจำแนกกลุ่ม ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความไว (Recall) และ F1-Score ดังตารางที่ 4.25

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.25 ค่าตัววัดประสิทธิภาพของตัวแบบจำแนกกลุ่มด้วยวิธีเอดาบูทร่วมกับเทคนิค SMOTE

Class	Precision	Recall	F1-Score	Accuracy	AUC
0	100%	99%	100%	99%	89.13%
1	44%	79%	57%		
Weighted-Avg.	99%	99%	99%		

จากตารางที่ 4.25 ทำนายธุรกรรมที่ถูกต้อง เท่ากับร้อยละ 99 มีค่าความแม่นยำร้อยละ 100 ค่าความไวร้อยละ 99 ค่าความถ่วงดุลร้อยละ 100 ค่าพื้นที่ใต้โค้งมีค่าเท่ากับร้อยละ 89.13 และธุรกรรมฉ้อโกงเท่ากับร้อยละ 1 มีค่าความแม่นยำร้อยละ 44 ค่าความไวร้อยละ 79 ค่าความถ่วงดุลร้อยละ 57



รูปที่ 4.13 ค่า AUC ของวิธีเอดาบูทร่วมกับเทคนิค SMOTE

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.6 ผลการเปรียบเทียบการจำแนกรายการธุรกรรมฉ้อโกงระหว่างเทคนิคการจำแนกกลุ่มที่ใช้การปรับความไม่สมดุล และไม่ใช้การปรับความไม่สมดุล

ตารางที่ 4.26 การจำแนกรายการธุรกรรมฉ้อโกงระหว่างเทคนิคการจำแนกกลุ่มที่ไม่ใช้การปรับความไม่สมดุล

Classifier	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	6%	81%	11%	100%	89.13%
Decision Tree	97%	70%	81%	100%	87.37%
Random Forest	99%	69%	82%	100%	89.13%
Extreme Gradient Boosting	97%	69%	81%	100%	82.22%
AdaBoost	90%	76%	82%	100%	85.96%

เนื่องจากชุดข้อมูลได้รับผลกระทบจากจากความไม่สมดุลสูงระหว่างสองคลาส ตัวแยกประเภททั้งหมดจึงมีความแม่นยำสูงในตารางที่ 4.26 ดังนั้น เพื่อให้ได้ความแม่นยำที่แท้จริงรวมไปถึงค่าอื่น ๆ ด้วยเราจึงปรับสมดุลชุดข้อมูลโดยใช้เทคนิค SMOTE จากนั้นเราปรับใช้อีกครั้งสำหรับการจำแนกกลุ่มแต่ละวิธี ผลลัพธ์ของการจำแนกกลุ่มแต่ละวิธีหลัง SMOTE แสดงอยู่ในตารางที่ 4.27

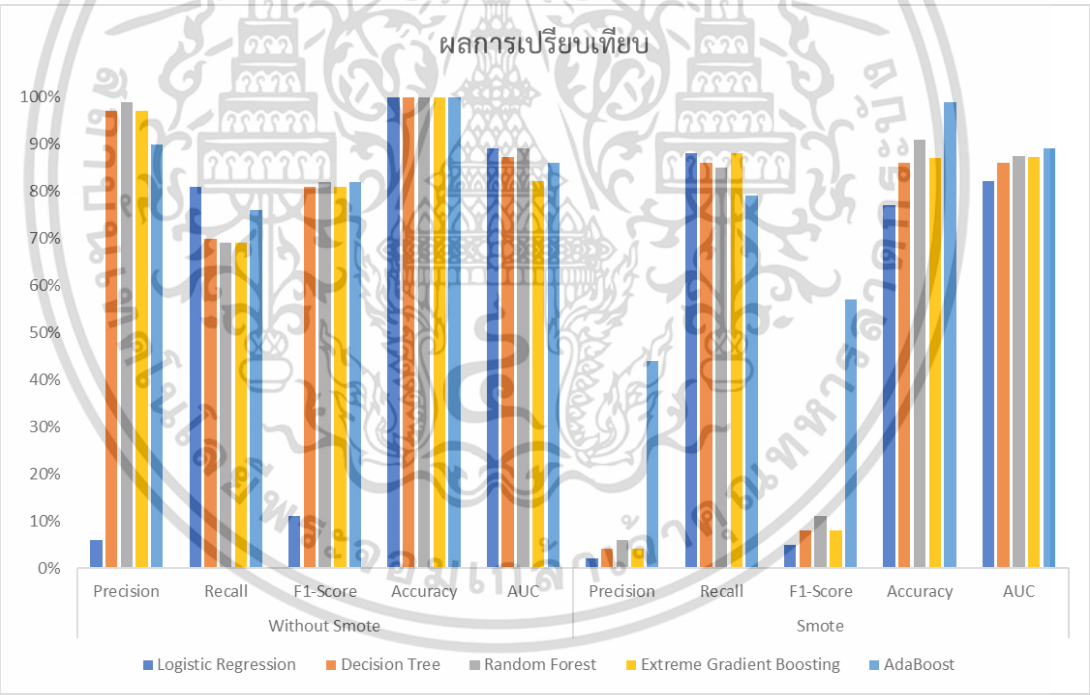
ตารางที่ 4.27 การจำแนกรายการธุรกรรมฉ้อโกงระหว่างเทคนิคการจำแนกกลุ่มที่ใช้การปรับความไม่สมดุล

Classifier	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	2%	88%	5%	77%	82.22%
Decision Tree	4%	86%	8%	86%	86.04%
Random Forest	6%	85%	11%	91%	87.53%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Classifier	Precision	Recall	F1-Score	Accuracy	AUC
Extreme Gradient Boosting	4%	88%	8%	87%	87.37%
AdaBoost	44%	79%	57%	99%	89.13%

การใช้เทคนิค SMOTE กับชุดข้อมูล จะสามารถสังเกตเห็นได้ว่า ประสิทธิภาพของตัวแบบที่ดีที่สุดเรียงลำดับดังต่อไปนี้ ได้แก่วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง วิธีการถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ และวิธีเอดาบูท โดยพิจารณาจากค่าความไว คือ 88%, 88%, 85%, 86%, 79% ตามลำดับ



รูปที่ 4.14 เปรียบเทียบเทคนิคการจำแนกกลุ่มที่ใช้การปรับความไม่สมดุลโดย SMOTE และไม่ใช้การปรับความไม่สมดุล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.7 อภิปรายผล

งานวิจัยครั้งนี้ได้ศึกษาการจำแนกกลุ่มสำหรับการตรวจจับการฉ้อโกงบัตรเครดิตข้อมูลที่ใช้ในการศึกษาประกอบด้วย รายการธุรกรรมการใช้จ่ายบัตรเครดิตและร้านค้า รวมทั้งผลการตรวจสอบการฉ้อโกงบัตรเครดิต จากข้อมูลที่ได้จากธนาคารพาณิชย์ไทยแห่งหนึ่งได้ตัวแปรอิสระ 10 ตัว และตัวแปรตาม 1 ตัว แต่สามารถนำตัวแปรอิสระมาวิเคราะห์ได้ 8 ตัว ได้แก่ Amount, ECOM, Group Merchant, Credit limit, Country, Date, Time, State และทำการการวิเคราะห์ความสำคัญของตัวแปรพบว่า ECOM เท่ากับ 0 นั่นคือไม่มีความสำคัญจึงตัดตัวแปร ECOM ออกจึงทำให้เหลือตัวแปรอิสระอยู่ 7 ตัว เพื่อใช้ในการจำแนกกลุ่มด้วยเทคนิคการเรียนรู้ของเครื่องสำหรับการตรวจจับการฉ้อโกงบัตรเครดิตและเปรียบเทียบประสิทธิภาพของตัวแบบต่าง ๆ เพื่อให้ได้การจำแนกกลุ่มรายการที่ฉ้อโกงโดยผู้วิจัยเลือกใช้เทคนิคการจำแนกกลุ่ม 5 วิธี ได้แก่ วิธีการถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท ผลการศึกษาพบว่า ผลลัพธ์ที่เราได้รับจากการจำแนกกลุ่มทั้งหมดที่ไม่ใช้เทคนิค SMOTE ไม่เป็นที่น่าพอใจ เพราะชุดข้อมูลมีปัญหาความไม่สมดุลสูงมาก โดยร้อยละของข้อมูลทุจริตคือ 0.67% ของชุดข้อมูล ตัวแบบที่มีการใช้เทคนิค SMOTE ร่วมด้วย มีประสิทธิภาพสูงขึ้น วิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง เป็นวิธีที่มีความเหมาะสมมากที่สุด สำหรับการทำนายตัวแบบ เนื่องจาก มีค่า Precision 4%, Recall 88%, F1-Score 8%, Accuracy 87% และ AUC 87.37% เมื่อใช้การปรับความไม่สมดุลโดยเทคนิค SMOTE

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วัตถุประสงค์ของงานวิจัยนี้เพื่อศึกษาการจำแนกกลุ่มสำหรับการตรวจจับการฉ้อโกงบัตรเครดิต โดยมีวิธีการจำแนกกลุ่มทั้งหมด 5 วิธี คือ วิธีการถดถอยลอจิสติกทวิภาค วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท จากการสร้างตัวแบบในบทที่ 3 และ 4 ผู้วิจัยเลือกตัวแบบที่เหมาะสม โดยทำการเปรียบเทียบประสิทธิภาพ ค่าความไว ของแต่ละวิธีและเลือกตัวแบบจากวิธีที่ให้ค่าความไวมากที่สุด

5.1.1 สรุปผลและตรวจสอบประสิทธิภาพของตัวแบบไม่ใช้การปรับความไม่สมดุล

จากการตรวจสอบประสิทธิภาพของตัวแบบไม่ใช้การปรับความไม่สมดุล โดยใช้วิธีการจำแนกกลุ่มทั้งหมด 5 วิธี คือ วิธีการถดถอยลอจิสติกทวิภาค วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท พบว่าวิธีการถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท มีค่าความถูกต้องเท่ากันหมดที่ร้อยละ 100 ค่าความไววิธีการถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท มีค่าเท่ากับ 81%, 70%, 69%, 69% และ 76% ตามลำดับ

5.1.2 สรุปผลและตรวจสอบประสิทธิภาพของตัวแบบใช้การปรับความไม่สมดุล

จากการตรวจสอบประสิทธิภาพของตัวแบบใช้การปรับความไม่สมดุล โดยใช้วิธีการจำแนกกลุ่มทั้งหมด 5 วิธี คือ วิธีการถดถอยลอจิสติกทวิภาค วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท พบว่าวิธีการถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท มีค่าความถูกต้องมีค่าเท่ากับ 77%, 86%,91%, 87% และ 99% ตามลำดับ ค่าความไว วิธีการถดถอยลอจิสติก วิธีต้นไม้การตัดสินใจ วิธีการสุ่มป่าไม้ วิธีเอ็กซ์ทรีมกาเดียนบูทติ้งและวิธีเอดาบูท มีค่าเท่ากับ 88%, 86%, 85%, 88% และ 79% ตามลำดับ จากผลลัพธ์ค่าความไวเพิ่มขึ้นจากการไม่ใช้การปรับความไม่สมดุล ดังนั้น ตัวแบบที่เหมาะสมที่สุด คือ ตัวแบบวิธีเอ็กซ์ทรีมกาเดียนบูทติ้ง เนื่องจากมีค่าความไวมากที่สุด มีค่าเท่ากับ 88% และค่า AUC เท่ากับ 87.37%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.2 ข้อเสนอแนะ

1. เนื่องจากข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลชุดข้อมูลการตรวจจับการฉ้อโกงบัตรเครดิตที่ได้จากธนาคารพาณิชย์ไทยแห่งหนึ่ง ซึ่งข้อมูลส่วนบุคคล เช่น เพศ รายได้ อายุ เป็นต้น ไม่ได้นำเข้ามาวิเคราะห์ด้วย ทำให้ข้อมูลที่มาทำนายอาจจะยังไม่ได้ประสิทธิภาพสูงสุด ทั้งนี้เพื่อเพิ่มประสิทธิภาพการทำนาย ควรเพิ่มตัวแปด้านข้อมูลส่วนบุคคล อาทิเช่น อายุ เพศ เป็นต้น

2. ผู้วิจัยอาจวิเคราะห์ข้อมูลด้วยวิธีการจำแนกวิธีอื่น ๆ เช่น วิธีนาอีฟเบส์ วิธีเพื่อนบ้านใกล้สุด k ตัว เป็นต้น

3. ควรศึกษาวิธีการปรับความไม่สมดุลของข้อมูลวิธีอื่น ๆ เช่น Cost Sensitive Learning

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

Arora, S. and Kumar, D. 2017. “Selection of optimal credit card fraud detection models using a coefficient sum approach”, 4 8 2 - 4 8 7 . International Conference on Computing, Communication and Automation (ICCCA).

Taha, A.A. and Malebary, S.J. 2020. “An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine”, IEEE Access 8 25579–25587.

Makki, S. Assaghir, Z., Taher, Y. Haque, R. Hacid, M. and Zeineddine, H. 2020. “An experimental study with imbalanced classification approaches for credit card fraud detection”, IEEE Access 7 93010–93022.

Asha, R.B. Suresh Kumar, K.R. 2021. “Credit card fraud detection using artificial neural network.” Global Transitions Proceedings. 2(1):35-41

Kumar, M.S. Soundarya, V. Kavitha, S. Keerthika, E.S. and Aswini, E. 2019. “Credit card fraud detection using random forest algorithm”, 149–153. in: Proceedings of the 3rd International Conference on Computing and Communications Technologies (ICCCT).

Abdulghani, A.Q. UCAN, O. N. and Alheeti, K. M. 2021 “Credit Card Fraud Detection Using XGBoost Algorithm”, 487-492. International Conference on Developments in eSystems Engineering (DeSE).

Nitesh, V. Chawla, Kevin, W. and Hall, W. 2002 “SMOTE: Synthetic Minority Over-sampling Technique,Journal of Artificial Intelligence”, 321357 Research16.

Kumar, P. and Iqbal, F. 2019 “Credit card fraud identification using machine learning approaches”, 1-4 in: Proceedings of the 1st International Conference on Innovations in Infor- mation and Communication Technology (ICIICT).

Y. Lucas et al., 2020 “Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs”, 393-402. Future. Genera. Computer. System.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Chawla, N.B. and Kegelmeyer, W. (2002). SMOTE: “Synthetic Minority Over-sampling Technique”, 321-357 Res. (JAIR).16.

ธนาคารแห่งประเทศไทย. 2564**. การดำเนินการกรณีการตัดเงินที่ผิดปกติผ่านบัตรเครดิตและบัตรเดบิต.** [Online].
Available : https://www.bot.or.th/Thai/AboutBOT/Activities/Pages/JointPress_23102021.aspx.

ธนาคารแห่งประเทศไทย. 2559. **ธนาคารแห่งประเทศไทย (ข้อมูลสถิติ) จำนวนบัตรพลาสติก** [Online]. Available : https://www.bot.or.th/App/BTWS_STAT/statistics.

สมาคมโปรแกรมเมอร์ไทย. 2561. **Deep learning คืออะไร** [Online]. Available : https://www.thaiprogrammer.org/2018/12/deep-learn

กัลยาวานิชย์ บัญชา. 2548. **สถิติสำหรับงานวิจัย** จุฬาลงกรณ์มหาวิทยาลัย.

กัลยา วานิชย์บัญชา. 2548. **การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows**. พิมพ์ครั้งที่ 4. ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

กัลยา วานิชย์บัญชา. 2551. **การวิเคราะห์ข้อมูลหลายตัวแปร**. พิมพ์ครั้งที่ 3. ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

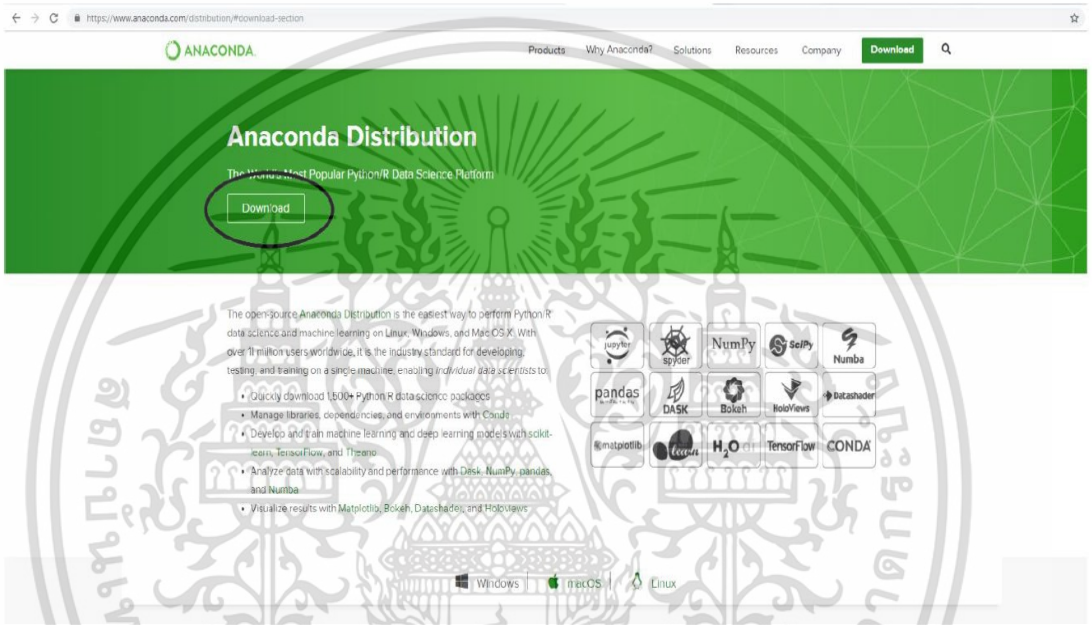
ภาคผนวก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ก

วิธีติดตั้งโปรแกรมไพธอน

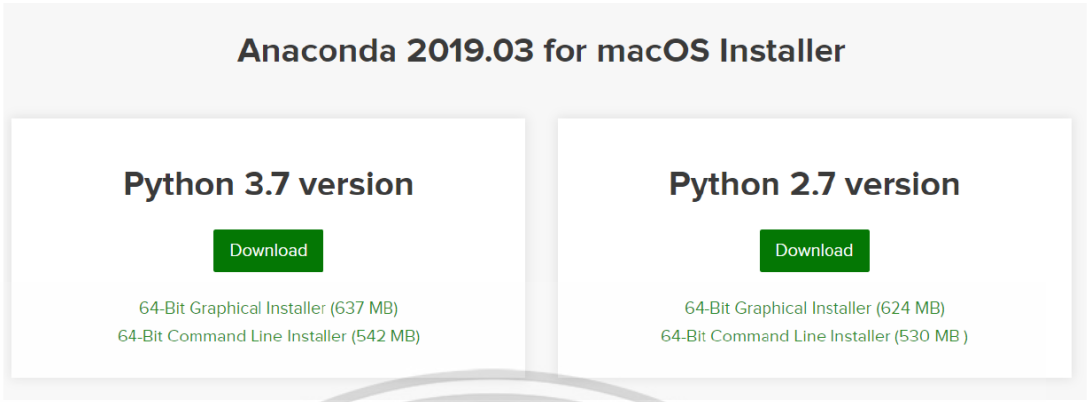
1. เปิดเบราว์เซอร์ของคุณและไปที่เว็บไซต์ Anaconda (https://www.anaconda.com/distribution/) เพื่อดาวน์โหลดและติดตั้ง Anaconda แล้วคลิกที่ดาวน์โหลด



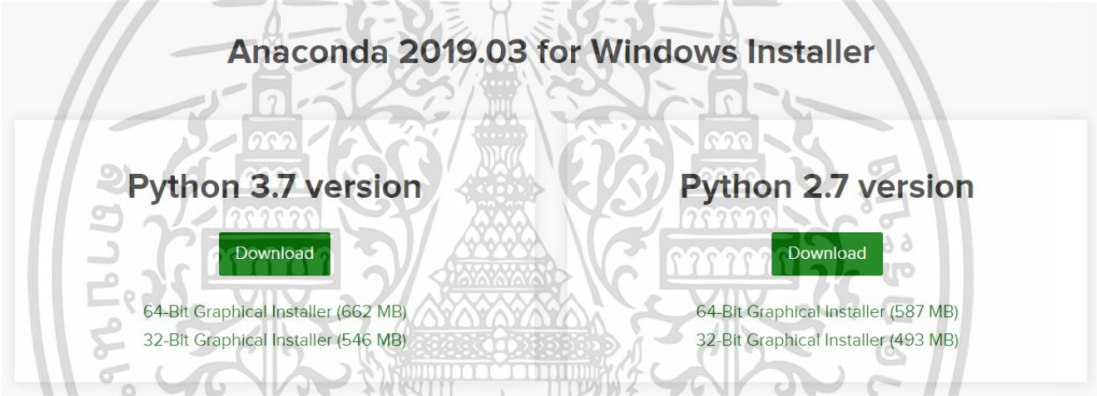
2. ปรากฏขึ้นตามค่าเริ่มต้น Anaconda จะแสดงลิงก์ดาวน์โหลดสำหรับระบบปฏิบัติการ Mac และ Windows

ระบบระบบปฏิบัติการ Mac คลิก" ตัวติดตั้งกราฟิก 64 บิต" ภายใต้เวอร์ชัน Python 3.7 เพื่อเริ่มดาวน์โหลดไฟล์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

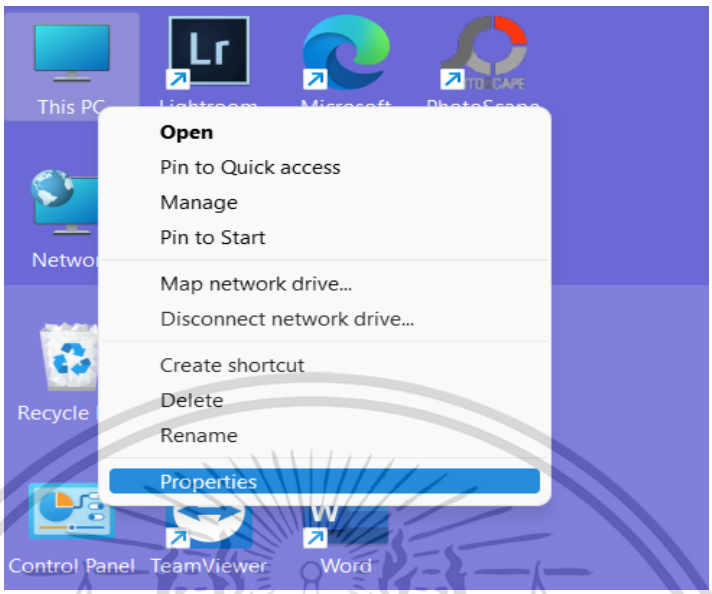


ระบบระบบปฏิบัติการ Windows ตัวติดตั้งกราฟิก 64 บิต" ภายใต้เวอร์ชัน Python 3.7 เพื่อเริ่มดาวน์โหลดไฟล์



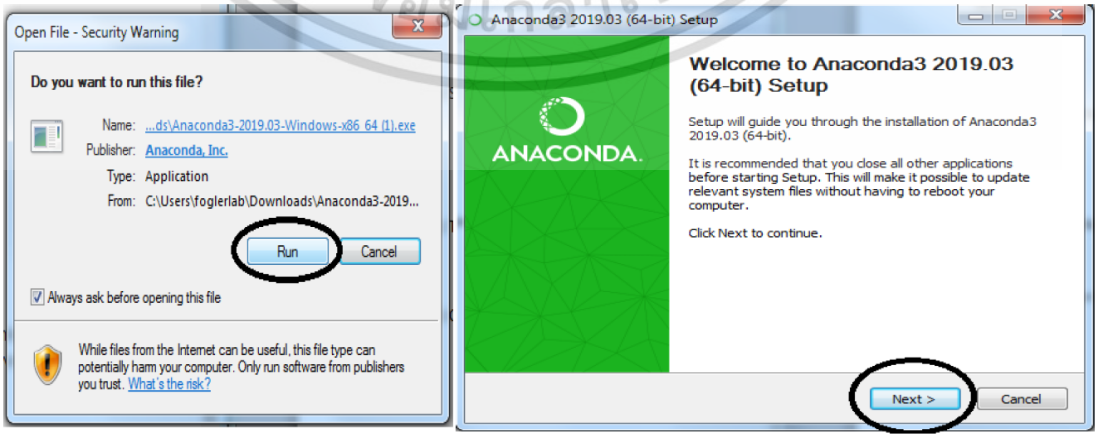
สำหรับ Windows: 64 บิตและ 32 บิต ค้นหาว่าระบบของคุณ คือ 64-Bit หรือ 32-Bit และต้องเลือกไฟล์ให้เหมาะสมสำหรับระบบของคุณ จากนั้นคลิกดาวน์โหลด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



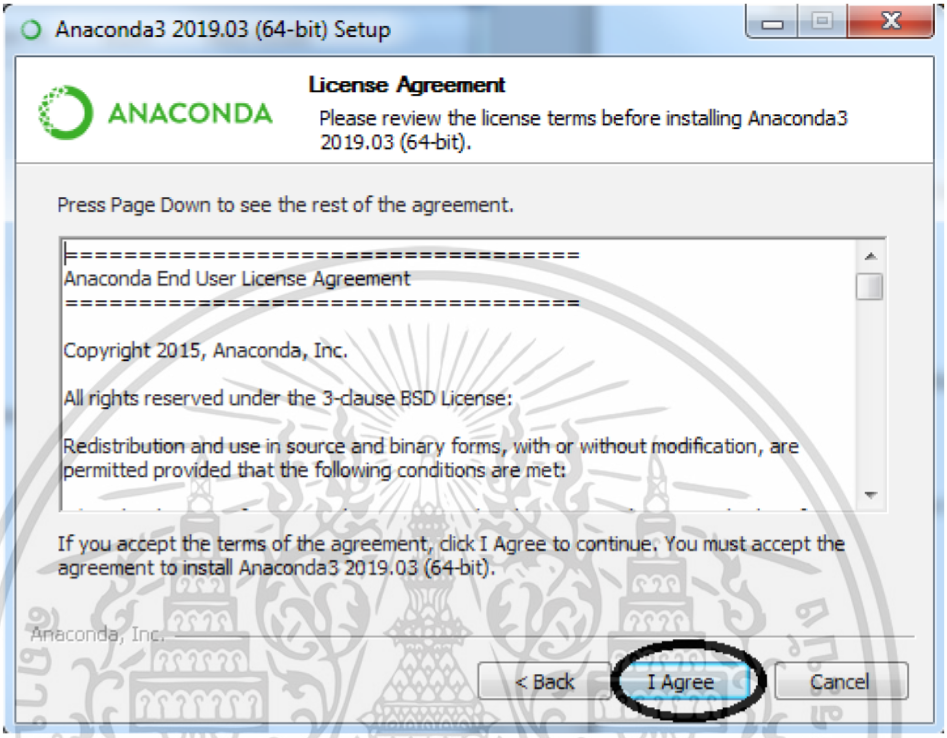
Device name	DESKTOP-EACAVSE
Processor	11th Gen Intel(R) Core(TM) i5-11300H @ 3.10GHz 2.61 GHz
Installed RAM	8.00 GB (7.74 GB usable)
Device ID	171C7377-1ECF-40CF-B1E2-4DCF4DB68C07
Product ID	00327-36337-65322-AAOEM
System type	64-bit operating system, x64-based processor

3. เมื่อดาวน์โหลดไฟล์อย่างสมบูรณ์ให้คลิกที่ไฟล์ที่ดาวน์โหลดมา คุณจะเห็นว่าหน้าต่างต่อไปนี้ปรากฏขึ้น คลิกที่ปุ่ม 'Run' จากนั้นคลิกปุ่ม 'Next'



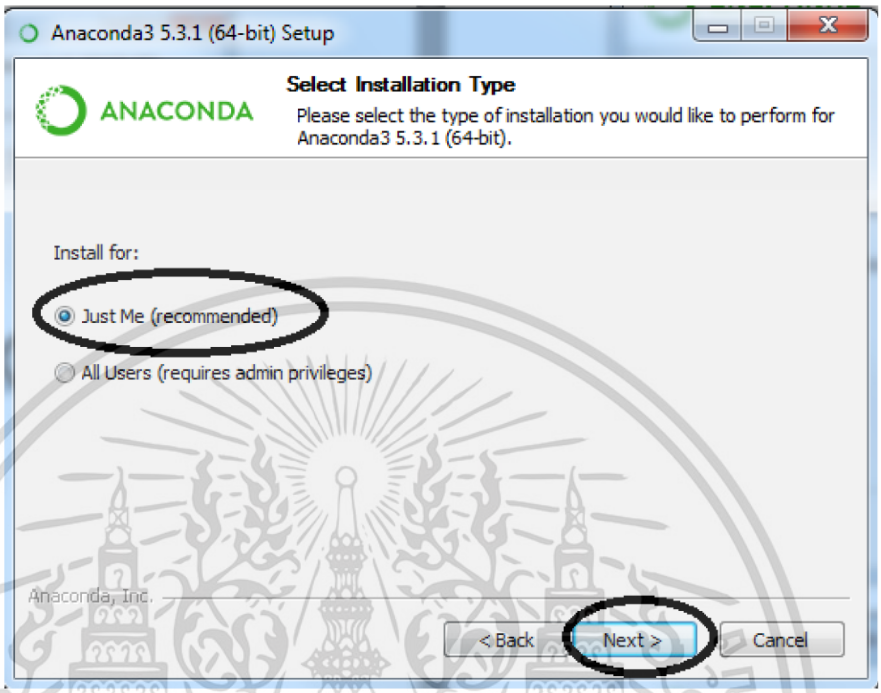
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. หน้าต่างใหม่จะปรากฏขึ้นเพื่อขอให้คุณยอมรับเงื่อนไขของข้อตกลงเลือก "I Agree"

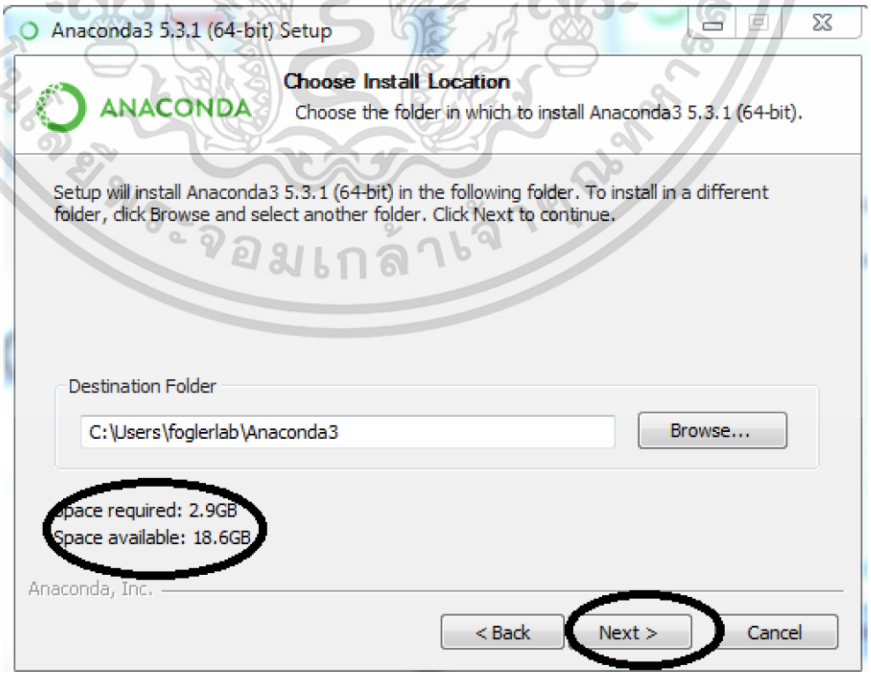


เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. คลิกเลือก “Just Me” แล้วกด Next

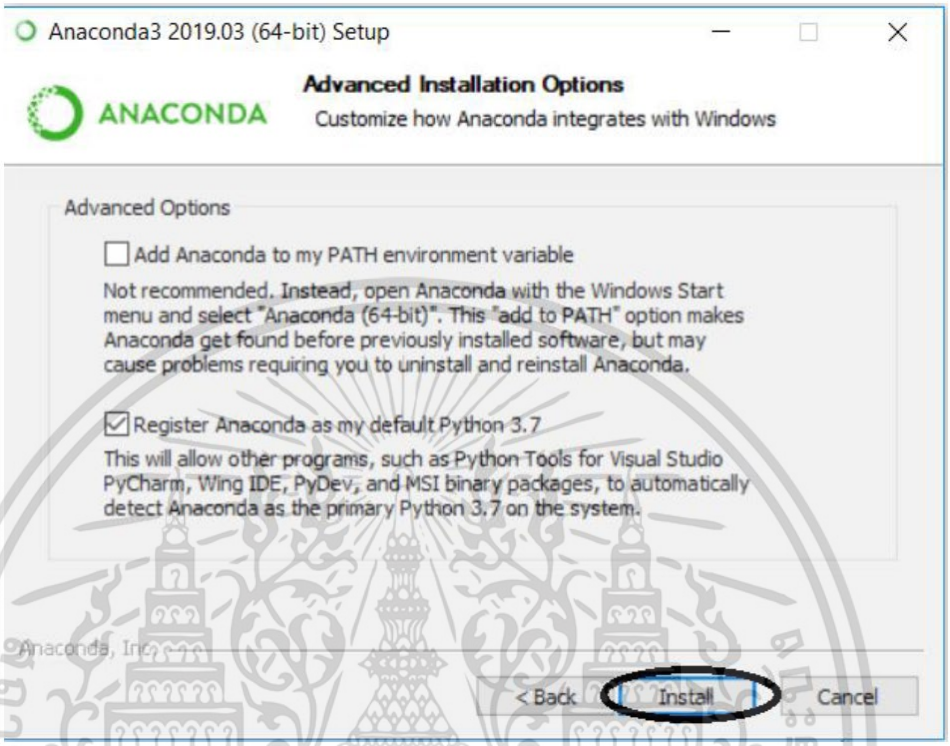


6. ตรวจสอบว่าพื้นที่ที่จะติดตั้งเพียงพอหรือไม่ ถ้าเพียงพอกด Next



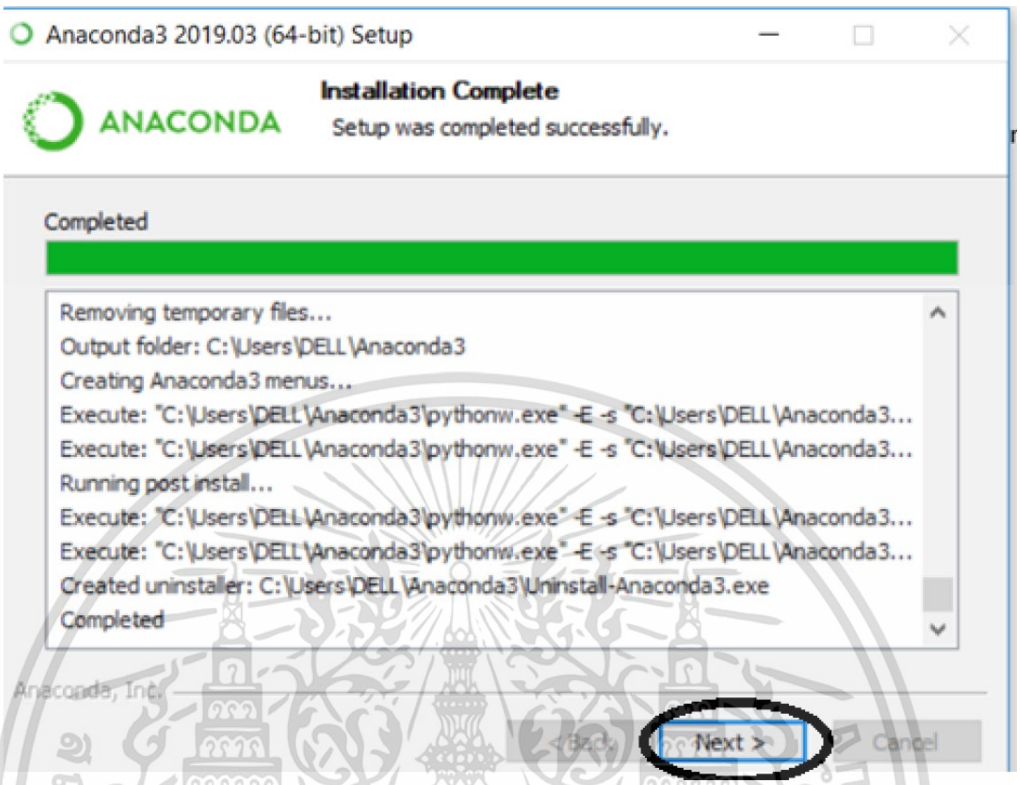
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7. แถบหน้าต่างนี้ปรากฏขึ้นให้กด Install

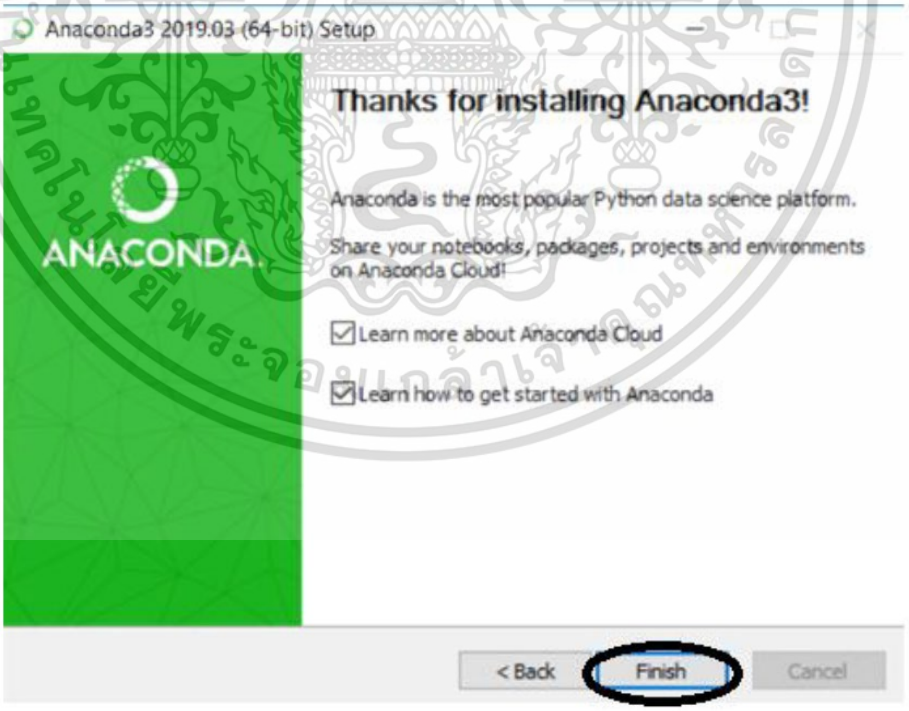


8. แถบหน้าต่างนี้ปรากฏขึ้นให้กด Next

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

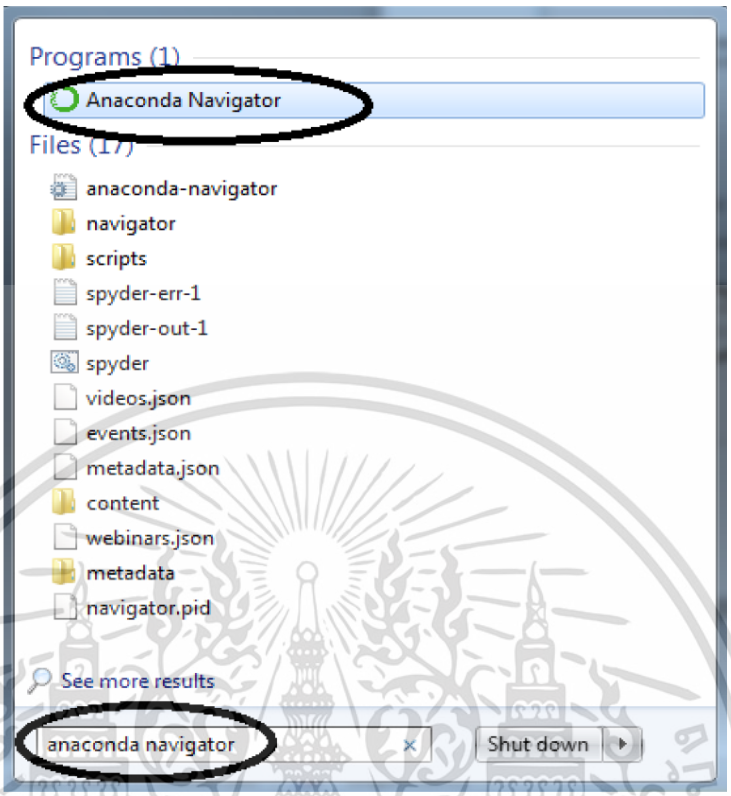


9. ถ้าติดตั้งสมบูรณ์แล้ว ให้กด Finish

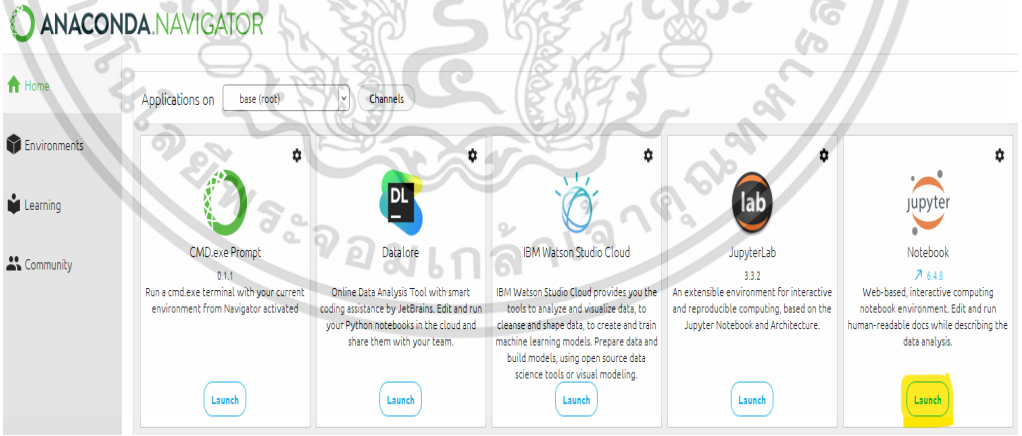


10. ค้นหา Anaconda Navigator

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



11. เลือก jupyter แล้วกด Launch



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ข

ชุดคำสั่ง

```
Import sklearn
!pip install feature_engine imbalanced_learn
!pip install genism
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import LabelEncoder
from sklearn.utils import shuffle
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.patches import Rectangle
import numpy as np
from pprint import pprint as pp
import csv
from pathlib import Path
import seaborn as sns
from itertools import product
import string
from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import BorderlineSMOTE
from imblearn.pipeline import Pipeline
from sklearn.metrics import r2_score, classification_report, confusion_matrix,
accuracy_score, roc_auc_score, roc_curve, precision_recall_curve,
average_precision_score
from sklearn.metrics import homogeneity_score, silhouette_score
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import MiniBatchKMeans, DBSCAN
from sklearn.model_selection import train_test_split, GridSearchCV
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30,
random_state = 0)
from sklearn.linear_model import LogisticRegression
model = LogisticRegression (solver = 'saga', penalty = 'l2',
class_weight = ’balanced’)
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier (criterion = 'entropy', max_depth = 7,
min_samples_split = 3, splitter = 'best')
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(criterion = 'entropy', max_depth = 8,
min_samples_leaf = 3)
from xgboost import XGBClassifier
model = XGBClassifier(, max_depth = 5)
form sklearn.ensemble.AdaBoostClassifier
model = AdaBoostClassifier (base_estimator = (criterion = 'entropy', max_depth = 7,
min_samples_split = 3, splitter = 'best'))
print('classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred)
```

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นางสาวณัฏฐา ดวงศรี
วัน เดือน ปีเกิด	29 กันยายน พ.ศ. 2539
ที่อยู่ปัจจุบัน	เขตคันนายาว จังหวัดกรุงเทพมหานคร
ประวัติการศึกษา	(2560) วิทยาศาสตรบัณฑิต สาขาวิชาสถิติประยุกต์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

แบบฟอร์มการตรวจสอบการคัดลอกผลงานทางวิชาการ

ข้าพเจ้า (นาย/นาง/นางสาว) ณัฏฐา ดวงศรี รหัสนักศึกษา 63605123

ระดับ (ปริญญาตรี/ปริญญาโท/ปริญญาเอก) สาขาวิชา สถิติและการวิเคราะห์ธุรกิจ

ภาควิชา สถิติประยุกต์ คณะ วิทยาศาสตร์

ได้เสนอ

- ☐ วิทยานิพนธ์
- ☐ โครงงานพิเศษ
- ☐ สหกิจศึกษา
- ☑ การค้นคว้าอิสระ
- ☐ ปัญหาพิเศษ
- ☐ เทียบเท่า ระบุ.....
- ☐ ปริญญานิพนธ์
- ☐ การศึกษาอิสระ

หัวข้อเรื่อง

(ไทย) การตรวจจับการฉ้อโกงบัตรเครดิตโดยเทคนิคการเรียนรู้ของเครื่อง

(อังกฤษ) Credit Card Fraud Detection Using Machine Learning Techniques

ได้ตรวจเช็คตรวจผลงานวิชาการข้างต้นแล้ว ในภาคเรียนที่ 2 วันที่ 18 เดือน ตุลาคม ปี 2565

โดยใช้โปรแกรม

- ☑ อักขราวิสุทธิ์
- ☐ TURNITIN

ทั้งนี้ ตรวจสอบพบความเหมือนของเนื้อหา 6.93 % โดยอาจารย์ที่ปรึกษายอมรับได้ว่าไม่ได้คัดลอกข้อความที่มีสาระสำคัญจากผลงานของผู้อื่น

ลายมือชื่อนักศึกษา ณัฏฐา ดวงศรี

(นางสาวณัฏฐา ดวงศรี)

วันที่ 18 ตุลาคม 2565

ได้รับความเห็นชอบจากอาจารย์ที่ปรึกษา

ลายมือชื่ออาจารย์ที่ปรึกษา พรพิมล ชัยวุฒิศักดิ์

(ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์)

วันที่ 18 ตุลาคม 2565

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้