การอนุมัติสินเชื่อลูกค้าใหม่โดยใช้การเรียนรู้เชิงลึก

NEW CUSTOMER LOAN APPROVAL USING DEEP LEARNING

จิรายุ โพธิสาร

JIRAYU PHOTISAN

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2566

KMITL-2023-SC-M-017-034

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

NEW CUSTOMER LOAN APPROVAL USING DEEP LEARNING

JIRAYU PHOTISAN

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIRMENT
OF THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS
KMITL DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCIENCE SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2023

KMITL-2023-SC-M-017-034

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2023

SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อการค้นคว้าอิสระ การอนุมัติสินเชื่อลูกค้าใหม่โดยใช้การเรียนรู้เชิงลึก
ชื่อนักศึกษา นายจิรายุ โพธิสาร
รหัสประจำตัว 64605024
ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์)
ภาควิชา ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ. 2566
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ รองศาสตราจารย์ ดร.ละออ บุญเกษม

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อสร้างตัวแบบเพื่อใช้ในการอนุมัติสินเชื่อของลูกค้าใหม่ ด้วยวิธีการเรียนรู้เชิงลึก โดยเปรียบเทียบกับการใช้วิธีการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะและเปรียบเทียบกับการใช้วิธีการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะและจัดการข้อมูลที่มีความผิดปกติ ข้อมูลที่ใช้ฝึกตัวแบบจำนวน 44,605 ตัวอย่าง ภาษาที่ใช้ในการเขียนโปรแกรมคือภาษา python กระบวนการเริ่มด้วยการจัดเตรียมข้อมูลเพื่อการประมวลผลโดยได้นำ Mutual information มาช่วยในการเลือกคุณลักษณะและจัดการข้อมูลที่มีค่าผิดปกติ (Anomaly) โดยใช้ Isolation forest แบ่งข้อมูลสำหรับการฝึก 80% และสำหรับการทดสอบ 20% แล้วจึงนำข้อมูลที่เตรียมไว้สำหรับฝึกไปสร้างตัวแบบและมีกระบวนการวัดผลตัวแบบ โดยใช้ค่าเรียกคืน (Recall) ค่าความแม่น (Accuracy) และค่าความเที่ยง (Precision) ในการวัดผลตัวแบบ และประสิทธิภาพของตัวแบบมีค่าเรียกคืน (Recall) เท่ากับ 70% ค่าความแม่น (Accuracy) เท่ากับ 70% และค่าความเที่ยง (Precision) อยู่ที่ 64% โดยวิธีการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะและการจัดการข้อมูลที่มีความผิดปกติมีประสิทธิภาพสูงที่สุด

คำสำคัญ : การเรียนรู้เชิงลึก, ค่าผิดปกติ, สินเชื่อ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Independent Study Title	New Customer Loan Approval Using Deep Learning
Student Name	Mr. Jirayu Photisan
Student ID	64605024
Degree	Master of Science (Data Science and Analytics)
	KMITL Digital Analytics and Intelligence Center
Year	2023
Independent Study Advisor	Assoc.Prof. Dr. Laor Boongasame

Abstract

This study aims to create a deep learning model that could be used to approve new customer loans. It compares deep learning with feature selection and deep learning with both feature selection and anomaly detection. The data used to train the model consists of 44,605 samples. The study was developed using Python as a programming language. The process begins by preparing the data for processing, using Mutual information to help select features and manage abnormal data with Isolation Forest. The data is split into training (80%) and testing (20%). Evaluate the model using recall, accuracy, and precision. The performance of Recall is 70%, Accuracy is 70%, and Precision is 64%. In conclusion, deep learning with feature selection and anomaly detection is more effective.

Keywords: Deep learning, Anomaly, Loan

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

งานวิจัยเรื่อง การอนุมัติสินเชื่อลูกค้าใหม่โดยใช้การเรียนรู้ของเครื่อง สามารถดำเนินการจนประสบความสำเร็จลุล่วงไปได้ด้วยดีเนื่องจากได้รับความความอนุเคราะห์เป็นอย่างยิ่งจาก รศ.ดร.ละออ บุญเกษม อาจารย์ที่ปรึกษาการค้นความอิสระและ ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์ ที่ได้กรุณาเสียสละเวลาให้คำปรึกษาและให้ความรู้แนวคิดในการทำงานวิจัยและชี้ให้เห็นข้อบกพร่องพร้อมแนวทางการแก้ไขจนกระทั่งงานวิจัยครั้งนี้สำเร็จไปได้ด้วยดี

ขอขอบคุณอาจารย์และเพื่อนคณะวิทยาศาสตร์สาขาวิทยาการข้อมูลและการวิเคราะห์ทุกท่านที่ให้คำแนะนำและให้การช่วยเหลือในการทำวิจัยครั้งนี้

ขอบคุณครอบครัวที่ให้การสนับสนุนงบประมาณและเวลาในการทำวิจัยครั้งนี้

สุดท้ายนี้ผู้วิจัยหวังว่างานวิจัยฉบับนี้จะเป็นประโยชน์สำหรับผู้ที่สนใจ

นายจิรายุ โพธิสาร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ก
บทคัดย่อภาษาอังกฤษ	ข
กิตติกรรมประกาศ	ค
สารบัญ	ง
สารบัญตาราง	ฉ
สารบัญรูป	ช
บทที่ 1 **บทนำ**	1
1.1 ความเป็นมาและความสำคัญของงานวิจัย/ปัญหา	1
1.2 วัตถุประสงค์ของงานวิจัย	1
1.3 ขอบเขตของงานวิจัย	1
1.4 ประโยชน์ที่คาดว่าจะได้รับ	1
บทที่ 2 **ทฤษฎีและงานวิจัยที่เกี่ยวข้อง**	2
2.1 การเรียนรู้ของเรื่อง (Machine Learning)	2
2.2 โครงข่ายประสาทเทียม (Artificial Neural Network)	3
2.2.1 Binary Cross-Entropy (Loss Function)	4
2.2.2 Sigmoid Function (Activation Function)	5
2.2.3 อัตราการเรียนรู้ (Learning Rate)	6
2.2.4 การแพร่กระจายย้อนกลับ (Backpropagation)	6
2.3 การเรียนรู้เชิงลึก (Deep Learning)	7
2.4 Isolation Forest	8
2.5 การเลือกคุณลักษณะ (Feature Selection)	8
2.5.1 ข้อมูลร่วม (Mutual Information)	8
2.5.2 วิธีการเลือกแบบก้าวหน้า (Forward Selection)	9
2.5.3 วิธีการเลือกแบบถอยหลัง (Backward Selection)	9
2.5.4 วิธีการเลือกแบบลำดับขั้น (Stepwise Selection)	9
2.6 งานวิจัยที่เกี่ยวข้อง	10
บทที่ 3 **วิธีการดำเนินงานวิจัย**	13
3.1 ชุดข้อมูล	13
3.2 การจัดเตรียมข้อมูลเพื่อการประมวลผล	15
3.2.1 จัดการข้อมูลสูญหาย (Missing Value)	16

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2.2 ตรวจสอบความถูกต้องของข้อมูล 16
3.2.3 ทำให้เป็นค่ามาตรฐาน (Standardization) 17
3.2.4 การเข้ารหัสข้อมูลเชิงกลุ่ม (Categorical Data Encoding) 17
3.2.5 เลือกคุณลักษณะ 17
3.2.6 จัดการข้อมูลที่มีความผิดปกติ 18
3.3 การสร้างตัวแบบ (Modelling) 19
3.3.1 จัดการข้อมูลที่มีลักษณะไม่สมดุล (Imbalance) 19
3.3.2 แบ่งข้อมูลสำหรับฝึกและทดสอบ 20
3.3.3 การสร้างตัวแบบการเรียนรู้เชิงลึก (Deep Learning Model) 20
3.3.4 การหาค่า Hyperparameters ที่เหมาะสม 23
3.4 การวัดผลตัวแบบ (Model Evaluation) 23
บทที่ 4 **ผลการวิจัยและการอภิปรายผล** 24
4.1 กำหนดโครงสร้างของการเรียนรู้เชิงลึก 24
4.1.1 การหาโครงสร้างของการเรียนรู้เชิงลึก 24
4.1.2 การปรับค่า Hyperparameters 24
4.2 เลือกคุณลักษณะด้วย Mutual Information 25
4.3 จัดการข้อมูลที่มีความผิดปกติ 26
4.4 ผลการศึกษาและการวัดประสิทธิภาพ 27
4.5 ปัญหาที่พบจากการศึกษา 27
บทที่ 5 **สรุปผลการวิจัยและข้อเสนอแนะ** 28
เอกสารอ้างอิง 29
ประวัติผู้เขียน 30

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่	หน้า
2.1 แสดงประสิทธิภาพของ model สำหรับข้อมูล German Credit Data set	11
2.2 แสดงประสิทธิภาพของ model สำหรับข้อมูล Australian Credit Approval.	11
2.3 แสดงประสิทธิภาพของ model สำหรับข้อมูล Kaggle Credit Data	11
2.4 แสดงการเปรียบเทียบงานวิจัย	12
3.1 แสดงคุณลักษณะของชุดข้อมูล	13
3.2 แสดงวิธีการแก้ไขคุณลักษณะที่ข้อมูลสูญหาย	16
3.3 แสดงความผิดปกติของข้อมูลของแต่ละคุณลักษณะที่พบ	16
3.4 แสดงตัวอย่างผลลัพธ์ของการทำให้ค่าเป็นมาตรฐาน	17
3.5 แสดงตัวอย่างผลลัพธ์ของการเข้ารหัสด้วยวิธีการ One hot Encoding ของคุณลักษณะ BUY_ACCESSORY	17
3.6 แสดงโครงสร้างของตัวเข้ารหัสอัตโนมัติที่มีชั้นซ่อน 3 ชั้น	21
3.7 แสดงโครงสร้างของตัวเข้ารหัสอัตโนมัติที่มีชั้นซ่อน 5 ชั้น	22
3.8 แสดงโครงสร้างของตัวเข้ารหัสอัตโนมัติที่มีชั้นซ่อน 7 ชั้น	23
3.9 แสดงค่า Hyperparameters ที่ใช้ในการหาโครงสร้างของตัวแบบ	23
4.1 แสดงค่าประสิทธิภาพของตัวแบบของตัวแบบการเรียนรู้เชิงลึกของแต่ละชั้นซ่อน	24
4.2 เปรียบเทียบประสิทธิภาพของตัวแบบในการใช้ Mutual Information	26
4.3 แสดงประสิทธิภาพของตัวแบบของการลบข้อมูลแต่ละแบบ	26
4.4 เปรียบเทียบประสิทธิภาพของวิธีการเรียนรู้เชิงลึกร่วมกับ Mutual Information และ Isolation Forest	28

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่	หน้า
2.1 แสดงโครงสร้างของเซลล์ประสาทเทียม	4
2.2 แสดงสมการ sigmoid	5
2.3 แสดงอัตราการเรียนรู้แต่ละแบบ	6
2.4 แสดงการเรียนรู้เชิงลึก	7
2.5 เปรียบเทียบระหว่าง Machine Learning และ Deep learning	7
2.6 แสดงต้นไม้ตัดสินใจที่ใช้หาความผิดปกติของข้อมูล	8
3.1 แสดงขั้นตอนการดำเนินงานวิจัย	13
3.2 แสดงขั้นตอนการเตรียมข้อมูล	15
3.3 แสดงค่า Mutual Information ของคุณลักษณะที่มีค่าสูงสุด 10 คุณลักษณะ	18
3.4 แสดงตัวแบบการเรียนรู้เชิงลึก	19
3.5 แสดงตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะโดยใช้ Mutual Information	19
3.4 แสดงตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะโดยใช้ Mutual Information และการจัดการข้อมูลผิดปกติด้วย Isolation Forest	20
3.5 แสดงตัวอย่างของข้อมูลก่อนและหลังการทำ Undersampling	20
3.6 แสดงตัวอย่างของข้อมูลก่อนแหละหลังการทำ Oversampling	21
4.1 แสดงค่าสูญเสียของการฝึก 0 - 500 รอบ เปรียบเทียบกับ 0 - 100 รอบ	25
4.2 แสดงค่าการสูญเสียของการฝึกระหว่าง Learning rate 0.001 และ 0.00001	25
4.3 แสดงค่าการสูญเสียของการฝึกระหว่างการใช้ Mutual Information และ ไม่ใช้ Mutual Information	26

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

นับจากปี พ.ศ. 2562 เกิดการระบาดของไวรัส โควิด19 ทำให้ประชากรส่วนใหญ่ต้องทำงานที่บ้านและนักเรียนนักศึกษาต้องเรียนออนไลน์ ทำให้จำนวนของความต้องการอุปกรณ์ด้านไอทีสำหรับการทำงานและการเรียนเพิ่มมากขึ้นด้วย แต่เนื่องจากราคาของอุปกรณ์ไอทีที่มีราคาสูงทำให้โอกาสของการเข้าถึงอุปกรณ์เหล่านี้ โดยเฉพาะสำหรับนักเรียนนักศึกษาที่มีกำลังทรัพย์ไม่มาก การปล่อยสินเชื่อสำหรับกลุ่มนักเรียนนักศึกษาจึงเป็นการเปิดโอกาสให้พวกเขาได้เข้าถึงเครื่องมือสำหรับการเรียนรู้ได้เพิ่มมากยิ่งขึ้น ทางผู้จัดจำหน่ายสินค้าเองก็สามารถทำยอดการขายได้เพิ่มมากขึ้นเช่นกัน

อย่างไรก็ตามการปล่อยสินเชื่อให้กับนักเรียนนักศึกษา จากสถิติพบว่ามีความเสี่ยงสูงที่จะมีโอกาสผิดชำระ และมีแนวโน้มที่จะมีผู้ที่มีการผิดนัดชำระเพิ่มขึ้นอย่างต่อเนื่อง นอกจากนี้ข้อมูลที่ใช้ในการประเมินสินเชื่อมีข้อมูลเกี่ยวกับการเงินไม่มาก แตกต่างจากข้อมูลการขออนุมัติสินเชื่อของธนาคาร จึงเป็นปัญหาที่ท้าทายสำหรับการทำนายลูกค้าที่ควรได้รับการอนุมัติสินเชื่อ

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาวิธีการเลือกคุณลักษณะ (Feature Selection) ที่เหมาะสมสำหรับการทำนายลูกค้าที่มีโอกาสได้รับการอนุมัติสินเชื่อ
2. เพื่อพัฒนาตัวแบบการทำนายการอนุมัติสินเชื่อลูกค้าใหม่ด้วยวิธีการเรียนรู้เชิงลึก
3. เพื่อเปรียบเทียบประสิทธิภาพการทำนายการอนุมัติสินเชื่อลูกค้าใหม่ด้วยวิธีการเรียนรู้เชิงลึกกับวิธีการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะและวิธีการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะและการจัดการข้อมูลที่มีความผิดปกติ

1.3 ขอบเขตของงานวิจัย

1. ทำการศึกษาคุณลักษณะ (Feature) ที่มีผลต่อการทำให้เกิดอนุมัติสินเชื่อลูกค้าใหม่
2. สร้างตัวแบบ (Model) โดยใช้ภาษา Python

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ทราบถึงคุณลักษณะ (Feature) ใดที่มีผลต่อการทำนายลูกค้าที่มีโอกาสได้รับการอนุมัติสินเชื่อ
2. ทราบถึงวิธีการพัฒนาตัวแบบสำหรับการทำนายการอนุมัติสินเชื่อลูกค้าใหม่ด้วยวิธีการเรียนรู้เชิงลึก
3. ทราบถึงประสิทธิภาพของวิธีการเรียนรู้เชิงลึกเมื่อเปรียบเทียบกับวิธีอื่น ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การเรียนรู้ของเครื่อง (Machine learning)

การเรียนรู้ของเครื่อง (Machine learning) คือศาสตร์แขนงหนึ่งที่ทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้และตัดสินใจด้วยตนเองโดยการเรียนรู้ข้อมูลและทำนายข้อมูลได้ ทำงานได้โดยปราศจากการลำดับของโปรแกรม สามารถจำแนกได้ 3 ประเภทหลักๆ คือ

1. การเรียนรู้แบบมีผู้สอน (Supervised Learning) คือ การใช้ชุดข้อมูลสำหรับการฝึก (Training Dataset) และชุดข้อมูลและข้อมูลที่เป็นผลลัพธ์ (Label) จากนั้นใช้ชุดข้อมูลเหล่านี้มาใช้ในการสร้างตัวแบบ (model) เมื่อได้ model แล้วจึงสำข้อมูลใหม่ที่ model ไม่เคยรู้จักนำเป็นเป็นข้อมูล input เพื่อให้ model ได้ทำนายผลลัพธ์และได้เป็น output ออกมา
2. การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการเรียนรู้ที่ทำให้คอมพิวเตอร์สามารถหาวิธีการเรียนรู้ได้ด้วยตนเอง โดยอาศัยมนุษย์ระบุความต้องการลงไป มีเพียงข้อมูลขาเข้าสู่ระบบ (Input) เพียงเท่านั้น ตัวอย่างของการเรียนรู้แบบไม่มีผู้สอนเช่น การจัดกลุ่ม (Clustering) , การลดมิติข้อมูล (Dimension reduction)
3. การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) โดยทำการเรียนรู้จากการลองผิดลองถูกในสถานการณ์ในอดีตหรือแบบจำลอง หาผลลัพธ์ออกมาดีก็จะมีการให้คะแนนและลงโทษหากผลลัพธ์ออกมาไม่ดี ตัวอย่างของการประยุกต์ใช้การเรียนรู้แบบเสริมกำลัง เช่น AlphaGo ที่สามารถเล่นชนะผู้เล่นระดับโลกได้

การเรียนรู้ของเครื่องยังมีความแตกต่างกันของชุดข้อมูลที่ใช้สำหรับสร้าง model เช่น

1. ชุดข้อมูลสำหรับฝึก (Training Dataset) คือชุดข้อมูลที่ไว้สำหรับทำการฝึก model
2. ชุดข้อมูลสำหรับการตรวจสอบความถูกต้อง (Validation Dataset) คือชุดข้อมูลสำหรับวัดประสิทธิภาพของ model หลังการปรับ parameter แต่ละครั้ง
3. ชุดข้อมูลสำหรับการทดสอบ (Test Dataset) คือชุดข้อมูลสำหรับทดสอบ model ว่ามีประสิทธิภาพเป็นอย่างไรซึ่งเป็นชุดข้อมูลที่ตัวแบบไม่เคยเห็นมาก่อน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2 โครงข่ายประสาทเทียม (Artificial Neural Network)

แนวคิดโครงข่ายประสาทเทียมเริ่มต้นมาจากโครงข่ายประสาทชีววิทยา (Biological Neural Network) ของมนุษย์แล้วนำมาเป็นต้นแบบในการสร้างตัวประมวลผล โดย Martin, Harward and Mark (1996) ได้กล่าวไว้ว่า องค์ประกอบพื้นฐานของโครงข่ายประสาทของมนุษย์ประกอบไปด้วย เซลล์ประสาท (Neuron) และ จุดประสานประสาท (Synapse) แต่ละเซลล์ประสาทประกอบไปด้วย

1. เดนไดรท์ (Dendrite) คือส่วนที่ทำหน้าที่รับสัญญาณที่ถูกส่งมาจากเซลล์อื่นๆ
2. ตัวเซลล์ (Cell Body) ทำหน้าที่รับสัญญาณที่ส่งมาจาก Dendrite
3. แอคซอน (Axon) เหมือน output ของ เซลล์ทำหน้าที่ส่งสัญญาณที่รับมาจาก Cell Body ไปยัง โครงข่ายอื่นๆ

โครงข่ายประสาทเทียมมีกระบวนการคล้ายกับการส่งผ่านสัญญาณประสาทในสมองของมนุษย์ คือมีความสามมารถในการรวบรวมความรู้ผ่านกระบวนการเรียนรู้ (Learning process) และความรู้จะถูกกับอยู่ในโครงข่ายในรูปของน้ำหนัก (weight) ซึ่งสามารถเปลี่ยนไปได้เมื่อมีการเรียนรู้ใหม่ๆ เข้ามา

การประมวลผลเกิดขึ้นในหน่วยย่อยที่เรียกว่า โหนด (node) ซึ่งจำลองการทำงานของเซลล์ และการส่งสัญญาณระหว่างโหนดที่เชื่อมต่อกันจำลองการทำงานของเดนไดรท์และแอคซอน ภายในโหนดจะมีฟังก์ชันกระตุ้น (activation function) หรือฟังก์ชันการแปลง (transfer function) โดยโครงข่ายประสาทเทียมประกอบไปด้วย 5 องค์ประกอบดังนี้

1. **ข้อมูลเข้า (input)** เป็นข้อมูลที่เป็นตัวเลขหรือข้อมูลเชิงปริมาณถ้าหากเป็นข้อมูลเชิงคณภาพต้องแปลงให้อยู่ในรูปในเชิงปริมาณที่โครงข่ายประสาทเทียมยอมรับได้
2. **ข้อมูลออก (output)** เป็นผลลัพธ์ที่ได้จากกระบวนการเรียนรู้ของโครงข่ายประสาเทียม
3. **ค่าน้ำหนัก (weights)** คือสิ่งทีได้จากการเรียนรู้ของโครงข่ายประสาทเทียมหรือ ค่าความรู้ (knowledge) ซึ่งจะถูกเก็บเป็นทักษะเพื่อใช้ในการจดจำข้อมูลอื่นๆ ที่อยู่ในรูปแบบเดียวกัน
4. **ฟังก์ชันกระตุ้น (activation function)** หรือฟังก์ชันการแปลง (transfer function) จะทำหน้าที่ตัดสินใจว่าจะส่งสัญญาณใดออกไปในรูปแบบใด เป็นการคำนวนการจำลองการทำงานของโครงข่ายประสาทเทียม เช่น ฟังก์ชันไฮเปอร์โบลิคแทนเจนต์ (hyperbolic tangen function) , ซิกมอยด์ฟังก์ชัน (sigmoid function) เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5. **ฟังก์ชันผลรวม (summation function: S)** เป็นผลรวมของข้อมูลเข้า (a_i) และค่าน้ำหนัก (w_i)

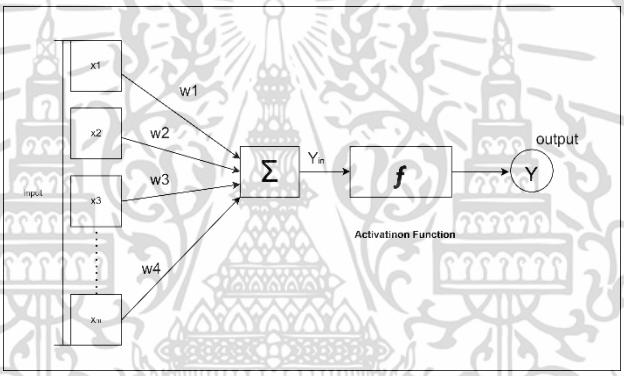
$$S = \sum_{i=0}^{n} a_i w_i$$

โดยที่

n คือ จำนวนข้อมูลทั้งหมด

a_i คือ ข้อมูลขาเข้า

w_i คือ ค่าน้ำหนัก



รูปที่ 2.1 แสดงโครงสร้างของเซลล์ประสาทเทียม

2.2.1 Binary Cross-Entropy (Loss function)

การที่ตัวแบบจะต้องรู้ว่าปรับค่าน้ำหนักอย่างไรให้ถูกต้องซึ่งเราสามารถดูได้จากค่าค่าผิดพลาด (Error) หรือซึ่งเราเรียกการคำนวณค่าค่าผิดพลาดของโครงข่ายประสาทเทียมว่าฟังก์ชันการสูญเสีย (Loss function) ซึ่งฟังก์ชันการสูญเสียมีการคำนวณได้หลากหลายแบบที่แพร่หลายคือ ค่าเฉลี่ยนผิดพลาดกำลังสอง (MSE) ซึ่งอีกตัวที่เราสนใจคือ Binary Cross Entropy ที่เกิดจากการแจกแจงความน่าจะเป็น 2 แบบ คือ การแจกแจงความน่าจะเป็นที่เราอยากได้ (Actual) กับการแจกแจงความน่าจะเป็นที่ถูกประมาณโดย Model (Predicted) โดยจะมีค่าเพียงสองคลาสคือ ไม่ 0 ก็ 1 เพื่อทำให้ผลลัพธ์ที่ได้บอกว่ามีโอกาสกี่เปอร์เซ็นต์ที่จะมีโอกาสเป็นคลาส 1 จะต้องใช้ฟังก์ชันกระตุ้นใน output layer จะต้องเป็น Sigmoid เสมอ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$Loss = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log \hat{y}_i + (1 + y_i) \cdot \log(1 - \hat{y}_i)$$

โดยที่

n คือ จำนวนที่ใช้

y_i คือ ค่าจริงที่ i ใด ๆ

$\hat{y}_i$ คือ ค่าที่ได้จากการทำนายที่ i ใด ๆ

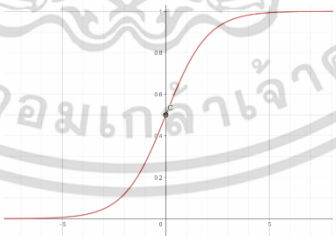
2.2.2 Sigmoid function (Activation function)

Sigmoid function เป็นฟังก์ชันที่เป็นรูป S curve เนื่องจาก output ของ sigmoid function มีค่าระหว่าง 0-1 ซึ่งเหมาะสมอย่างยิ่งที่จะใช้ในงานที่ต้องการ output ที่มีความน่าจะเป็น (Probability) 0,1 หรือ yes, no

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

x คือ ข้อมูลเข้าของ ซิกมอยด์ ฟังก์ชัน

e คือ จำนวนของออยเลอร์



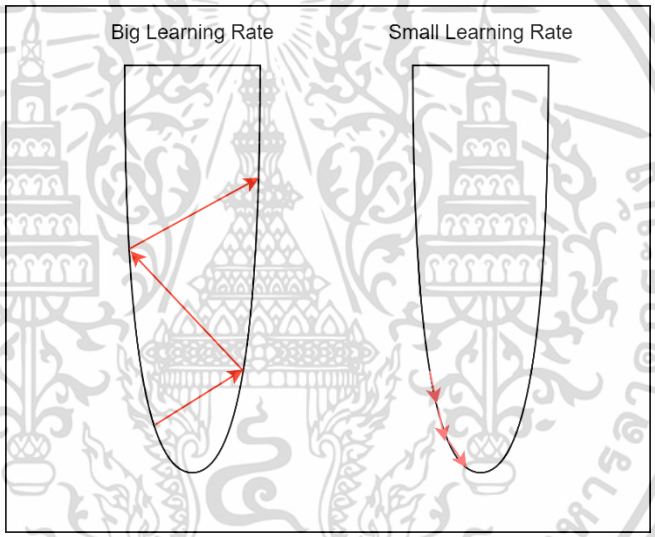
รูปที่ 2.2 แสดงสมการ sigmoid

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2.3 อัตราการเรียนรู้ (Learning Rate)

Learning Rate คือ ไฮเปอร์พารามิเตอร์ (Hyperparameter) ที่สำคัญตัวหนึ่ง ที่มีหน้าที่ปรับขนาดของ error ในแต่ละครั้งของการปรับปรุงค่าน้ำหนัก

1. อัตราการเรียนรู้มีค่าน้อยทำให้ค่าน้ำหนักของตัวแบบเปลี่ยนแปลงน้อย การ ฝึกตัวแบบจะค่อย ๆ ไปทีละน้อยโอกาสเกิดค่าความผิดพลาดมีน้อยดังรูปที่ 2.3 ซึ่งการที่มีค่าน้อยทำให้ความเร็วของการฝึกตัวแบบ (Train model) ช้าแต่ก็จะมีโอกาสเข้าใกล้จุดที่มีค่าความผิดพลาดได้ต่ำที่สุด
2. อัตราการเรียนรู้ มีค่ามากทำให้ค่าน้ำหนักของตัวแบบเปลี่ยนแปลงมาก การฝึกตัวแบบ จะขยับไปมาก ค่า loss ก็จะมีมากขึ้นดังรูปที่ 2.3 ซึ่งการที่มีค่ามากทำให้ความเร็วของการฝึกตัวแบบเร็วจำนวนรอบของการฝึก (Epoch) น้อยแต่ก็จะมีโอกาสข้ามจุดที่มีค่าความผิดพลาดที่ต่ำที่สุดได้ง่าย



รูปที่ 2.3 แสดงอัตราการเรียนรู้แต่ละแบบ

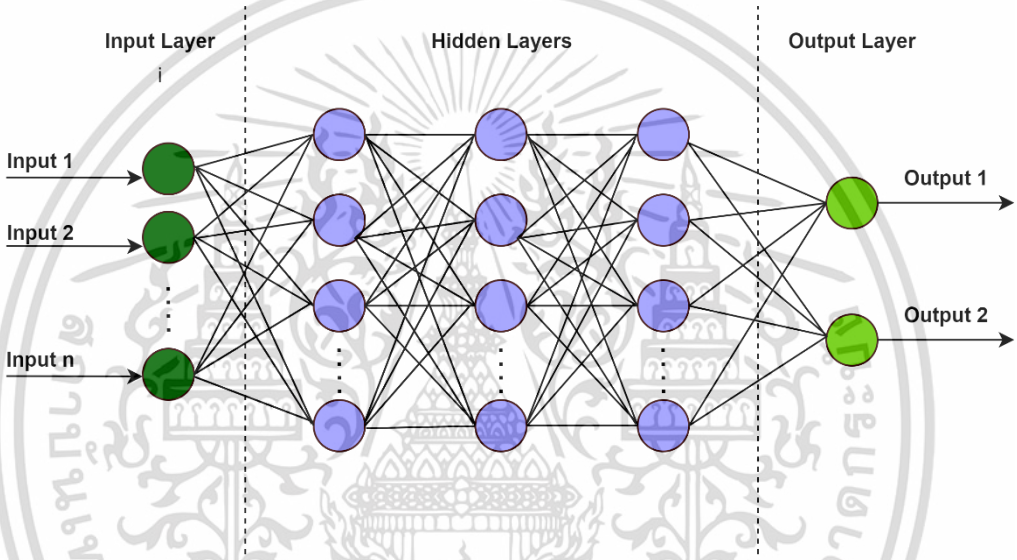
2.2.4 การแพร่กระจายย้อนกลับ (Backpropagation)

ขั้นตอนการเรียนรู้แบบแพร่กระจายย้อนกลับเป็นวิธีที่นิยมใช้ในโครงข่ายประสาทเทียมวิธีหนึ่ง นิยมใช้ในโครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron) หรือ MLP เพื่อปรับค่าน้ำหนักในเส้นเชื่อมระหว่างโหนดให้เหมาะสมโดยจะคำนวณจากความแตกต่างของค่าข้อมูลออกที่คำนวณได้กับค่าข้อมูลออกที่ต้องการ[5]

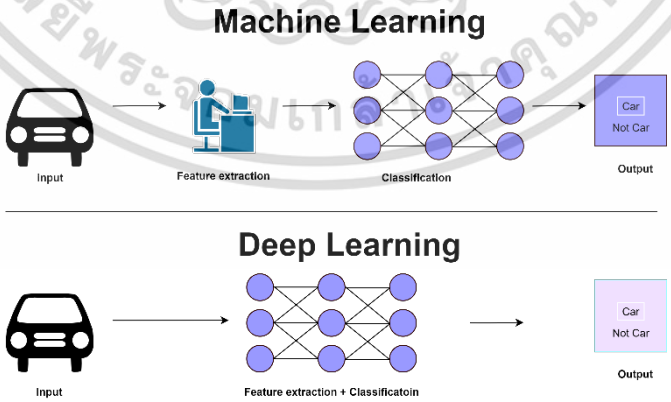
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึกเป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง (Machine learning) มีพื้นฐานมาจากโครงข่ายประสาทเทียม (Artificial Neural Network) [4] หรือ ANN ที่เป็นโหนดหลาย ๆ ชั้น ตามรูปที่ 2.4 ซึ่งประกอบไปด้วย ชั้นข้อมูลเข้า (input layer) และชั้นข้อมูลออก (output layer) ส่วนชั้นข้อมูลที่อยู่ระหว่างทั้งสองเรียกว่า ชั้นซ่อน (hidden layer) โดยที่ข้อมูลจะถูกป้อนเข้าสู่ชั้นข้อมูลเข้าและชั้นข้อมูลออกจะส่งข้อมูลไปยังชั้นข้อมูลเข้าของ layer ถัดไปตาม รูปที่ 2.4 และจุดเด่นของ Deep learning อีกหนึ่งจุดคือการสกัดคุณลักษณะ (Feature Extraction) หรือเป็นการคัดแยกคุณลักษระเด่นออกมาจากข้อมูล ดังรูปที่ 2.5



รูปที่ 2.4 แสดงการเรียนรู้เชิงลึก [10]

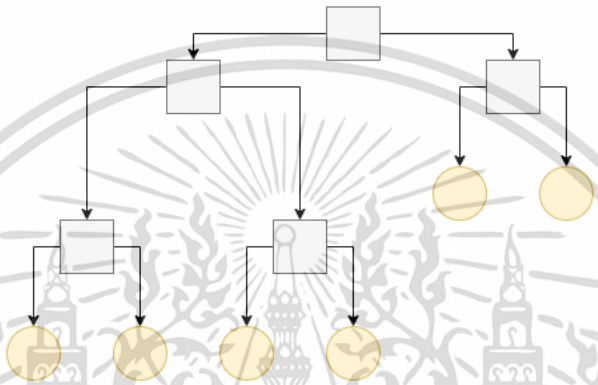


รูปที่ 2.5 เปรียบเทียบระหว่าง Machine Learning และ Deep learning

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.4 Isolation Forest

Isolation Forest เป็นการเรียนรู้แบบไม่มีผู้สอน พัฒนาโดย Fei Tony Liu [6] ใช้ตรวจจับค่าผิดปกติ (Anomaly Detection) [6] โดยใช้โครงสร้างต้นไม้ (tree) ซึ่งจะแบ่งข้อมูลออกเป็นกลุ่มต่างๆ อย่างสุ่ม เริ่มต้นจากการสุ่มคุณลักษณะ (Attribute) และแบ่งข้อมูล (Partition) ระหว่างค่าต่ำสุดและค่าสูงสุดเพื่อ แยกตัวอย่างและจะแบ่งข้อมูลไปเรื่อย ๆ จนกระทั่งข้อมูลแต่ละตัวแยกกันอย่างสมบูรณ์



รูปที่ 2.6 แสดงต้นไม้ตัดสินใจที่ใช้หาความผิดปกติของข้อมูล

จากรูปที่ 2.6 จะเห็นว่าต้นไม้ถูกแบ่งออกเป็นสองกิ่งคือกิ่งซ้ายและกิ่งขวา แล้วทำการแบ่งข้อมูลไปเรื่อย ๆ จนไม่สามารถแบ่งได้อีกหลังจากนั้นพบว่าข้อมูลที่ปกติจะใช้ชั้นของต้นไม้ที่ลึกมากในการแบ่งข้อมูลให้อิสระจากกัน ส่วนข้อมูลที่ผิดปกตินั้นจะมีความลึกที่น้อยกว่าแล้วใช้ความลึกของต้นไม้นี้ในการคำนวณคะแนนความผิดปกติ (Anomaly score) เพื่อใช้ในการแยกข้อมูล โดยมีค่าระหว่าง 0 – 1 โดยที่ 1 หมายถึงข้อมูลมีแนวโน้มที่จะผิดปกติ ส่วนข้อมูลที่มีคะแนนน้อยกว่า 0.5 ถือว่าเป็นข้อมูลที่มีความปกติ [7]

2.5 การเลือกคุณลักษณะ (Feature selection)

การจำแนกประเภทข้อมูล (Classification) พบว่าในหลายๆ ครั้งคุณลักษณะมีจำนวนมาก บางคุณลักษณะไม่ได้มีความสำคัญในการจำแนกประเภทข้อมูล ดังนั้นจึงจำเป็นต้องคัดเลือกคุณลักษณะที่สำคัญมาใช้ในการฝึกตัวแบบ ขั้นตอนนี้เรียกว่าการเลือกคุณลักษณะ (Feature selection) ซึ่งสามารถทำได้หลายวิธี ในที่นี้ขอนำเสนอวิธีการเลือกคุณลักษณะ 4 วิธีดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.5.1 ข้อมูลร่วม (Mutual Information)

Mutual Information หรือข้อมูลร่วม คือวัดการแปรผันร่วมหรือข้อมูลร่วมระหว่างสองตัวแปรสุ่ม ถ้าค่า Mutual Information มีค่าสูงแสดงว่ามีการแปรผันร่วมระหว่างสองตัวแปรสุ่มที่มาก นั่นคือ ตัวแปรหนึ่งสามารถสรุปข้อมูลเกี่ยวกับตัวแปรอื่นได้มาก ในทางกลับกัน ถ้า Mutual Information มีค่าต่ำ ตัวแปรสองตัวแปรนี้อาจจะไม่มีความสัมพันธ์กันมากนัก [9]

2.5.1 วิธีการเลือกแบบก้าวหน้า (Forward Selection)

เป็นวิธีการเลือกคุณลักษณะเข้าสู่สมการหรือตัวแบบทีละตัวตามลำดับความสัมพันธ์กับตัวแปรตาม Y จากคุณลักษณะที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดจะถูกเลือกเข้าก่อน โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) สูงสุด โดยเมื่อคุณลักษณะถูกเลือกมาแล้ว จะทำการทดสอบว่าคุณลักษณะที่เข้ามาทำให้ตัวแบบมีประสิทธิภาพในการทำนายดีขึ้นหรือไม่ หากมีประสิทธิภาพดีขึ้น จะทำการเก็บคุณลักษณะไว้ หากทำให้ประสิทธิภาพแย่ลงจะเลือกคุณลักษณะที่มีความสัมพันธ์ถัดไปแล้วทำการทดสอบอีกครั้ง ทำเช่นนี้ไปจนกว่าจะไม่มีคุณลักษณะใดเข้าทำการทดสอบได้อีก จึงหยุดการคัดเลือกคุณลักษณะถือว่าคุณลักษณะที่ได้ มีความเหมาะสมกับตัวแบบที่ใช้ แต่จุดบกพร่องของวิธีนี้คือกรณีมีคุณลักษณะจำนวนมากจะใช้เวลาในการทดสอบนานและไม่ได้ทดสอบประสิทธิภาพของการทำนายระหว่างคุณลักษณะที่ตัดออกไปแล้วก่อนหน้านี้กับคุณลักษณะที่เพิ่มเข้ามา [10]

2.5.2 วิธีการเลือกแบบถอยหลัง (Backward Selection)

เป็นวิธีการเลือกคุณลักษณะออกจากสมการหรือตัวแบบทีละตัวตามลำดับความสัมพันธ์กับตัวแปรตาม Y จากคุณลักษณะที่มีความสัมพันธ์กับตัวแปรตามน้อยที่สุดจะถูกเลือกออกก่อน โดยพิจารณาจากค่าสถิติ Partial F ต่ำที่สุด เมื่อคุณลักษณะถูกเลือกมาแล้วจะทำการทดสอบว่าคุณลักษณะที่ตัดออกไป ทำให้ตัวแบบมีประสิทธิภาพในการทำนายดีขึ้นหรือไม่ หากมีประสิทธิภาพดีขึ้นจะทำการเก็บคุณลักษณะไว้และหากทำให้ประสิทธิภาพแย่ลง จะเลือกคุณลักษณะที่มีความสัมพันธ์ถัดไปแล้วทำการทดสอบอีกครั้ง ทำเช่นนี้ไปจนกว่าจะไม่มีคุณลักษณะใดเข้าทำการทดสอบได้อีกจึงหยุดการคัดเลือกคุณลักษณะ ถือว่าคุณลักษณะที่ได้นั้นมีความเหมาะสมกับตัวแบบที่ใช้ จุดบกพร่องของวิธีนี้คือไม่สามารถทราบว่าคุณลักษณะที่ใช้แต่ละตัวสามารถทำนายตัวแปรตามได้มากหรือไม่บอกเพียงคุณลักษณะมีผลต่อการทำนายและหากมีคุณลักษณะจำนวนมากจะใช้เวลาในการประมวลผลมาก [10]

2.5.3 การคัดเลือกแบบลำดับขั้น (Stepwise Selection)

เป็นวิธีการที่ผสมผสานระหว่างวิธีการเลือกแบบก้าวหน้า (Forward Selection) และวิธีการเลือกแบบถอยหลัง (Backward Selection) เข้าด้วยกัน ในขั้นแรกจะเลือกคุณลักษณะที่มีค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายสูงที่สุดเข้าสมการหรือตัวแบบก่อน จากนั้นจะทดสอบคุณลักษณะที่ไม่ได้อยู่ในสมการว่าจะมีคุณลักษณะใดบ้างมีสิทธิ์เข้ามาอยู่ในสมการด้วยวิธีการคัดเลือกบแบบก้าวหน้า (Forward Selection) ขณะเดียวกันก็จะทดสอบคุณลักษณะที่อยู่ในสมการแล้วด้วยว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

คุณลักษณะใดที่มีโอกาสถูกขจัดออกจากสมการด้วยวิธีการคัดเลือกบแบถอยหลัง (Backward Selection) โดยจะทำการคัดเลือกผสมทั้งสองวิธีนี้จนไม่มีคุณลักษณะใดถูกคัดออกและไม่มีคุณลักษณะใดถูกนำเข้า กระบวนการก็จะยุติ จุดด้อยของวิธีนี้คือหากมีคุณลักษณะจำนวนมากจะใช้เวลาในการประมวลผลมาก แต่วิธีนี้จะแก้จุดด้อยของวิธีการคัดเลือกแบบก้าวหน้า (Forward Selection) และวิธีการคัดเลือกแบบถอยหลัง (Backward Selection) [10]

2.6 งานวิจัยที่เกี่ยวข้อง

ในส่วนนี้จะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับการใช้ deep learning สำหรับการอนุมัติสินเชื่อเพื่อนำองค์ความรู้มาประยุกต์ใช้ ผู้พัฒนาได้เลือกหัวข้อวิจัยที่สนใจทั้งหมด 3 โครงงานได้แก่ Ensemble Learning or Deep Learning Application to Default Risk Analysis, Credit Score Prediction System using Deep Learning and K-Means Algorithms และ Deep Learning Techniques for Credit Scoring.

2.6.1 Ensemble Learning or Deep Learning Application to Default Risk Analysis

งานวิจัยชิ้นนี้เสนอโดย Shigeyuki Hamori , Minami Kawai, Takahiro Kume, Yuji Murakami และ Chikara Watanabe จาก Kobe University ได้ศึกษาเกี่ยวกับการใช้ ensemble method ประกอบด้วย bagging, random forest, และ boosting เปรียบเทียบกับ neural network ในการวัดประสิทธิภาพในการวิเคราะห์ผู้ที่มีโอกาสผิดนัดชำระหนี้โดยพบว่า ensemble มีประสิทธิภาพดีกว่า neural network โดย ที่ ensemble มีค่า accuracy อยู่ที่ 80.12% ส่วน neural network อยู่ที่ 70.64% และได้นำเสนอว่าถ้าใช้ neural network กับข้อมูลที่มีปริมาณน้อยจะทำให้ความแม่นจำในการทำนายมีเปอร์เซ็นต์ที่น้อย

2.6.2 Credit Score Prediction System using Deep Learning and K-Means Algorithms

งานวิจัยชิ้นนี้เสนอโดย Ashwani Kumar, D. Shanthi, Pronaya Bhattacharya แห่ง Department of Computer Science and Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, India และ Department of Computer Science and Engineering, Institute of Technology, Nirma University, India ได้ศึกษาเกี่ยวกับการ การใช้ Deep neural network และ k-mean ในการทำนายการลูกค้าที่มีโอกาสผิดนัดชำระหนี้ โดยได้ค่า accuracy ที่ดีและสามารถใช้ทำนายลูกค้าเก่าได้อีกด้วย

2.6.4 Deep Learning Techniques for Credit Scoring

งานวิจัยชิ้นนี้เสนอโดย Le Quy Tai และ Giang Thi Thu Huyen จาก Banking Academy of Vietnam ได้ทดลองใช้ deep neural network และ Convolutional neural network ในการทำนาย credit scoring ซึ่งมีค่าเป็น good หรือ bad และได้ใช้ k-Nearest Neighbor (KNN),

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Classification and Regression Tree (CART) , Naïve Bayes (NB) และ Support Vector Machine (SVM) มาใช้เปรียบเทียบด้วย โดยใช้ข้อมูลจาก 3 แหล่งข้อมูลคือ

1. German Credit Data set โดยมีตัวอย่าง 1000 ตัวอย่าง good credit 700 ตัวอย่างและ bad credit 300 ตัวอย่าง
2. Australian Credit Approval มีตัวอย่าง 960 ตัวอย่าง
3. Kaggle Credit Data มีจำนวนตัวอย่าง 250,000 ตัวอย่าง

ได้ผลลัพธ์ดังต่อไปนี้

ตารางที่ 2.1 แสดงประสิทธิภาพของ model สำหรับข้อมูล German Credit Data set

	KNN	CART	NB	SVM	DNSS	CNN
Accuracy	67.71%	64.71%	70.57%	69.42%	75.90%,	**76.00%**
Precision	73.74%	74.97%	80.89%	70.43%	**81.00%**	79.00%
Recall	83.36%	74.64%	75.05%	**96.30%**	82.00%	80.00%
F1	78.01%	75.01%	77.77%	81.25%	**81.50%**	79.50%

ตารางที่ 2.2 แสดงประสิทธิภาพของ model สำหรับข้อมูล Australian Credit Approval

	KNN	CART	NB	SVM	DNSS	CNN
Accuracy	68.51%	83.44%	80.51%	56.53%	**87.54%**	81.86%
Precision	69.14%	81.11%	**85.43%**	15.00%	85.03%	81.84%
Recall	51.03%	82.63%	67.30%	0.98%	85.48%	**86.77%**
F1	57.92%	81.84%	74.53%	1.83%	**83.47%**	82.13%

ตารางที่ 2.3 แสดงประสิทธิภาพของ model สำหรับข้อมูล Kaggle Credit Data

	KNN	CART	NB	SVM	DSNN	CNN
Accuracy	92.87%	89.20%	92.88%	93.01%	**93.63%**	93.18%
Precision	30.31%	25.15%	35.85%	36.70%	**54.00%**	35.90%
Recall	1.72%	27.86%	2.34%	28.40%	**47.70%**	37.90%
F1	3.25%	26.44%	4.38%	32.05%	**50.10%**	32.00%

ทั้งนี้ผู้วิจัยได้สรุปว่า Deep Neural Network มีประสิทธิภาพที่ดีเมื่อเทียบกับวิธีอื่นๆ และจะยิ่งมีประสิทธิภาพดีขึ้นเมื่อข้อมูลมขนาดใหญ่ขึ้น ทั้งยังสามารถทำงานได้ดีกับข้อมูลที่ไม่สมดุล (unbalance dataset) อีกด้วย

ตารางที่ 2.4 แสดงการเปรียบเทียบงานวิจัย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ลำดับ	ชื่อเรื่อง	วัตถุประสงค์	เทคนิคที่ใช้
1	Ensemble Learning or Deep Learning Application to Default Risk Analysis	การวัดประสิทธิภาพในการวิเคราะห์ผู้ที่มีโอกาสผิดนัดชำระหนี้	- ensemble - neural network
2	Credit Score Prediction System using Deep Learning and K-Means Algorithms	การใช้ Deep neural network และ k-mean ในการทำนายการลูกค้าที่มีโอกาสผิดนัดชำระหนี้	- Deep neural network - K-mean
3	Deep Learning Techniques for Credit Scoring	ทดลองใช้ deep neural network และ Convolutional neural network ในการทำนาย credit scoring และเปรียบเทียบกับเทคนิคอื่นๆ	- Deep neural network - Convolutional neural network - K-Nearest Neighbor - CART - Naïve Bayes - Support Vector Machine

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3

วิธีการดำเนินงานวิจัย

ในบทนี้จะนำเสนอถึงวิธีการดำเนินงานวิจัยของการอนุมัติสินเชื่อลูกค้าใหม่ โดยที่การศึกษานี้คือการประยุคเอาการเรียนรู้เชิงลึกซึ่งเป็นการเรียนรู้แบบมีผู้สอน มาใช้ในการทำนายผู้ที่ผิดนัดชำระเพื่อใช้ในการตัดสินใจในการอนุมัติสินเชื่อสำหรับลูกค้าใหม่ โดยกระบวนการวิจัยได้แบ่งออกเป็น 4 ขั้นตอนหลักได้แก่ การเก็บข้อมูล การจัดเตรียมข้อมูลเพื่อการประมวลผล การสร้างตัวแบบและการวัดผลตัวแบบดังรูปที่ 3.1

รูปที่ 3.1 แสดงขั้นตอนการดำเนินงานวิจัย

3.1 ชุดข้อมูล (Dataset)

ชุดข้อมูลนี้เป็นชุดข้อมูลการนัดชำระหนี้ ประกอบไปด้วยข้อมูล 44,605 ตัวอย่างซึ่งมีคุณลักษณะทั้งหมด 25 ตัวและจะมีฉลากคำตอบที่ระบุมีการผิดนัดชำระหรือไม่ ดังที่แสดงในตารางที่ 3.1

ตารางที่ 3.1 แสดงคุณลักษณะของชุดข้อมูล

ชื่อ	ชนิดของข้อมูล	รายละเอียด
NUMBER_OF_INSTALLMENT	number	จำนวนงวด
PRODUCT_CATEGORY	text	ประเภทของสินค้า
DOWN_AND_TAX	number	ผลรวมของเงินดาวน์และภาษี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.1 แสดงคุณลักษณะของชุดข้อมูล (ต่อ)

ชื่อ	ชนิดของข้อมูล	รายละเอียด
INTEREST	number	ดอกเบี้ยตลอดอายุสัญญา
INSTALLMENT	number	ค่างวดรวม
HIRE_PURCHASE_VAT	number	ราคาหลังหักดาวน์แล้ว
DUEDATE_1_OR_16	number	วันที่ชำระเงิน ประกอบด้วย 2 ค่าคือ 1 และ 16
CURRENT_PROVINCE	text	ที่อยู่ปัจจุบัน
RESIDENCE_PROVINCE	text	ที่อยู่ตามทะเบียนบ้าน
DOCUMENT_PROVINCE	text	ที่อยู่จัดส่งเอกสาร
CURRENT_PROVINCE_OWNER_TYPE	text	ประเภทเจ้าของที่อยู่ปัจจุบัน
RESIDENCE_PROVINCE_OWNER_TYPE	text	ประเภทเจ้าของที่อยู่ตามทะเบียนบ้าน
DOCUMENT_PROVINCE_OWNER_TYPE	text	ประเภทเจ้าของที่อยู่จัดส่งเอกสาร
AGE	number	อายุ
LEVEL_TYPE	number	ระดับปริญญา
U_LEVEL	number	ระดับชั้นปี
MAIN_INCOME	number	รายได้
UNIVERSITY	text	มหาวิทยาลัย
FACULTY	text	สาขา
DOWN_PERCENT	text	ร้อยละของเงินดาวน์
GENDER	text	เพศ
RELATION_REF	text	ข้อมูลบุคคลอ้างอิง
LIVE_AT_HOME	number	อาศัยอยู่บ้านผู้ปกครอง โดย 1 คือใช่ 0 คือไม่ใช่
BUY_INSURE	text	ซื้อประกันเพิ่มหรือไม่ โดย N แทนไม่ซื้อ, Y แทนซื้อ
BUY_ACCESSORY	text	ซื้อสินค้าเพิ่มหรือไม่ โดย N แทนไม่ซื้อ, Y แทนซื้อ

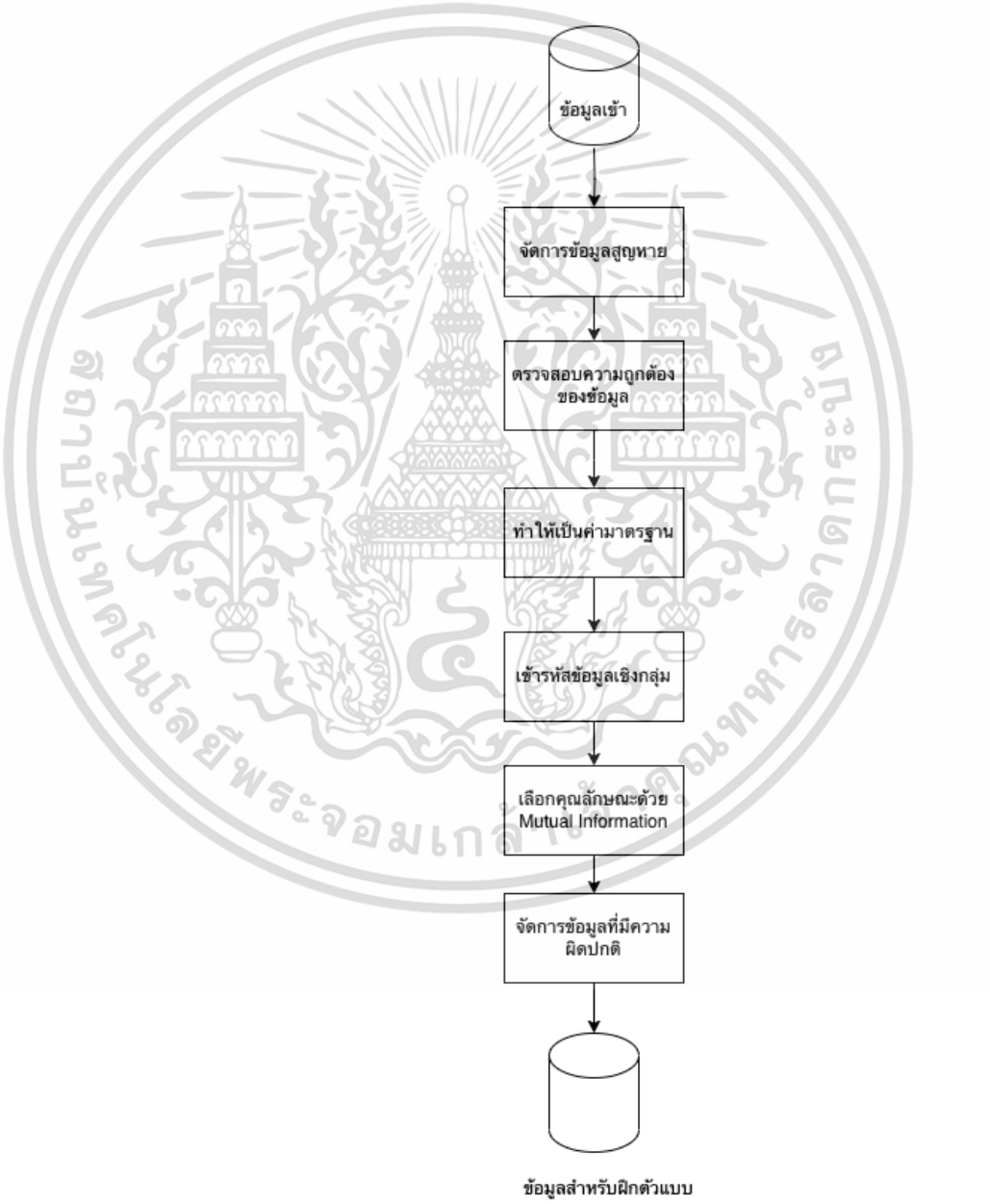
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.1 แสดงคุณลักษณะของชุดข้อมูล (ต่อ)

OVERDUE	number	ผิดนัดชำระหรือไม่ โดย 1 แทนผิดนัดชำระ , 0 แทน ไม่ผิดนัดชำระ

3.2 การจัดเตรียมข้อมูลเพื่อการประมวลผล

ในขั้นตอนนี้ผู้วิจัยได้แบ่งการดำเนินงานเป็น 5 ขั้นตอนดังรูปที่ 3.2



รูปที่ 3.2 แสดงขั้นตอนการเตรียมข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2.1 จัดการข้อมูลสูญหาย (Missing Value)

ในการวิจัยครั้งนี้พบว่าในชุดข้อมูลนี้มีในบางแถวของข้อมูลมีคุณลักษณะบางคุณลักษณะที่ข้อมูลสูญหาย (Missing Value) จึงต้องจัดการข้อมูลเหล่านี้ด้วยกระบวนการต่าง ๆ ตามลักษณะของข้อมูลโดยรายละเอียดการแก้ไขมีดังต่อไปนี้ ดังตารางที่ 3.2

ตารางที่ 3.2 แสดงวิธีการแก้ไขคุณลักษณะที่ข้อมูลสูญหาย

ตัวแปร	วิธีการ
NUMBER_OF_INSTALLMENT	แทนค่าที่ว่างด้วยค่าของค่ากลาง (median)
DUEDATE_NUM	แทนค่าว่างด้วยค่าฐานนิยม (mode)
RELATION_REF	แทนค่าว่างด้วยคำว่า “ไม่ระบุ”
LEVEL_TYPE	แทนค่าว่างโดยการพิจารณาจาก 2 ค่า คือ U_LEVELและ AGE
U_LEVEL	แทนค่าว่างโดยการพิจารณาจาก 2 ค่าคือ AGE และ LEVEL_TYPE
PRODUCT_CATEGORY	ลบแถวที่ไม่มีข้อมูล
MAIN_INCOME	ลบแถวที่ไม่มีข้อมูล
DOCUMENT_PROVINCE_OWNER_TYPE	ลบแถวที่ไม่มีข้อมูล

3.2.2 ตรวจสอบความถูกต้องของข้อมูล

การตรวจสอบความถูกต้องของข้อมูล ทำให้เกิดความมั่นใจในข้อมูลก่อนนำมาใช้วิเคราะห์ เพื่อลดความผิดพลาดที่อาจจะเกิดขึ้นจากข้อมูลที่ไม่สมบูรณ์ซึ่งอาจจะก่อให้เกิดความคลาดเคลื่อนของตัวแบบ พบว่าคุณลักษณะ AGE และ MAIN_INCOME มีความผิดปกติและได้ทำการแก้ไขดังตารางที่ 3.3

ตารางที่ 3.3 แสดงความผิดปกติของข้อมูลของแต่ละคุณลักษณะที่พบ

ตัวแปร	วิธีการ
AGE	ลบแถวที่มีข้อมูลมีค่า เกิน 35
MAIN_INCOME	ลบแถวที่มีข้อมูลมีค่า เกิน 100,000 และน้อยกว่า 0
LEVEL_TYPE	ลบแถวที่มีข้อมูล ที่มีค่าเป็น มัธยมต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2.3 การทำให้ค่าเป็นมาตรฐาน (Standardization)

คือการปรับช่วงขอบเขตของข้อมูลแต่ละคุณลักษณะให้อยู่ในช่วงเดียวกัน เพื่อให้เหมาะสมสำหรับการนำไปประมวลผลต่อ ในการศึกษานี้เลือกวิธีการทำให้เป็นมาตรฐานแซด (Z-Score Normalization) มาใช้ปรับขอบเขตของข้อมูล โดยตัวอย่างของการทำให้ค่าเป็นมาตรฐานได้ โดยตัวอย่างของการทำให้เป็นค่ามาตรฐานได้แสดงตามตารางที่ 3.4

ตารางที่ 3.4 แสดงตัวอย่างผลลัพธ์ของการทำให้ค่าเป็นมาตรฐาน

ชื่อคุณลักษณะ	ค่าต้นฉบับ	ค่าหลังจากการทำมาตรฐาน
DOWN_AND_TAX	7290.0	0.718794
INTEREST	0.499718	1.732135
MAIN_INCOME	3000.0	-1.199517

3.2.4 การเข้ารหัสข้อมูลเชิงกลุ่ม (Categorical Data Encoding)

บางคุณลักษณะในชุดข้อมูลเป็นข้อมูลเชิงคุณภาพ ก่อนที่จะนำคุณลักษณะเหล่านี้ไปผ่านการเรียนรู้ของตัวแบบ ต้องทำการเข้ารหัสก่อน ในการทดลองครั้งนี้เลือกใช้วิธี One Hot Encoding มาใช้ในการเข้ารหัส ตัวอย่าง คุณลักษณะ BUY_ACCESSORY มีข้อมูลสองกลุ่มคือ Y (ซื้อสินค้าเพิ่มเติม) และ N (ไม่ซื้อสินค้าเพิ่มเติม) ซึ่งหลังจากการทำ One hot Encoding แล้วจะได้คุณลักษณะใหม่ชื่อว่า BUY_ACCESSORY_Y ซึ่งจะมีค่าเป็น 1 กรณีมีการซื้อสินค้าเพิ่มเติมและจะมีอีกคุณลักษณะชื่อว่า BUY_ACCESSORY_N จะมีค่าเป็น 0 โดยตัวอย่างการเข้ารหัสได้แสดงในตารางที่ 3.5

ตารางที่ 3.5 แสดงตัวอย่างผลลัพธ์ของการเข้ารหัสด้วยวิธีการ One Hot Encoding ของคุณลักษณะ BUY_ACCESSORY

Feature ที่สร้างขึ้นใหม่	**ค่าของข้อมูล**
BUY_ACCESSORY_Y	1
BUY_ACCESSORY_N	0

3.2.5 เลือกคุณลักษณะ

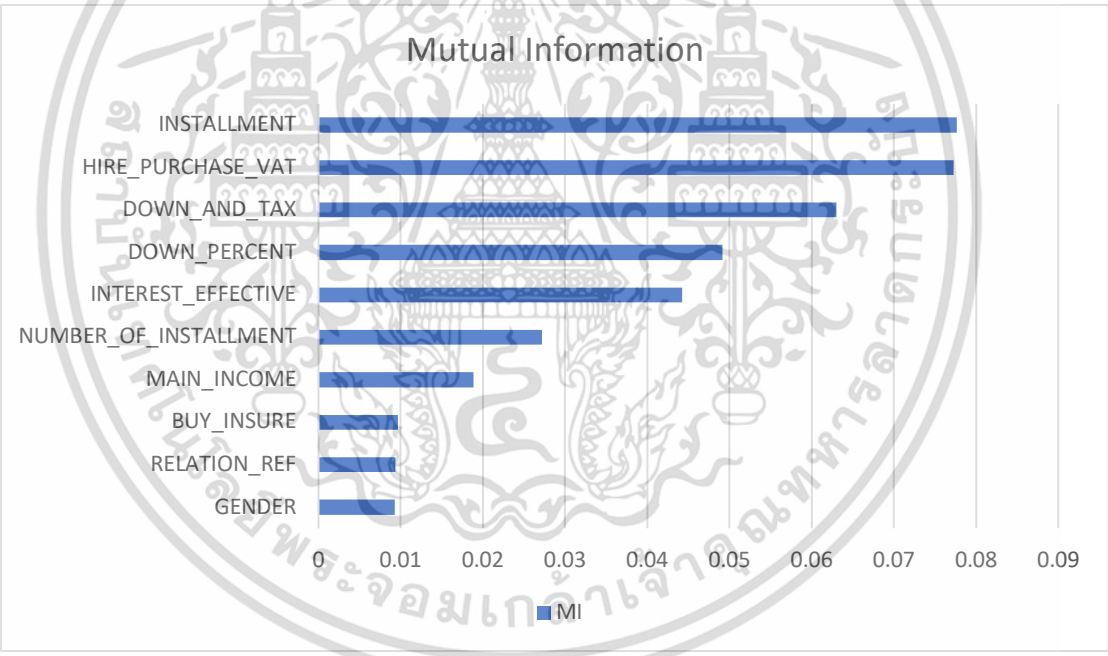
จากวิธีการเลือกคุณลักษณะทั้ง 4 วิธีที่ได้นำเสนอมา วิธีข้อมูลร่วม (Mutual information) เป็นวิธีที่ใช้เวลาน้อยที่สุด เนื่องจากเป็นการหาความสัมพันธ์ระหว่างคุณลักษณะไม่ต้องนำคุณลักษณะที่เลือกไปฝึกตัวแบบก่อนแล้วจึงนำผลที่ได้มาตัดสินใจว่าจะทำการเลือกคุณลักษณะนั้นๆ หรือไม่ เหมือนวิธีการเลือกตัวแบบอีก 3 วิธีและยังใช้ตัวแบบการเรียนรู้เชิงลึก (Deep learning) ซึ่งมีจุดด้อยในเรื่องการใช้เวลาในการฝึกตัวแบบค่อนข้างมาก ทำให้วิธีการเลือกคุณลักษณะแบบ วิธีการเลือกแบบก้าวหน้า, วิธีการเลือกแบบถอยหลังและวิธีการเลือกแบบลำดับขั้น ใช้เวลาในการคัดเลือกตัวแบบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค่อนข้างมาก ดังนั้นทางผู้วิจัยจึงขอเลือกวิธีข้อมูลร่วม (Mutual information) ใช้ในการเลือกคุณลักษณะ

Mutual Information หรือข้อมูลร่วม คือวัดการแปรผันร่วมหรือข้อมูลร่วมระหว่างสองคุณลักษณะที่สุ่มขึ้นมาถ้าค่า Mutual Information มีค่าสูงแสดงว่ามีการแปรผันร่วมระหว่างสองคุณลักษณะที่มาก นั่นคือคุณลักษณะหนึ่งสามารถสรุปข้อมูลเกี่ยวกับคุณลักษณะอื่นได้มาก ในทางกลับกัน ถ้า Mutual Information มีค่าต่ำ คุณลักษณะสองตัวนี้อาจจะไม่มีความสัมพันธ์กันมากนัก ซึ่งชุดข้อมูลนี้ประกอบไปด้วยคุณลักษณะหลายคุณลักษณะ ซึ่งบางคุณลักษณะมีผลทำให้ตัวแบบมีประสิทธิภาพดีขึ้นและบางคุณลักษณะก็อาจจะทำให้ประสิทธิภาพของตัวแบบลดลงจึงต้องมีการตัดบางคุณลักษณะออกไปโดยใช้ค่า Mutual Information ซึ่งในการทดลองครั้งนี้ใช้ค่า Mutual Information ที่มากกว่า 0.01 คือ ตั้งแต่คุณลักษณะ MAIN_INCOME เนื่องจากมีความต่างจากคุณลักษณะ BUY_INSURE ค่อนข้างมาก ดังรูปที่ 3.3

รูปที่ 3.3 แสดงค่า Mutual Information ของคุณลักษณะที่มีค่าสูงสุด 10 คุณลักษณะ



3.2.6 จัดการข้อมูลที่มีความผิดปกติ

เนื่องจากข้อมูลเข้าอนุญาตให้ผู้ใช้งานนำเข้าข้อมูลแบบอิสระและไม่ได้รับการตรวจสอบความถูกต้องที่ดีพอ จึงมีโอกาสที่มีข้อมูลที่มีความผิดปกติและจะส่งผลทำให้ประสิทธิภาพของตัวแบบลดลงได้ ทางผู้วิจัยได้ใช้ Isolation Forest มาช่วยในการหาความผิดปกติของข้อมูล โดย Isolation Forest จะหาคะแนนความผิดปกติ (Anomaly Score) ของข้อมูลหลังจากนั้นเราจะทำการลบข้อมูลที่มีความผิดปกติทิ้งไปแต่การลบข้อมูลทิ้งย่อมกระทบกับประสิทธิภาพของตัวแบบโดยเฉพาะการเรียนรู้เชิงลึก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

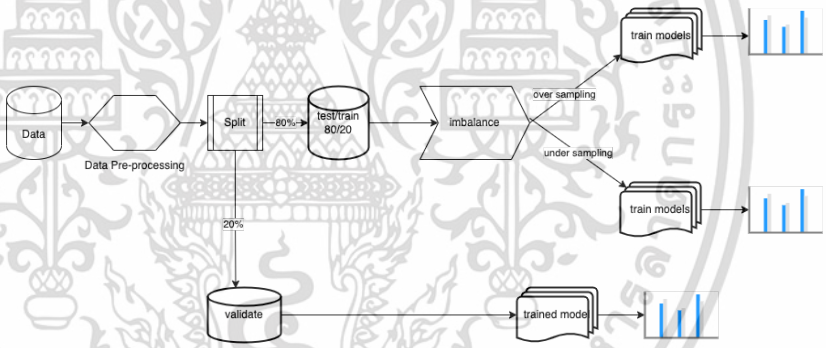
ซึ่งต้องการข้อมูลในการฝึกค่อนข้างมากฉะนั้นควรลบข้อมูลเท่าที่จำเป็นและน้อยที่สุด โดยการทดลองครั้งนี้จะลบข้อมูลที่มีความผิดปกติออกไปที่ 1% , 5% ,10% และ 15%

3.3 การสร้างตัวแบบ (Modelling)

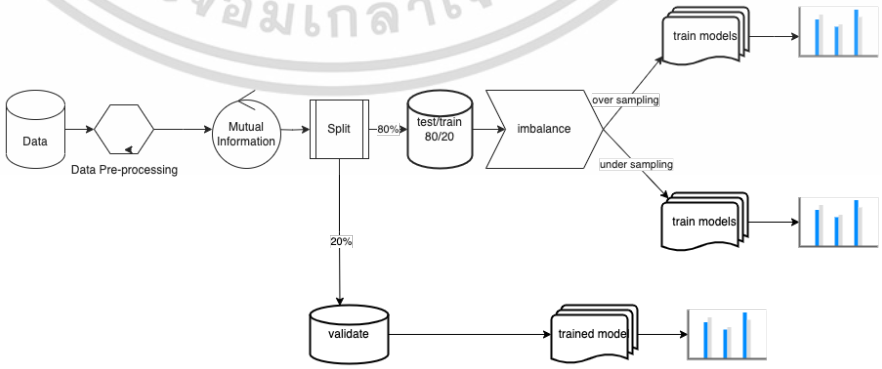
หลังจากที่ได้ทำการจัดเตรียมข้อมูลเพื่อการประมวลผล (Data Pre-processing) แล้ว หลังจากนั้นแบ่งตัวแบบออกเป็น 3 แบบคือ

- ตัวแบบการเรียนรู้เชิงลึก
- ตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะโดยใช้ Mutual Information
- ตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะโดยใช้ Mutual Information และการจัดการข้อมูลผิดปกติด้วย Isolation Forest

หลังจากนั้นแบ่งข้อมูลสำหรับฝึกและข้อมูลสำหรับทดสอบในที่นี้ได้แบ่งข้อมูลสำหรับการฝึก 80% และข้อมูลสำหรับการทดสอบ 20% จากนั้นจัดการกับข้อมูลที่มีความไม่สมดุล (Imbalance) โดยวิธีการจัดการกับข้อมูลที่ไม่สมดุลนี้ทางผู้วิจัยเลือกใช้สองวิธีคือ Oversampling กับ Undersampling หลังจากนั้นจึงนำข้อมูลไปฝึกในตัวแบบการเรียนรู้เชิงลึกแล้ววัดผล ดังรูปที่ 3.4 – 3.6

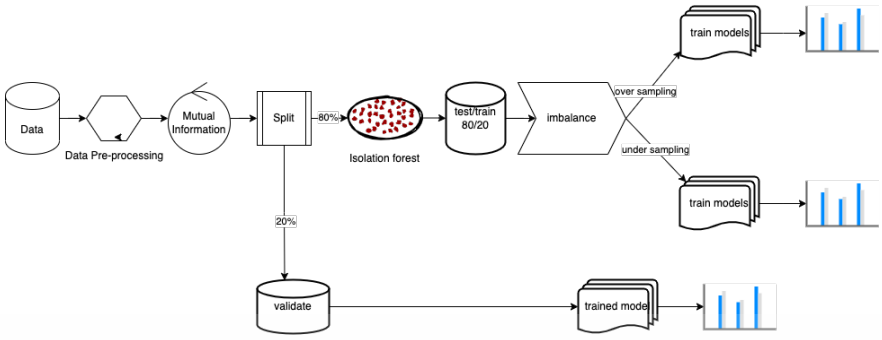


รูปที่ 3.4 แสดงตัวแบบการเรียนรู้เชิงลึก



รูปที่ 3.5 แสดงตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะโดยใช้ Mutual Information

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

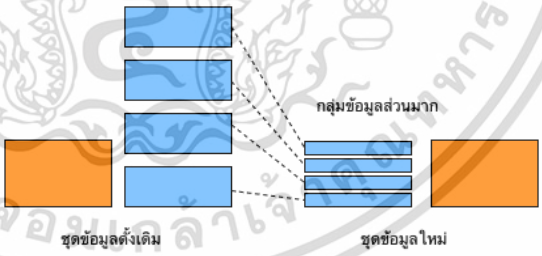


รูปที่ 3.6 แสดงตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะโดยใช้ Mutual Information และการ จัดการข้อมูลผิดปกติด้วย Isolation Forest

3.3.1 จัดการข้อมูลที่มีลักษณะไม่สมดุล (Imbalance)

การจัดการกับข้อมูลที่ไม่สมดุลนั้นมีหลากหลายวิธี ในการวิจัยครั้งนี้ จะใช้วิธีการจัดการกับการข้อมูลที่ไม่สมดุลสองวิธีดังนี้

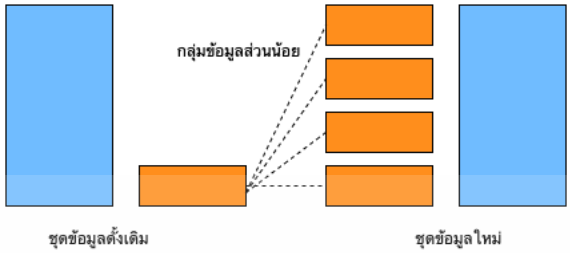
Undersampling คือกระบวนการลดข้อมูลของกลุ่มข้อมูลส่วนมาก ลดลงมาให้มีจำนวนใกล้เคียงหรือเท่ากันกับข้อมูลที่อยู่ในกลุ่มส่วนน้อยเพื่อทำให้ข้อมูลเกิดความสมดุลมากขึ้นแล้วจึงนำข้อมูลไปทำการฝึกตัวแบบ โดยตัวอย่างการทำ Undersampling ได้แสดงในรูปที่ 3.7



รูปที่ 3.7 แสดงตัวอย่างของข้อมูลก่อนและหลังการทำ Undersampling

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Oversampling คือกระบวนการเพิ่มข้อมูลของกลุ่มข้อมูลส่วนน้อยให้มีประมาณใกล้เคียงหรือเท่ากันข้อมูลที่อยู่ในกลุ่มข้อมูลส่วนมาก เพื่อทำให้ข้อมูลเกิดความสมดุลมากขึ้นแล้วจึงนำข้อมูลไปทำการฝึกตัวแบบ โดยตัวอย่างการทำ Oversampling ได้แสดงในรูปที่ 3.8



รูปที่ 3.8 แสดงตัวอย่างของข้อมูลก่อนแหละหลังการทำ Oversampling

3.3.2 แบ่งข้อมูลสำหรับฝึกและทดสอบ

แบ่งข้อมูลสำหรับฝึกและทดสอบได้แบ่งออกเป็น 2 ส่วนคือชุดข้อมูลสำหรับการฝึกและทดสอบจำนวน 80% และชุดข้อมูลสำหรับการตรวจสอบ (validation set) จำนวน 20% ซึ่งชุดข้อมูลสำหรับการฝึกและทดสอบจะแบ่งย่อยเป็น 2 ส่วนคือข้อมูลสำหรับฝึก (training set) 80% และข้อมูลสำหรับการทดสอบ (testing set) จำนวน 20%

3.3.3 การสร้างตัวแบบการเรียนรู้เชิงลึก (Deep Learning Model)

ในการสร้างตัวแบบของการเรียนรู้เชิงลึกประกอบไปด้วยชั้นข้อมูลเข้า ชั้นข้อมูลออกและชั้นซ่อน ฟังก์ชั้นกระตุ้นใช้สองฟังก์ชั่นประกอบด้วย relu และ sigmoid function สำหรับฟังก์ชันการสูญเสีย (loss function) ใช้ค่าเฉลี่ยครอสเอนโทปีแบบทวิภาค (Binary Cross-entropy) และ optimizer เป็น Adam ในส่วนของชั้นซ่อนนั้นสามารถระบุได้ว่าต้องมีจำนวนใด ดังนั้นการทดลองนี้จึงได้ทำการทดลองแบ่งจำนวนชั้นซ่อนออกเป็น 3 แบบดังตารางที่ 3.6 – 3.8

ตารางที่ 3.6 แสดงโครงสร้างของตัวเข้ารหัสอัตโนมัติที่มีชั้นซ่อน 3 ชั้น

ลำดับที่	จำนวนนิวรอน	ฟังก์ชั่นกระตุ้น	ประเภทชั้น
1	6	relu	ชั้นข้อมูลเข้า
2	8	relu	ชั้นซ่อน
3	4	relu	ชั้นซ่อน
4	2	relu	ชั้นซ่อน
5	1	sigmoid	ชั้นข้อมูลออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.7 แสดงโครงสร้างของตัวเข้ารหัสอัตโนมัติที่มีชั้นซ่อน 5 ชั้น

ลำดับที่	จำนวนนิวรอน	ฟังก์ชั่นกระตุ้น	ประเภทชั้น
1	6	relu	ชั้นข้อมูลเข้า
2	32	relu	ชั้นซ่อน
3	16	relu	ชั้นซ่อน
4	8	relu	ชั้นซ่อน
5	4	relu	ชั้นซ่อน
6	2	relu	ชั้นซ่อน
7	1	sigmoid	ชั้นข้อมูลออก

ตารางที่ 3.8 แสดงโครงสร้างของตัวเข้ารหัสอัตโนมัติที่มีชั้นซ่อน 7 ชั้น

ลำดับที่	จำนวนนิวรอน	ฟังก์ชั่นกระตุ้น	ประเภทชั้น
1	6	relu	ชั้นข้อมูลเข้า
2	128	relu	ชั้นซ่อน
3	64	relu	ชั้นซ่อน
4	32	relu	ชั้นซ่อน
5	16	relu	ชั้นซ่อน
6	8	relu	ชั้นซ่อน
7	4	relu	ชั้นซ่อน
8	2	relu	ชั้นซ่อน
9	1	sigmoid	ชั้นข้อมูลออก

โดยจะกำหนด Hyperparameters ที่ใช้สำหรับการเปรียบเทียบโครงสร้างของตัวแบบดังตารางที่ 3.9 ซึ่งค่าของ Hyperparameter แต่ละประเภทที่ถูกเลือกมานั้นมีประสิทธิภาพเพียงพอสำหรับการฝึกตัวแบบการเรียนรู้เชิงลึก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.9 แสดงค่า Hyperparameters ที่ใช้ในการหาโครงสร้างของตัวแบบ

ชื่อ	ค่าที่กำหนด
Epoch	100
Batch size	64
Optimizer	Adam
Loss function	Binary crossentropy
Learning rate	0.00001

3.3.4 หาค่า Hyperparameters ที่เหมาะสม

Hyperparameters ที่ต้องทดลองหาค่าที่เหมาะสมและเกี่ยวข้องกับการฝึกตัวแบบมีดังนี้

1. Epoch เป็นจำนวนรอบของการฝึกสอน
2. Batch size เป็นจำนวนชุดข้อมูลที่จะให้ Optimizer คำนวณใน 1 ครั้ง
3. Loss function เป็นฟังก์ชันสำหรับคำนวณค่าการสูญเสีย
4. Learning rate เป็นค่าที่ใช้กำหนดว่าการปรับค่าน้ำหนักในแต่ละรอบการฝึก

3.4 การวัดผลตัวแบบ (Model Evaluation)

ในการทดลองครั้งนี้ได้แบ่งข้อมูลสำหรับฝึก 80% และสำหรับการทดสอบอีก 20% โดยการวัดประสิทธิภาพของตัวแบบจะใช้ค่าความแม่น ความเที่ยง และค่าเรียกคืน ผู้วิจัยต้องการให้มีค่าผลบวกเท็จ (False Positive) ที่ต่ำ จึงขอเน้นประสิทธิภาพของค่าเรียกคืน (Recall) เป็นหลัก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

ผลการวิจัยและการอภิปรายผล

งานวิจัยเรื่องนี้มีวัตถุประสงค์เพื่อศึกษาการอนุมัติสินเชื่อลูกค้าใหม่โดยใช้การเรียนรู้ของเครื่อง โดยทางผู้วิจัยได้นำเทคนิคการเรียนรู้เชิงลึกมาใช้ในการวิจัยครั้งนี้ ซึ่งในบทนี้จะกล่าวถึงโครงสร้างของตัวแบบการเรียนรู้เชิงลึกผลของการศึกษาและผลของการทดสอบประสิทธิภาพของตัวแบบและปัญหาที่พบจากการศึกษา

4.1 กำหนดโครงสร้างของการเรียนรู้เชิงลึก

จากการทดลองหาค่าโครงสร้างของตัวแบบการเรียนรู้เชิงลึกเพื่อที่จะได้โครงสร้างที่มีประสิทธิภาพสูงสุดในการจำแนกข้อมูล การทดลองแบ่งเป็นสองส่วนคือ ส่วนของการหาโครงสร้างการเรียนรู้เชิงลึกและส่วนของการหาค่า Hyperparameters

4.1.1 การหาโครงสร้างของการเรียนรู้เชิงลึก

จากการทดลองหาจำนวนชั้นซ่อนที่ทำให้ประสิทธิภาพของตัวแบบมีประสิทธิภาพสูงสุดพบว่า จากการทดลองใช้ชั้นซ่อน 3 ,5 และ 7 ชั้น พบว่า โครงสร้างการเรียนรู้เชิงลึกที่มีชั้นซ่อน 5 ชั้น มีประสิทธิภาพสูงสุดดังตารางที่ 4.1

ตารางที่ 4.1 แสดงค่าประสิทธิภาพของตัวแบบของตัวแบบการเรียนรู้เชิงลึกของแต่ละชั้นซ่อน

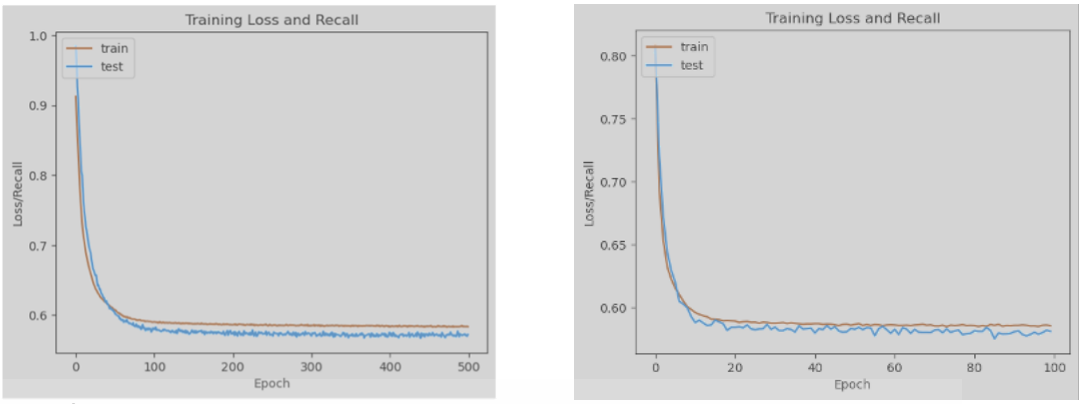
ชั้นซ่อน	Recall	Accuracy	Precision
3	69%	69%	64%
5	70%	70%	64%
7	68%	69%	63%

4.1.2 การปรับค่า Hyperparameters

1. Epoch

   จากการทดลองด้วยการฝึก 500 รอบพบกว่าค่าสูญเสียลดลงตามรูปที่ 4.1 ดูจากแนวโน้มพบว่าค่าสูญเสียไม่ได้ปรับตัวลดลงอย่างมีนัยยะสำคัญผู้วิจัยจึงทำการเลือกค่า 100 รอบเพื่อความเร็วในการฝึกตัวแบบ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



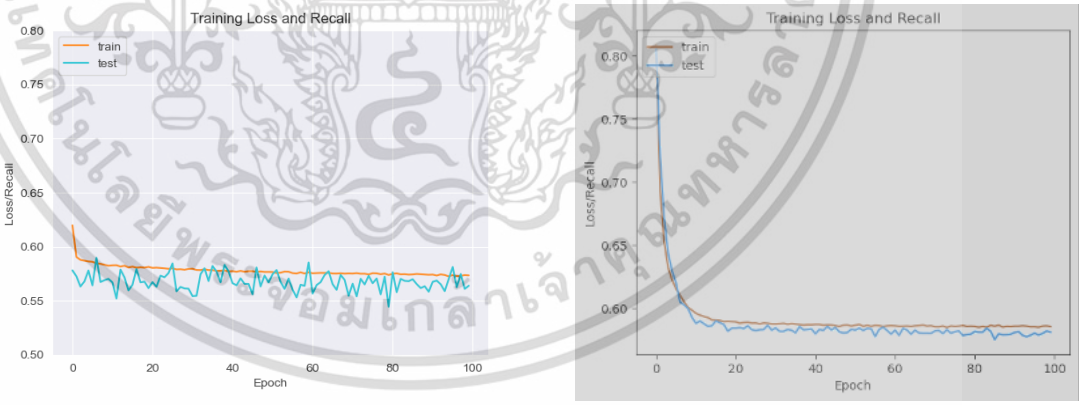
รูปที่ 4.1 แสดงค่าสูญเสียของการฝึก 0 - 500 รอบ เปรียบเทียบกับ 0 - 100 รอบ

2. Batch size และ Loss function

   จากการทดลองพบว่า ค่า Batch size ไม่ได้มีผลต่อการฝึกตัวแบบอย่างมีนัยสำคัญ เพียงแต่ช่วยลดระยะเวลาในการฝึกสอนให้น้อยลงและเนื่องจากตัวแปรตาม Y มีลักษณะวิภาค(Binary) ส่วนของ Loss function จึงต้องใช้ค่าเฉลี่ยครอสเอนโทปีแบบทวิภาค (Binary Cross-entropy)

3. Learning rate

   จากการทดลองพบว่า เมื่อใช้ค่า 0.001 และ 0.00001 พบว่า ค่า 0.00001 ค่ามีความแปรปรวนน้อยกว่า ดังภาพที่ 4.2



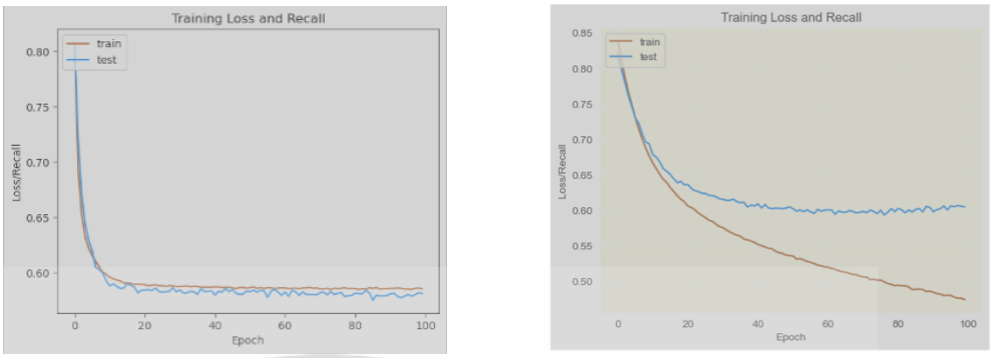
รูปที่ 4.2 แสดงค่าการสูญเสียของการฝึกระหว่าง Learning rate 0.001 (รูปซ้าย) และ 0.00001 (รูปขวา)

4.2 เลือกคุณลักษณะด้วย Mutual Information

ในการวิจัยครั้งนี้ได้ใช้ Mutual Information มาช่วยในการเลือกคุณลักษณะก่อนทำการฝึกตัวแบบโดยเลือกเฉพาะคุณลักษณะที่มีคะแนนมากกว่า 0.01 มาใช้ในการฝึกตัวแบบ หลังจากการตัด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวแบบที่เหลือออกไปแล้วพบว่าการฝึกตัวแบบลู่เข้าได้เร็วและมีตัวแบบมีประสิทธิภาพดีขึ้น ดังรูปที่ 4.3 และมีประสิทธิภาพของค่าเรียกคืนมากขึ้นจาก 60% เป็น 70% ดังตารางที่ 4.2



รูปที่ 4.3 แสดงค่าการสูญเสียของการฝึกระหว่างการใช้ Mutual Information (รูปซ้าย) และ ไม่ใช้ Mutual Information (รูปขวา)

ตารางที่ 4.2 เปรียบเทียบประสิทธิภาพของตัวแบบในการใช้ Mutual Information

ใช้ Mutual Information	Recall	Accuracy	Precision
ใช้	69%	66%	63%
ไม่ใช้	60%	69%	62%

4.3 จัดการข้อมูลที่มีความผิดปกติ

จากการทดลองครั้งนี้ได้นำ Isolation Forest มาช่วยหาคะแนนความผิดปกติ (Anomaly score) ของข้อมูล หลังจากนั้นทำการลบข้อมูลแถวที่มีคะแนนความผิดปกติสูงสุด โดยการทดลองครั้งนี้เลือกทางผู้วิจัยเลือกลบข้อมูลที่มีคะแนนความผิดปกติสูงสุด 1% , 5% , 10% และ 15% หลังจากการลบข้อมูลที่มีความผิดปกติออกแล้วนำข้อมูลไปฝึกตัวแบบพบว่า การลบข้อมูลที่มีคะแนนความผิดปกติออก 10% ได้ผลดีที่สุด ดังตารางที่ 4.3 ทั้งนี้ผู้วิจัยใช้ Isolation Forest ร่วมกับ Mutual Information

ตารางที่ 4.3 แสดงประสิทธิภาพของตัวแบบจากการลบข้อมูลแต่ละแบบ

Percent	Recall	Accuracy	Precision
1%	70	68	64
5%	70	68	64
10%	70	70	64
15%	69	70	64

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.4 ผลการศึกษาและการวัดประสิทธิภาพ

จากการสร้างตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะด้วย Mutual Information และจัดการข้อมูลผิดปกติด้วย Isolation forest ได้ประสิทธิภาพโดยรวมสูงที่สุด โดยที่ได้ค่าความแม่น(Accuracy) 70% ค่าความเที่ยง (Precision) 64% และค่าเรียกคืน (Recall) 70% ดังตารางที่ 4.4 จากประสิทธิภาพของตัวแบบที่ได้แสดงให้เห็นว่า ตัวแบบที่เราได้สามารถทำนายภาพรวมได้ถูกต้อง 70% กรณีที่มีลูกค้าที่ผิดนัดชำระ 100 คนตัวแบบสามารถทำนายว่าเป็นลูกค้าผิดนัดชำระได้ถึง 70 คน

ตารางที่ 4.4 เปรียบเทียบประสิทธิภาพของวิธีการเรียนรู้เชิงลึกรวมกับ Mutual Information (MI) และ Isolation forest

Model	Recall	Accuracy	Precision
Deep learning	60%	69%	62%
Deep learning + MI	69%	66%	63%
Deep learning + MI + Isolation forest	70%	70%	64%

4.5 ปัญหาที่พบจากการศึกษา

จากการทดลองพบว่า การทดลองช่วงแรกประสิทธิภาพของตัวแบบที่ใช้วิธีการเรียนรู้เชิงลึกมีประสิทธิภาพด้อยกว่าวิธีอื่น ๆ ที่ใช้ในการเปรียบเทียบ ดังตารางที่ 4.5 โดยเฉพาะค่า Recall ซึ่งผู้วิจัยต้องการให้ค่า Recall มีค่าสูง ดังนั้นผู้วิจัยจึงได้นำเทคนิคอื่น ๆ มาช่วยในการจัดการกับข้อมูลเพื่อให้ตัวแบบมีประสิทธิภาพเพิ่มมากขึ้น ณ ที่นี้ได้นำ Isolation forest มาใช้แยกข้อมูลที่ผิดปกติโดย Isolation forest สามารถหาคะแนนความผิดปกติ (Anomaly score) ของข้อมูลได้ หลังจากนั้นจึงทำการลบข้อมูลที่มีคะแนนความผิดปกติไปทั้งสิ้น 10% และใช้ Mutual Information มาใช้เลือกคุณลักษณะโดยการเลือกเฉพาะกลุ่มคุณลักษณะที่มีค่า Mutual Information สูงแล้วจึงนำไปฝึกตัวแบบพบว่าประสิทธิภาพของตัวแบบดีขึ้นดังตารางที่ 4.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพของการอนุมัติสินเชื่อลูกค้าใหม่โดยใช้วิธีการเรียนรู้เชิงลึก โดยใช้ข้อมูล 44,605 ตัวอย่าง หลังจากการจัดเตรียมข้อมูลเพื่อการประมวลผลและนำข้อมูลไปฝึกตัวแบบ สามารถสรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1.1 จากการใช้ตัวแบบการเรียนรู้เชิงลึกเพียงอย่างเดียว พบว่าประสิทธิภาพโดยรวมต่ำที่สุด โดยมีค่าความแม่นเท่ากับ 69% ค่าความเที่ยงเท่ากับ 62% และค่าเรียกคืนเท่ากับ 60%

5.1.2 จากการใช้ตัวแบบการเรียนรู้เชิงลึกและใช้ Mutual information ในการช่วยเลือกคุณลักษณะพบว่าประสิทธิภาพของค่าเรียกคืนดีกว่าการใช้ตัวแบบการเรียนรู้เชิงลึกเพียงอย่างเดียว เนื่องจากได้เลือกเฉพาะคุณลักษณะที่มีผลที่ดีกับการฝึกตัวแบบ โดยมีค่าความแม่นเท่ากับ 66% ค่าความเที่ยงเท่ากับ 63% และค่าเรียกคืนเท่ากับ 69%

5.1.3 จากการใช้ตัวแบบการเรียนรู้เชิงลึกร่วมกับ Mutual information และใช้ Isolation forest มาช่วยในการกำจัดข้อมูลที่มีความผิดปกติพบว่า ประสิทธิภาพโดยรวมดีที่สุด แบบ โดยมีค่าความแม่นเท่ากับ 70% ค่าความเที่ยงเท่ากับ 64% และค่าเรียกคืนเท่ากับ 70% เนื่องจากได้มีการจัดการข้อมูลที่มีความผิดปกติออกไปก่อนการฝึกตัวแบบ ทำให้ประสิทธิภาพในการทำนายดีขึ้น

จากการศึกษาครั้งนี้พบว่าประสิทธิภาพของตัวแบบการเรียนรู้เชิงลึกร่วมกับการเลือกคุณลักษณะด้วย Mutual information และจัดการข้อมูลที่ผิดปกติด้วย Isolation forest ได้ตัวแบบที่มีประสิทธิภาพที่ดี โดยที่ประสิทธิภาพค่า Recall อยู่ที่ 70% ค่า Accuracy อยู่ที่ 70% และค่า Precision อยู่ที่ 64% ตัวแบบที่ได้สามารถทำนายภาพรวมได้ถูกต้อง 70% กรณีที่มีลูกค้าที่ผิดนัดชำระ 100 คนตัวแบบสามารถทำนายว่าเป็นลูกค้าผิดนัดชำระได้ถึง 70 คน

5.2 ข้อเสนอแนะ

1. การกำจัดข้อมูลที่มีความผิดปกติสามารถช่วยให้ตัวแบบมีประสิทธิภาพที่ดีขึ้นได้

2. การปรับ Hyperparameters ที่ละเอียดช่วยให้ตัวแบบมีประสิทธิภาพที่ดีขึ้นได้ แต่อาจจะใช้เวลาในการปรับและการทดสอบเพิ่มมากขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

[4] Jian Huang , Junyi Chai, Stella Cho. 2020 **. Deep learning in finance and banking: A literature review and classification** [Online]. Available doi: 10.1186/s11782-020-00082-6

[5] Varakul Kanchanakanho, Sirapat Chiewchanwattana, Khamron Sunat. 2019. **Selection of †the Attributes Describing Artificial NeuralNetwork Models for Violent Crimes Forecasting** [Online]. Available https://ph02.tci-thaijo.org/index.php/gskku/article/view/22516/19279

[10] F. Bre, J. M. Gimenez, and V. D. Fachinotti. 2018, **Prediction of wind pressure coefficients on building surfaces using artificial neural networks**, Energy and Buildings, vol. 158, pp. 1429-1441, Jan 2018, doi: 10.1016/j.enbuild.2017.11.045

[6] Fei Tony Liu, Kai Ming Ting, **Isolation Forest,** [Online] Available https://cs.nju.edu.cn/zhouzh/ zhouzh.files/publication/icdm08b.pdf

[7] อาณัตชัย เตชะวิเศษชัย. 2564. **Anomaly Detection with Isolation Forest: แยกข้อมูลผิดปกติง่าย ๆ** ด้วย Isolation Forest. [Online]. https://bigdataexperience.org.

[9] Daniel B. Russakof, Carlo Tomasi, Torsten Rohlfing, Calvin R. Maurer Jr. 2004. **Image Similarity Using Mutual Information of Regions** [Online]. Available https://link.springer.com/content/pdf/10.1007/978-3-540-24672-5_47.pdf

[10] รัฐพงศ์ ชัยเอิก, ปริม ชูคากร, วรรณพร จันโทภาส. 2560. **วิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ.** [Online] http://sc2.kku.ac.th/stat/statweb/images/Eventpic/60/Seminar/01_11_.pdf

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นายจิรายุ โพธิสาร
วัน เดือน ปีเกิด	05 เมษายน 2531
ที่อยู่ปัจจุบัน	42/182 บ้านกลางเองรัตนาธิเบศร์ ต.บางกระสอ อ.เมืองนนทบุรี จ.นนทบุรี 11000
ประวัติการศึกษา	(2554) วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ เกรดเฉลี่ย 3.14 (มหาวิทยาลัยแม่โจ้)
ทุนการศึกษาที่ได้รับ	ไม่มี
ผลงานทางวิชาการ	ไม่มี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้