การทำนายและวิเคราะห์การอนุมัติสินเชื่อส่วนบุคคล

PREDICTION AND ANALYSIS OF PERSONAL CREDIT APPROVAL

พรรษมน มะโนมัย

PASSAMON MANOMAI

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2567

KMITL-2024-SC-M-017-032

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

PREDICTION AND ANALYSIS OF PERSONAL CREDIT APPROVAL

PASSAMON MANOMAI

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR
THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS
KMITL DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2024

KMITL-2024-SC-M-017-032

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2024
SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อการค้นคว้าอิสระ การทำนายและวิเคราะห์การอนุมัติสินเชื่อส่วนบุคคล
ชื่อนักศึกษา นางสาวพรรษมน มะโนมัย
รหัสประจำตัว 65056065
ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์)
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ. 2567
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ รองศาสตราจารย์ ดร.ฉัฐไชย์ ลีนาวงศ์

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายการอนุมัติสินเชื่อส่วนบุคคล รวมถึงศึกษาระดับความสำคัญของปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อ โดยใช้ข้อมูลจากแหล่งข้อมูลแคกเกิล (Kaggle) เป็นชุดข้อมูลการอนุมัติเงินกู้ซึ่งรวบรวมบันทึกทางการเงินและข้อมูลที่เกี่ยวข้องที่ใช้เพื่อกำหนดคุณสมบัติของบุคคล เลือกใช้แบบจำลองการเรียนรู้ของเครื่อง 3 วิธี ได้แก่ วิธีป่าสุ่ม (Random Forest), วิธีต้นไม้ตัดสินใจ (Decision Tree) และการวิเคราะห์การถดถอยลอจิสติก (Logistic Regression) การวิจัยเริ่มจากการจัดเตรียมข้อมูลและเพิ่มข้อมูลให้สมดุล จากนั้นสร้างแบบจำลองและเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง เกณฑ์ที่ใช้ในการพิจารณาคือ ค่าความถูกต้อง ค่าความแม่นยำ และค่าความไว นอกจากนี้ได้แสดงค่าฟีเจอร์ อิมพอร์แทนเซส (Feature Importances) และค่าเพอร์มิวเทชัน อิมพอร์แทนซ์ (Permutation Importance) เพื่อศึกษาปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อ ผลการวิจัยสรุปได้ว่าแบบจำลองวิธีป่าสุ่มเป็นแบบจำลองที่มีประสิทธิภาพที่สุดในการทำนายการอนุมัติสินเชื่อ สำหรับปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อมากที่สุด 3 อันดับแรกคือ คะแนนเครดิต ระยะเวลาเงินกู้ และอัตราส่วนหนี้สิน

คำสำคัญ : การเรียนรู้ของเครื่อง การอนุมัติสินเชื่อส่วนบุคคล คะแนนเครดิต ระยะเวลาเงินกู้ อัตราส่วนหนี้สิน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Independent Study Title	Prediction and Analysis of Personal Credit Approval
Student Name	Passamon Manomai
Student ID	65056065
Degree	Master of Science (Data Science and Analytics)
	KMITL Digital Analytics and Intelligence Center
Year	2024
Independent Study Advisor	Assoc. Prof. Dr. Chartchai Leenawong

Abstract

The objective of this research was to study and compare the performance of machine learning models in predicting personal loan approval. It also examines the significance of factors influencing loan approval using data from Kaggle's loan approval dataset, which includes financial records and relevant information used to assess individual eligibility. Three machine learning methods are employed: Random Forest, Decision Tree, and Logistic Regression. The research begins with data preparation and balancing, followed by model construction and performance comparison. The evaluation criteria include accuracy, precision, and recall. Feature importances and Permutation Importance are also presented to investigate the factors affecting loan approval. The findings indicate that the Random Forest model is the most effective in predicting loan approval. The top three factors influencing loan approval are credit score, loan duration, and debt-to-income ratio.

Keywords: Machine learning, Personal loan approval, Credit score, Loan duration, Debt-to-income ratio

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

งานวิจัยเรื่อง การทำนายการอนุมัติสินเชื่อส่วนบุคคลด้วยวิธีการเรียนรู้ของเครื่องสามารถดำเนินการจนประสบความสำเร็จลุล่วงไปด้วยดี เนื่องจากได้รับความกรุณาจาก รศ.ดร.ฉัฐไชย์ ลีนาวงศ์ อาจารย์ที่ปรึกษาการค้นคว้าอิสระที่ให้คำปรึกษา ข้อเสนอแนะ เอื้อเฟื้อเอกสารต่างๆที่ใช้เป็นแนวทางในการวิเคราะห์ข้อมูล รวมถึงตรวจทานแก้ไขความถูกต้องตลอดจนติดตามงานทุกๆขั้นตอนของการดำเนินงาน จนกระทั่งงานวิจัยครั้งนี้สำเร็จเรียบร้อยด้วยดี ผู้วิจัยขอขอบพระคุณด้วยความเคารพอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ ผศ.ดร.พรรณทิพา วาณิชย์จิรัฐติกาล และดร.กิตติมศักดิ์ ในจิต คณะกรรมการการค้นคว้าอิสระที่ให้คำปรึกษา และคำแนะนำเพื่อความสมบูรณ์ยิ่งขึ้น

ขอขอบคุณคณาจารย์สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และช่วยเหลือให้คำแนะนำในเรื่องต่างๆมาโดยตลอด

ขอขอบพระคุณบิดามารดา ที่คอยให้กำลังใจและสนับสนุนผู้จัดทำการค้นคว้าอิสระมาโดยตลอด และขอขอบคุณเพื่อนคณะวิทยาศาสตร์สาขาวิทยาการข้อมูลและการวิเคราะห์ทุกท่านที่ให้ความรู้และช่วยเหลือในการวิจัย

ขอขอบคุณข้อมูลจากแคกเกิล (Kaggle) แหล่งข้อมูลคุณภาพที่เผยแพร่ข้อมูลสาธารณะที่ดีและมีประโยชน์จนทำให้งานวิจัยนี้สำเร็จ

สุดท้ายนี้ผู้วิจัยหวังว่างานวิจัยฉบับนี้จะเป็นประโยชน์สำหรับหน่วยงานที่เกี่ยวข้องและผู้สนใจศึกษาต่อไป

นางสาวพรรษมน มะโนมัย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทคัดย่อ	ก
Abstract	ข
กิตติกรรมประกาศ	ค
สารบัญ	ง
สารบัญตาราง	ฉ
สารบัญรูป	ช
บทที่ 1 บทนำ	**1**
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของงานวิจัย	2
1.3 ขอบเขตของงานวิจัย	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	**3**
2.1 แนวคิดของสินเชื่อส่วนบุคคล	3
2.2 การจัดเตรียมข้อมูล (Data Preprocessing)	4
2.1.1 การเพิ่มข้อมูล (Data Augmentation)	4
2.1.2 การแปลงข้อมูล (Data Transformation)	5
2.1.3 สัมประสิทธิ์สหสัมพันธ์สเปียร์แมน (Spearman correlation coefficient)	5
2.2 การเรียนรู้ของเครื่อง (Machine Learning)	5
2.2.1 เทคนิคต้นไม้ตัดสินใจ	6
2.2.2 เทคนิควิธีป่าสุ่ม	6
2.2.3 การถดถอยลอจิสติก	7
2.3 การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation)	9
2.3.1 ค่าความถูกต้อง (Accuracy)	10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
2.3.2 ค่าความแม่นยำ (Precision)	10
2.3.3 ค่าความระลึก (Recall)	10
2.4 การหาความสัมพันธ์ของตัวแปรที่ส่งผลต่อการทำนายของแบบจำลอง	11
2.4.1 ค่าฟีเจอร์ อิมพอร์แทนเซส (Feature Importances)	11
2.4.2 ค่าเพอร์มิวเทชัน อิมพอร์แทนซ์ (Permutation Importance)	11
2.5 งานวิจัยที่เกี่ยวข้อง	11
บทที่ 3 วิธีการดำเนินงานวิจัย	**14**
3.1 แผนผังขั้นตอนการทำงาน	14
3.2 การจัดเตรียมข้อมูล (Data Preprocessing)	15
3.2.1 การเก็บรวบรวมข้อมูล (Data Collection)	15
3.1.2 การจัดการกับความไม่สมดุลของข้อมูล (Imbalanced data)	16
3.1.3 การหาความสัมพันธ์ระหว่างข้อมูล (Correlation)	17
3.2 การสร้างแบบจำลอง	18
3.3 การเปรียบเทียบประสิทธิภาพแบบจำลอง	18
3.4 การหาความสัมพันธ์ของตัวแปรที่ส่งผลต่อการทำนายของแบบจำลอง	18
บทที่ 4 ผลการวิจัยและการอภิปรายผล	**19**
4.1 ผลการวิเคราะห์ของการเรียนรู้ของเครื่อง	19
4.2 ผลการวิเคราะห์ของการหาความสำคัญของตัวแปร	22
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ	**24**
5.1 สรุปผลการวิจัย	24
5.2 ข้อเสนอแนะ	25
เอกสารอ้างอิง	26
ประวัติผู้เขียน	28

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

หน้า

ตารางที่ 2.1 ตารางเมทริกซ์ความสับสน 9

ตารางที่ 3.1 ตัวแปร ชื่อหัวข้อ คำอธิบาย และประเภทของข้อมูล 15

ตารางที่ 4.1 ตัวอย่างผลการทำนายของแบบจำลอง 20

ตารางที่ 4.2 ผลการทำนายการอนุมัติสินเชื่อส่วนบุคคลของแบบจำลอง 21

ตารางที่ 4.3 เมทริกซ์ความสับสนของแบบจำลองวิธีป่าสุ่ม 21

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

หน้า

รูปที่ 2.1 หลักการทำงานวิธีต้นไม้ตัดสินใจ 6
รูปที่ 2.2 หลักการทำงานวิธีป่าสุ่ม 7
รูปที่ 2.3 ฟังก์ชันลอจิสติกมาตรฐาน 8
รูปที่ 2.4 เส้นโค้งการถดถอยลอจิสติก 9
รูปที่ 3.1 แผนผังขั้นตอนการทำงาน 14
รูปที่ 3.2 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ลำดับที่สเปียร์แมน 17
รูปที่ 4.1 ค่าความสำคัญของตัวแปรที่แสดงด้วยฟีเจอร์ อิมพอร์แทนเซส 22
รูปที่ 4.2 ค่าความสำคัญของตัวแปรที่แสดงด้วยเพอร์มิวเทชัน อิมพอร์แทนซ์ 22

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

สังคมปัจจุบันต้องเผชิญกับความท้าทายทางเศรษฐกิจ เนื่องจากวิกฤตการณ์โควิด-19 ได้ส่งผลกระทบอย่างหนักต่อระบบเศรษฐกิจทำให้เศรษฐกิจทั่วโลกต้องหยุดชะงักและในปัจจุบันสภาพเศรษฐกิจยังไม่ได้กลับสู่ภาวะปกติอย่างเต็มที่ส่งผลให้มีผู้คนจำนวนมากได้รับผลกระทบ ไม่ว่าจะเป็นกลุ่มคนที่มีรายได้ไม่เพียงพอกับภาระค่าใช้จ่ายภายในครอบครัว กลุ่มคนที่เกิดเหตุฉุกเฉินต้องใช้เงินในการจ่ายค่ารักษาพยาบาล หรือต้องการนำเงินไปพัฒนาตนเองเพิ่มโอกาสทางการศึกษา ด้วยสาเหตุต่างๆที่กล่าวมานี้ล้วนเป็นสาเหตุที่ทำให้เกิดการตัดสินใจใช้บริการสินเชื่อส่วนบุคคลกับธนาคารเพื่อเพิ่มความสมดุลและสภาพคล่องในการประกอบชีวิต

ธนาคารเป็นหน่วยงานที่ทำหน้าที่รับฝากเงินและเป็นแหล่งให้เงินกู้ยืมหรือกล่าวได้ว่าธนาคารเป็นตัวกลางที่รับเงินออมจากผู้ที่ต้องการออมเงินมาให้กู้กับผู้ที่ต้องการกู้ยืม อย่างไรก็ตามการฝากเงินกับธนาคารทางผู้ออมเงินจะได้ผลตอบแทนเป็นดอกเบี้ยเงินฝาก ส่วนผู้ที่มากู้ยืมเงินกับธนาคารจะต้องมีการเสียดอกเบี้ยให้กับธนาคารเรียกว่าดอกเบี้ยเงินกู้ ซึ่งในส่วนนี้ส่งผลให้เกิดส่วนต่างระหว่างดอกเบี้ยเงินฝากกับดอกเบี้ยเงินกู้ ดังนั้นดอกเบี้ยจากเงินกู้จึงถือเป็นรายได้หลักของธนาคาร ในการปล่อยสินเชื่อแต่ละประเภทธนาคารจะมีการกำหนดหลักเกณฑ์ไว้เพื่อใช้ในการประกอบการพิจารณาการอนุมัติสินเชื่อซึ่งส่วนใหญ่ธนาคารจะใช้หลัก 5Cs ประกอบด้วย คุณลักษณะและความน่าเชื่อถือของผู้กู้ ความสามารถในการชำระหนี้ สินทรัพย์หรือเงินฝากของผู้กู้ หลักประกันของผู้กู้ และปัจจัยภายนอกที่จะส่งผลต่อการชำระหนี้ของผู้กู้ สาเหตุที่ต้องมีการกำหนดหลักเกณฑ์เนื่องจากหากการอนุมัติสินเชื่อไม่มีประสิทธิภาพมากพอก็จะส่งผลให้เกิดผลเสีย เพราะการอนุมัติสินเชื่อของธนาคารจะต้องมีค่าใช้จ่ายกันสำรองไว้ซึ่งมีไว้รองรับกรณีลูกหนี้ไม่สามารถชำระหนี้ได้หรือเรียกว่าเป็นหนี้เสีย หากธนาคารมีการกันเงินในส่วนนี้ไว้มากจากการปล่อยสินเชื่อผู้กู้ที่มีความเสี่ยงสูงจะส่งผลให้ธนาคารมีกำไรลดลง ดังนั้นการบริหารจัดการสินเชื่อที่ดีแต่แรกย่อมลดความเสี่ยงในการเกิดหนี้เสียได้ดีที่สุด การวิเคราะห์ในการอนุมัติสินเชื่อ วิเคราะห์แนวโน้มความสามารถในการจ่ายสินเชื่อของผู้กู้จึงเป็นสิ่งสำคัญ

จากที่กล่าวมาข้างต้น ผู้วิจัยได้ตระหนักถึงปัจจัยต่าง ๆ ที่จะส่งผลต่อการอนุมัติสินเชื่อ รวมถึงการทำนายการอนุมัติสินเชื่อส่วนบุคคล จึงนำการเรียนรู้ของเครื่องมาช่วยในการวิเคราะห์ให้เกิดประสิทธิภาพมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อนำการเรียนรู้ของเครื่องมาใช้ในการทำนายการอนุมัติสินเชื่อส่วนบุคคล
2) เพื่อศึกษาระดับความสำคัญของปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อส่วนบุคคล
3) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในการอนุมัติสินเชื่อส่วนบุคคล

1.3 ขอบเขตของงานวิจัย

ข้อมูลจากแคกเกิลเป็นชุดข้อมูลการอนุมัติเงินกู้ซึ่งรวบรวมบันทึกทางการเงินและข้อมูลที่เกี่ยวข้องที่ใช้เพื่อกำหนดคุณสมบัติของบุคคลในการได้รับเงินกู้จากสถาบันให้กู้ยืม ประกอบด้วยปัจจัยต่าง ๆ เช่น คะแนนเครดิต รายได้ สถานะการจ้างงาน ระยะเวลาเงินกู้ จำนวนเงินกู้ มูลค่าทรัพย์สิน และสถานะเงินกู้

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้ผลการทำนายการอนุมัติสินเชื่อส่วนบุคคลที่แม่นยำมากขึ้น
2) ทราบถึงปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อ รวมถึงเป็นแนวทางให้ผู้กู้ได้ทราบถึงปัจจัยที่สำคัญ
3) ทราบถึงประสิทธิภาพของแบบจำลอง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการศึกษา ประกอบด้วย 5 ส่วนคือ นิยามของสินเชื่อ การจัดเตรียมข้อมูล การเรียนรู้ของเครื่อง มาตรวัดประสิทธิภาพ และงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดของสินเชื่อส่วนบุคคล

สินเชื่อส่วนบุคคล หมายถึง สินเชื่ออเนกประสงค์ที่สถาบันทางการเงินหรือธนาคารจะปล่อยกู้ให้ยืมเอาไปใช้เป็นค่าใช้จ่ายต่างๆ ในชีวิตประจำวัน โดยจะมีการคิดอัตราดอกเบี้ย รวมไปถึงรูปแบบการชำระคืน ตามเงื่อนไขที่ธนาคารต่างๆ เป็นผู้กำหนด ภายใต้กฎหมายในประเทศนั้นๆ โดยการปล่อยสินเชื่อจะช่วยให้ปริมาณการบริโภคในระดับครัวเรือนเพิ่มมากขึ้น กระตุ้นการขยายตัวของอุปสงค์ ในรูปของตัวเงินที่มีต่อสินค้าและบริการ อีกทั้งช่วยรักษากระแสหมุนเวียนของรายได้ แต่ในการปล่อยสินเชื่อก็ถือเป็นการเพิ่มความเสี่ยงให้กับสถาบันการเงินจากการขาดความสามารถในการชำระหนี้ของผู้กู้ การปล่อยสินเชื่อจึงต้องมีการกำหนดหลักเกณฑ์เพื่อพิจารณาความสามารถ ความพร้อมที่ผู้กู้ยืมจะจ่ายคืนเงินที่กู้ยืมตามเงื่อนไขสัญญา โดยสถาบันการเงินจะใช้หลัก 5Cs ดังนี้

1. คุณสมบัติของผู้กู้ (Character) หมายถึง เป็นบุคคลที่มีอาชีพและรายได้แน่นอน มีอายุครบ 20 ปี บริบูรณ์ และเมื่อรวมอายุผู้กู้กับระยะเวลาที่ชำระเงินกู้ต้องไม่เกิน 65 ปี ซึ่งแสดงให้เห็นถึงความต้องการ ความตั้งใจในการชำระหนี้คืน และความรับผิดชอบต่อภาระหนี้ สินที่ผูกพันของผู้กู้ที่จะชำระคืนเมื่อถึงกำหนด ตลอดจนการพิจารณาถึงการปฏิบัติต่อหนี้ ในอดีต
2. ความสามารถในการชำระหนี้ (Capacity) หมายถึง ความสามารถในการหารายได้ของผู้กู้เพื่อนำมาชำระหนี้ ปัจจัยที่แสดงให้เห็นว่าผู้กู้หรือลูกหนี้ มีความสามารถในการชำระหนี้ได้แก่
    - รายได้ หากผู้กู้มีรายได้หรือมีความสามารถในการหารายได้สูงและสม่ำเสมอก็จะทำให้ผู้กู้สามารถชำระหนี้ได้ตรงกำหนดและเงื่อนไข
    - รายจ่าย ผู้กู้มีภาระที่ต้องใช้จ่ายมากน้อยเพียงใด ถ้ามีรายได้และรายจ่ายมาก ความสามารถของผู้กู้จะมีน้อย ถ้าผู้กู้เป็นหนี้เพิ่มขึ้นอาจจะมีปัญหาในการจ่ายเงินคืนกรณีที่ปล่อยสินเชื่อเพิ่มนั้นจะมีอัตราเสี่ยงสูงขึ้น และส่งผลให้หนี้เดิมที่มีอยู่จะมีอัตราเสี่ยงเพิ่มสูงขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- ระยะเวลา เป็นปัจจัยหนึ่งที่ใช้ในการพิจารณาความสามารถในการชำระหนี้ ผู้กู้บางรายอาจไม่สามารถชำระหนี้ ตามเงื่อนไขได้ถ้าให้ระยะเวลาสั้นไป แต่ถ้าหากระยะเวลานานขึ้น ผู้กู้ก็อาจจะมีความสามารถในการชำระหนี้ได้ตามเงื่อนไขโดยไม่มีปัญหาผิดนัด

3. เงินทุนในการดำเนินงาน (Capital) เป็นปัจจัยสำคัญในการจัดอันดับความสามารถในการชำระคืนได้อย่างดีโดยพิจารณาจากงบทางการเงิน เพื่อดูว่าทรัพย์สินสุทธิ (ทรัพย์สินทั้งหมด – หนี้สินทั้งหมด) มีจำนวนเท่าใด ทรัพย์สินที่มีอยู่มีสภาพคล่องและความเสี่ยงมากน้อยเพียงใด
4. หลักทรัพย์ค้ำประกันหนี้ (Collateral) การให้เครดิตในปัจจุบันจำเป็นต้องมีหลักประกันในการชำระคืน หลักประกันที่ดีควรเป็นหลักทรัพย์ที่ตีตราได้ง่าย และเปลี่ยนเป็นเงินสดได้รวดเร็ว เช่น ที่ดิน อาคารโรงงาน เครื่องมือเครื่องจักร หุ้นกู้ หุ้นสามัญของธุรกิจที่ซื้อขายกัน
5. สถานการณ์ทั่วไป (Condition) เป็นการพิจารณาถึงเหตุการณ์ที่อยู่นอกเหนือการควบคุม เช่น ภาวะทางด้านเศรษฐกิจ การเมืองและสังคม นโยบายรัฐ ปัญหาแรงงานหรือภาวะสงคราม ที่จะมีผลทำให้ฐานะของลูกหนี้เปลี่ยนแปลงไปในทางที่ดีหรือแย่ลงได้

2.2 การจัดเตรียมข้อมูล (Data Preprocessing)

2.1.1 การเพิ่มข้อมูล (Data Augmentation)

การเพิ่มข้อมูลเป็นเทคนิคที่ใช้เพื่อเพิ่มจำนวนข้อมูลในชุดข้อมูล โดยการเพิ่มข้อมูลจะทำในกรณีที่ข้อมูลไม่สมดุล (Imbalance Data) โดยวิธีที่นิยมใช้ในการเพิ่มข้อมูล มีดังนี้

1. การสุ่มตัวอย่างเพิ่มจากข้อมูลเริ่มต้นอย่างสุ่ม (Random Over-sampling: ROS) เป็นการสุ่มตัวอย่างจากข้อมูลกลุ่มน้อยและจะทำซ้ำข้อมูล ซึ่งส่งผลให้มีการกระจายคลาสที่สมดุล ในวิธีนี้ตัวอย่างของกลุ่มตัวอย่างกลุ่มน้อยจะถูกสุ่มเลือก ส่งผลให้วิธีนี้มีข้อจำกัด คือการเพิ่มข้อมูลลักษณะนี้อาจจะทำให้เกิดปัญหาข้อมูลถูกรบกวนได้ง่าย (Batarseh, Chong, and Kulkarni, 2020)
2. วิธีการสุ่มตัวอย่างสังเคราะห์ของข้อมูลกลุ่มน้อย (Synthetic Minority Over-sampling Technique: SMOTE) เป็นวิธีการเพิ่มตัวอย่างข้อมูลโดยการสุ่มสร้างตัวอย่างข้อมูลขึ้นมาใหม่ ด้วยการนำตัวอย่างข้อมูลจากกลุ่มตัวอย่างที่มีจำนวนน้อยมาพิจารณาทีละตัวจนครบทุกตัว หลักการคือกำหนดจำนวนด้วยอัลกอริทึมเพื่อนบ้านที่ใกล้เคียงที่สุด (K-Nearest Neighbor: KNN) จำนวน เค ตัว แล้วทำการสุ่มสร้างข้อมูลขึ้นมาใหม่บนทางที่เชื่อมโยง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ระหว่างจุดข้อมูลที่กำลังพิจารณาและจุดของข้อมูลเพื่อนบ้านที่ใกล้ที่สุด (Kevin, Lawrence, Nitesh, and W., 2022)

2.1.2 การแปลงข้อมูล (Data Transformation)

การทำงานของแบบจำลองส่วนมากทำงานบนตัวเลข ดังนั้นเมื่อมีข้อมูลที่เป็นเชิงคุณภาพทั้งที่มีลำดับหรือไม่มีลำดับ ต้องเปลี่ยนให้อยู่ในรูปแบบของค่าทวิภาค (Binary Values) ที่มีค่า 0 หรือ 1 เท่านั้น การแปลงข้อมูลแบบนี้จะเรียกว่า One-Hot Encoding (Sasiwut, 2020)

วิธี One-Hot Encoding เป็นกระบวนการแปลงข้อมูลจากรูปแบบหนึ่งเป็นรูปแบบอื่นโดยเฉพาะอย่างยิ่งในการแปลงข้อมูลจำแนกประเภท (Categorical Data) เช่น การแปลงค่าของข้อมูลสีจากชื่อสีเป็นตัวเลข โดยที่แต่ละค่าของข้อมูลจำแนกประเภท จะถูกแปลงเป็นเวกเตอร์ (Vector) ของ 0 และ 1

2.1.3 สัมประสิทธิ์สหสัมพันธ์สเปียร์แมน (Spearman correlation coefficient)

ค่าสัมประสิทธิ์สหสัมพันธ์สเปียร์แมนเป็นการหาความสัมพันธ์ระหว่างตัวแปร โดยจะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 ส่วนการแปลความหมาย หากค่าสัมประสิทธิ์สหสัมพันธ์สเปียร์แมนใกล้ -1.0 หมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากค่า +1.0 หมายความว่าตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมากในทิศทางเดียวกัน และหากค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน (สถาบันข้อมูลขนาดใหญ่ (องค์การมหาชน), 2564) โดยค่าสัมประสิทธิ์สหสัมพันธ์สเปียร์แมนจะเหมาะสำหรับข้อมูลลำดับ (Ordinal data) ข้อมูลช่วง (Interval Data) อัตราส่วน (Ratio Data) ที่ไม่ได้เป็นไปตามการแจกแจงปกติ (Non-Normal Distribution) และข้อมูลที่มีความสัมพันธ์เชิงไม่เชิงเส้น (Non-linear Relationships)

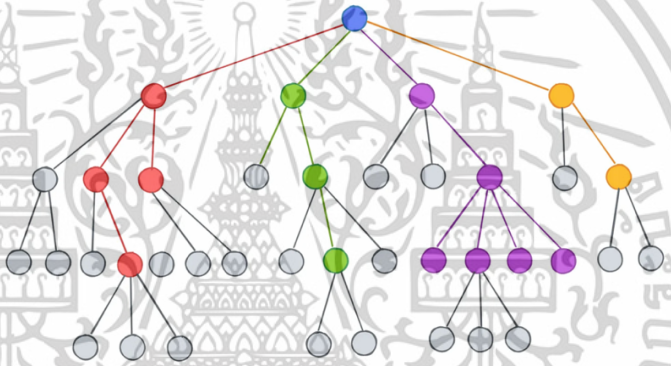
2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง คือ การออกแบบโปรแกรมให้สามารถเรียนรู้และพัฒนาตนเองได้จากประสบการณ์ หลักการของการเรียนรู้ของเครื่องคือการนำข้อมูลชุดฝึกสอนและผลลัพธ์มาป้อนเข้าไปให้กับคอมพิวเตอร์เพื่อสอนให้คอมพิวเตอร์เรียนรู้และทำให้เกิดการพัฒนาประสบการณ์ของตัวโปรแกรม เป็นการสร้างแบบจำลองการเรียนรู้ให้คอมพิวเตอร์สามารถทำนายหรือตัดสินใจได้ด้วยตนเองอย่างอัตโนมัติคล้ายมนุษย์ (อรพิน, 2564)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.2.1 เทคนิคต้นไม้ตัดสินใจ

เป็นการเรียนรู้โดยการจำแนกประเภทข้อมูลออกเป็นกลุ่ม โดยใช้คุณสมบัติของข้อมูลเป็นตัวกำหนด ซึ่งประกอบไปด้วย โหนดภายใน กิ่ง และโหนดใบ วิธีการวิเคราะห์แบบต้นไม้ตัดสินใจเป็นการค้นหาจากบนลงล่าง โดยเริ่มจากการเลือกคุณสมบัติที่ดีที่สุดมาเป็นโหนดราก และวนสร้างโหนดลูกและเส้นเชื่อมไปเรื่อยๆ จนกว่าข้อมูลที่ได้จะถูกจัดไว้เป็นกลุ่มเดียวกันถึงจะหยุดสร้างต้นไม้ จากการคำนวณเกนความรู้ (Information Gain: IG) โดยเลือกคุณสมบัติที่มีค่าเกนความรู้สูงที่สุด หรือการคำนวณค่าดัชนีจีนี (Gini Index) ซึ่งเป็นค่าที่บ่งบอกว่าตัวแปรใดสมควรนำมาใช้ในการแบ่ง โดยค่าดัชนีจีนีที่น้อยที่สุดคือค่าที่ดีที่สุด จะนำตัวแปรนั้นมาเป็นโหนดราก ดังรูปประกอบที่ 2.1 (ชณิดาภา บุญประสม และ จรัญ แสนราช, 2561)



รูปที่ 2.1 หลักการทำงานวิธีต้นไม้ตัดสินใจ

(ที่มา https://medium.com/intro-to-artificial-intelligence/decision-tree-learning-e153b5b4ecdf)

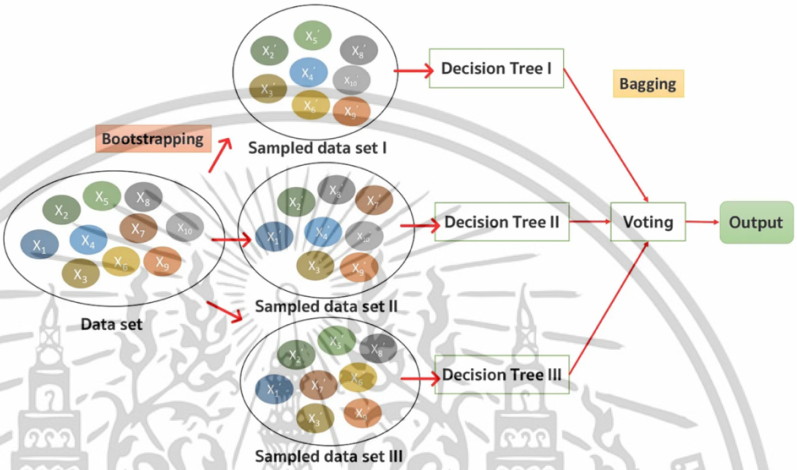
จากรูปที่ 2.1 แสดงถึงการจำแนกประเภทข้อมูลออกเป็นกลุ่มโดยใช้การวิเคราะห์แบบต้นไม้ตัดสินใจ ซึ่งเป็นการวิเคราะห์จากบนลงล่าง เริ่มจากการเลือกคุณสมบัติที่ดีที่สุดมาเป็นโหนดราก (Root Node) ซึ่งเป็นจุดเริ่มต้นของการแบ่งข้อมูล วนสร้างโหนดลูกและเส้นเชื่อม (branches) ไปเรื่อยๆ โดยการเลือกคุณสมบัติที่ดีที่สุดในแต่ละขั้นตอน การแบ่งข้อมูลจะดำเนินต่อไปจนกว่าข้อมูลที่ได้จะถูกจัดไว้เป็นกลุ่มเดียวกันทั้งหมดหรือจนกว่าจะถึงเงื่อนไขหยุดการแบ่ง

2.2.2 เทคนิควิธีป่าสุ่ม

เป็นการฝึกตัวแบบทำนายที่เหมือนกันหลายๆ ครั้ง บนข้อมูลชุดเดียวกัน โดยแต่ละครั้งของการฝึกข้อมูลจะเลือกส่วนของข้อมูลที่ฝึกไม่เหมือนกัน จากนั้นเอาการตัดสินใจของตัวแบบทำนายเหล่านั้นมา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โหวตกันว่า คลาสไหนถูกเลือกมากที่สุด ซึ่งจะช่วยแก้ปัญหาค่าความแปรปรวนสูง (High Variance) ที่เกิดจากต้นไม้แต่ละต้นได้ ซึ่งจะใช้ตัวแบบทำนายจากต้นไม้ตัดสินใจหลายๆ ต้น ตั้งแต่ 10 ถึงมากกว่า 1000 ตัวแบบทำนาย ซึ่งแต่ละตัวแบบทำนายจะได้รับชุดข้อมูลที่ไม่เหมือนกัน และจะทำนายแยกกันออกมา หลังจากนั้นจะเลือกผลการทำนายสุดท้าย จากค่าการทำนายที่ได้รับการโหวตมากที่สุด ดังรูปประกอบที่ 2.2 (ปรเมษฐ์ ธันวานนท์, ชัยกรยิ่ง เสรี, วรพล พงษ์เพ็ชร และ ธนภัทร ฆังคะจิตรสกสรรค์, 2560)



รูปที่ 2.2 หลักการทำงานวิธีป่าสุ่ม

(ที่มา https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c)

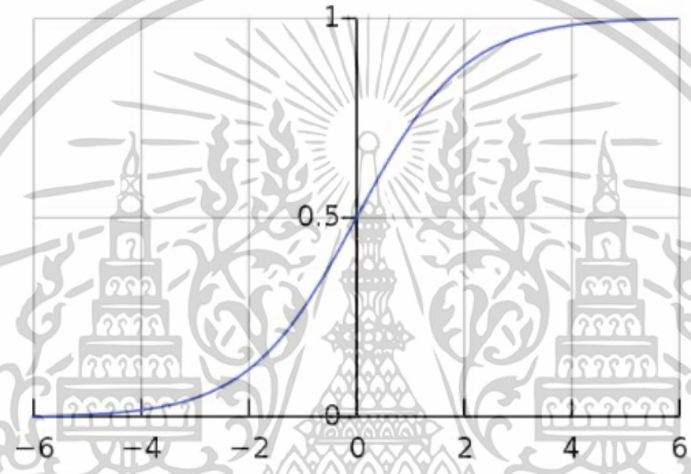
จากรูปที่ 2.2 เป็นการแสดงหลักการทำงานของวิธีป่าสุ่ม โดยเริ่มจากการสุ่มตัวอย่างข้อมูลจากชุดข้อมูลต้นฉบับหลายครั้ง ให้ได้ข้อมูลออกมา เอ็น (n) ชุด ที่ไม่เหมือนกัน หลังจากนั้นนำชุดข้อมูลที่สุ่มตัวอย่างมาใช้ในการฝึกแบบจำลองต้นไม้ตัดสินใจหลายต้น ในรูปนี้แสดงให้เห็นการฝึกต้นไม้ตัดสินใจจำนวน 3 ต้น โดยใช้ชุดข้อมูลที่สุ่มตัวอย่างต่างกัน 3 ชุด ได้แก่ Sampled data set I, Sampled data set II, และ Sampled data set III จากนั้นนำผลการทำนายจากแต่ละต้นมาโหวต ผลการทำนายสุดท้ายจะถูกเลือกจากค่าการทำนายที่ได้รับการโหวตมากที่สุด ผลลัพธ์สุดท้ายที่ได้จากการรวมการทำนายหลายๆ ต้นจะถูกนำมาใช้เป็นผลลัพธ์ของวิธีป่าสุ่ม

2.2.3 การถดถอยลอจิสติก

การวิเคราะห์ความถดถอยลอจิสติก เป็นเทคนิคการวิเคราะห์สถิติเชิงคุณภาพใช้สำหรับหาความสัมพันธ์ระหว่างตัวแปรอิสระ (มากกว่าหรือเท่ากับ 1 ตัว) กับตัวแปรตาม และพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ เมื่อแบ่งตามระดับการวัดของตัวแปร สามารถแบ่งออกเป็น 2 ประเภทคือ การ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

วิเคราะห์ความถดถอยลอจิสติกทวิ (Binary logistic regression) และการวิเคราะห์ลอจิสติกแบบพหุกลุ่ม (Multinomial logistic regression) ซึ่งการวิเคราะห์ความถดถอยลอจิสติกดังกล่าวแตกต่างกันในด้านของตัวแปรตาม การวิเคราะห์ความถดถอยลอจิสติกทวิใช้กับตัวแปรตามที่มีค่าได้เพียง 2 ค่า คือ 0 กับ 1 เช่น กลุ่มที่มีเหตุการณ์กับกลุ่มที่ไม่มีเหตุการณ์ เป็นต้น ส่วนการวิเคราะห์ลอจิสติกแบบพหุกลุ่มใช้กับตัวแปรตามที่มีมากกว่า 2 ค่า เช่น ตัวแปรตาม หมายถึง รูปแบบการขนส่งของแต่ละบุคคล ได้แก่ เครื่องบิน รถยนต์ส่วนบุคคล และรถโดยสารสาธารณะ เป็นต้น ดังรูปประกอบที่ 2.3 และ ดังรูปประกอบที่ 2.4 (ยุทธ ไกยวรรณ์, 2555)



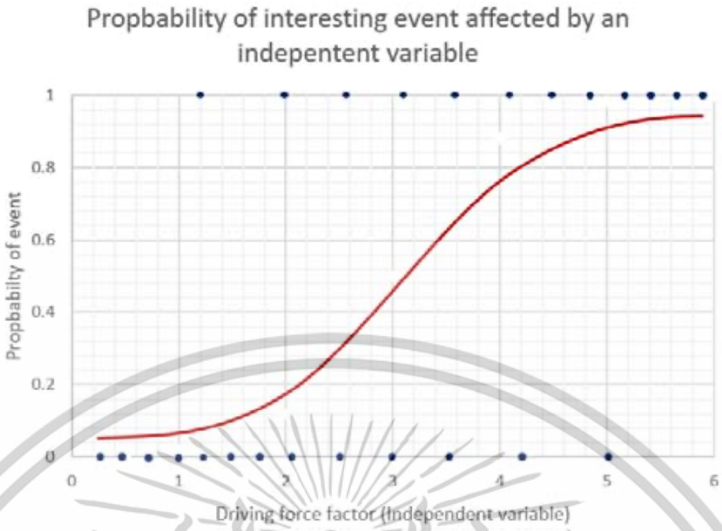
รูปที่ 2.3 ฟังก์ชันลอจิสติกมาตรฐาน

(ที่มา https://forest-admin.forest.ku.ac.th/304xxx/?q=system/files/book/5%282018%29%20Logistic %20Regression.pdf)

จากรูปที่ 2.3 เป็นกราฟเส้นโค้งลอจิสติกแสดงความสัมพันธ์ระหว่างตัวแปรอิสระและค่าความน่าจะเป็นที่อยู่ในช่วง 0 ถึง 1 รูปแบบของเส้นโค้งนี้มีลักษณะเป็น S-curve ซึ่งบ่งบอกถึงความสัมพันธ์ที่ไม่เป็นเชิงเส้น (non-linear) เมื่อค่าของตัวแปรอิสระเพิ่มขึ้น ค่าความน่าจะเป็นจะเพิ่มขึ้นจาก 0 เข้าใกล้ 1 ตามสมการลอจิสติก

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

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.4 เส้นโค้งการถดถอยลอจิสติก

จากรูปที่ 2.4 เป็นกราฟแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ (Driving force factor) และความน่าจะเป็นของเหตุการณ์ที่สนใจ เส้นโค้งสีแดงแสดงการประมาณค่าความน่าจะเป็นของเหตุการณ์ที่สนใจด้วยการวิเคราะห์ความถดถอยลอจิสติก จุดสีน้ำเงินบนกราฟแสดงค่าจริงของเหตุการณ์ที่เกิดขึ้น (0 หรือ 1) การวิเคราะห์ความถดถอยลอจิสติกใช้ในการประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ โดยพิจารณาจากค่าของตัวแปรอิสระ เส้นโค้งที่ได้แสดงถึงความสัมพันธ์ที่ไม่เป็นเชิงเส้นระหว่างตัวแปรอิสระและค่าความน่าจะเป็น

2.3 การประเมินประสิทธิภาพของแบบจำลอง (Model Evaluation)

การประเมินประสิทธิภาพของแบบจำลองสามารถคำนวณได้จากเมทริกซ์ความสับสน ซึ่งเป็นวิธีการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องที่นิยมใช้อย่างมากในการจำแนกประเภทของข้อมูลแบบไบนารี และใช้สำหรับการวัดผลของการเรียนรู้ของเครื่อง โดยค่าที่ได้จากการทำนายจากแบบจำลองสามารถอธิบายในตารางที่ 2.1

ตารางที่ 2.1 ตารางเมทริกซ์ความสับสน

	ข้อมูลจริงอนุมัติ	ข้อมูลจริงไม่อนุมัติ
ทำนายอนุมัติ	TP (True Positive)	FP (False Positive)
ทำนายไม่อนุมัติ	FN (False Negative)	TN (True Negative)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดยที่

ผลบวกจริง (True Positive: TP) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ข้อมูลจริงอนุมัติและทำนายอนุมัติ

ผลบวกเท็จ (False Positive: FP) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น ข้อมูลจริงไม่อนุมัติและทำนายอนุมัติ

ผลลบจริง (True Negative: TN) คือ สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ข้อมูลจริงไม่อนุมัติและทำนายไม่อนุมัติ

ผลลบเท็จ (False Negative: FN) คือ สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น ข้อมูลจริงอนุมัติและทำนายไม่อนุมัติ

2.3.1 ค่าความถูกต้อง (Accuracy)

เป็นการวัดความแม่นยำของแบบจำลองโดยรวม กล่าวคือแบบจำลองทำนายถูกกี่ครั้งจากจำนวนการทำนายทั้งหมด แสดงดังสมการ

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

2.3.2 ค่าความแม่นยำ (Precision)

การวัดความแม่นยำของแบบจำลองโดยการพิจารณาแยกทีละคลาส แสดงดังสมการ

$$\text{Precision} = \frac{TP}{TP+FP}$$

2.3.3 ค่าความระลึก (Recall)

เป็นการวัดความถูกต้องของแบบจำลองโดยการพิจารณาแยกทีละคลาส ตัวอย่างเช่น การวัดว่าผลการทำนายอนุมัติความถูกต้องเท่าไหร่เมื่อเทียบกับข้อมูลอนุมัติจริงทั้งหมด แสดงดังสมการ

$$\text{Recall} = \frac{TP}{TP+FN}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.4 การหาความสัมพันธ์ของตัวแปรที่ส่งผลต่อการทำนายของแบบจำลอง

2.4.1 ค่าฟีเจอร์ อิมพอร์แทนเซส (Feature Importances)

เป็นการวัดความสำคัญของตัวแปรแต่ละตัวในแบบจำลองการเรียนรู้ของเครื่อง การวัดนี้แสดงถึงระดับที่ตัวแปรแต่ละตัวมีผลต่อการทำนายของแบบจำลอง ซึ่งค่าเหล่านี้มักคำนวณจากแบบจำลองที่ใช้ต้นไม้การตัดสินใจ หรือแบบจำลองที่มีพื้นฐานจากต้นไม้ เช่น วิธีป่าสุ่มและเกรเดียนท์ บูสติ้ง (Gradient Boosting) โดยการคำนวณค่าฟีเจอร์ อิมพอร์แทนเซส พิจารณาจากการที่ตัวแปรช่วยลดความผิดพลาดในการแยกข้อมูลภายในต้นไม้ตัดสินใจได้มากน้อยเพียงใด

2.4.2 ค่าเพอร์มิวเทชัน อิมพอร์แทนซ์ (Permutation Importance)

เป็นวิธีการวัดความสำคัญของตัวแปรโดยไม่ขึ้นกับประเภทของแบบจำลองที่ใช้ วิธีการนี้ประเมินผลกระทบของการสุ่มสับเปลี่ยนค่าของตัวแปรที่มีต่อประสิทธิภาพของแบบจำลอง ซึ่งแสดงถึงความสำคัญของตัวแปรต่อการทำนายของแบบจำลอง วิธีการนี้ไม่มีอคติต่อตัวแปรที่มีสเกลค่าต่างกันและสามารถใช้กับแบบจำลองทุกประเภท ขั้นตอนการคำนวณเพอร์มิวเทชัน อิมพอร์แทนซ์ประกอบด้วย

1. วัดประสิทธิภาพของแบบจำลองด้วยชุดข้อมูลทดสอบ
2. สุ่มสับเปลี่ยนค่าของตัวแปรที่ต้องการวัด
3. วัดประสิทธิภาพของแบบจำลองอีกครั้งด้วยข้อมูลที่ถูกสับเปลี่ยน
4. ประเมินการลดลงของประสิทธิภาพของแบบจำลอง ซึ่งการลดลงจะแสดงถึงความสำคัญของตัวแปรนั้น

2.5 งานวิจัยที่เกี่ยวข้อง

Mahdi, Mohammad, and Muhammed (2021) ได้ศึกษาเกี่ยวกับประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องสำหรับระบบอนุมัติสินเชื่อ โดยทำการสำรวจจากวรรณกรรมจำนวน 52 บทความครอบคลุมตั้งแต่ปี ค.ศ. 2000 – 2021 จากการศึกษาพบว่าการวิเคราะห์การถดถอยลอจิสติกเป็นแบบจำลองที่ทำนายแม่นยำที่สุด ตามด้วยแบบจำลองต้นไม้การตัดสินใจ และ โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น

Bhavana, Lakshmi, Loukya, Saibaba, Sunanda, and Tumuluru (2022) ได้ศึกษาการประเมินความเสี่ยงในการขอสินเชื่อโดยใช้อัลกอริทึมการเรียนรู้ของเครื่อง วิเคราะห์ข้อมูลลูกค้าในส่วนของ อายุ รายได้ จำนวนเงินกู้ และระยะเวลาในการทำงาน ซึ่งอัลกอริทึมที่ใช้ประกอบด้วย วิธีป่าสุ่ม, อัลกอริทึมซัพพอร์ตวอกเตอร์แมชชีน (Support Vector Machine: SVM), วิธีเพื่อนบ้านใกล้สุด (K-Nearest Neighbor), และ การวิเคราะห์การถดถอยลอจิสติก ผลการศึกษาพบว่าวิธีป่าสุ่มมีค่าความ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ถูกต้องสูงที่สุดคือ 0.81 รองลงมาคือวิธีการวิเคราะห์การถดถอยลอจิสติก อัลกอริทึมซัพพอร์ตวอกเตอร์แมชชีนและวิธีเพื่อนบ้านใกล้สุด ตามลำดับ

Kavya, Maddumala, Ramya, Tejaswini, and Triveni (2020) ได้ศึกษาการคาดการณ์การอนุมัติเงินกู้ที่โดยใช้อัลกอริทึม การถดถอยโลจิสติก ต้นไม้การตัดสินใจและวิธีป่าสุ่ม ผลการศึกษาพบว่าวิธีต้นไม้การตัดสินใจเป็นวิธีที่มีความแม่นยำมากที่สุด

Yengejeh (2023) ได้ศึกษาเกี่ยวกับการอนุมัติบัตรเครดิตโดยใช้อัลกอริทึมต้นไม้การตัดสินใจ ผลการศึกษาพบว่าวิธีต้นไม้การตัดสินใจมี ค่าความถูกต้องอยู่ที่ 0.86 ค่าความแม่นยำ 0.83 ค่าความไว 0.89 และ ค่าเอฟ1-สกอร์ (F1-Score) 0.86 สำหรับปัจจัยที่ส่งผลต่อการอนุมัติบัตรเครดิตมากที่สุดสามอันดับ คือ ประวัติในการชำระหนี้ สถานะหนี้ และสถานการณ์จ้างงาน

Chavan, Diwate, and Chavan (2021) ได้ศึกษาการทำนายการอนุมัติเงินกู้โดยใช้การเรียนรู้ของเครื่องเพื่อลดความเสี่ยงในการปล่อยกู้ของธนาคาร โดยใช้อัลกอริทึมซัพพอร์ตวอกเตอร์แมชชีน และต้นไม้การตัดสินใจ ซึ่งผลการศึกษาพบว่า คะแนนเครดิตที่น้อยลงมีแนวโน้มที่จะได้รับการอนุมัติเงินกู้น้อยลง ผู้สมัครที่มีรายได้สูงขึ้นและจำนวนเงินกู้ต่ำมีแนวโน้มที่จะได้รับการอนุมัติมากขึ้น เพศและสถานภาพสมรสไม่มีผลต่อการอนุมัติเงินกู้อย่างมีนัยสำคัญ

Bhadola, Khan, Kumar, and Singh (2021) ได้เปรียบเทียบความแม่นยำและประสิทธิภาพของแบบจำลอง การถดถอยลอจิสติก, ต้นไม้การตัดสินใจและวิธีป่าสุ่มสำหรับการทำนายการอนุมัติเงินกู้ ซึ่งจากการศึกษาพบว่า การถดถอยลอจิสติก, ต้นไม้การตัดสินใจและวิธีป่าสุ่ม ได้ค่าความถูกต้องเท่ากับ 0.81, 0.94, และ 0.83 ตามลำดับ ในขณะที่ ครอส-วาลิเดชั่น (Cross-Validation) มีค่าเท่ากับ 0.81, 0.72, และ 0.80 ตามลำดับ แสดงให้เห็นว่าสำหรับชุดข้อมูลที่กำหนด ค่าความถูกต้องของโมเดลที่ใช้ต้นไม้ตัดสินใจสูงที่สุด แต่วิธีป่าสุ่มดีกว่าในการทำนายทั่วไป ถึงแม้ครอส-วาลิเดชั่นของวิธีป่าสุ่มจะไม่สูงมากเท่าการถดถอยลอจิสติก

สุเมธ จุฑาจันทร์ และสมพร ปั่นโภชา (2564) ได้ศึกษาประสิทธิภาพของแบบจำลองการอนุมัติสินเชื่อจาก 3 แบบจำลอง ได้แก่ วิธีเพื่อนบ้านใกล้สุด การถดถอยโลจิสติก และต้นไม้การตัดสินใจแบบ CART โดยใช้ข้อมูลของเจอมันเครดิต (German Credit ที่มาจาก UC Irvine Machine Learning Reposity) และใช้โปรแกรมอาร์เป็นเครื่องมือในการสร้างระเบียบวิธีการเรียนรู้ของเครื่อง ซึ่งจากการศึกษาพบว่าแบบจำลองวิธีเพื่อนบ้านใกล้สุดมีประสิทธิภาพสูงที่สุด มีค่าความถูกต้อง 0.993 ค่าความแม่นยำ 0.99 ค่ารีคอล 1 และค่าเอฟ1-สกอร์ 0.995 รองลงมาคือการถอถอยลอจิสติก ต้นไม้การตัดสินใจแบบ CART ตามลำดับ

ปารดา ศัสตุระ (2565) ได้ศึกษาปัจจัยที่ไม่ก่อให้เกิดรายได้ และการประยุกต์เหมืองข้อมูลในการพยากรณ์สถานะหนี้เสียของลูกค้าสินเชื่อ โดยรวบรวมข้อมูลจากลูกค้าที่ได้รับสินเชื่อโควิด-19 ของ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ธนาคารออมสิน สาขามหาวิทยาลัย เชียงใหม่ ตั้งแต่เดือน กรกฎาคม 2564 ถึง ธันวาคม 2564 จำนวน 7,215 ราย ซึ่งแบบจำลองที่ใช้ประกอบด้วย ต้นไม้การตัดสินใจ นาอีฟเบย์ (Naive Bayes) และวิธีป่าสุ่ม ผลการศึกษาพบว่าแบบจำลองต้นไม้การตัดสินใจมีค่าความถูกต้องสูงที่สุดคือ 0.96 รองลงมาคือวิธีป่าสุ่ม และนาอีฟเบย์ ตามลำดับ สำหรับปัจจัยที่มีผลต่อการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ ได้แก่ อาชีพ สถานะภาพ ระดับการศึกษา รายได้ เพศ

ชุภาดา ผิวเผือก (2562) ได้ศึกษาปัจจัยที่ใช้ในการวิเคราะห์สินเชื่อของธนาคารพาณิชย์ในประเทศไทย โดยกลุ่มตัวอย่างที่ใช้ในงานวิจัยคือ เจ้าหน้าที่สินเชื่อระดับปฏิบัติการของธนาคารพาณิชย์ เครื่องมือที่ใช้คือแบบสอบถาม ผลการศึกษาพบว่าด้านหลักประกัน ประกอบด้วย ประเภทของหลักประกัน กรรมสิทธิ์ในหลักประกันและมูลค่าหลักประกันมีความสำคัญมากที่สุด รองลงมาคือ ด้านความสามารถในการชำระหนี้ ประกอบด้วย สัดส่วนการขายลูกค้ารายใหญ่ สัดส่วนการซื้อกับผู้ขายรายใหญ่ สถานประกอบการ (ดูกรรมสิทธิ์) ยอดขายของกิจการ ต้นทุนขาย/ยอดขาย ค่าใช้จ่ายในการขายและบริหาร ยอดหนี้รวมของกิจการ และทรัพย์สินรวมของกิจการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3

วิธีการดำเนินงานวิจัย

งานวิจัยนี้ศึกษาการทำนายและวิเคราะห์การอนุมัติสินเชื่อส่วนบุคคล โดยใช้การเรียนรู้ของเครื่อง 3 วิธี คือ วิธีต้นไม้ตัดสินใจ วิธีป่าสุ่ม และวิธีการวิเคราะห์การถดถอยลอจิสติก ซึ่งใช้ค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึกในการเปรียบเทียบประสิทธิภาพของแบบจำลอง

3.1 แผนผังขั้นตอนการทำงาน

แผนขั้นตอนการทำงานสามารถเขียนสรุปเป็นแผนผังได้ ดังรูปที่ 3.1

รูปที่ 3.1 แผนผังขั้นตอนการทำงาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 การจัดเตรียมข้อมูล (Data Preprocessing)

3.2.1 การเก็บรวบรวมข้อมูล (Data Collection)

ผู้วิจัยได้นำข้อมูลจากแคกเกิล ซึ่งเป็นชุดข้อมูลการอนุมัติเงินกู้ที่รวบรวมบันทึกทางการเงินและข้อมูลที่เกี่ยวข้องเพื่อกำหนดคุณสมบัติของบุคคลในการได้รับเงินกู้จากสถาบันให้กู้ยืม โดยชุดข้อมูลมีจำนวนทั้งสิ้น 4,269รายการ ประกอบด้วยตัวแปรทั้งหมด 13 ตัว คือ ตัวแปรตาม 1 ตัว และตัวแปรอิสระ 12 ตัว ซึ่งตัวแปรตามเป็นข้อมูลประเภทเชิงคุณภาพ ส่วนตัวแปรอิสระเป็นข้อมูลเชิงคุณภาพ 3 ตัว ข้อมูลเชิงปริมาณ 9 ตัว และหลังจากผู้วิจัยได้ศึกษางานวิจัยต่างๆ และทฤษฎีบทที่เกี่ยวข้องจึงได้กำหนดตัวแปรอิสระเพิ่ม 1 ตัว เป็นข้อมูลเชิงปริมาณซึ่งเป็นตัวแปรที่แสดงถึงความสามารถในการชำระหนี้ ขั้นตอนต่อมาคือการแปลงข้อมูลที่เป็นข้อมูลเชิงคุณภาพให้อยู่ในรูปแบบของค่าทวิภาค ที่มีค่า 0 หรือ 1 เท่านั้นโดยรายละเอียดของข้อมูลต่างๆ ดังตารางที่ 3.1

ตารางที่ 3.1 ตัวแปร ชื่อหัวข้อ คำอธิบาย และประเภทของข้อมูล

ตัวแปร	ชื่อหัวข้อ	คำอธิบาย	ประเภท
Y	loan_status	สถานะเงินกู้ 0 = ไม่อนุมัติ 1 = อนุมัติ	คุณภาพ
	loan_id	ไอดีเงินกู้	คุณภาพ
X_1	no_of_dependents	จำนวนผู้อยู่ในอุปการะของผู้สมัคร	ปริมาณ
X_2	education	การศึกษาของผู้สมัคร 0 = ไม่ผ่านการศึกษาขั้นบังคับ 1 = ผ่านการศึกษาขั้นบังคับ	คุณภาพ
X_3	self_employed	สถานการณ์จ้างงานของผู้สมัคร 0 = ไม่มีงานทำ 1 = มีงานทำ	คุณภาพ
X_4	income_annum	รายได้ประจำปีของผู้สมัคร	ปริมาณ
X_5	loan_amount	จำนวนเงินกู้	ปริมาณ
X_6	loan_term	ระยะเวลาเงินกู้เป็นปี	ปริมาณ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.1 (ต่อ) ตัวแปร ชื่อหัวข้อ คำอธิบาย และประเภทของข้อมูล

ตัวแปร	ชื่อหัวข้อ	คำอธิบาย	ประเภท
X_7	cibil_score	คะแนนเครดิต	ปริมาณ
X_8	residential_assets_value	มูลค่าสินทรัพย์ที่อยู่อาศัย	ปริมาณ
X_9	commercial_assets_value	มูลค่าสินทรัพย์เชิงพาณิชย์	ปริมาณ
X_{10}	luxury_assets_value	มูลค่าสินทรัพย์หรูหรา	ปริมาณ
X_{11}	bank_asset_value	มูลค่าสินทรัพย์ในธนาคาร	ปริมาณ
X_{12}	debt ratio	สัดส่วนระหว่างหนี้สินกับสินทรัพย์รวม	ปริมาณ

จากตารางที่ 3.1 แสดงตัวแปรทั้งหมด 14 ตัว โดยตัวแปรตามเป็นข้อมูลเชิงคุณภาพ และตัวแปรอิสระ 13 ตัว แบ่งเป็นข้อมูลเชิงคุณภาพ 3 ตัว ข้อมูลเชิงปริมาณ 10 ตัว ซึ่งเป็นตัวแปรเพิ่มเติม 1 ตัว นั่นคือ ตัวแปรสัดส่วนระหว่างหนี้สินกับสินทรัพย์รวม ตัวแปรเพิ่มเติมนี้เป็นตัวแปรที่แสดงถึงความสามารถในการชำระหนี้ของผู้กู้ โดยนำจำนวนเงินกู้มาหารด้วยผลรวมของมูลค่าสินทรัพย์ที่อยู่อาศัย มูลค่าสินทรัพย์เชิงพาณิชย์ มูลค่าสินทรัพย์หรูหรา มูลค่าสินทรัพย์ในธนาคาร และได้ตัดตัวแปรออก 1 ตัวแปร นั่นคือ ตัวแปรไอดีเงินกู้ เนื่องจากเป็นตัวแปรที่แสดงลำดับที่จึงไม่ได้มีความสำคัญต่อการอนุมัติเงินกู้ ดังนั้นข้อมูลชุดนี้จะมีตัวแปรทั้งหมด 13 ตัว ตัวแปรตาม 1 ตัว และตัวแปรอิสระ 12 ตัว แบ่งเป็นข้อมูลเชิงคุณภาพ 2 ตัว ข้อมูลเชิงปริมาณ 10 ตัว

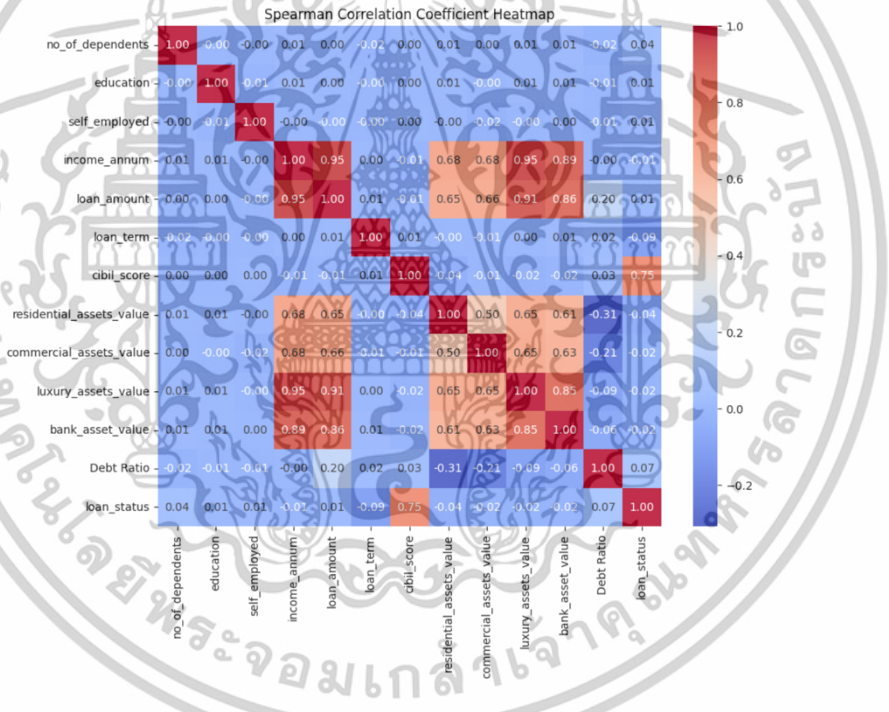
3.1.2 การจัดการกับความไม่สมดุลของข้อมูล (Imbalanced data)

เนื่องจากข้อมูลที่ใช้ในการทำนายมีความไม่สมดุล โดยมีจำนวนข้อมูลสถานะอนุมัติเงินกู้มากกว่าข้อมูลสถานะไม่อนุมัติเงินกู้ ซึ่งอาจทำให้การทำนายไม่แม่นยำเท่าที่ควร ดังนั้นเพื่อเพิ่มประสิทธิภาพในการทำนายจึงได้ดำเนินการสร้างข้อมูลเพิ่ม การเพิ่มข้อมูลนี้สร้างขึ้นโดยการสุ่มข้อมูลแบบสุ่มจากข้อมูลที่มีอยู่แล้ว โดยมีการสุ่มค่าจากคอลัมน์ 3 คอลัมน์ ได้แก่ "education", "self_employed", และ "loan_amount" เพื่อสร้างข้อมูลใหม่ การเพิ่มข้อมูลจะเพิ่มทีละ 200 รายการ จากนั้นใช้วิธีการคำนวณเพื่อนบ้านใกล้สุด (K-nearest neighbors) เพื่อซ่อมแซมข้อมูลที่หายไป โดยนำข้อมูลที่มีอยู่เป็นพื้นฐานในการคำนวณหาค่าของข้อมูลที่หายไป ซึ่งทำขั้นตอนนี้จำนวน 5 ครั้ง เพื่อให้ได้ข้อมูลจำนวน 1,000 รายการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.1.3 การหาความสัมพันธ์ระหว่างข้อมูล (Correlation)

การหาความสัมพันธ์ระหว่างข้อมูลเป็นการดูว่าตัวแปรมีความสัมพันธ์กันในระดับใด และมีความสัมพันธ์ในทิศทางใด โดยสัมประสิทธิ์สหสัมพันธ์จะมีค่าอยู่ระหว่าง -1 ถึง 1 หากค่าเข้าใกล้ 1 จะหมายความว่าตัวแปรทั้งสองมีความสัมพันธ์กันโดยตรงอย่างมาก หากมีค่าเข้าใกล้ -1 จะหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม และหากมีค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน โดยทางผู้วิจัยได้หาความสัมพันธ์ของข้อมูลเพื่อลดจำนวนตัวแปรที่ไม่มีความสัมพันธ์กับตัวแปรสถานะเงินกู้ เพื่อให้แบบจำลองมีประสิทธิภาพในการทำนายมากขึ้น ซึ่งทางผู้วิจัยหาความสัมพันธ์ระหว่างข้อมูลโดยใช้วิธีสัมประสิทธิ์สหสัมพันธ์ลำดับที่สเปียร์แมน (Spearman rank correlation coefficient) และสร้างแผนภูมิที่แสดงข้อมูลค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรแต่ละตัวออกมา ดังรูปที่ 3.2



รูปที่ 3.2 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ลำดับที่สเปียร์แมน

จากรูปที่ 3.2 เป็นแผนภูมิที่แสดงค่าสัมประสิทธิ์สหสัมพันธ์ลำดับที่สเปียร์แมนของตัวแปรแต่ละตัว ซึ่งจะเห็นว่ามีตัวแปรทุกตัวส่งผลต่อการอนุมัติเงินกู้ทั้งหมด ดังนั้นจะไม่ตัดตัวแปรใดออก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามพบว่า ตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรต้นมาก คือ คะแนนเครดิต ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ 0.75 รองลงมาคือ ระยะเวลาเงินกู้ ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ -0.09

3.2 การสร้างแบบจำลอง

นำชุดข้อมูลมาสร้างแบบจำลอง โดยแบ่งชุดข้อมูลออกเป็น ชุดข้อมูลการเรียนรู้ (Training set) และข้อมูลที่ใช้ในการทดสอบ (Test set) โดยแบ่งเป็นอัตราส่วน 80:20 จากนั้นนำข้อมูลมาสร้างแบบจำลอง โดยแบบจำลองที่เลือก นำมาจากการศึกษาวิจัยที่เกี่ยวข้องและเลือกวิธีที่มีประสิทธิภาพที่สุดมาใช้ในการทดลอง 3 แบบจำลอง ดังนี้

1. วิธีการต้นไม้ตัดสินใจ
2. วิธีป่าสุ่ม
3. วิธีการวิเคราะห์การถดถอยลอจิสติก

3.3 การเปรียบเทียบประสิทธิภาพแบบจำลอง

การเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด 3 แบบจำลอง ด้วยมาตรวัดประสิทธิภาพจำนวน 3
ค่า คือ ค่าความถูกต้อง ค่าความแม่นยำและค่าความไวที่มีค่าสูงสุดหรือเข้าใกล้ 1

3.4 การหาความสัมพันธ์ของตัวแปรที่ส่งผลต่อการทำนายของแบบจำลอง

การหาความสัมพันธ์ของตัวแปรที่ส่งผลต่อการทำนายของแบบจำลอง เป็นการหาเพื่อตรวจสอบว่าตัวแปรไหนส่งผลต่อการอนุมัติสินเชื่อมากที่สุด 3 อันดับแรก โดยเทคนิคที่เลือกใช้คือค่าฟีเจอร์ อิมพอร์แทนเซส (Feature Importances) และเพอร์มิวเทชัน อิมพอร์แทนซ์ (Permutation Importance) ซึ่งทั้งสองค่าเป็นค่าที่แสดงถึงความสำคัญของตัวแปรว่ามีอิทธิพลต่อผลลัพธ์ของแบบจำลองมากน้อยเพียงใด โดยตัวแปรที่มีค่าสูงแสดงว่ามีความสำคัญต่อผลลัพธ์ของแบบจำลองมาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

ผลการวิจัยและการอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการทำนายการอนุมัติสินเชื่อส่วนบุคคล โดยวิธีการเรียนรู้ของเครื่อง นำข้อมูลมาจากแคกเกิล เป็นชุดข้อมูลการอนุมัติเงินกู้ซึ่งรวบรวมบันทึกทางการเงินและข้อมูลที่เกี่ยวข้องที่ใช้เพื่อกำหนดคุณสมบัติของบุคคลในการได้รับเงินกู้จากสถาบันให้กู้ยืม จากนั้นทำการเพิ่มข้อมูลเนื่องจากข้อมูลมีความไม่สมดุลเพื่อให้แบบจำลองมีประสิทธิภาพและทำนายได้แม่นยำขึ้น และทำการเตรียมข้อมูล เพื่อจะไปสู่ขั้นตอนของการสร้างแบบจำลอง โดยแบบจำลองเป็นแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีต้นไม้ตัดสินใจ และวิธีการวิเคราะห์การถดถอยลอจิสติก ขั้นตอนสุดท้ายทำการเปรียบเทียบประสิทธิภาพแบบจำลอง โดยผู้วิจัยกำหนดใช้ค่าความถูกต้อง ค่าความแม่นยำ และค่าความไว

4.1 ผลการวิเคราะห์ของการเรียนรู้ของเครื่อง

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องในการทำนายการอนุมัติสินเชื่อส่วนบุคคล โดยใช้ข้อมูลทั้งหมด 5,269 รายการ แบ่งชุดข้อมูลออกเป็นชุดข้อมูลการเรียนรู้จำนวน 4,215 รายการ คิดเป็นร้อยละ 80 และข้อมูลที่ใช้ในการทดสอบจำนวน 1,054 รายการ คิดเป็นร้อยละ 20 ได้แสดงผลการทำนายไว้ดังตารางที่ 4.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.1 ตัวอย่างผลการทำนายของแบบจำลอง

No of Dependents	Education	Self Employed	Income Annum	Loan Amount	Loan Term	Cibil Score	Residential Assets	Commercial Assets	Luxury Assets	Bank Asset	Debt Ratio	Actual	Predicted Decision Tree	Predicted Random Forest	Predicted Logis
3	1	0	3,156,000	8,500,000	11	516	5,824,000	3,728,000	11,336,000	2,852,000	0.36	0	0	0	1
0	1	1	800,000	2,200,000	20	782	1,300,000	800,000	2,800,000	600,000	0.4	1	1	1	1
3	1	0	1,100,000	3,100,000	14	359	1,500,000	1,000,000	4,000,000	1,000,000	0.41	0	0	0	1
3	1	0	3,500,000	13,300,000	4	798	7,400,000	5,300,000	12,100,000	3,600,000	0.47	1	1	1	1
2	0	1	7,000,000	23,800,000	10	689	8,400,000	5,400,000	15,700,000	7,800,000	0.64	1	0	1	1
1	1	0	1,800,000	6,500,000	6	540	2,500,000	3,200,000	5,000,000	1,100,000	0.55	0	0	0	1
1	0	1	1,100,000	2,700,000	16	346	1,900,000	300,000	2,200,000	800,000	0.52	0	0	0	1
4	0	1	9,100,000	35,200,000	6	341	1,000,000	300,000	23,100,000	8,200,000	1.08	0	0	0	0
4	1	0	4,800,000	12,000,000	6	441	8,400,000	1,300,000	18,200,000	3,200,000	0.39	0	0	0	0
2	1	1	8,900,000	32,500,000	8	663	12,340,000	7,560,000	23,940,000	7,400,000	0.63	0	1	0	0
0	1	0	8,000,000	29,800,000	12	736	600,000	15,700,000	30,800,000	8,400,000	0.54	1	1	1	1
5	0	1	8,600,000	26,300,000	10	843	23,300,000	9,800,000	18,100,000	11,300,000	0.42	1	1	1	1
1	1	1	1,800,000	6,300,000	12	749	4,200,000	200,000	3,800,000	2,700,000	0.58	1	1	1	1
1	0	1	5,200,000	20,500,000	10	407	1,100,000	3,400,000	18,200,000	3,800,000	0.77	0	0	0	1
1	1	0	2,600,000	9,900,000	6	594	2,900,000	4,300,000	6,700,000	2,500,000	0.6	1	1	1	1
3	0	1	4,900,000	12,700,000	8	409	2,000,000	2,500,000	19,300,000	4,700,000	0.45	0	0	0	0
1	1	0	700,000	1,400,000	6	759	1,000,000	500,000	2,600,000	700,000	0.29	1	1	1	1
2	0	0	1,176,000	3,100,000	15	512	1,088,000	1,376,000	3,520,000	1,316,000	0.42	0	0	0	1

จากตารางที่ 4.1 เป็นตารางแสดงตัวอย่างผลการทำนายของแบบจำลองวิธีต้นไม้ตัดสินใจ วิธีป่าสุ่ม และวิธีการถดถอยลอจิสติก ซึ่งจากตารางจะเห็นว่าวิธีป่าสุ่มสามารถทำนายได้ตรงกับค่าจริงมากที่สุด รองลงมาจะเป็นวิธีต้นไม้ตัดสินใจ ส่วนวิธีการถดถอยลอจิสติกจะมีความผิดพลาดค่อนข้างสูง

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องในการทำนายการอนุมัติสินเชื่อส่วนบุคคล นอกจากจะแสดงตัวอย่างผลการทำนายแล้วจะแสดงในส่วนของค่าความถูกต้อง ความแม่นยำ และค่าความไว เพื่อเป็นการเปรียบเทียบประสิทธิภาพของแบบจำลองดังตารางที่ 4.2

ตารางที่ 4.2 ผลการทำนายการอนุมัติสินเชื่อส่วนบุคคลของแบบจำลอง

	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความไว
วิธีต้นไม้ตัดสินใจ	93.93%	94%	95%
วิธีป่าสุ่ม	95.92%	94%	98%
วิธีการวิเคราะห์การถดถอยลอจิสติก	65.28%	62%	81%

จากตารางที่ 4.2 แสดงผลการทำนายการอนุมัติสินเชื่อส่วนบุคคลของแบบจำลองวิธีป่าสุ่ม วิธีต้นไม้ตัดสินใจและวิธีการวิเคราะห์การถดถอยลอจิสติก พบว่าแบบจำลองวิธีป่าสุ่มมีค่าความถูกต้องอยู่ที่ 95.92% ค่าความแม่นยำอยู่ที่ 94% และค่าความไวอยู่ที่ 98% รองลงมาคือวิธีต้นไม้ตัดสินใจ มีค่าความถูกต้องอยู่ที่ 93.93% ค่าความแม่นยำอยู่ที่ 94% และค่าความไวอยู่ที่ 95% ส่วนวิธีการวิเคราะห์การถดถอยลอจิสติก มีค่าความถูกต้องอยู่ที่ 65.28% ค่าความแม่นยำอยู่ที่ 62% และค่าความไวอยู่ที่ 81% จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีป่าสุ่มให้ค่าประสิทธิภาพการทำนายดีที่สุด โดยแสดงเมทริกซ์ความสับสน (Confusion Matrix) เพิ่มเติม ดังตารางที่ 4.3

ตารางที่ 4.3 เมทริกซ์ความสับสนของแบบจำลองวิธีป่าสุ่ม

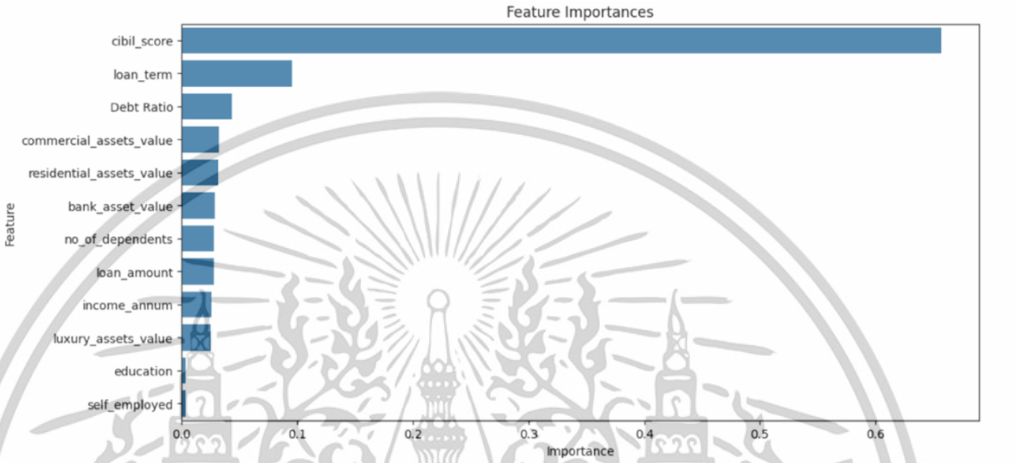
	ทำนายอนุมัติสินเชื่อ	ทำนายไม่อนุมัติสินเชื่อ
อนุมัติสินเชื่อ	519	11
ไม่อนุมัติสินเชื่อ	34	490

จากตารางที่ 4.3 เป็นตารางเมทริกซ์ความสับสน แสดงความถี่ของผลลัพธ์ที่เกิดขึ้นจริงและผลลัพธ์ที่ได้จากค่าทำนายซึ่งเป็นตารางที่แสดงถึงประสิทธิภาพของแบบจำลองวิธีป่าสุ่ม โดยในการอนุมัติสินเชื่อแบบจำลองสามารถทำนายได้ถูกต้อง 519 ราย จาก 530 ราย แบบจำลองทำนายผิดพลาด 11 ราย จาก 530 ราย คิดเป็นร้อยละ 2.1 ส่วนการไม่อนุมัติสินเชื่อแบบจำลองสามารถทำนายได้ถูกต้อง 490 ราย จาก 524 ราย แบบจำลองทำนายผิดพลาด 34 ราย จาก 524 ราย คิดเป็นร้อยละ 6.5 ซึ่งจะเห็นว่าแบบจำลองมีความผิดพลาดในการทำนายน้อยมากเมื่อเทียบกับการทำนายที่ถูกต้อง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

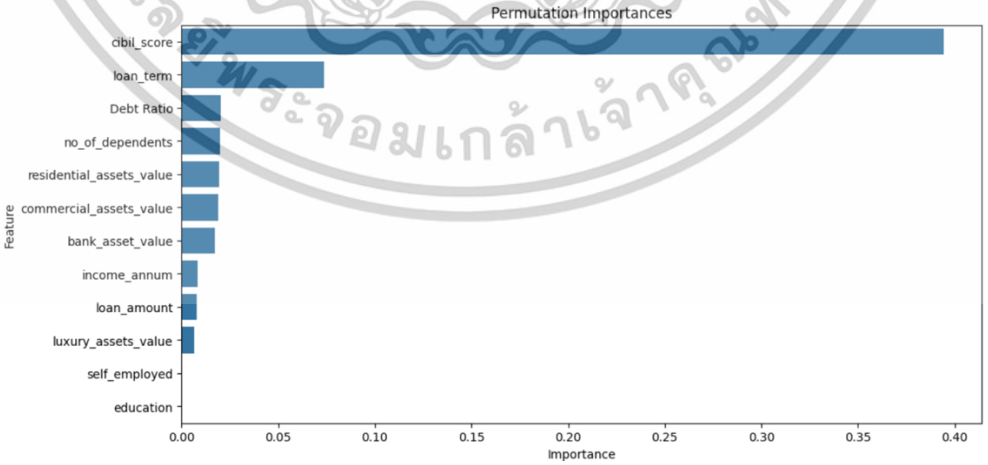
4.2 ผลการวิเคราะห์ของการหาความสำคัญของตัวแปร

จากการนำแบบจำลองการเรียนรู้ของเครื่องทั้ง 3 วิธีไปทำนายการอนุมัติสินเชื่อส่วนบุคคล พบว่าแบบจำลองวิธีป่าสุ่มมีประสิทธิภาพสูงที่สุด ทางผู้วิจัยจึงได้ศึกษาต่อเกี่ยวกับตัวแปรที่มีความสำคัญต่อผลของแบบจำลอง โดยใช้ค่าฟีเจอร์ อิมพอร์แทนเซส แสดงดังรูปที่ 4.1 และค่าเพอร์มิวเทชัน อิมพอร์แทนซ์ แสดงดังรูปที่ 4.2



รูปที่ 4.1 ค่าความสำคัญของตัวแปรที่แสดงด้วยฟีเจอร์ อิมพอร์แทนเซส

จากรูปที่ 4.1 แสดงค่าฟีเจอร์ อิมพอร์แทนเซสของตัวแปรแต่ละตัว ยิ่งค่ามากแสดงว่ามีความสำคัญต่อแบบจำลองมาก ซึ่งจากรูปจะเห็นว่าตัวแปรที่มีความสำคัญต่อแบบจำลองมากที่สุด 3 อันดับแรกคือ คะแนนเครดิต ระยะเวลาเงินกู้ และสัดส่วนระหว่างหนี้สินกับสินทรัพย์รวม ตามลำดับ จึงสรุปได้ว่าปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อส่วนบุคคลมากที่สุด คือ คะแนนเครดิต



รูปที่ 4.2 ค่าความสำคัญของตัวแปรที่แสดงด้วยเพอร์มิวเทชัน อิมพอร์แทนซ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 4.2 แสดงค่าเพอร์มิวเทชัน อิมพอร์แทนซ์ของตัวแปรแต่ละตัว จากรูปจะเห็นว่าตัวแปรที่มีค่ามากที่สุดที่แสดงถึงความสำคัญต่อแบบจำลองมากที่สุด 3 อันดับแรกคือ คะแนนเครดิต ระยะเวลาเงินกู้ และสัดส่วนระหว่างหนี้สินกับสินทรัพย์รวม ตามลำดับ จึงสรุปได้ว่าปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อส่วนบุคคลมากที่สุด คือ คะแนนเครดิตเช่นเดียวกับการหาค่าด้วยค่าฟีเจอร์ อิมพอร์แทนเซส

จากการหาปัจจัยที่มีความสำคัญต่อแบบจำลองมากที่สุดด้วยค่า 2 ค่า พบว่าทั้งค่าฟีเจอร์ อิมพอร์แทนเซส และค่าเพอร์มิวเทชัน อิมพอร์แทนซ์ มีตัวแปรที่มีความสำคัญต่อแบบจำลองตรงกัน 3 อันดับแรก นั่นก็คือ คะแนนเครดิต ระยะเวลาเงินกู้ และสัดส่วนระหว่างหนี้สินกับสินทรัพย์รวม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาการทำนายการอนุมัติสินเชื่อส่วนบุคคลโดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง โดยงานวิจัยมีวัตถุประสงค์ดังนี้ เพื่อนำการเรียนรู้ของเครื่องมาใช้ในการทำนายการอนุมัติสินเชื่อส่วนบุคคล เพื่อศึกษาระดับความสำคัญของปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อโดยใช้ค่าฟีเจอร์ อิมพอร์แทนเซส และเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในการอนุมัติสินเชื่อโดยใช้ค่าความถูกต้อง ค่าความแม่นยำ และค่าความไว

การศึกษาครั้งนี้ได้นำข้อมูลมาจากแคกเกิล เป็นชุดข้อมูลการอนุมัติเงินกู้ซึ่งรวบรวมบันทึกทางการเงินและข้อมูลที่เกี่ยวข้องที่ใช้เพื่อกำหนดคุณสมบัติของบุคคลในการได้รับเงินกู้จากสถาบันให้กู้ยืม และทำการเพิ่มข้อมูลเนื่องจากข้อมูลมีความไม่สมดุล และทำการเตรียมข้อมูล เพื่อจะไปสู่ขั้นตอนของการสร้างแบบจำลอง จากนั้นทำการเปรียบเทียบประสิทธิภาพแบบจำลอง และหาปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อ ซึ่งสามารถแสดงการสรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1 สรุปผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพการทำนายการอนุมัติเงินกู้ของแบบจำลองทั้งหมด 3 วิธี พบว่าแบบจำลองวิธีป่าสุ่มมีประสิทธิภาพการทำนายสูงที่สุด ซึ่งมีค่าความถูกต้องอยู่ที่ 95.92% ความแม่นยำอยู่ที่ 94% และความไวอยู่ที่ 98% ดังนั้นแบบจำลองวิธีป่าสุ่มมีความเหมาะสมที่จะนำมาใช้ในการทำนายการอนุมัติสินเชื่อ และผลการศึกษาปัจจัยที่ส่งผลต่อการอนุมัติสินเชื่อ 3 อันดับ พบว่าคะแนนเครดิตเป็นปัจจัยที่มีระดับความสำคัญมากที่สุด เนื่องจากคะแนนเครดิตสะท้อนถึงความน่าเชื่อถือทางการเงินของบุคคล ไม่ว่าจะเป็นประวัติการชำระหนี้ จำนวนหนี้สิน ประวัติการใช้บัตรเครดิต ปัจจัยที่มีระดับความสำคัญรองลงมาคือ ระยะเวลาเงินกู้ กล่าวได้ว่ายิ่งมีระยะเวลาเงินกู้น้อยจะเพิ่มโอกาสในการอนุมัติเงินกู้มากขึ้น ปัจจัยที่มีระดับความสำคัญลำดับที่ 3 คือ อัตราส่วนหนี้สิน ซึ่งเป็นสัดส่วนระหว่างหนี้สินกับสินทรัพย์รวมของบุคคล โดยค่านี้จะแสดงถึงความสามารถในการชำระหนี้ของผู้กู้ จึงนับเป็นอีกหนึ่งตัวแปรที่มีความสำคัญต่อการอนุมัติสินเชื่อ

จากการนำแบบจำลองทั้ง 3 วิธีมาเปรียบเทียบประสิทธิภาพพบว่าวิธีป่าสุ่มเป็นวิธีที่มีประสิทธิภาพสูงสุดซึ่งสอดคล้องกับวิจัยของ Bhavana et al., (2022) ที่ศึกษาการประเมินความเสี่ยงในการขอสินเชื่อโดยใช้วิธีป่าสุ่ม, อัลกอริทึมซัพพอร์ตวอกเตอร์แมชชีน, วิธีเพื่อนบ้านใกล้สุด, และการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

วิเคราะห์การถดถอยลอจิสติก ผลการศึกษาพบว่าวิธีป่าสุ่มมีค่าความถูกต้องสูงที่สุดและวิจัยของ Bhadola et al., (2021) ที่เปรียบเทียบความแม่นยำและประสิทธิภาพของแบบจำลองการถดถอยลอจิสติก, ต้นไม้การตัดสินใจและวิธีป่าสุ่ม สำหรับการทำนายการอนุมัติเงินกู้ ผลการศึกษาพบว่าวิธีป่าสุ่มมีค่าความถูกต้องสูงที่สุด แต่ขัดแย้งกับวิจัยของ Kavya et al., (2020) เนื่องจากวิจัยเล่มนี้ได้ศึกษาการคาดการณ์การอนุมัติเงินกู้โดยใช้แบบจำลองการถดถอยโลจิสติก ต้นไม้การตัดสินใจและวิธีป่าสุ่ม แต่ผลการศึกษาพบว่าวิธีต้นไม้การตัดสินใจเป็นวิธีที่มีความแม่นยำมากที่สุด จากความขัดแย้งดังกล่าวอาจมีปัจจัยหลายประการที่ส่งผลต่อผลลัพธ์ของการวิจัยดังนี้

1. ความแตกต่างของชุดข้อมูลที่ใช้ในการทดสอบ กล่าวได้ว่าข้อมูลที่มีความซับซ้อนและมีลักษณะเฉพาะอาจทำให้แบบจำลองบางแบบทำงานได้ดีกว่า
2. การปรับจูนพารามิเตอร์ของโมเดล การปรับค่าพารามิเตอร์อย่างเหมาะสมสามารถส่งผลต่อประสิทธิภาพของแบบจำลองได้
3. ขนาดของข้อมูล ข้อมูลขนาดเล็กหรือใหญ่สามารถส่งผลต่อการเลือกใช้แบบจำลองที่เหมาะสม

5.2 ข้อเสนอแนะ

1. เนื่องจากข้อมูลที่ใช้ในการวิจัยนี้มาจากแคกเกิล ซึ่งเป็นแหล่งข้อมูลที่มีคุณภาพแต่ยังคงมีข้อจำกัดในด้านความหลากหลายและความสมบูรณ์ของข้อมูล ดังนั้นหากมีการใช้ข้อมูลจากสถาบันการเงินก็จะส่งผลให้ข้อมูลมีความสมบูรณ์และสามารถเพิ่มประสิทธิภาพในการทำนายได้ อีกทั้งยังทำให้สามารถปรับตัวต่อสภาพเศรษฐกิจที่เปลี่ยนแปลงได้ดียิ่งขึ้น

2. ข้อมูลมีจำนวนค่อนข้างน้อย จึงทำให้ข้อมูลสำหรับใช้ในการเรียนรู้ไม่เพียงพอ ดังนั้นหากสามารถเพิ่มจำนวนข้อมูลอาจเพิ่มประสิทธิภาพในการทำนายผลการเรียนของแบบจำลองให้ดียิ่งขึ้น

3. ศึกษาแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกเพิ่มเติม เพื่อเพิ่มตัวเลือกในการเลือกแบบจำลองและทำนายการอนุมัติเงินกู้ให้มีประสิทธิภาพยิ่งขึ้น

4. เลือกใช้แบบจำลองให้เหมาะสมกับชุดข้อมูล และเลือกเฉพาะตัวแปรที่มีความสำคัญต่อแบบจำลองเพื่อให้การทำนายมีความแม่นยำมากที่สุด จากการศึกษาครั้งนี้พบว่าแบบจำลองการถดถอยลอจิสติกมีค่าความถูกต้องน้อยมาก สาเหตุมาจากแบบจำลองมีจำนวนตัวแปรมากเกินไปส่งผลให้แบบจำลองทำนายไม่แม่นยำเท่าที่ควร ดังนั้นควรมีการศึกษาแบบจำลองต่างๆ เพิ่มเติมเพื่อให้เหมาะสมกับชุดของข้อมูลมากขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

ชณิดาภา บุญประสม และจรัญ แสนราช. (2561). “การวิเคราะห์การทำนายการลาออกกลางคันของนักศึกษา.”วารสารวิชาการครุศาสตร์อุตสาหกรรม พระจอมเกล้าพระนครเหนือ, ฉบับที่ 9, 143-150.

ชุภาดา ผิวเผือก. (2562). “ปัจจัยที่ใช้ในการวิเคราะห์สินเชื่อของธนาคารพาณิชย์ในประเทศไทย” วิทยานิพนธ์มหาบัณฑิต, สาขาวิชาการบัญชี, คณะบัญชี, มหาวิทยาลัยธุรกิจบัณฑิตย์.

ปารดา ศัสตุระ. (2565). “การประยุกต์เหมืองข้อมูลในการพยากรณ์สถานะ NPLs ของลูกค้าสินเชื่อกรณีศึกษา ธนาคารออมสิน สาขามหาวิทยาลัยเชียงใหม่” วิทยานิพนธ์มหาบัณฑิต คณะบริหารธุรกิจ, มหาวิทยาลัยหอการค้าไทย.

ปรเมษฐ์ ธันวานนท์, ชัยกรยิ่ง เสรี, วรพล พงษ์เพ็ชร, และธนภัทร ฆังคะจิตรสกสรรค์. (2560). “การประยุกต์ใช้โมเดลการเรียนรู้แบบรวมกลุ่มเพื่อพยากรณ์แนวโน้มของราคาหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทย”. JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY, 7, 12-21.

ยุทธ ไกยวรรณ์. (2555). “หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย.” วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิจัย, ฉบับที่ 1, 1-12.

สถาบันข้อมูลขนาดใหญ่ (องค์การมหาชน). (2564). **การวิเคราะห์ความสัมพันธ์กับข้อมูลขนาดใหญ่.** (ออนไลน์) เข้าถึงได้ที่ https://bdi.or.th/big-data-101/correlation-analysis-in-big-data/

สุเมธ จุฑาจันทร์ และสมพร ปั่นโภชา. (2564). “เปรียบเทียบผลลัพธ์ของการอนุมัติสินเชื่อด้วย 3 แบบจำลองของระเบียบวิธี Machine Learning โดยใช้โปรแกรมอาร์” วิทยานิพนธ์มหาบัณฑิต สาขาวิศวกรรมการเงิน, มหาวิทยาลัยหอการค้าไทย.

อรพิน ประวัติบริสุทธิ์. (2564). **Python สำหรับงาน Data Science Data Visualization และ Machine Learning**, โปรวิชั่น, กรุงเทพฯ.

Bhadola, E., Khan, A., Kumar, A., & Singh, N., (2021). Loan Approval Prediction Model A Comparative Analysis. Advances and Applications in Mathematical Sciences, 20(3), 427-435.

Bhavana, S., Lakshmi, B., Loukya, M., Saibaba, C. M. H., Sunanda, N., & Tumuluru, P., (2022). Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. Proceedings of the Second International Conference on Artificial Intelligence and Smart Energy (ICAIS-2022), DOI: 10.1109/ICAIS53314.2022.9742800.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrance O., & Kegelmeyer, W., Philip. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.)

Chavan, P., Diwate, Y., & Chavan, P., (2021). Loan Approval Prediction Using Machine Learning. International Research Journal of Engineering and Technology (IRJET), 08(05), 1741-1745.

Kavya, M., Maddumala, V., Ramya, R., Tejaswini, J., & Triveni, P., (2020). Journal of Engineering Science, 11(04), 523-532.

Kulkarni, A., Feras, A., Batarseh & Chong, D. (2020). Foundations of data imbalance and solutions for a data democracy. Data Democracy at the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering, 83-106.

Mahdi, H., Mohammad, A., & Muhammed, J.A. (2021). Credit Approval System Using Machine Learning. Challenges and Future Directions Computer Science and Engineering International Islamic University Chittagong, Bangladesh.

Sasiwut Chaiyadecha. (2020). **One-Hot Encoding สร้างตัวแปร Dummies สำหรับ Classification Model.** [Online]. Available: https://lengyi.medium.com/one-hot-encoding-737c66e5b1bd.

Witchapong Daroontham. (2023). **เจาะลึก Random Forest Part 2 of รู้จัก Decision Tree, Random Forest, และ XGBoost.** [Online]. Available: https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f4 1a1c1c

Yengejeh, A., (2023). **Analysis of Credit Approval by Decision Tree Analysis of Credit Approval by Decision Tree**. Data Science and Data Mining, University of Central Florida, May 2023. [Online]. Available: https://stars.library.ucf.edu/data-science-mining/5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นางสาวพรรษมน มะโนมัย
วัน เดือน ปีเกิด	09 มิถุนายน 2542
ที่อยู่ปัจจุบัน	3763/48 ซอยเจริญราษฎร์ 7 ถนนจันทน์ แขวงบางโคล่ เขตบางคอแหลม กรุงเทพ 10120
ประวัติการศึกษา	(2564) เศรษฐศาสตรบัณฑิต สาขาเศรษฐศาสตร์ เกรดเฉลี่ย 3.49 (มหาวิทยาลัยศรีนครินทรวิโรฒ) (2566) วิทยาศาสตร์มหาบัณฑิต สาขาวิทยาการข้อมูลและการวิเคราะห์ (สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง) เกรดเฉลี่ย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้