ระบบทำนายการจัดชั้นสินเชื่อที่อยู่อาศัยโดยใช้แบบจำลองการเรียนรู้ของเครื่อง

ประพลเวท บุญประเสริฐ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2564

HOUSING-LOAN CLASSIFICATION SYSTEM USING MACHINE LEARNING MODELS

PRAPONWET BUNPRASERT

A Thematic Paper Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

2021

ใบรับรองงานสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อสารนิพนธ์ ระบบทำนายการจัดชั้นสินเชื่อที่อยู่อาศัยโดยใช้แบบจำลองการเรียนรู้ของเครื่อง

เสนอโดย นายประพลเวท บุญประเสริฐ

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.ธนภัทร ฆังคะจิตร

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

.....ประธานกรรมการ

(ดร.สรรพฤทธิ์ มฤคทัต)

.....กรรมการและอาจารย์ที่ปรึกษา

(ดร.ธนภัทร ฆังคะจิตร)

.....กรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ดร.ชัยพร เขมะภาตะพันธ์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ ...31... เดือน ...กรกฎาคม... พ.ศ. ...2564...

หัวข้อสารนิพนธ์	ระบบทำนายการจัดชั้นสินเชื่อที่อยู่อาศัยโดยใช้แบบจำลองการเรียนรู้ของเครื่อง
ชื่อผู้เขียน	ประพลเวท บุญประเสริฐ
อาจารย์ที่ปรึกษา	ดร. ธนภัทร ฆังคะจิตร
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2563

บทคัดย่อ

เนื่องจากรายได้หลักของสถาบันการเงินมาจากเงินให้สินเชื่อ ดังนั้นความสามารถในการชำระหนี้ของผู้ขอสินเชื่อย่อมมีความสำคัญต่อผลประกอบการของธนาคาร การมีเครื่องมือบริหารความเสี่ยงสินเชื่อที่เหมาะสม สำหรับใช้ในการประเมินความสามารถในการชำระหนี้ของผู้ขอสินเชื่อ จึงเป็นสิ่งที่ธนาคารต้องให้ความสำคัญ ทั้งนี้เกณฑ์ที่ใช้ต้องสอดคล้องกับมาตรฐานการกำกับดูแลจาก ธปท. ในฐานะผู้กำกับดูแลสถาบันการเงิน ดังนั้นการพัฒนาโมเดลเพื่อเป็นเครื่องมือในการวิเคราะห์คุณภาพสินเชื่อที่อยู่อาศัยจากการเปลี่ยนแปลงสถานะการจัดชั้น จากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ซึ่งถือเป็นสินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL) จะเป็นตัวช่วยในการวิเคราะห์คุณภาพของสินเชื่อให้สอดคล้องกับเป้าหมายในการรักษาเสถียรภาพของระบบสถาบันการเงินไทย ผู้เขียนจึงได้พัฒนาระบบทำนายการจัดชั้นสินเชื่อที่อยู่อาศัย โดยใช้ Machine Learning การเรียนรู้ได้จากตัวอย่างด้วยตนเอง แบบ Ensemble Machine learning (Voting Classifier) เพื่อใช้เป็น Smart Indicator ในการติดตามคุณภาพของสินเชื่อที่อยู่อาศัย ซึ่งในงานศึกษานี้ใช้ข้อมูลสินเชื่อประเภทที่อยู่อาศัย ในการพัฒนา Finance Classification Model เพื่อให้สามารถแยกแยะสินเชื่อดี – เสียออกจากกันได้ ธนาคารจะสามารถใช้ Smart Indicator นี้ ร่วมกับเครื่องมือบริหารความเสี่ยงอื่นๆ เพื่อจัดการความเสี่ยงได้ก่อนที่จะเกิดเหตุการณ์การผิดนัดชำระหนี้ และ/หรือ เพื่อประกอบการตัดสินใจในการดำเนินนโยบายบริหารความเสี่ยงที่เหมาะสม ผลจากการทดสอบความแม่นยำพบว่าระบบที่นำเสนอให้ความถูกต้องในการทำนายพฤติกรรมที่เป็นสินเชื่อที่ไม่ก่อเกิดรายได้ (NPL) โดยให้ผล Recall เฉลี่ยมากถึง 95 % จะส่งผลให้ใช้ติดตามและประเมินความสามารถในการชำระหนี้ของสินเชื่อเพื่อทำให้คุณภาพสินเชื่อประเภทที่อยู่อาศัยของธนาคารได้ดียิ่งขึ้น

Thematic Paper Title	HOUSING-LOAN CLASSIFICATION SYSTEM USING MACHINE LEARNING MODELS
Author	Praponwet Bunprasert
Thesis Advisor	Dr. Thanapat Kangkachit
Department	Big Data Engineering
Academic Year	2020

ABSTRACT

This is because loans are the primary source of income for financial institutions. As a result, the ability of the loan applicant to repay debt is critical to the bank's performance. Having appropriate credit risk management tools for use in assessing loan applicants' repayment ability. As a result, the Bank must pay close attention to this issue. The criteria used must be consistent with the BOT's regulatory standards as a financial institution regulator. As a result of the change in classification status, the development of a model to be used as a tool for analyzing mortgage loan quality. From normal class debtor to default debtor, which is considered a non-performing loan (NPL), credit quality will be analyzed in accordance with the goals of stabilizing the Thai financial institution system. As a result, the authors created a mortgage classification prediction system using machine learning, self-learning from the Ensemble Machine learning (Voting Classifier) model, to be used as a smart indicator for monitoring mortgage loan quality. The data from home loans was used in this study. To create a Finance Classification Model in order to distinguish between good and bad loans. This Smart Indicator will be used by the bank in conjunction with other risk management tools. To manage risks prior to the occurrence of a debt default and/or to make decisions about the implementation of appropriate risk management policies. The accuracy test results showed that the proposed system is accurate in predicting non-performing credit behavior (NPL) with an average recall of 95%. The ability to repay loans in order to improve the quality of mortgage loans.

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ โดยการให้ความช่วยเหลือ ของ ดร.ธนภัทร ฆังคะจิตรซึ่งเป็นอาจารย์ที่ปรึกษาสารนิพนธ์ ที่ได้กรุณาให้คำแนะนำ ติดตาม ตรวจสอบ ให้กำลังใจ และแก้ไขข้อบกพร่องต่าง ๆ มาโดยตลอด ไม่ถอดใจ เป็นห่วง ผลักดัน สละเวลาดึกดื่นอันมีค่าของท่าน เพื่อให้สารนิพนธ์ฉบับนี้สมบูรณ์ ผู้เขียนจึงขอกราบขอบพระคุณอย่างสูงไว้ ณ โอกาสนี้

ผู้เขียนขอกราบขอบพระคุณ การให้ความช่วยเหลือ ผู้ช่วยศาสตราจารย์ ดร.วรพล พงษ์เพ็ชร ดร.เอกสิทธิ์ พัชรวงศ์ศักดา ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น อ.อ้อม อ.แก้ว ซึ่งเป็นอาจารย์ที่ได้กรุณาให้ความรู้ คำแนะนำ กำลังใจ คำปรึกษาต่างๆ มาโดยตลอด ทำผู้เขียนดำเนินแนวทางในการใช้ความรู้ความสามารถเพื่อให้สารนิพนธ์ฉบับนี้สมบูรณ์ จึงขอกราบขอบพระคุณไว้ ณ โอกาสนี้

ผู้เขียนขอกราบขอบพระคุณ นางสาวกุลธิดา รอดบุญ รวมถึงเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิตย์ เพื่อนนักศึกษาทุกท่านที่ให้ความสนิทสนม แบ่งปันมิตรภาพ ความรู้รอยยิ้มที่แลกเปลี่ยนซึ่งกันและกันให้กับผู้เขียน ทำให้การจัดทำสารนิพนธ์ของผู้เขียนในครั้งนี้สำเร็จลุล่วงไปด้วยดี

ผู้เขียนขอกราบขอบพระคุณ หลักสูตรวิศวกรรมข้อมูลขนาดใหญ่ของมหาวิทยาลัยธุรกิจบัณฑิตย์ และมหาวิทยาลัยธุรกิจบัณฑิตย์ ครอบครัวของผู้เขียน ที่สร้างสภาพแวดล้อมต่างๆ ทำให้กับผู้เขียนมีพลังที่จะลุก และผลักดันตนเองให้สามารถจัดทำสารนิพนธ์ของผู้เขียนในครั้งนี้ได้สำเร็จ

ประพลเวท บุญประเสริฐ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

รายได้ดอกเบี้ยจากเงินให้สินเชื่อนับเป็นแหล่ง รายได้หลักของธนาคารพาณิชย์ไทย จากข้อมูลผลการ ดำเนินงานของระบบธนาคารพาณิชย์ไทย พบว่า ประมาณร้อยละ 56 ของรายได้ทั้งหมดมาจาก รายได้ดอกเบี้ยจากเงินให้สินเชื่อ ซึ่งแสดงให้เห็นถึง ความสำคัญของความสามารถในการชำระหนี้ที่ มีต่อผลประกอบการของธนาคารดังนั้น การมีเครื่องมือบริหารความเสี่ยงที่เหมาะสมเพื่อใช้ในการ ติดตามและประเมินความสามารถในการชำระหนี้ของลูกค้าจึงเป็นสิ่งที่ธนาคารต้องให้ความสำคัญ (ที่มา สายกำกับสถาบันการเงิน และกลุ่มงานดาต้าอนาไลติกส์ ธนาคารแห่งประเทศไทย)

งานสารนิพนธ์นี้จึงมีวัตถุประสงค์เพื่อพัฒนาโมเดลในการเป็นผู้ช่วยการวิเคราะห์คุณภาพสินเชื่อที่อยู่อาศัยจากการเปลี่ยนแปลง สถานการณ์จัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ซึ่งสินเชื่อที่ไม่ก่อให้เกิดรายได้)Non-performing loan: NPL) ถือเป็นเครื่องชี้หลักสำหรับวิเคราะห์และ ติดตามคุณภาพสินเชื่อของสถาบันการเงินที่ใช้กัน อย่างแพร่หลาย โดยการติดตามคุณภาพสินเชื่อในเบื้องต้น ผลที่ได้จากงานสารนิพนธ์นี้จะทำให้สามารถพยากรณ์พฤติกรรมที่สินเชื่อที่อยู่อาศัย จากสถานการณ์จัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ทำให้สามารถลดโอกาสเกิดภาระสินเชื่อที่ไม่ก่อให้เกิดรายได้ อีกทั้งยังเป็นเครื่องมือในการการติดตามคุณภาพสินเชื่อของสถาบันการเงินได้ทันท่วงที ภายใต้นิยามของการผิดนัดชำระให้สอดคล้องกับสถานการณ์และลักษณะเฉพาะของ สินเชื่อที่ต่างประเภทกันได้

1.2 วัตถุประสงค์ของงานสารนิพนธ์

เพื่อนำเสนอวิธีการในการพยากรณ์พฤติกรรมสินเชื่อที่อยู่อาศัย จากสถานะการจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้

1.3 ขอบเขตงานสารนิพนธ์

1.3.1 สินเชื่อสำหรับประเภทที่อยู่อาศัยของธนาคาร

1.3.2 พยากรณ์พฤติกรรมสินเชื่อประเภทที่อยู่อาศัย จากสถานการณ์จัดชั้นจากลูกหนี้ชั้นปกติ ไปเป็นลูกหนี้ผิดนัดชำระหนี้ภายใน 1 เดือนข้างหน้า

1.3.3 เป็นส่วนประกอบหนึ่งของระบบงานที่ใช้สำหรับการเฝ้าระวังการตกชั้นของสินเชื่อประเภทที่อยู่อาศัย

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 เพื่อช่วยติดตาม ป้องกันสินเชื่อที่อยู่อาศัย จากสถานการณ์จัดชั้นจากลูกหนี้ชั้นปกติ ไปเป็นลูกหนี้ผิดนัดชำระหนี้ภายใน 1 เดือนข้างหน้า

1.5.2 เพื่อเป็นเครื่องมือในการช่วยเพิ่ม ประสิทธิภาพและพัฒนาแนวทางการบริหารสินเชื่อที่อยู่อาศัยของสถาบันการเงิน

1.5 นิยามศัพท์

1.5.1 **NPL** หมายถึง สินเชื่อที่ไม่ก่อให้เกิดรายได้

1.5.2 **Finance classification** หมายถึง วิธีปฏิบัติเกี่ยวกับการจัดชั้นหนี้ ประกอบด้วย

1.5.2.1 ลูกหนี้จัดชั้นปกติ (ระยะเวลาค้างชำระ < 1 เดือน)

1.5.2.2 สินเชื่อจัดชั้นกล่าวถึงเป็นพิเศษ (1 เดือน < ระยะเวลาค้างชำระ < 3 เดือน)

1.5.2.3 สินเชื่อจัดชั้นต่ำกว่ามาตรฐาน (3 เดือน < ระยะเวลาค้างชำระ < 6 เดือน)

1.5.2.4 สินเชื่อจัดชั้นสงสัย (6 เดือน < ระยะเวลาค้างชำระ < 12 เดือน)

1.5.2.5 สินเชื่อจัดชั้นสงสัยจะสูญ (ระยะเวลาค้างชำระ > 12 เดือน)

1.5.2.6 สินเชื่อจัดชั้นสูญ

1.5.3 **Imbalanced Data** หมายถึง ข้อมูลคำตอบของแต่ละคลาสมีจำนวนไม่เท่ากัน

1.5.4 **Under-sampling** หมายถึง การสุ่มข้อมูลจาก majority class ให้มีจำนวนน้อยลง

1.5.5 **Over-sampling** หมายถึง การสร้างข้อมูลของ minority class ให้มีจำนวนมากขึ้น

1.5.6 **MSE (Mean squared error)** หมายถึง ค่าเฉลี่ยของความคลาดเคลื่อนยกกำลังสอง

1.5.7 **Accuracy** คือค่าความแม่นยำของโมเดลที่ใช้ในการพยากรณ์

1.5.8 **Precision** คือค่าที่บอกว่าโมเดลพยากรณ์ได้ว่า จริง ถูกต้องเท่าไหร่

1.5.9 **Recall** คือค่าที่โมเดลพยากรณ์ได้ว่า จริง เป็นอัตราส่วนเท่าไหร่เทียบกับของจริงทั้งหมด

บทที่ 2
ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

งานสารนิพนธ์เรื่องนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลในการเป็นผู้ช่วยการวิเคราะห์คุณภาพสินเชื่อที่อยู่อาศัยจากการเปลี่ยนแปลง สถานะการจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ซึ่งสินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL) ด้วยการประยุกต์ใช้ระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเอง (Machine Learning) โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังรายการต่อไปนี้

2.1 Machine Learning การเรียนรู้ได้จากตัวอย่างด้วยตนเอง

2.2 Logistic Regression เทคนิคการวิเคราะห์สถิติเชิงคุณภาพ

2.3 Decision Tree เทคนิคต้นไม้ตัดสินใจ

2.4 Random Forest เทคนิคการใช้ต้นไม้ตัดสินใจหลายตัว ทำนายพร้อมกัน

2.5 AdaBoost Forest เทคนิคการใช้ต้นไม้ตัดสินใจหลายตัว ทำนายต่อกัน

2.6 SVM เทคนิคการใช้เส้นแบ่งในการทำนาย

2.7 KNN เทคนิคการใช้เพื่อนบ้านใกล้ที่สุด

2.8 Naive Bayes เทคนิคจัดหมวดหมู่โดยใช้หลักความน่าจะเป็น

2.9 VotingClassifier เทคนิคการการเลือกคำทำนายจากหลาย Model แบบ Vote

2.10 ตัววัดประสิทธิภาพของโมเดล (Confusion Matrix)

2.11 Finance Classification วิธีปฏิบัติเกี่ยวกับการจัดชั้นหนี้

2.12 งานวิจัยที่เกี่ยวข้อง

2.1 Machine Learning

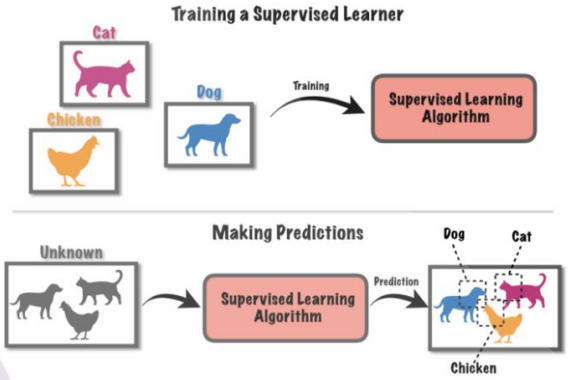
การเรียนรู้ได้จากตัวอย่างด้วยตนเองของเครื่อง ซึ่งจะสามารถถูกแบ่งออกเป็นการเรียนรู้ได้ 2 แบบใหญ่ ๆ ได้แก่ การเรียนรู้แบบผู้สอน(Supervised Learning) และ การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยมีรายละเอียดในแต่ละการเรียนรู้ดังนี้

- Supervised Learning

มีประเภทของ Supervised learning อยู่ 2 ประเภท

a. การแบ่งแยกประเภท (Classification)

b. การถดถอย (Regression)

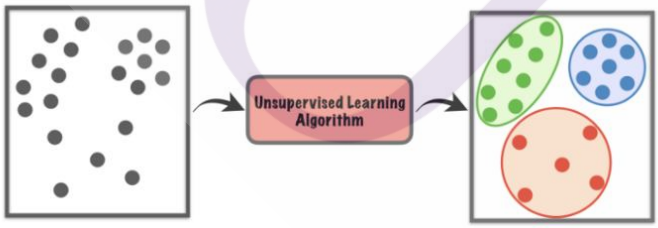


ภาพที่ 2.1 ตัวอย่างของแบบจำลองการเรียนรู้แบบ Supervised Learning

ที่มา: https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242

- Unsupervised learning

อัลกอริทึมจะตรวจสอบเฉพาะข้อมูลที่ป้อนเข้ามาเท่านั้น โดยปราศจากการให้ผลลัพธ์ที่จะเกิดขึ้น (เช่น การสำรวจข้อมูลประชากรเพื่อแบ่งกลุ่มของข้อมูลนั้น)

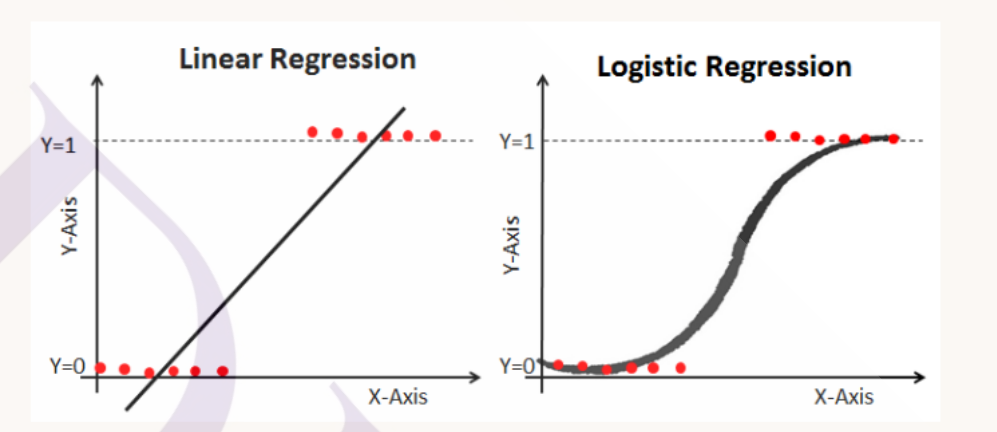


ภาพที่ 2.2 ตัวอย่างของแบบจำลองการเรียนรู้แบบ Unsupervised learning

ที่มา: https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242

2.2 Logistic Regression Classification

เป็นการสร้างสมการคณิตศาสตร์เพื่อแบ่งแยก (classify) ข้อมูลออกเป็น 2 กลุ่มคำตอบ เทคนิคนี้เป็นอีกเทคนิคที่นิยมให้เนื่องจากแปลความโมเดลได้ง่ายและแสดงให้เห็นถึงความสำคัญของตัวแปร (หรือ Feature) ดังภาพ



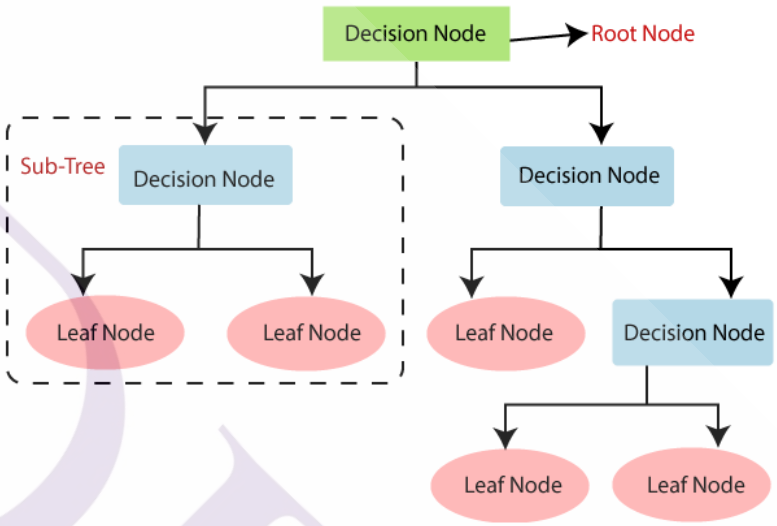
ภาพที่ 2.3 ตัวอย่างของเทคนิค Logistic Regression Classification

ที่มา: https://datacubeth.ai/machine-learning-model-comparison/

https://nonthakon.medium.com/

2.3 Decision Tree Classification

เป็นเทคนิคที่สร้างโมเดลแยกข้อมูลลงมาเสมือนเป็นต้นไม้และกิ่งก้าน ตามลำดับชั้นต่างๆ จนถึงชั้นล่างสุดถึงเป็นคลาสคำตอบ ดังภาพ



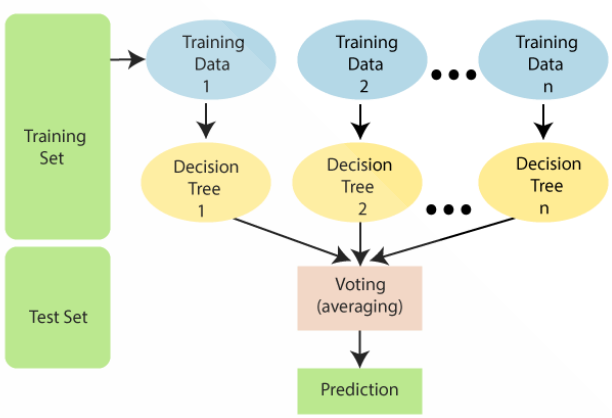
ภาพที่ 2.4 ตัวอย่างของเทคนิค Decision Tree Classification

ที่มา: https://datacubeth.ai/machine-learning-model-comparison/

https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

2.4 Random Forest Classification

เป็นเทคนิคที่สร้างโมเดล Decision Tree ขึ้นมาหลายๆ ต้นจากการสุ่ม (random) ข้อมูล ซึ่งเป็นเทคนิคหนึ่งของการสร้างโมเดลแบบ Ensemble ดังภาพ



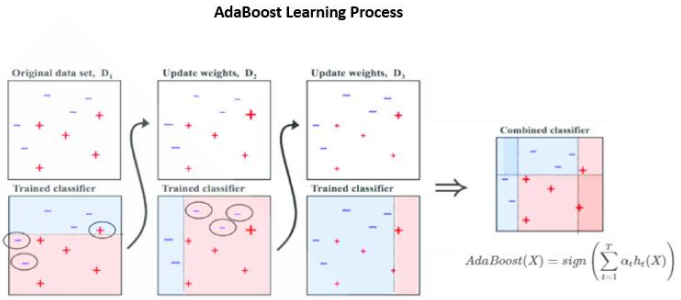
ภาพที่ 2.5 ตัวอย่างของเทคนิค Random Forest Classification

ที่มา: https://datacubeth.ai/machine-learning-model-comparison/

https://www.javatpoint.com/machine-learning-random-forest-algorithm

2.5 AdaBoost Forest Classification

เป็นโมเดลทางคณิตศาสตร์ที่เป็น **sequential ensemble method** ที่มีการ combine weak learner หลายๆตัวเข้าด้วยกัน แล้วสร้างเป็น strong learner ทำให้ประสิทธิภาพของการทำนายเพิ่มขึ้นและถูกเปลี่ยนการให้น้ำหนัก ซึ่งจะให้ความสำคัญกับจุดข้อมูลที่ ทำนายไม่ถูกเพื่อให้มีโอกาสที่จะถูกทำนายในรอบถัดไป โดยจะมีการปรับเพิ่มน้ำหนัก ของจุดข้อมูลที่ ทำนายไม่ถูกและปรับลดน้ำหนักของจุดข้อมูลที่ทำนายถูกแล้วนำจุดข้อมูลเหล่านี้ไปทดสอบ และสร้าง weak learner ตัวใหม่ ขั้นตอนเหล่านี้จะทำแบบ sequential และจะมีการปรับค่าน้ำหนักในทุกๆรอบ



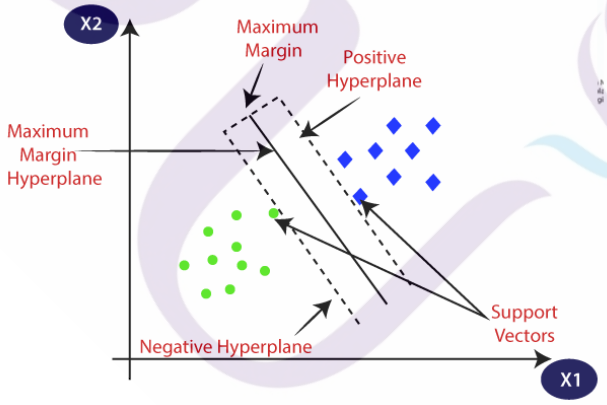
ภาพที่ 2.6 ตัวอย่างของเทคนิค AdaBoost Classification

ที่มา :https://www.pluralsight.com/guides/ensemble-methods:-bagging-versus-boosting

sirawichjaichuen.medium.com/adaboost-algorithm

2.6 SVM Classification

เป็นโมเดลทางคณิตศาสตร์ที่ค่อนข้างซับซ้อนแต่แนวคิดคือการปรับข้อมูลไปให้อยู่ในมิติ (Dimension) ที่สูงขึ้นเพื่อทำให้สามารถแบ่งแยกข้อมูลได้ง่ายขึ้น ดังภาพ



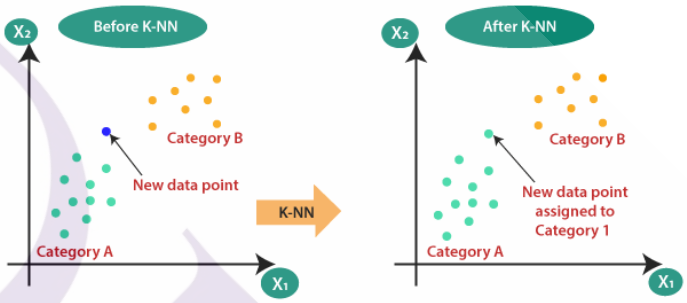
ภาพที่ 2.7 ตัวอย่างของเทคนิค SVM Classification

ที่มา: http://dataminingtrend.com/2014/support-vector-machine-svm/

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

2.7 KNN Classification

เป็นการสร้างโมเดลโดยมีแนวคิดว่าข้อมูลที่มีลักษณะคล้ายกันน่าจะอยู่ในกลุ่ม (หรือคลาสเดียวกัน) การจำแนกข้อมูลด้วยวิธีการค้นหาเพื่อนบ้านใกล้สุดจะเป็นการเรียนรู้โดยการเปรียบเทียบกันระหว่าง ข้อมูลที่ต้องการจำแนกและต้องการทำนายกับข้อมูลทั้งหมดในชุดข้อมูลสอนที่มีลักษณะเหมือนกันหรือใกล้เคียงกันด้วยการพิจารณาข้อมูลต่างๆ โดยข้อมูลเรคคอร์ดหนึ่งๆจะสามารถถูกมอง ว่าเป็นจุดหนึ่งในระนาบ n มิติ (เมื่อ n คือจำนวนแอทริบิวทั้งหมด)ดังภาพ



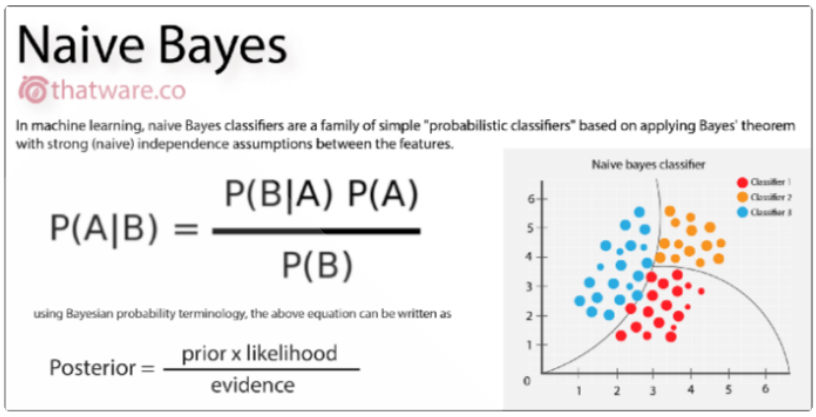
ภาพที่ 2.8 ตัวอย่างของเทคนิค KNN Classification

ที่มา: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

https://datacubeth.ai/machine-learning-model-comparison/

2.8 Naive Bayes Classification

เป็นเทคนิคที่สร้างโมเดลโดยการคำนวณค่าความน่าจะเป็น (probability) ต่างๆ ในข้อมูล ดังภาพ



ภาพที่ 2.9 ตัวอย่างของเทคนิค Naïve Bays Classification

ที่มา: https://datacubeth.ai/machine-learning-model-comparison/

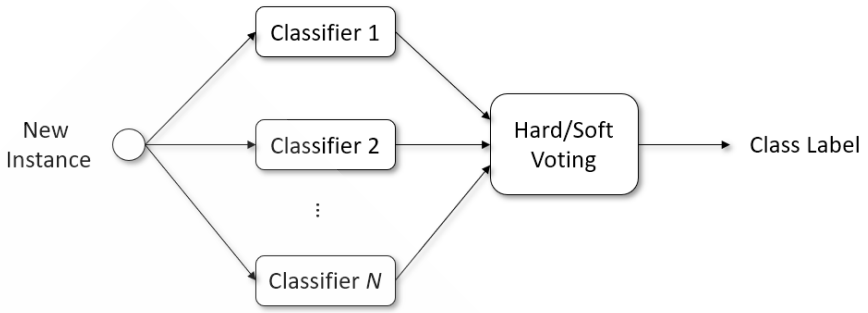
https://thatware.co/naive-bayes/

2.9 VotingClassifier

เป็นโมเดลทางคณิตศาสตร์ที่เป็น Ensemble method ที่มีการใช้เทคนิคโมเดลหลายๆตัวเข้าด้วยกันและนำผลลัพท์มาโหวตร่วมกันเพื่อทำนายค่าสุดท้าย ซึ่งสามารถเลือกโหวตได้สองแบบ คือ Hard และ Soft

แบบ Hard คือการเอาผลลัพท์สุดท้ายที่โมเดลแต่ละตัวทำนายมาโหวตกันตรงๆ เช่นในกรณี Binary classification ผลลัพท์จะเป็น 0 หรือ 1 ถ้าเรามี 3 โมเดลคือ A, B, C และแต่ละตัวทำนายผลเป็น 0, 0, 1 ก็จะโหวตผลสุดท้ายเป็น 0

แบบ soft คือการเอาค่าความน่าจะเป็น ที่โมเดลคำนวนความน่าจะเป็นของคำตอบมาใช้โหวตกัน ซึ่งถ้าค่าความน่าจะเป็น > 0.5 ผลลัพท์สุดท้ายก็จะเป็น 1 ถ้าเรามีโมเดล A, B, C ทำนายค่าความน่าจะเป็น = 0.4, 0.9, 0.8 ค่าเฉลี่ยคือ 0.7 แสดงว่าผลลัพท์สุดท้ายจะเป็น 1



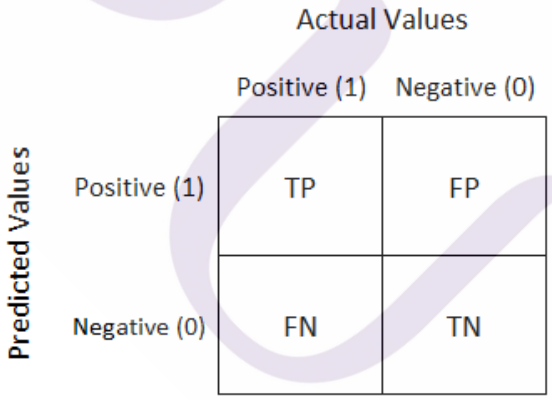
ภาพที่ 2.10 ตัวอย่างของเทคนิค Voting Classifier

ที่มา: https://www.quora.com/Can-we-use-neural-networks-in-ensemble-learning

https://medium.com/@supachaic/

2.10 Confusion Matrix

Confusion Matrix คือการวัดประสิทธิภาพของผลลัพธ์การทำนาย เปรียบเทียบกับผลลัพธ์จริง



ภาพที่ 2.11 ตัวอย่างเมตริกซ์การวัดประสิทธิภาพสำหรับการจำแนกประเภทข้อมูล 2 กลุ่ม

ที่มา (Sarang Narkhede-2018)

True Positive (TP) คือ สิ่งที่โปรแกรมทำนายว่าจริง และผลลัพธ์เป็นจริง

True Negative (TN) คือ สิ่งที่โปรแกรมทำนายว่าไม่จริง และผลลัพธ์ไม่จริง

False Positive (FP) คือ สิ่งที่โปรแกรมทำนายว่าจริง แต่ผลลัพธ์ไม่จริง

False Negative (FN) คือ สิ่งที่โปรแกรมทำนายว่าไม่จริง แต่ผลลัพธ์เป็นจริง

Accuracy คือค่าความแม่นยำของโมเดลที่ใช้ในการพยากรณ์

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision คือค่าที่บอกว่าโมเดลพยากรณ์ได้ว่า จริง ถูกต้องเท่าไร

Precision = TP/(TP+FP)

Recall คือ ค่าที่บอกว่าโมเดลพยากรณ์ได้ว่าจริง เป็นอัตราส่วนเท่าไรของจริงทั้งหมด

Recall = TP/(TP+FN)

2.11 Finance Classification

หมายถึง วิธีปฏิบัติเกี่ยวกับการจัดชั้นหนี้ ประกอบด้วย

1. ลูกหนี้จัดชั้นปกติ (ระยะเวลาค้างชำระ < 1 เดือน)
2. สินเชื่อจัดชั้นกล่าวถึงเป็นพิเศษ (1 เดือน < ระยะเวลาค้างชำระ < 3 เดือน)
3. สินเชื่อจัดชั้นต่ำกว่ามาตรฐาน (3 เดือน < ระยะเวลาค้างชำระ < 6 เดือน)
4. สินเชื่อจัดชั้นสงสัย (6 เดือน < ระยะเวลาค้างชำระ < 12 เดือน)
5. สินเชื่อจัดชั้นสงสัยจะสูญ (ระยะเวลาค้างชำระ > 12 เดือน)
6. สินเชื่อจัดชั้นสูญ

โดยข้อที่ 1 – 2 จะถูกจัดกลุ่มเป็นสินเชื่อที่ก่อให้เกิดรายได้ (Performing loan:PL) ส่วนข้อที่ 3 - 6 จะถูกจัดกลุ่มเป็นสินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL)

2.12 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวกับการการวิเคราะห์คุณภาพสินเชื่อที่อยู่อาศัยจากการเปลี่ยนแปลงสถานะการจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ซึ่งสินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL) มีจำนวนมาก แต่การนำเทคนิคการเรียนรู้ด้วยตนเองด้วยตนเองของการเงินอิสลามในประเทศไทยยังมีไม่มากพอ ซึ่งผู้วิจัยได้ศึกษางานวิจัยที่ได้รับการตีพิมพ์เหล่านั้นและสรุปได้ดังนี้

1. อโนทัย พุทธารี 2. จิตรภณ หรูเจริญพรพานิช 3. เพ็ญสิริ บำรุงเชาว์เกษม 4. ชินวัฒน์ เทพหัสดิน ณ อยุธยา(2018) CREDIT SCORING MODEL :ในงานวิจัยนี้ได้นำเสนอ เครื่องมือใน

การประเมินคุณภาพสินเชื่อ แนวคิดและการประยุกต์ใช้ Credit risk indicator 2 ประเภท ได้แก่ 1) การใช้ Default rate ในการพิจารณาคุณภาพสินเชื่อควบคู่กับ NPL ratio และ 2) การใช้ Credit score เพื่อเป็นเครื่องมือในการ ติดตามและพยากรณ์ความเสี่ยงที่อาจเกิดขึ้นใน อนาคตอันใกล้และการบริหารความเสี่ยงหรือกำกับ ดูแลเชิงรุก

ธนาคารแห่งประเทศไทย : ประกาศ ธปท. เรื่อง หลักเกณฑ์การจัดชั้นและการกันเงินสำรอง ของสถาบันการเงินเฉพาะกิจ

1. ญาณินี สิทธิสาร 2. ดร.พธุษดี ศิริแสงตระกูล ในงานวิจัยนี้ได้นำเสนอโมเดลเพื่อทำนายความสามารถในการชำระหนี้ของการกู้เงินของสวัสดิการครู (ช.พ.ค) ของจังหวัดอุดรธานี โดยใช้ Decisiontree / Bayesian Belief Network / Artificialneuralnetwork

1. วชิญว์สิฐ เกษรสิทธิ์ 2. ดร.วชิต หล่อจีระชุณห์ 3. ดร.จิราวัลย์ จิตรถเวช (2018) ในงานวิจัยนี้ได้นำเสนอการแก้ปัญหาข้อมูลไม่สมดุลของข้อมูลสำหรับการจำแนกผู้ป่วยโรคเบาหวาน โดยการแกปัญหาข้อ มูลไม่สมดุลที่ระดับข้อมูลจำนวน 4 วิธีคือวิธีสุ่มเกิน วิธีสุ่มลด วิธีผสมผสาน และวิธีสังเคราะห์ข้อมูลใหม่ (SMOTE) โดยใช้ เทคนิคการจำแนกคือวิธีการถดถอยโลจิสติกแบบมัลติโนเมียลและวิธีต้นไม้การตัดสินใจ

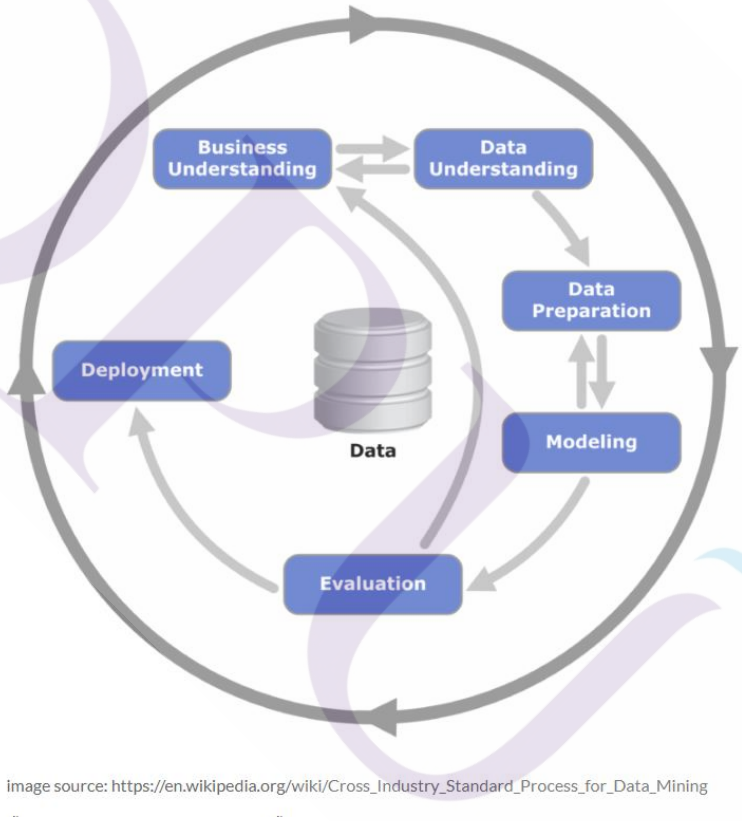
Shivani Parekh (2020) เป็นบทความได้นำเสนอ Ensemble modelling เพื่อเพิ่มประสิทธิภาพในการทำนายของ Model

Jason Brownlee (2020) เป็นบทความได้นำเสนอการ Evoluation Model โดยการหาค่า Threshold ที่เหมาะสมในการปรับค่าการทำนาย เพื่อให้ค่าความแม่นยำมากยิ่งขึ้น ทำอย่างไรในการ optimal Threshold ในพื้นที่ใต้กราฟ ROC

ดร.เอกสิทธิ์ พัชรวงศ์ศักดา หนังสือการวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิงเบื้องต้น (An Introduction to Data Mining Techniques) การแนะนำเทคนิคการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิงเบื้องต้นสำหรับนักศึกษาและผู้สนใจ

บทที่ 3
วิธีวิจัย

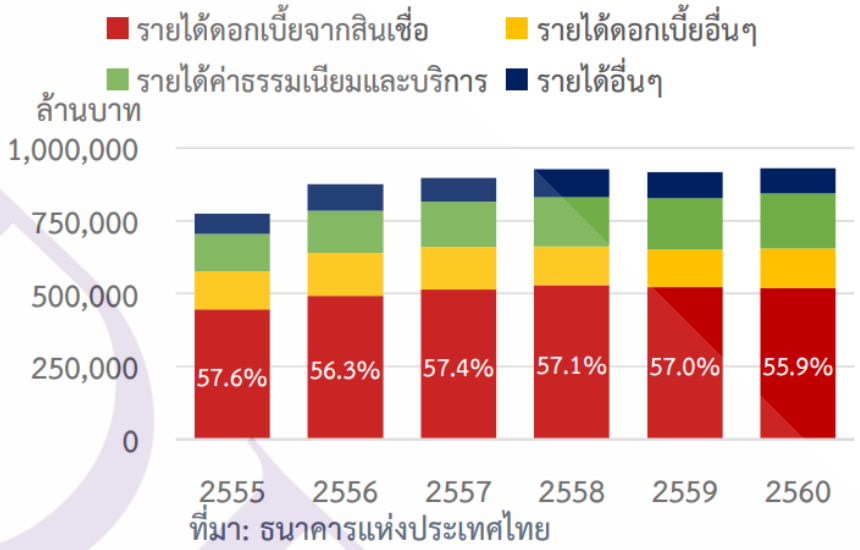
การศึกษางานครั้งนี้เป็นการนำเสนอเทคนิคการทำนายการจัดชั้นสินเชื่อที่อยู่อาศัยโดยใช้ MachineLearning การเรียนรู้ได้จากตัวอย่างด้วยตนเอง โดยมีแนวทางการวิจัยโดยกระบวนการ CRISP-DM ประกอบด้วย 6 ขั้นตอนดังภาพ



ภาพที่ 3.1 กระบวนการ CRISP-DM

3.1 ความเข้าใจทางธุรกิจ (Business Understanding)

สินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL) ถือเป็นเครื่องชี้หลักสำหรับวิเคราะห์และติดตามคุณภาพสินเชื่อของสถาบันการเงินที่ใช้กันอย่างแพร่หลาย



ภาพที่ 3.2 สัดส่วนรายได้ดอกเบี้ยจากเงินให้สินเชื่อต่อรายได้ทั้งหมด

โดยสินเชื่อที่ไม่ก่อให้เกิดรายได้ หมายถึง เงินให้สินเชื่อที่ถูกจัดชั้นต่ำกว่ามาตรฐาน จัดชั้นสงสัย จัดชั้นสงสัยจะสูญ และจัดชั้นสูญโดยการติดตามคุณภาพสินเชื่อในเบื้องต้นสามารถทำได้โดยการ พิจารณาสัดส่วนของสินเชื่อที่ไม่ก่อให้เกิดรายได้ ณ เวลาใดเวลาหนึ่ง นอกจากนั้นสัดส่วนของ NPL ยังสามารถสะท้อนคุณภาพสินเชื่อทำให้สามารถปรับเปลี่ยนหรือบริหารการผิดนัดชำระให้สอดคล้องกับสถานการณ์และลักษณะเฉพาะของสินเชื่อต่างประเภทกันได้ ซึ่งในงานศึกษาวิจัยนี้จะศึกษาเฉพาะข้อมูลสินเชื่อที่อยู่อาศัยเป็นหลัก

3.2 ความเข้าใจข้อมูลที่ใช้ในงาน (Data Understanding)

3.2.1 กระบวนการเก็บข้อมูลสินเชื่อที่อยู่อาศัย

โดยใช้ฐานข้อมูลสินเชื่อที่อยู่อาศัยปี 2019 ที่ยังมีสถานะ Active ส่วนตัวแปรที่ ใช้การจัดทำ Model ประกอบด้วยตัวแปร 3 กลุ่ม ได้แก่

1) ข้อมูลทั่วไปของลูกหนี้ เช่น ข้อมูลประเภทลูกค้า, สถานะสมรส, เพศ, ศาสนา, การศึกษา, อายุ, พื้นที่อยู่อาศัย, ช่วงของรายได้ เป็นต้น

2) ลักษณะสินเชื่อจากฐานข้อมูลสินเชื่อที่อยู่อาศัยรายตัว (Finance characteristics) ที่มีสถานะ Active ในแต่ละเดือน เช่น ข้อมูลยอดการขอสินเชื่อ , ข้อมูลงวดการชำระสินเชื่อ , ข้อมูลยอดคงเหลือของสินเชื่อ เป็นต้น

3) อัตราส่วนข้อมูลสินเชื่อที่อยู่อาศัยกับหลักประกันสินเชื่อ เช่น LTV, ประเภทหลักประกัน, มูลค่าหลักประกัน เป็นต้น

รายละเอียดตามตารางที่ 3.1 – 3.3 โดยตัวแปรที่ใช้ในการทดสอบมีทั้งสิ้น 35 ตัวแปร ครอบคลุมข้อมูลรายเดือนตั้งแต่ 2562/01/01 – 2562/07/31

ตารางที่ 3.1 ข้อมูลตัวแปรทั่วไปของลูกหนี้

ลำดับ	ตัวแปร	ตัวอย่างข้อมูล
1	ประเภทลูกค้า (Customer Type)	เช่น บุคคลธรรมดา / นิติบุคคล
2	ประเภทการชำระเงิน (Payment method)	เช่น 1_AFT
3	สถานภาพการสมรส (Marital Status)	เช่น โสด / สมรส /หม้าย
4	เพศ (Gender)	เช่น ชาย / หญิง
5	ศาสนา (Religion)	เช่น พุทธ / อิสลาม /คริสต์
6	กลุ่มศาสนาหลัก (Core religion)	เช่น อิสลาม / อื่นๆ
7	ระดับการศึกษา (Degree)	เช่น ปริญญาตรี / ปริญญาโท
8	อายุ (Age)	เช่น 20 / 30 / 40
9	อาชีพ (Occupation)	เช่น รับราชการ / พนักงาน /เจ้าของกิจการ
10	ช่วงรายได้ต่อเดือน (Salary Level)	เช่น <10,000 / 10,000 - 14,999
11	ที่มาของรายได้ (Salary Source)	เช่น เงินเดือน /รายได้จากธุรกิจ
12	จังหวัดที่อยู่ (State)	เช่น กรุงเทพ / ยะลา /นนทบุรี
13	ภูมิภาค (Region)	เช่น กรุงเทพมหานครและปริมณฑล

ตารางที่ 3.2 ข้อมูลตัวแปรลักษณะสินเชื่อจากฐานข้อมูลสินเชื่อที่อยู่อาศัย

ลำดับ	ตัวแปร	ตัวอย่างข้อมูล
1	บัญชีสินเชื่อ (Account)	เช่น 57015864
2	ราคาขาย (Selling Price)	เช่น 10,000,000 บาท
3	วงเงินสินเชื่อ (Credit Limit)	เช่น 5,000,000 บาท
4	ยอดเงินต้นคงเหลือ (Ledger Balance)	เช่น 3,000,000 บาท
5	อัตรากำไรสินเชื่อ(Prof/Div Rate)	เช่น 3.5 / 4.5
6	งวดการชำระสินเชื่อ (Term)	เช่น 120 / 360
7	งวดที่ชำระสินเชื่อ (MonthOnbook)	เช่น 115 / 330
8	ยอดเงินต้นคงค้าง (Uncollected Principal)	เช่น 3,000 บาท
9	ยอดอัตรากำไรคงค้าง (Uncollected Profit)	เช่น 100 บาท
10	ค่างวด (Payment)	เช่น 6,900 บาท
11	ค่าปรับผิดนัดชำระ (Late Charge Due)	เช่น 35.23 บาท
12	จำนวนเงินที่ชำระล่าสุด (Last Payment Amount)	เช่น 6,850 บาท
13	ยอดเงินต้นครบกำหนดชำระ(Total Principal Due)	เช่น 8,900 บาท
14	ยอดเงินอัตรากำไรครบกำหนดชำระ(Total Profit Due)	เช่น 1,900 บาท
15	ยอดเงินรวมตามกำหนด (Grand Total Due)	เช่น 11,900 บาท
16	จำนวนวันคงค้างกำหนดรับชำระ (Day Past Due)	เช่น 30 /60 /90
17	บัญชีสินเชื่อเคยปรับปรุงโครงสร้างหนี้ (Restructure Flag)	เช่น Yes / No
18	บัญชีสินเชื่อเหลืออายุกี่ปี	เช่น 3_> 1 <= 5 Years
19	จัดชั้นหนี้สินเชื่อ (Finance Classification)	เช่น 1 / 2 / 3 /4 / 5

ตารางที่ 3.3 ข้อมูลตัวแปรอัตราส่วนข้อมูลสินเชื่อที่อยู่อาศัยกับหลักประกันสินเชื่อ

ลำดับ	ตัวแปร	ตัวอย่างข้อมูล
1	ประเภทหลักประกัน (Customer Type)	เช่น ที่ดินพร้อมสิ่งปลูกสร้าง / คอนโด
2	ราคาประเมินหลักประกัน (Appraised value)	เช่น 1,000,000 บาท
3	ราคาจำนองหลักประกัน (Pledged value)	เช่น 1,000,000 บาท
4	สัดส่วนของบัญชีสินเชื่อเพื่อที่อยู่อาศัยที่มีมูลค่าสินเชื่อต่อมูลค่าหลักประกัน (LTV)	เช่น 80 / 90 /100

3.2.2 ทำความเข้าใจภาพรวมจำนวนบัญชีสินเชื่อที่อยู่อาศัยที่มีสถานะ Active ตั้งแต่ 2562/01/01 – 2562/12/31 แบ่งเป็นชุดข้อมูล PL และ NPL

3.3 ตรวจสอบความถูกต้องของข้อมูลและเตรียมข้อมูลที่ใช้ในงาน (Data Preparation)

3.3.1 กระบวนการตรวจสอบข้อมูลเพื่อดูความถูกต้องของข้อมูล(Data Exploration)

ตารางที่ 3.4 การตรวจสอบกลุ่มข้อมูลทั่วไปของลูกหนี้

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
1	ประเภทลูกค้า (Customer Type)	100 %
2	ประเภทการชำระเงิน (Payment method)	100 %
3	สถานภาพการสมรส (Marital Status)	Missing < 20 Records
4	เพศ (Gender)	Missing < 20 Records
5	ศาสนา (Religion)	Missing < 20 Records
6	กลุ่มศาสนาหลัก (Core religion)	Missing < 20 Records
7	ระดับการศึกษา (Degree)	100 %
8	อายุ (Age)	Missing < 20 Records
9	อาชีพ (Occupation)	Missing < 20 Records
10	ช่วงรายได้ต่อเดือน (Salary Level)	Missing < 500 Records NPL=78 , PL=410

ตารางที่ 3.4 (ต่อ)

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
11	ที่มาของรายได้ (Salary Source)	Missing < 500 Records
12	จังหวัดที่อยู่ (State)	Missing < 5 Records
13	ภูมิภาค (Region)	Missing < 10 Records

ตารางที่ 3.5 การตรวจสอบกลุ่มข้อมูลลักษณะสินเชื่อจากฐานข้อมูลสินเชื่อที่อยู่อาศัย

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
1	บัญชีสินเชื่อ (Account)	100%
2	ราคาขาย (Selling Price)	100%
3	วงเงินสินเชื่อ (Credit Limit)	100%
4	ยอดเงินต้นคงเหลือ (Ledger Balance)	100%
5	อัตรากำไรสินเชื่อ(Prof/Div Rate)	100%
6	งวดการชำระสินเชื่อ (Term)	Missing < 5 Records
7	งวดที่ชำระสินเชื่อ (MonthOnbook)	Missing < 5 Records
8	ยอดเงินต้นคงค้าง (Uncollected Principal)	100%
9	ยอดอัตรากำไรคงค้าง (Uncollected Profit)	100%
10	ค่างวด (Payment)	Missing < 25 Records
11	ค่าปรับผิดนัดชำระ (Late Charge Due)	100%
12	จำนวนเงินที่ชำระล่าสุด (Last Payment Amount)	100%
13	ยอดเงินต้นครบกำหนดชำระ(Total Principal Due)	100%
14	ยอดเงินอัตรากำไรครบกำหนดชำระ(Total Profit Due)	100%
15	ยอดเงินรวมตามกำหนด (Grand Total Due)	100%
16	จำนวนวันคงค้างกำหนดรับชำระ (Day Past Due)	100%
17	บัญชีสินเชื่อเคยปรับปรุงโครงสร้างหนี้ (Restructure Flag)	100%
18	บัญชีสินเชื่อเหลืออายุกี่ปี	100%
19	จัดชั้นหนี้สินเชื่อ (Finance Classification)	100%

ตารางที่ 3.6 การตรวจสอบกลุ่มอัตราส่วนข้อมูลสินเชื่อที่อยู่อาศัยกับหลักประกันสินเชื่อ

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
1	ประเภทหลักประกัน (Customer Type)	100%
2	ราคาประเมินหลักประกัน (Appraised value)	100%
3	ราคาจำนองหลักประกัน (Pledged value)	100%
4	สัดส่วนของบัญชีสินเชื่อเพื่อที่อยู่อาศัยที่มีมูลค่าสินเชื่อต่อมูลค่าหลักประกัน (LTV)	100%

3.3.2 กระบวนการความสะอาดข้อมูล (Data Cleansing)

เนื่องจากข้อมูลตัวแปรแต่ละตัว ถ้ามีความผิดปกติหรือไม่ครบถ้วนสมบูรณ์ก็จะทำให้การทำนายผล ผิดพลาด หรือถูกบิดเบือนไปจากความเป็นจริง จำเป็นต้องหาวิธีในการเติมเต็มข้อมูล หรือตัดข้อมูลบางส่วนออกไป โดยพิจารณาจากฐานข้อมูลสินเชื่อที่อยู่อาศัย ดังต่อไปนี้

- เติมข้อมูลด้วยค่า MEDIAN กรณีเป็นตัวแปรที่เป็น Number และไม่ใช่ข้อมูล Transaction เช่น อายุของผู้ขอสินเชื่อ เป็นต้น
- เติมข้อมูลด้วยค่า MODE ของตัวแปรที่เป็น Categorical เช่น ช่วงรายได้ของผู้ขอสินเชื่อ , อาชีพของผู้ขอสินเชื่อ ,สถานะของผู้ขอสินเชื่อ เป็นต้น
- เติมข้อมูล Based on Business logic เช่น ค่า LTV ไม่ควรเกิน 100 % หรือ ราคาประเมินหลักประกันไม่ควรมีค่าเท่ากับ 0 เป็นต้น
- ลบข้อมูลที่ Records ไม่สมบูรณ์บางตัวอย่าง เช่น ข้อมูลไม่มีงวดที่จ่ายเป็นต้น

3.3.3 กระบวนการเปลี่ยนแปลงข้อมูล (Data Transformation)

3.3.3.1 Normalizationเนื่องจากข้อมูลตัวแปรแต่ละตัว มีความหลากหลาย ทั้งชนิดข้อมูล รูปแบบข้อมูล และ ช่วงของข้อมูล เช่น ข้อมูลทั่วไปของลูกหนี้ Feature Age อายุ [10, 20,70], ข้อมูลลักษณะสินเชื่อจากฐานข้อมูลสินเชื่อที่อยู่อาศัยรายตัว Feature Credit Limit / Ledger Balance เป็นต้น เพื่อที่จะทำให้ประสิทธิภาพของ การทำนายสำหรับอัลกอริทึม Machine Learning จำเป็นที่ต้องทำ Normalization ก่อนที่จะป้อนข้อมูลให้กับ Model เช่นวิธีการต่างๆ ดังนี้

- Rescaling (Min-Max Normalization)หรือ Min-Max Normalization เป็นวิธีที่ง่ายที่สุดที่จะปรับช่วงข้อมูล ให้เป็นอยู่ในช่วง [0, 1] ด้วยการนำ Feature / Column นั้น ๆ ลบด้วยค่าที่น้อยที่สุด (Min) ของมัน แล้วหารด้วยช่วงของข้อมูลนั้น (Max – Min)

- Standardization หรือ Z-Score Normalization คือ การนำข้อมูล Feature / Column มาปรับให้ Mean = 0 และ Standard Deviation = 1 (Unit Variance)

- Mean Normalization คล้ายกับ Rescaling ด้านบน แตกต่างกันที่ใช้ Mean แทน Min ทำให้ช่วงของ Output [-0.5, 0.5] มีทั้งบวกและลบ Balance กัน ตรงเลข 0 (ขยับ Mean มาตรง 0)

3.3.3.2 One Hot Encoder การทำข้อมูลที่ถูกเก็บในลักษณะ Categorical ทั้งในลักษณะที่มีลำดับ (Ordinal number) และไม่มีลำดับ (Nominal number) เปลี่ยนให้อยู่ในรูปแบบของ Binary values ที่มีค่า 0 หรือ 1 เท่านั้น เช่น ข้อมูลทั่วไปของลูกหนี้ Feature Gender [F,M] หรือ Feature Religion เป็นต้น แสดงการ Transformation ใน Feature ต่างๆ ดังตารางด้านล่าง

ตารางที่ 3.7 กระบวนการ Transformation กลุ่มข้อมูลทั่วไปของลูกหนี้

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
1	ประเภทลูกค้า (Customer Type)	One hot encoder
2	ประเภทการชำระเงิน (Payment method)	One hot encoder
3	สถานภาพการสมรส (Marital Status)	One hot encoder
4	เพศ (Gender)	One hot encoder
5	ศาสนา (Religion)	One hot encoder
6	กลุ่มศาสนาหลัก (Core religion)	One hot encoder
7	ระดับการศึกษา (Degree)	One hot encoder
8	อายุ (Age)	Standardization
9	อาชีพ (Occupation)	One hot encoder
10	ช่วงรายได้ต่อเดือน (Salary Level)	One hot encoder
11	ที่มาของรายได้ (Salary Source)	One hot encoder
12	จังหวัดที่อยู่ (State)	One hot encoder
13	ภูมิภาค (Region)	One hot encoder

ตารางที่ 3.8 กระบวนการ Transformation กลุ่มข้อมูลลักษณะข้อมูลสินเชื่อที่อยู่อาศัย

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
1	บัญชีสินเชื่อ (Account)	Key
2	ราคาขาย (Selling Price)	Standardization
3	วงเงินสินเชื่อ (Credit Limit)	Standardization
4	ยอดเงินต้นคงเหลือ (Ledger Balance)	Standardization
5	อัตรากำไรสินเชื่อ(Prof/Div Rate)	Standardization
6	งวดการชำระสินเชื่อ (Term)	Standardization
7	งวดที่ชำระสินเชื่อ (MonthOnbook)	Standardization
8	ยอดเงินต้นคงค้าง (Uncollected Principal)	Standardization
9	ยอดอัตรากำไรคงค้าง (Uncollected Profit)	Standardization
10	ค่างวด (Payment)	Standardization
11	ค่าปรับผิดนัดชำระ (Late Charge Due)	Standardization
12	จำนวนเงินที่ชำระล่าสุด (Last Payment Amount)	Standardization
13	ยอดเงินต้นครบกำหนดชำระ(Total Principal Due)	Standardization
14	ยอดเงินอัตรากำไรครบกำหนดชำระ(Total Profit Due)	Standardization
15	ยอดเงินรวมตามกำหนด (Grand Total Due)	Standardization
16	จำนวนวันคงค้างกำหนดรับชำระ (Day Past Due)	Standardization
17	บัญชีสินเชื่อเคยปรับปรุงโครงสร้างหนี้ (Restructure Flag)	One hot encoder
18	บัญชีสินเชื่อเหลืออายุกี่ปี	One hot encoder
19	จัดชั้นหนี้สินเชื่อ (Finance Classification)	PL = 1 ,NPL = 0

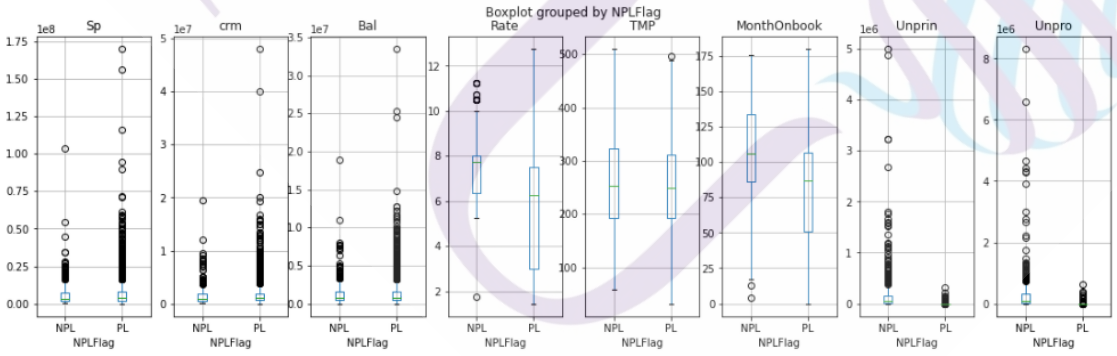
ตารางที่ 3.9 กระบวนการ Transformation กลุ่มข้อมูลสินเชื่อที่อยู่อาศัยกับหลักประกัน

ลำดับ	ตัวแปร	ความถูกต้องข้อมูล
1	ประเภทหลักประกัน (Customer Type)	One hot encoder
2	ราคาประเมินหลักประกัน (Appraised value)	Standardization
3	ราคาจำนองหลักประกัน (Pledged value)	Standardization
4	สัดส่วนของบัญชีสินเชื่อเพื่อที่อยู่อาศัยที่มีมูลค่าสินเชื่อต่อมูลค่าหลักประกัน (LTV)	Standardization

3.3.4 กระบวนการพิจารณาจะใช้ข้อมูลทั้งหมดหรือเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์ (Feature Selection)

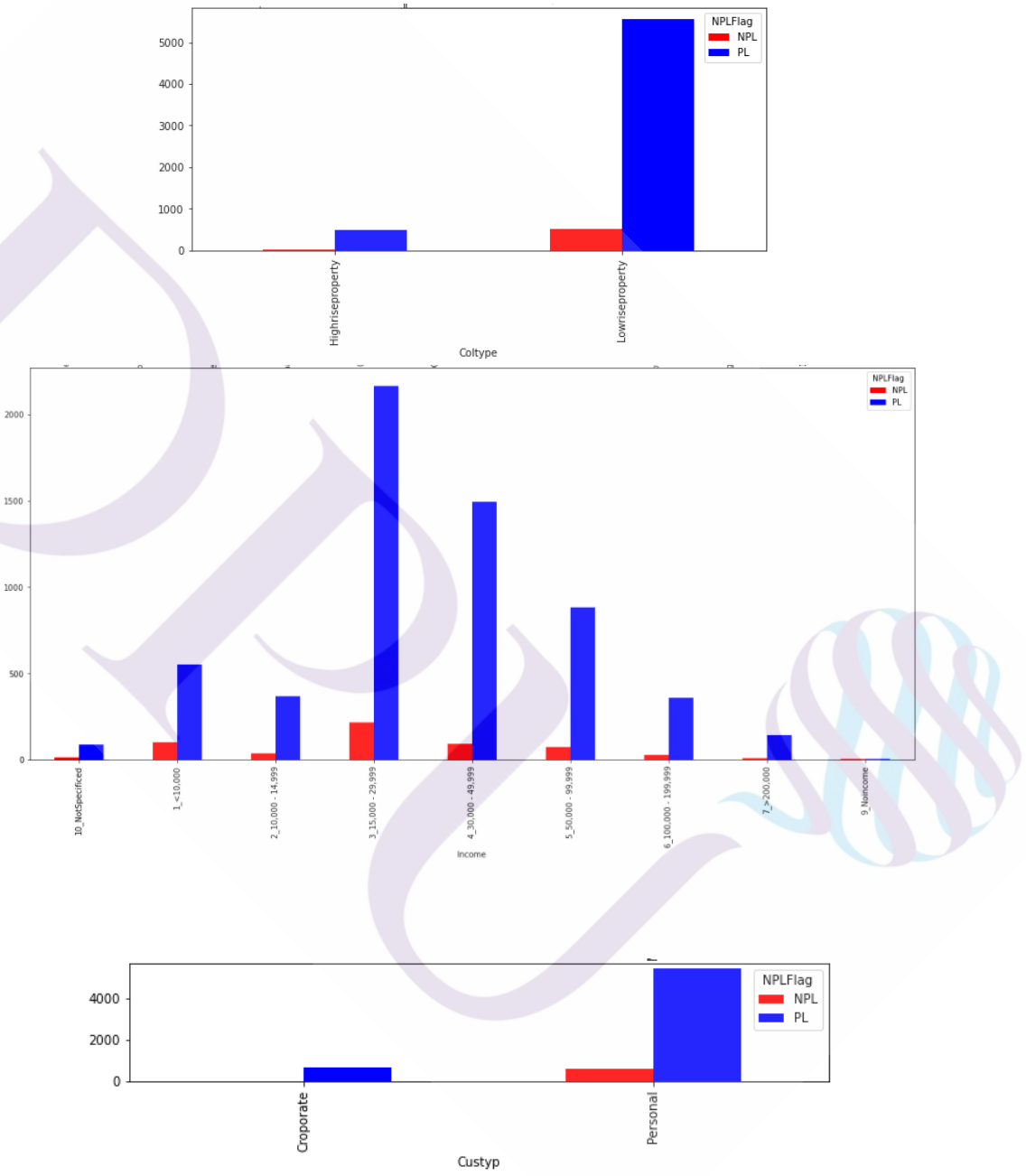
3.3.4.1 Visual exploration of relationship between variables การทำทดสอบ Plot graph ข้อมูลที่เป็น Feature ทั้งหมด เพื่อเปรียบเทียบกับ Target โดยจะแบ่งลักษณะเป็น 2 กลุ่มหลัก คือ

3.3.4.1.1 Categorical Target Variable Vs Continuous Predictor ทดสอบโดยการใช้ Box Plot ตัวอย่างดังภาพที่ 3.3



ภาพที่ 3.3 กราฟแสดงความสัมพันธ์ระหว่างตัว Feature & NPLFlag

3.3.4.1.2 Categorical Target Variable Vs Categorical Predictor ทดสอบโดยการใช้ Bar Plot ตัวอย่างดังภาพที่ 3.4



ภาพที่ 3.4 กราฟแสดงความสัมพันธ์ระหว่างตัว Feature & NPLFlag

3.3.4.2 Statistical measurement of relationship strength between variables

การทำทดสอบข้อมูลทางสถิติ โดยใช้ Anova / Chi-Square test ซึ่งจะช่วยให้เห็นความสัมพันธ์ของตัวแปรโดยละเอียดยิ่งขึ้น โดยได้ผลของการเลือก Feature ดังนี้

กลุ่มตัวแปรที่เป็น Categorical Target Variable Vs Continuous Predictor

Sp is NOT correlated with NPLFlag | P-Value: 0.09046372106108784
crm is NOT correlated with NPLFlag | P-Value: 0.4481568512285441
Bal is NOT correlated with NPLFlag | P-Value: 0.24317080894139234
Rate is correlated with NPLFlag | P-Value: 5.319820264439458e-66
TMP is NOT correlated with NPLFlag | P-Value: 0.6219402200281612
MonthOnbook is correlated with NPLFlag | P-Value: 4.4320149123715767e-44
Unprin is correlated with NPLFlag | P-Value: 7.905837694182945e-206
Unpro is correlated with NPLFlag | P-Value: 5.534300829962861e-229
DayPastDue is correlated with NPLFlag | P-Value: 0.0
Payment is NOT correlated with NPLFlag | P-Value: 0.4861712557468393
Lcd is correlated with NPLFlag | P-Value: 1.1959173470811203e-161
Lpa is correlated with NPLFlag | P-Value: 5.054014961455262e-28
Tprind is correlated with NPLFlag | P-Value: 5.734648103468045e-208
Tprod is correlated with NPLFlag | P-Value: 2.4038039603340593e-234
Gdt is correlated with NPLFlag | P-Value: 7.200186011329954e-234
Age is NOT correlated with NPLFlag | P-Value: nan
TotalAPP is NOT correlated with NPLFlag | P-Value: 0.24967350459160378
LTV is NOT correlated with NPLFlag | P-Value: 0.981984183606679

ชุดตัวแปรที่มีค่า P-Value < 0.05

['Rate',	'MonthOnbook',	'Unprin',	'Unpro',	'DayPastDue',	'Lcd',	'Lpa',	'Tprind',	'Tprod',	'Gdt']

กลุ่มตัวแปรที่เป็น Categorical Target Variable Vs Categorical Predictor

Rf is NOT correlated with NPLFlag | P-Value: 1.0
ProdType is NOT correlated with NPLFlag | P-Value: 1.0
pay_method is correlated with NPLFlag | P-Value: 4.0307702749685827e-237
Custyp is correlated with NPLFlag | P-Value: 2.5597454304881333e-14
Marsts is NOT correlated with NPLFlag | P-Value: 0.8414975464429585
Gender is correlated with NPLFlag | P-Value: 0.0005053896803007069
Religions is correlated with NPLFlag | P-Value: 7.868569716839309e-30
Edu is correlated with NPLFlag | P-Value: 1.2086381550707994e-33
OccEng is correlated with NPLFlag | P-Value: 8.470300487420019e-35
Income is correlated with NPLFlag | P-Value: 1.890187631904522e-12
Incsrc is correlated with NPLFlag | P-Value: 2.2108709127026183e-32
Coltype is correlated with NPLFlag | P-Value: 0.043695196334928674
Reg is correlated with NPLFlag | P-Value: 1.8610723733960762e-07
Mature is correlated with NPLFlag | P-Value: 1.0727246608459967e-69

ชุดตัวแปรที่มีค่า P-Value < 0.05

['pay_method',	'Custyp',	'Gender',	'Religions',	'Edu',	'OccEng',	'Income',	'Incsrc',	'Coltype',	'Reg',	'Mature']

หลังจากได้ชุดตัวแปรพร้อมทั้งหมดก็จะนำมาทำการตัดตัวแปรบางตัวอีกครั้งเนื่องจากมีความสัมพันธ์ซ้ำซ้อนกันเพื่อให้เหลือชุดตัวแปรที่น้อยที่สุดดังนี้

ตารางที่ 3.10 Feature Selection for Model

ลำดับ	ตัวแปร
1	ประเภทการชำระเงิน (Payment method)
2	ประเภทลูกค้า (Customer Type)
3	เพศ (Gender)
4	ช่วงรายได้ต่อเดือน (Salary Level)
5	ที่มาของรายได้ (Salary Source)
6	ประเภทหลักประกัน (Customer Type)
7	ภูมิภาค (Region)
8	บัญชีสินเชื่อเหลืออายุกี่ปี
9	อัตรากำไรสินเชื่อ(Prof/Div Rate)
10	งวดที่ชำระสินเชื่อ (MonthOnbook)
11	ยอดเงินต้นคงค้าง (Uncollected Principal)
12	ยอดอัตรากำไรคงค้าง (Uncollected Profit)
13	ค่าปรับผิดนัดชำระ (Late Charge Due)
14	จำนวนเงินที่ชำระล่าสุด (Last Payment Amount)
15	ยอดเงินต้นครบกำหนดชำระ(Total Principal Due)
16	ยอดเงินอัตรากำไรครบกำหนดชำระ(Total Profit Due)
17	ยอดเงินรวมตามกำหนด (Grand Total Due)
18	จำนวนวันคงค้างกำหนดรับชำระ (Day Past Due)

3.4 การพัฒนา Model (Modeling)

3.4.1 กระบวนการ Data Preprocessing for Model

3.4.1.1 การนำเข้าข้อมูลเพื่อทดสอบ Model จะใช้ข้อมูลสินเชื่อที่อยู่อาศัย 20190131 สถานะ Active ทั้งหมด และนำสินเชื่อดังกล่าวไปหาคำตอบ (PL/ NPL) จากข้อมูลสินเชื่อที่อยู่อาศัยในเดือน 20190228(ทำนายล่วงหน้า 1 เดือน)

โดยเลือกข้อมูลเฉพาะที่ Match เท่านั้น และเลือก Feature ที่ได้ทำการทดสอบมาใช้งาน

- Identify unique values: ["NCID"].nunique() ตรวจสอบเลขบัญชีสินเชื่อ
- Check target variable distribution: ["NPLFlag"].value_counts() ดูการกระจายตัวของตัวแปร Target
- Split the dataset into dependent and independent variables : TargetVariable='NPLFlag'
- Splitting the data into Training and Testing sample: Train size = 70, Test Size = 30
- Random Oversampling:
    - 1.RandomOverSampler = 0.25
    - 2.SMOT
- Remove Identifiers: ["NCID"]

3.4.1.2 การนำเข้าข้อมูลเพื่อทดสอบ Model จะใช้ข้อมูลสินเชื่อที่อยู่อาศัย 20190131 สถานะ Active ทั้งหมด และนำสินเชื่อดังกล่าวไปหาคำตอบ (PL/ NPL) จากข้อมูลสินเชื่อที่อยู่อาศัยในเดือน 20190430 (ทำนายล่วงหน้า 3 เดือน)

โดยเลือกข้อมูลเฉพาะที่ Match เท่านั้น และเลือก Feature ที่ได้ทำการทดสอบมาใช้งาน

- Identify unique values: ["NCID"].nunique() ตรวจสอบเลขบัญชีสินเชื่อ
- Check target variable distribution: ["NPLFlag"].value_counts() ดูการกระจายตัวของตัวแปร Target
- Split the dataset into dependent and independent variables : TargetVariable='NPLFlag'
- Splitting the data into Training and Testing sample: Train size = 70, Test Size = 30
- Random Oversampling:
    - 2.SMOT
- Remove Identifiers: ["NCID"]

3.4.2 กระบวนการ Model Selection ในกระบวนการนี้จะนำเข้าข้อมูลทดสอบกับ Classification Model ดังต่อไปนี้

3.4.2.1 กระบวนการ Model Selection ของการทำนายล่วงหน้า 1 เดือน

เปรียบเทียบ Model (การทำซ้ำครั้งที่ 1): ทดสอบ Model ผ่านชุดข้อมูลและประเมินความแม่นยำและคะแนนส่วนเบี่ยงเบนมาตรฐาน การทำนายได้ให้น้ำหนักกับ Class NPL ดังนั้นจึงจำเป็นต้องสนใจที่ค่า Precision / Recall และ f1-score เป็นตัววัดที่เหมาะสำหรับการเลือกแบบ Model มากกว่าดูจากค่า Accuracy เพียงอย่างเดียว

No	Model	Baseline
1	Logistic Regression	(solver='liblinear',class_weight='balanced')
2	Decision Trees	(max_depth=2,criterion='entropy')
3	Random Forest	(n_estimators=100, criterion = 'entropy')
4	AdaBoost	(n_estimators=500, max_depth=2,learning_rate=0.01)
5	KNN	(n_neighbors=10)
6	SVM	(kernel = 'linear') / (kernel = 'rbf')
7	Naive Bayes	GaussianNB()
8	VotingClassifier	ensemble (ทั้ง 7 Model)
Random = 42 /MinMaxScaler/RandomOverSample (minority)		

ภาพที่ 3.5 Classification Model

เปรียบเทียบ Model (การทำซ้ำครั้งที่ 2): ในการทำซ้ำครั้งที่สองของการเปรียบเทียบ โดยจะใช้พารามิเตอร์ที่เหมาะสมสำหรับModelพื้นฐานก่อนที่จะทำการวนซ้ำครั้งที่ 2 เพื่อเพิ่มประสิทธิภาพพารามิเตอร์และสรุปผลการประเมินสำหรับการเลือก Model ยกตัวอย่างดังนี้ เช่น ทำการปรับพารามิเตอร์ ของบาง Model เพื่อทดสอบสอบผลของ Class NPL (Precision / Recall และ f1-score)

No	Model	Baseline	f1-score	Class	NPL	PL
1	Logistic Regression	(C=1, penalty='l2', tol=0.01, solver='saga')	0.95	NPL	161	8
				PL	8	1805
2	Decision Trees	(max_depth=2,criterion='entropy')	0.95	NPL	167	2
				PL	17	1769
3	Random Forest	(max_depth=2, n_estimators=100,criterion='gini')	0.94	NPL	161	8
				PL	12	1801
4	AdaBoost	(max_depth=2,n_estimators=500, base_estimator=DTC ,learning_rate=0.01)	0.96	NPL	165	4
				PL	9	1804
5	KNN	(n_neighbors= 4)	0.7	NPL	125	44
				PL	62	1751
6	SVM	(kernel = 'rbf')	0.86	NPL	131	38
				PL	6	1807
7	Naive Bayes	GaussianNB()	0.78	NPL	134	35
				PL	40	1773
8	VotingClassifier	ensemble(Decision Trees , Random Forest , AdaBoost)	0.96	NPL	166	3
				PL	10	1803
Random = 42 /StandardScaler()/RandomOverSample (minority) ใช้ข้อมูลจำลองเดือน มกราคม โดยค่าตอบเป็น Class ของเดือน กุมภาพันธ์						

ภาพที่ 3.6 Optimal Classification Model

3.4.2.2 กระบวนการ Model Selection ของการทำนายล่วงหน้า 3 เดือน

การเปรียบเทียบ โดยจะใช้พารามิเตอร์ที่เหมาะสมสำหรับModel พื้นฐานก่อนที่จะทำการวนซ้ำเพื่อเพิ่มประสิทธิภาพพารามิเตอร์และสรุปผลการประเมินสำหรับการเลือก Model ยกตัวอย่างดังนี้

3 Month

No	Model	Baseline	f1-score	Class	NPL	PL
1	Decision Trees	(max_depth=2,criterion='entropy')	0.15	NPL	42	119
				PL	381	1440
2	Random Forest	(max_depth=2, n_estimators=1000,criterion='gini')	0.17	NPL	70	91
				PL	584	1237
3	AdaBoost	(max_depth=2,n_estimators=1000, base_estimator=DTC ,learning_rate=0.01)	0.13	NPL	26	135
				PL	215	1606
4	VotingClassifier	ensemble(Decision Trees , Random Forest , AdaBoost)	0.17	NPL	63	98
				PL	525	1296
Random = 42 /StandardScaler()/SMOTE ใช้ข้อมูลจำลองเดือน มกราคม โดยคำตอบเป็น Class ของเดือน เมษายน						

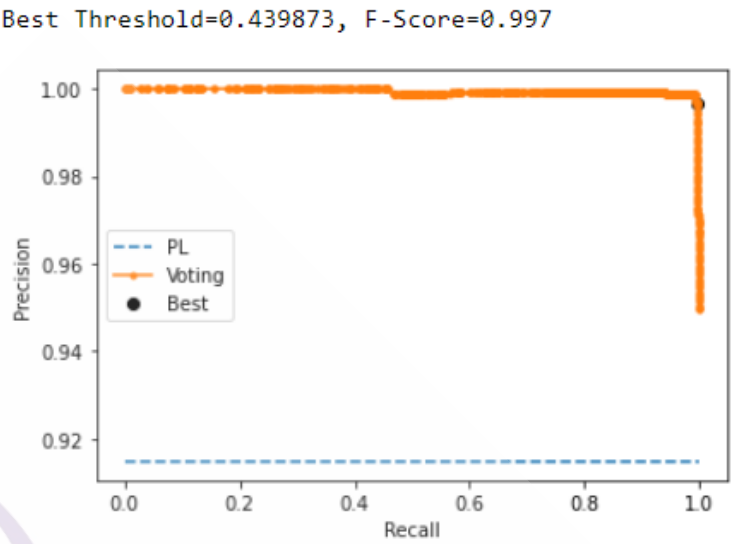
ภาพที่ 3.7 Optimal Classification Model ของการทำนายล่วงหน้า 3 เดือน

3.5 กระบวนการ Model Evaluation

3.5.1 กระบวนการ Model Evaluation ของการทำนายล่วงหน้า 1 เดือน

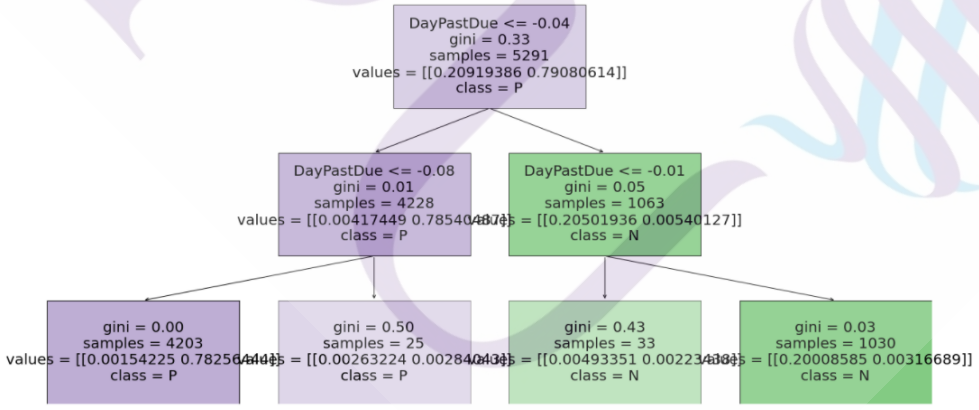
ในกระบวนการนี้จะนำเข้าข้อมูลทดสอบใส่ใน Classification Model เช่นการ Setup ดังต่อไปนี้

- ทดสอบการ Sampling ระหว่าง :
  - 1.RandomOverSampler = 0.25
  - 2.SMOT
- ทดสอบการ Set Threshold



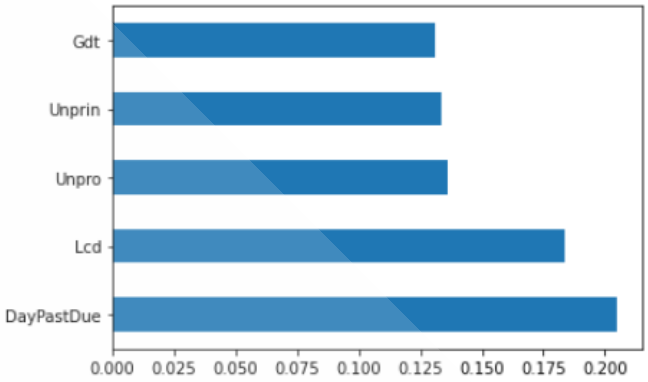
ภาพที่ 3.8 ตัวอย่าง Set Best Threshold

- ทดสอบการ Set random seed
- ทดสอบการ Set Max Depth ของ Tree Classification Model



ภาพที่ 3.9 ตัวอย่าง Optimal Max Depth Tree

- ทดสอบการ Set Model ภายใน Voting Classifier
- ทดสอบการหา Feature importance



ภาพที่ 3.10 ตัวอย่าง Test Feature importance

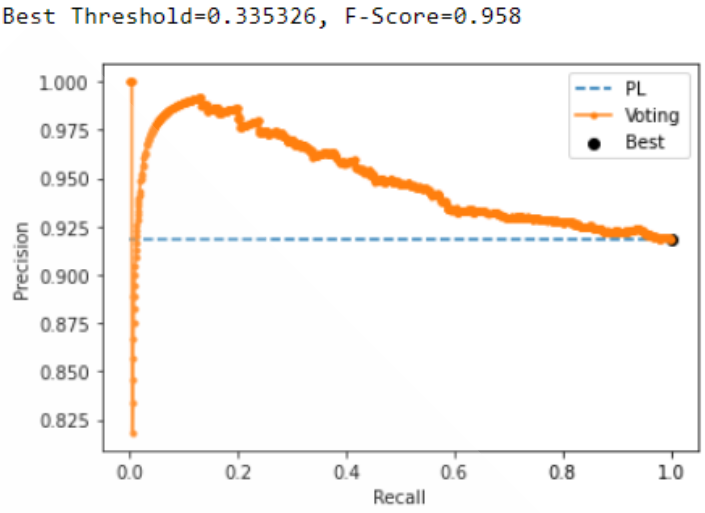
- วนการทดสอบในการปรับค่าต่างๆ ก่อนทำการ Save Model

ซึ่งจากการทดสอบปรับค่าต่างๆและผลจากการทดลองซ้ำ ได้ทำการเลือก Model Classification = Voting Classifier เป็น Model ที่ใช้ Deploy สำหรับ ข้อมูลที่จะทำนายจริงต่อไป

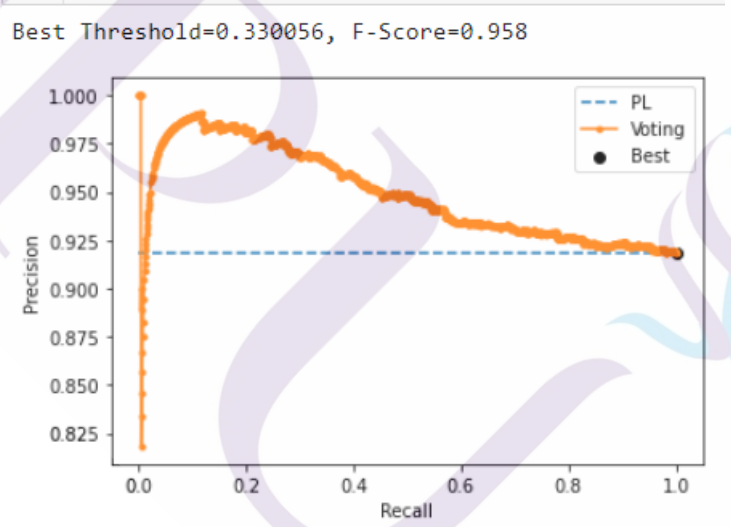
3.5.2 กระบวนการ Model Evaluation ของการทำนายล่วงหน้า 3 เดือน

ในกระบวนการนี้จะนำเข้าข้อมูลทดสอบใส่ใน Classification Model เช่นการ Setup ดังต่อไปนี้

- ทดสอบการ ระหว่างการใช้ Feature Day past Due และไม่ใช้ ซึ่ง Feature นี้เป็น Feature ที่สำคัญในการสร้าง Model
- ทดสอบการ Sampling ระหว่าง :
  - 1.RandomOverSampler = 0.25
  - 2.SMOT
- ทดสอบการ Set Threshold โดยแบ่งเป็น

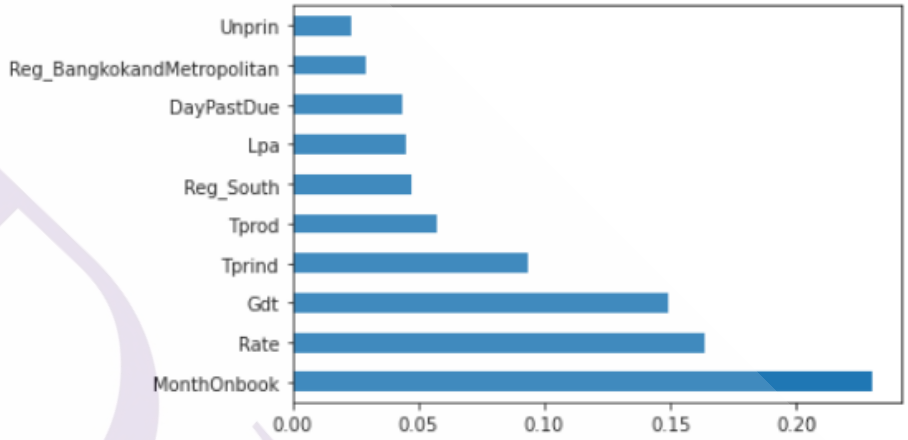


ภาพที่ 3.11 ตัวอย่าง Set Best Threshold ของการทำนายล่วงหน้า 3 เดือน โดยใช้ Feature Day past Due

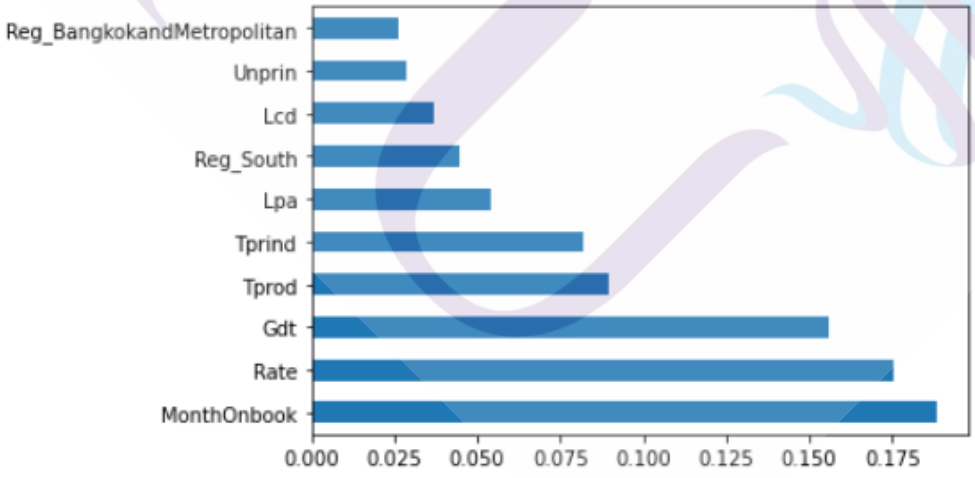


ภาพที่ 3.12 ตัวอย่าง Set Best Threshold ของการทำนายล่วงหน้า 3 เดือน โดยไม่ใช้ Feature Day past Due

- ทดสอบการ Set random seed
- ทดสอบการ Set Model ภายใน Voting Classifier
- ทดสอบการหา Feature importance



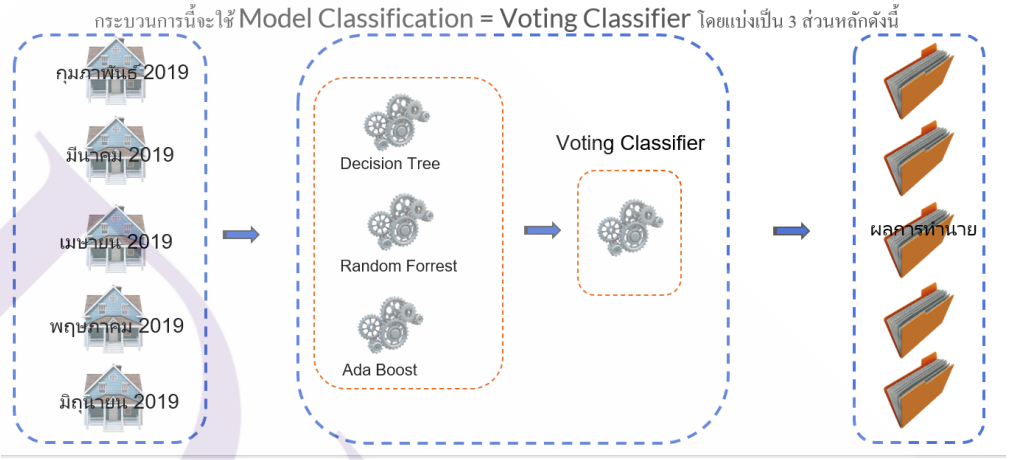
ภาพที่ 3.13 ตัวอย่าง ตัวอย่าง Test Feature importance ของการทำนายล่วงหน้า 3 เดือนโดยใช้ Feature Day past Due



ภาพที่ 3.14 ตัวอย่าง ตัวอย่าง Test Feature importance ของการทำนายล่วงหน้า 3 เดือนโดยไม่ใช้ Feature Day past Due

3.6 กระบวนการทำนายข้อมูล

ขั้นตอนการนำโมเดลไปใช้งานมีขั้นตอนการจัดเตรียมข้อมูลเหมือนกับ ขั้นตอนการสร้างโมเดลโดยจะแบ่งเป็น 3 ส่วนประกอบหลักดังภาพ



ภาพที่ 3.15 กระบวนการทำงานของ Model 3 ส่วนหลัก

3.7 เครื่องมือที่ใช้ในงานสารนิพนธ์

3.7.1 ภาษา Python

Python เป็นภาษาคอมพิวเตอร์ที่นิยมใช้ และเหมาะสำหรับการใช้งานด้านข้อมูลที่แพร่หลายรวมถึงมีชุมชน (Community) ให้ศึกษาแลกเปลี่ยนอีกทั้งภาษายังเข้าใจง่ายในการเริ่มศึกษาเพื่อนำมาพัฒนางาน

3.7.2 Program Rapid Miner Studio

Rapid Miner Studio เป็นเครื่องมือทางคอมพิวเตอร์ที่นิยม เหมาะสำหรับการใช้งานด้านการวิเคราะห์ข้อมูลและเข้าใจง่ายในการเริ่มศึกษาเพื่อนำมาพัฒนางาน

บทที่ 4
ผลการศึกษา

จากการทำสารนิพนธ์เพื่อพัฒนาโมเดลในการเป็นผู้ช่วยการวิเคราะห์คุณภาพสินเชื่อจากการเปลี่ยนแปลง สถานะการจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ซึ่งสินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL) ด้วยการประยุกต์ใช้ระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเอง (Machine Learning) Voting classifier Model มาใช้ในการสร้างโมเดล ซึ่งผลการทดสอบ และวิเคราะห์ข้อมูลมีรายละเอียดดังนี้

4.1 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล

4.1.1 ผลการวัดประสิทธิภาพความถูกต้องของโมเดลของการทำนายล่วงหน้า 1 เดือน

ข้อมูลที่ใช้ในการทดสอบในการสร้าง Voting classifier Model จำนวน 6605 Records ซึ่งเป็นชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยเดือนมกราคม และมีสถานะเป็น NPL/PL ในชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยเดือนกุมภาพันธ์ โดยมีการแบ่งสถานะเป็น NPL (Non-performing loan) 559 ชุดข้อมูล และ แบ่งสถานะเป็น PL (Performing loan) 6049 ชุดข้อมูล ซึ่งผลที่ได้ สามารถสรุปผลการเปรียบเทียบประสิทธิภาพได้ดังภาพที่ 4.1 และ ภาพที่ 4.2

VotingClassifier:

1.DecisionTreeClassifier(criterion='entropy',max_depth=2
2.RandomForestClassifier(max_depth=2)
3.AdaBoostClassifier(base_estimator=DecisionTreeClassifier(max_depth=2), learning_rate=0.01,n_estimators=500

Class	precision	recall	f1-score
NPL	0.94	0.98	0.96
PL	1	0.99	1

Accuracy of the model on Testing Sample Data: 0.99

	NPL	PL
NPL	166	3
PL	10	1803
Precision	0.94	1
Recall	0.98	0.99
F1-score	0.96	1

ภาพที่ 4.1 Voting Classification Model (Optimal)

VotingClassifier: Threshold

1.DecisionTreeClassifier(criterion='entropy',max_depth=2
2.RandomForestClassifier(max_depth=2)
3.AdaBoostClassifier(base_estimator=DecisionTreeClassifier(max_depth=2), learning_rate=0.01,n_estimators=500

Class	precision	recall	f1-score
NPL	0.92	0.99	0.95
PL	1	0.99	1

Accuracy of the model on Testing Sample Data: 0.99

	NPL	PL
NPL	167	2
PL	14	1799
Precision	0.92	1
Recall	0.99	0.99
F1-score	0.95	1

ภาพที่ 4.2 Voting Classification Model (Threshold + Optimal)

4.1.2 ผลการวัดประสิทธิภาพความถูกต้องของโมเดลของการทำนายล่วงหน้า 3 เดือน

ข้อมูลที่ใช้ในการทดสอบในการสร้าง Voting classifier Model จำนวน 6605 Records ซึ่งเป็นชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยเดือนมกราคม และมีสถานะเป็น NPL/PL ในชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยเดือนเมษายน โดยมีการแบ่งสถานะเป็น NPL (Non-performing loan) 559 ชุดข้อมูล และ แบ่งสถานะเป็น PL (Performing loan) 6049 ชุดข้อมูล ซึ่งผลที่ได้ สามารถสรุปผลการเปรียบเทียบประสิทธิภาพได้ดังภาพที่ 4.3

VotingClassifier: Threshold ทำนายล่วงหน้า 3 เดือน โดยใช้ Feature Day past Due

	NPL	PL
NPL	0	161
PL	1	1820
Precision	0	0.92
Recall	0	1
F1-score	0	0.96

ภาพที่ 4.3 Voting Classification Model (Threshold + Optimal) โดยใช้ Feature Day past Due

VotingClassifier: Threshold ทำนายล่วงหน้า 3 เดือน โดยไม่ใช้ Feature Day past Due

	NPL	PL
NPL	0	161
PL	2	1819
Precision	0	0.92
Recall	0	1
F1-score	0	0.96

ภาพที่ 4.4 Voting Classification Model (Threshold + Optimal) โดยไม่ใช้ Feature Day past Due

4.2 ผลการวัดประสิทธิภาพโดยใช้ชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัย

4.2.1 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล แบบทำนายล่วงหน้า 1 เดือน

โดยใช้ชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยของเดือน กุมภาพันธ์ - มิถุนายน

ซึ่งผลที่ได้ สามารถสรุปผลได้ดังภาพ

VotingClassifier: Optimal

	FEB		Mar		April		May		June	
	NPL	PL	NPL	PL	NPL	PL	NPL	PL	NPL	PL
NPL	558.00	36.00	561.00	31.00	569.00	29.00	586.00	24.00	594.00	26.00
PL	37.00	6,004.00	23.00	6,086.00	31.00	6,090.00	46.00	6,166.00	47.00	6,206.00

VotingClassifier:Optimal

	FEB			Mar			April			May			June		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
NPL	0.94	0.94	0.94	0.96	0.95	0.95	0.95	0.95	0.95	0.93	0.96	0.94	0.93	0.96	0.94
PL	0.99	0.99	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	1.00	0.99	0.99

Average	Precision	Recall	F-Score
NPL	0.94	0.95	0.94
PL	1.00	0.99	0.99

ภาพที่ 4.5 Confusion Metrics (Optimal)

VotingClassifier: Threshold

	FEB		Mar		April		May		June	
	NPL	PL	NPL	PL	NPL	PL	NPL	PL	NPL	PL
NPL	559.00	35.00	572.00	20.00	572.00	26.00	581.00	29.00	591.00	29.00
PL	23.00	6,018.00	25.00	6,084.00	25.00	6,096.00	22.00	6,190.00	24.00	6,229.00

VotingClassifier: Threshold

	FEB			Mar			April			May			June		
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
NPL	0.96	0.94	0.95	0.96	0.97	0.96	0.96	0.96	0.96	0.95	0.95	0.96	0.96	0.95	0.96
PL	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Average	Precision	Recall	F-Score
NPL	0.96	0.95	0.96
PL	1.00	1.00	1.00

ภาพที่ 4.6 Confusion Metrics (Threshold + Optimal)

4.2.2 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล แบบทำนายล่วงหน้า 3 เดือน

ซึ่งผลที่ได้ สามารถสรุปผลได้ดังภาพ

VotingClassifier: Threshold ทำนายล่วงหน้า 3 เดือน โดยใช้ Feature Day past Due

	Mar ทำนาย June	
	NPL	PL
NPL	533.00	86.00
PL	42.00	6,040.00

VotingClassifier: Threshold ทำนายล่วงหน้า 3 เดือน โดยไม่ใช้ Feature Day past Due

	Mar ทำนาย June	
	NPL	PL
NPL	515.00	104.00
PL	53.00	6,029.00

ภาพที่ 4.7 Confusion Metrics (Threshold + Optimal) แบบทำนายล่วงหน้า 3 เดือน

4.3 สรุปผลการเปรียบเทียบประสิทธิภาพ

4.3.1 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล แบบทำนายล่วงหน้า 1 เดือน

จากการทดลองข้อมูล Unseen โดยใช้ชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยของเดือน กุมภาพันธ์ – มิถุนายน ผลประสิทธิภาพของ Voting Classifier ที่มีการปรับค่า Threshold จะให้ผลการทดสอบที่มีค่าประสิทธิภาพโดยเฉลี่ยมากกว่า ดังภาพที่ 4.8

มีค่า Average ที่ดีกว่า Model แบบ optimal

VotingClassifier:Optimal				VotingClassifier: Threshold + Optimal		
Average	Precision	Recall	F-Score	Precision	Recall	F-Score
NPL	0.94	0.95	0.94	**0.96**	**0.95**	**0.96**
PL	1.00	0.99	0.99	**1.00**	**1.00**	**1.00**

ภาพที่ 4.8 เปรียบเทียบผลระหว่าง (Optimal) และ (Threshold + Optimal)

ซึ่งปัญหาของ Error ส่วนใหญ่จะประกอบด้วย 2 ปัญหาหลักคือ

4.3.1.1 ปัญหาของสถานะ NPL แต่ทำนายเป็น PL ที่พบเกิดพบเกิดจากปัญหาการทำ TDR ช่วงเวลาของ Monitor กับ Class หลัง Monitor ไม่สัมพันธ์กันจึงทำให้การทำนายจากชุดข้อมูลไม่ถูกต้องบางส่วน

4.3.1.2 ปัญหาของสถานะ PL แต่ทำนายเป็น NPL ที่พบเกิดจากปัญหาการทำ TDR หรือการทำ Override Class จึงทำให้การทำนายจากชุดข้อมูลไม่ถูกต้องบางส่วน หรือเนื่องจากการจัดสถานะชั้นหนี้ที่ต้องพิจารณาจากระดับลูกค้า ซึ่งมีบัญชีสินเชื่อประเภทอื่นๆเป็น NPL จึงทำให้ทุกบัญชีสินเชื่อของ ลูกค้านั้นๆ สถานะจะเป็น NPL ทั้งหมด

4.3.2 ผลการวัดประสิทธิภาพความถูกต้องของโมเดล แบบทำนายล่วงหน้า 3 เดือน

จากการทดลองข้อมูล Unseenโดยใช้ชุดข้อมูลจำลองสินเชื่อที่อยู่อาศัยของเดือนมีนาคม ทำนายผลเดือน มิถุนายน ผลประสิทธิภาพของ Voting Classifier ที่มีการปรับค่า Threshold ทั้ง แบบที่ใช้ Feature Day Past Due เทียบกับ แบบไม่ใช้ Feature Day Past Due จะให้ผลการทดสอบที่มีค่าประสิทธิภาพโดยมากกว่า ดังภาพที่ 4.9

VotingClassifier: Threshold ทำนายล่วงหน้า 3 เดือน โดยใช้ Feature Day past Due

	Precision	Recall	F-Score
NPL	0.93	0.86	0.89
PL	0.99	0.99	0.99

VotingClassifier: Threshold ทำนายล่วงหน้า 3 เดือน โดยไม่ใช้ Feature Day past Due

	Precision	Recall	F-Score
NPL	0.91	0.83	0.87
PL	0.98	0.98	0.98

ภาพที่ 4.9 เปรียบเทียบผลระหว่าง Voting Classifier ที่มีการปรับค่า Threshold ทั้ง แบบที่ใช้ Feature Day Past Due เทียบกับ แบบไม่ใช้ Feature Day Past Due

4.4 ผลการวัดความพึงพอใจของผู้ใช้งาน

วัดผลความพึงพอใจของผู้ใช้งาน กับกลุ่มเป้าหมายที่ทำงานเกี่ยวข้องกับการทำงานโดยใช้คำถามจำนวน 5 ข้อ และข้อเสนอแนะสำหรับผลิตภัณฑ์ อีก 1 ข้อ ผลที่ได้มีดังนี้

4.4.1 คำถาม (Questionnaire)

1. Model มีประโยชน์ต่อการควบคุม NPL มากน้อยเพียงใด

2. Model สามารถลดภาระในการพิจารณาควบคุมสินเชื่อที่เป็น NPL มากน้อยเพียงใด

3. Model จะช่วยลดสินเชื่อที่เป็น NPL มากน้อยเพียงใด

4. Model จะมีประโยชน์ต่อการติดตามสินเชื่อที่เป็น NPL มากน้อยเพียงใด

5. Model เหมาะสมที่จะนำมาใช้เป็น Early Warning สินเชื่อที่เป็น NPL มากน้อยเพียงใด

6.Model เหมาะสมที่จะนำมาประกอบกับเครื่องมืออื่นเพื่อควบคุมสินเชื่อที่เป็น NPL มากน้อยเพียงใด

4.4.2 กลุ่มเป้าหมาย (Target Users) ทั้งหมด 7 กลุ่มแบ่งเป็น

ฝ่ายบริหารความเสี่ยง

ฝ่ายกำกับและระเบียบกฎเกณฑ์ธนาคาร

ฝ่ายบริหารความเสี่ยงสินเชื่อรายย่อย

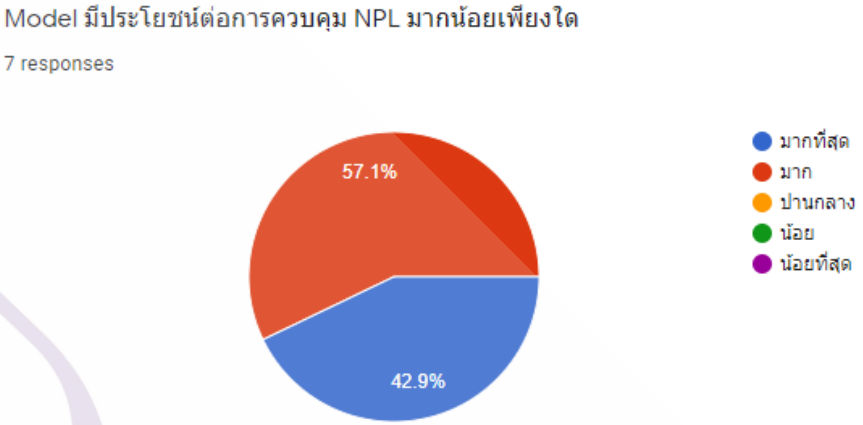
ฝ่ายกลยุทธ์ธนาคาร

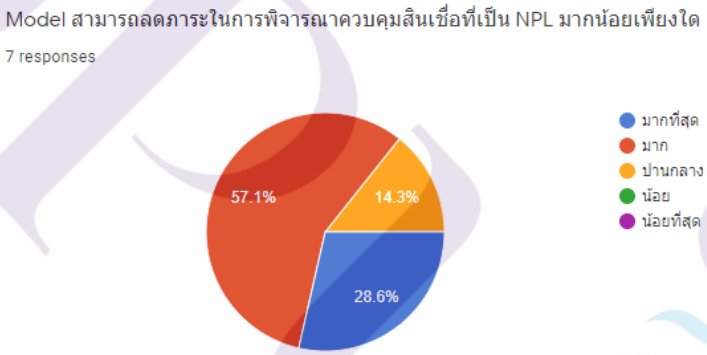
ฝ่ายพัฒนาผลิตภัณฑ์

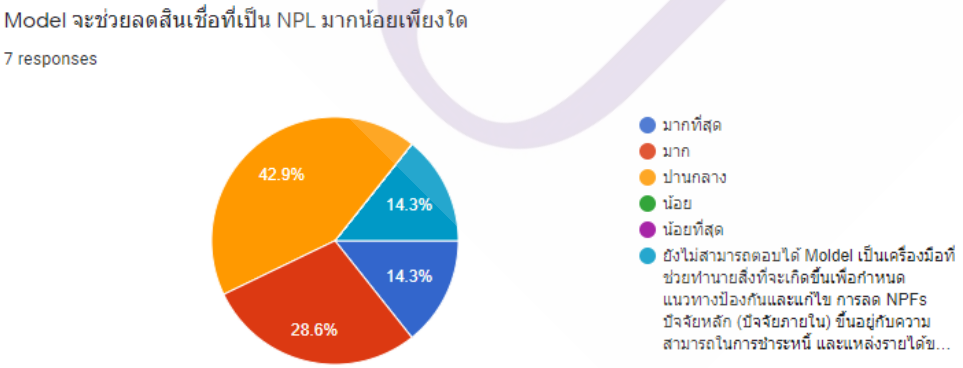
ฝ่ายบริหารและพัฒนาระบบเทคโนโลยีสารสนเทศ

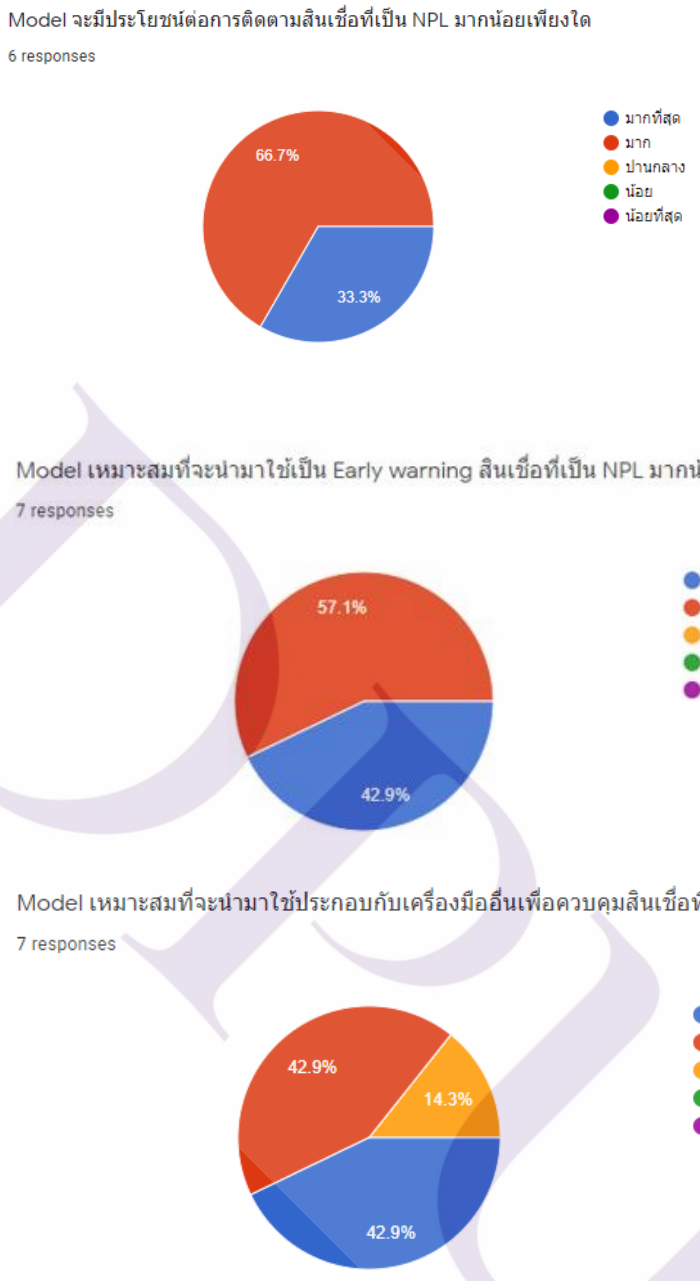
ส่วนรักษาความปลอดภัยทางเทคโนโลยี

4.4.3 ผลการประเมินความพึงพอใจ



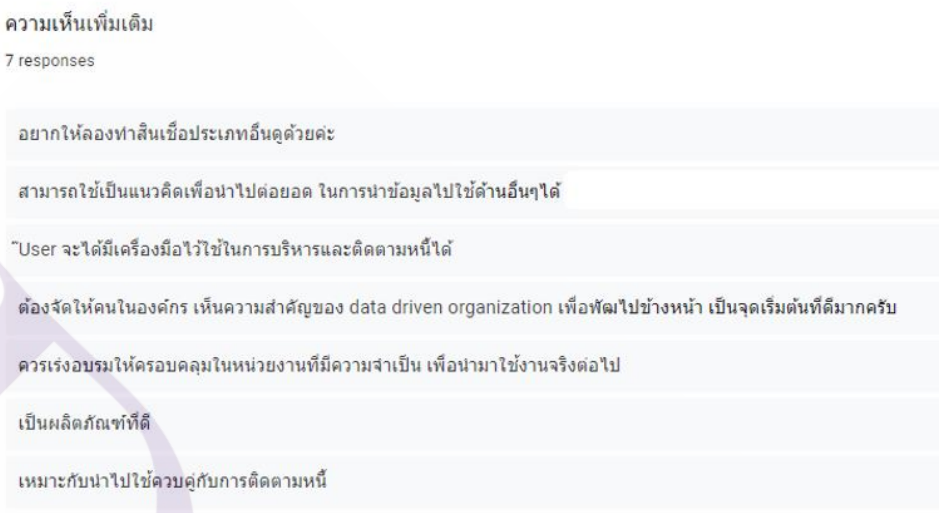






ภาพที่ 4.10 ผลการประเมินความพึงพอใจ

4.4.5 สรุปความคิดเห็นและข้อเสนอแนะสำหรับผลิตภัณฑ์ของผู้ใช้งาน



ภาพที่ 4.11 ความคิดเห็นและข้อเสนอแนะสำหรับผลิตภัณฑ์ของผู้ใช้งาน

บทที่ 5
บทสรุปและข้อเสนอแนะ

งานสารนิพนธ์เรื่องนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลสำหรับสนับสนุนการวิเคราะห์คุณภาพสินเชื่อที่อยู่อาศัยจากการเปลี่ยนแปลง สถานการณ์จัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ ซึ่งสินเชื่อที่ไม่ก่อให้เกิดรายได้ (Non-performing loan: NPL) ด้วยการประยุกต์ใช้ระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเอง (Machine Learning) โดยสรุปผลงานวิจัย ดังรายการต่อไปนี้

5.1 สรุปผลการศึกษา

5.1.1 ได้โมเดลที่มีประสิทธิภาพดีในการสถานการณ์ทำนายจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ โดยโมเดลประกอบด้วย 2 ขั้นตอนดังนี้

ขั้นตอนที่ 1 ทำนายจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้โดยใช้ Classification Model ประกอบด้วย

1. Decision Tree Model
2. Randomforest Model
3. Adaboost Model

ขั้นตอนที่ 2 Evolution model โดยปรับค่า Threshold ที่เหมาะสมในชุดข้อมูล

5.1.2 ผลการทดลองให้ความแม่นยำในการทำนายสถานการณ์ทำนายจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ แบบล่วงหน้า 1 เดือน โดยวัดค่า AVG ของ Recall ประมาณ 95% , Precision มากกว่า 95 % และ F-Score กกว่า 95 %

5.1.3 ผลการทดลองให้ความแม่นยำในการทำนายสถานการณ์ทำนายจัดชั้นจากลูกหนี้ชั้นปกติไปเป็นลูกหนี้ผิดนัดชำระหนี้ แบบล่วงหน้า 3 เดือน โดยผลการทดลองเปรียบเทียบแบบใช้ Feature Day Past Due และแบบไม่ใช้ Feature Day Past Due โดยวัดค่า ของ Recall ประมาณ 86 %, Precision มากกว่า 93 % และ F-Score กกว่า 89 %

5.1.4 ปัญหาที่พบในการทำนายเกิดจากปัญหาของสถานะ PL แต่ทำนายเป็น NPL ที่พบเกิดจากปัญหาการทำ TDR หรือการทำ Override Class จึงทำให้การทำนายจากชุดข้อมูลไม่ถูกต้องบางส่วน หรือเนื่องจากการจัดสถานะชั้นหนี้ที่ต้องพิจารณาจากระดับลูกค้า ซึ่งมีบัญชีสินเชื่อประเภทอื่นๆเป็น NPL จึงทำให้ทุกบัญชีสินเชื่อของ ลูกค้านั้นๆ สถานะจะเป็น NPL ทั้งหมด

5.1.5 ปัญหาของสถานะ NPL แต่ทำนายเป็น PL ที่พบเกิดพบเกิดจากปัญหาการทำ TDRช่วงเวลาของ Monitor กับ Class หลัง Monitor ไม่สัมพันธ์กันจึงทำให้การทำนายจากชุดข้อมูลไม่ถูกต้องบางส่วน

5.2 ข้อเสนอแนะ

5.2.1 การเพิ่มประสิทธิภาพของโมเดลให้สามารถทำงานผลิตภัณฑ์อื่นๆประกอบด้วย

5.2.2 เพิ่มโมเดลวิเคราะห์ ให้หลากหลาย และหา Feature ที่เป็น External source มาประกอบ เช่น สถานการณ์ของเศรษฐกิจในยุคปัจจุบัน

5.2.3 นำหลักการของ Business Model มาประยุกต์ใช้ให้ครอบคลุมในทุกๆผลิตภัณฑ์ และวัดผล เพื่อ evaluate model เป็นประจำ

5.2.4 สร้างความรู้ความเข้าใจ ให้กับผู้ใช้งานและผลักดันให้เกิด Data Driven Organization

บรรณานุกรม

บรรณานุกรม

Supachai(2017). เรื่องเล่าจาก TechJam by KBTG 2017 (Data Track)
https://medium.com/@supachaic/A-online-audition-techjam-2017-data-track-b7466d278301

Datacubeth(2020). ขั้นตอนการสร้างโมเดล Decision Tree
https://datacubeth.ai/decision-tree/

Datacubeth(2020). การประยุกต์ใช้ Predictive Modeling ในเชิงธุรกิจ (Business)
https://datacubeth.ai/predictive-modeling-for-business/

Datacubeth(2020). กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM
https://datacubeth.ai/crisp-dm/

Datacubeth(2020).เปรียบเทียบความถูกต้อง (Accuracy) กับการแปลความ (Explainability) ของโมเดลต่างๆ
https://datacubeth.ai/machine-learning-model-comparison/

Sirawich Jaichuen(2019). AdaBoost Algorithm
https://sirawichjaichuen.medium.com/adaboost-algorithm-cfe6b58e60fa

Alan Jeffares(2018). Supervised vs Unsupervised Learning in 3 Minutes
https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242

Java T Point(2020). Decision Tree Classification Algorithm
https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Java T Point(2020). Decision Tree Classification Algorithm
https://www.javatpoint.com/machine-learning-random-forest-algorithm

Java T Point(2020). Decision Tree Classification Algorithm
https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

Java T Point(2020). Decision Tree Classification Algorithm
https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Gaurav Singhal(2020). Ensemble Methods in Machine Learning: Bagging Versus Boosting
https://www.pluralsight.com/guides/ensemble-methods:-bagging-versus-boosting

Kostas Hatalis(2019). Can we use neural networks in ensemble learning?

https://www.quora.com/Can-we-use-neural-networks-in-ensemble-learning

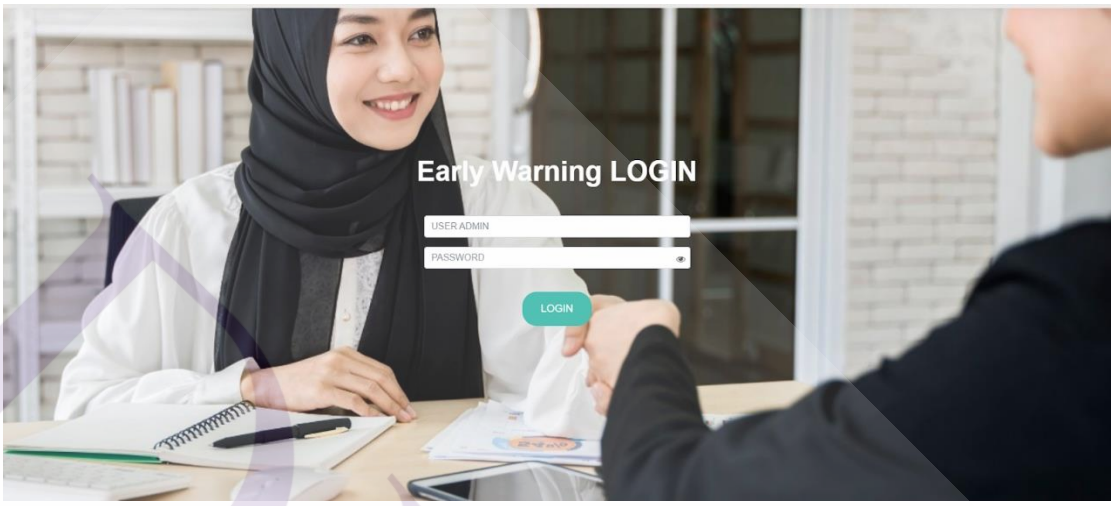
Tara Boyle(2019). Dealing with Imbalanced Data

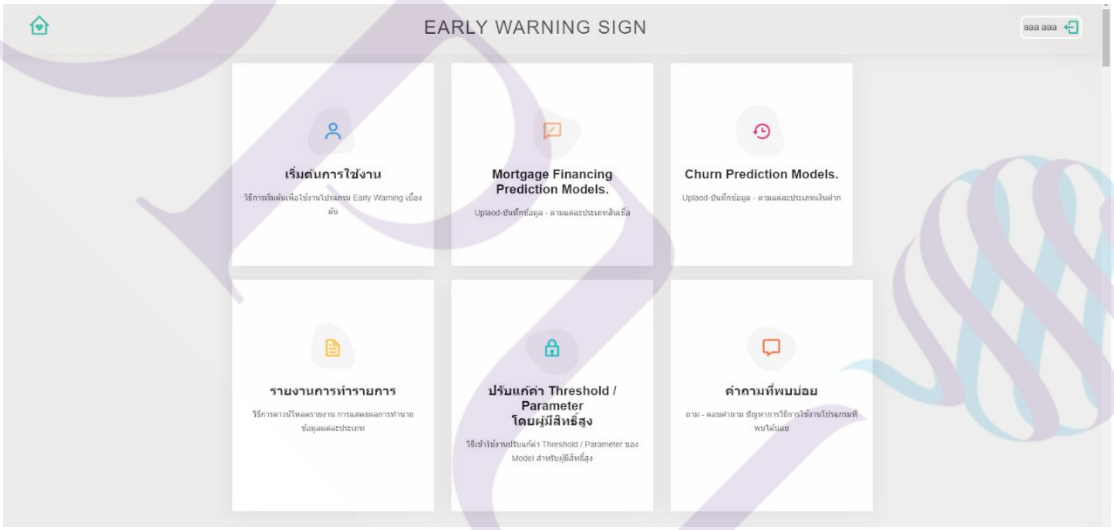
https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

ภาคผนวก

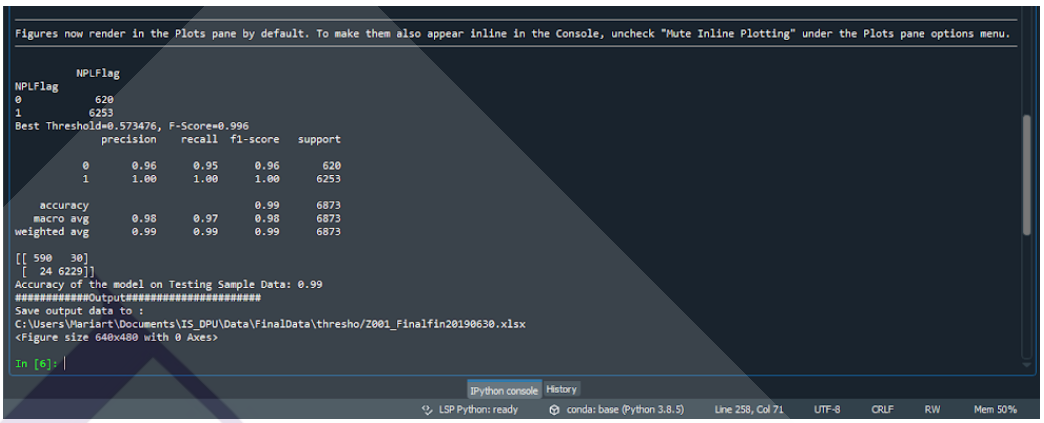
ภาคผนวก ก: ตัวอย่างหน้าจอการใช้งานระบบ

1. ตัวอย่างหน้าจอ Login และ การเข้าใช้งานระบบ



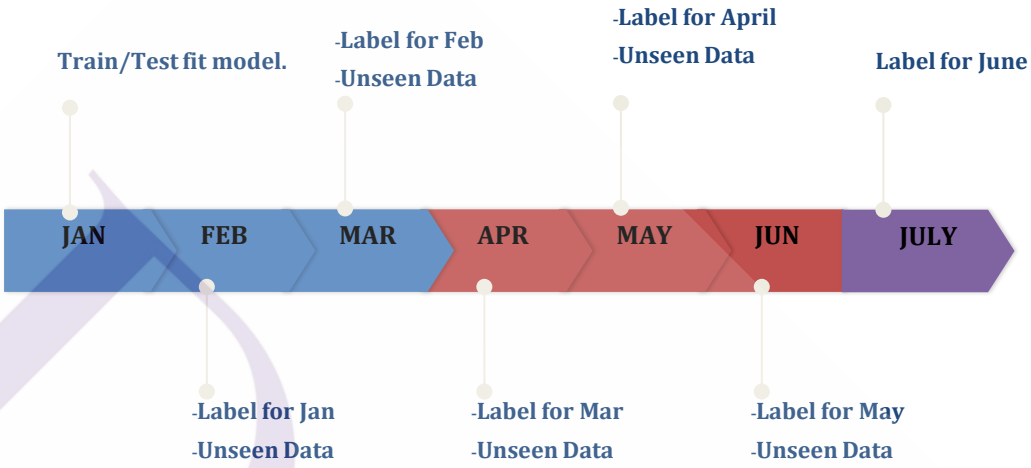


2. ตัวอย่างการ รัน Models แบบใช้ Command Line

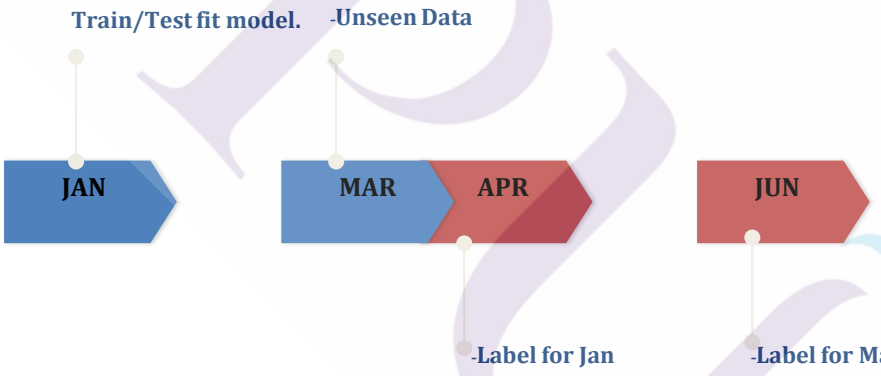


ภาคผนวก ข: ตัวอย่างการใช้งานข้อมูล

1.ตัวอย่างการใช้งานข้อมูลของการทำนายล่วงหน้า 1 เดือน



2. ตัวอย่างการใช้งานข้อมูลของการทำนายล่วงหน้า 3 เดือน



ประวัติผู้เขียน

ชื่อ-นามสกุล	ประพลเวท บุญประเสริฐ
ประวัติการศึกษา	อุตสาหกรรมศาสตร์บัณฑิต
	สาขาเทคโนโลยีอุตสาหกรรมคอมพิวเตอร์
	มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
	ปีการศึกษา 2547
ประวัติการทำงาน	ผู้จัดการอาวุโสบริหารส่วนระบบฐานข้อมูลธนาคาร