การจัดการความเสี่ยงในการผิดนัดชำระหนี้ส่วนบุคคลแบบออนไลน์

Risk Default Management in Online Peer to Peer Lending

อติวิชญ์ ชนินทร์โชดึก[1*] และ เอกรัฐ รัฐกาญจน์[2]

บทคัดย่อ

ธุรกิจปล่อยกู้เงินแบบออนไลน์ Peer to Peer Lending จากชุดข้อมูล Lending club จะเห็นว่ามีอัตราการผิดนัดชำระหนี้ของผู้กู้สูง เนื่องจากผู้กู้ไม่ต้องมีทรัพย์สินมาใช้เป็นหลักประกันความเสี่ยง การพิจารณาเครดิตของผู้กู้เป็นปัจจัยหนึ่งในการคัดเลือกปล่อยกู้ให้กับผู้กู้แต่ละราย ในงานวิจัยนี้ต้องการใช้เทคนิคการเรียนรู้ของเครื่องมาช่วยในการหาผู้กู้ที่มีแนวโน้มในการผิดนัดชำระหนี้ และหาปัจจัยที่ส่งผลกระทบต่อการผิดนัดชำระหนี้ของผู้กู้ได้ รวมถึงการใช้เทคนิค sampling ในการแก้ปัญหาข้อมูล Imbalance และเทคนิค feature selection ที่ช่วยเพิ่มประสิทธิภาพของโมเดลให้มีประสิทธิภาพเพิ่มมากขึ้น ซึ่งผลลัพธ์ที่ทำให้โมเดลมีค่ามากที่สุด คือ การรวมกันระหว่างเทคนิค IHT Under sampling และโมเดล Neural network ได้ประสิทธิภาพ recall 93.13% และค่า ROC Curve 59.2% และปัจจัยที่ส่งผลในการทำนายว่าผู้กู้มีแนวโน้มในการผิดนัดชำระหนี้หรือไม่ 5 อันดับแรกคือ อัตราดอกเบี้ย, คะแนน credit scoring, ระยะเวลาในการกู้, สถานะบ้านที่อยู่อาศัยของผู้กู้ และจำนวนเงินเฉลี่ยในบัญชีทั้งหมด

คำสำคัญ: สินเชื่อส่วนบุคคลแบบออนไลน์, การผิดนัดชำระหนี้, การเรียนรู้ของเครื่อง, ความไม่สมดุลของข้อมูล, การจำแนกข้อมูล

Abstract

Peer to Peer Lending from Lending Club dataset to show the borrowers default rate is high due to the borrowers don’t have any assets to use as collateral. Consideration of the borrower’s credit is one of the factors in selecting a loan for each borrower. This research use Machine Learning Techniques to help identify borrowers with a tendency to default and find factors that affect the borrower’s default. Including use Sampling Techniques to fix imbalance problems and Feature Selection Techniques to optimize model performance. As a result, the best performance of model is combination between the IHT Under-Sampling Technique and Neural Network Model. The result show that Recall efficiency is 93.13% and ROC Curve is 59.2%. Top 5 factors influencing the prediction of borrower’s tendency to default were interest rate, credit scoring, loan duration, borrower housing status, and average amount in all accounts

***Keywords*: Peer to Peer Lending, Default, Machine Learning, Imbalance, Classification**

---

[*] *Ativit.chan@stu.nida.ac.th*

[1,2] *คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์*

1. บทนำ (introduction)

ในปัจจุบัน เทคโนโลยีเข้ามามีบทบาทอย่างมากต่อการใช้ชีวิตประจำวัน ทำให้สะดวกสบายและรวดเร็วมากขึ้น เห็นได้จากพฤติกรรมการของผู้บริโภคที่ใช้บริการผ่านช่องทางออนไลน์ที่เพิ่มสูงขึ้น เช่น การซื้อสินค้าผ่านช่องทางออนไลน์ การจองโรงแรม การจองตั๋วหนังภาพยนตร์ การสั่งอาหารออนไลน์ รวมถึงธุรกรรมทางการเงินต่าง ๆ ที่เข้าถึงได้ง่าย Peer to Peer Lending คือ ธุรกรรมสินเชื่อระหว่างบุคคล (ผู้กู้) และบุคคล (นักลงทุน) ผ่านระบบหรือเครือข่ายอิเล็กทรอนิกส์ ซึ่งจะทำหน้าที่เป็นกลางทางการเงินแทนธนาคาร การกู้ยืมแบบดั่งเดิมที่ใช้ธนาคารเป็นตัวกลางทางการเงินจะมีต้นทุนที่สูง ซึ่งการขอสินเชื่อกับธนาคารจะต้องมีประวัติการเงินและสินทรัพย์ค้ำประกัน ทำให้ผู้กู้บางกลุ่มโดยเฉพาะกิจการขนาดเล็ก (SME) ไม่สามารถทำธุรกรรมทางการเงินได้ (พรชนก เทพขาม, 2019)

Lending Club เป็น Peer to peer platform หรือสินเชื่อระหว่างบุคคลที่ใหญ่ที่สุดในโลกเกิดขึ้นที่ประเทศสหรัฐอเมริกา เกิดขึ้นในช่วงกลางปี 2007 มีสินเชื่อบริการหลากหลายประเภท เช่น สินเชื่อบุคคล สินเชื่อบ้าน สินเชื่อรถยนต์ เป็นต้น และยังสามารถกู้แบบเดี่ยวและกู้ร่วมได้อีกด้วย Lending Club เป็นแพลตฟอร์มที่ทำหน้าเป็นตัวกลางในการตรวจสอบและปล่อยกู้ให้ผู้กู้กับนักลงทุนผ่าน Platform ของ Lending Club ซึ่งได้รับความนิยมเป็นอย่างมากและเติบโตขึ้นอย่างรวดเร็ว มีมูลค่าสูงถึง 53,722 ล้านดอลล่าร์ และมีส่วนแบ่งการตลาดถึง 71.9% ในประเทศสหรัฐอเมริกา ข้อดีของสินเชื่อระหว่างบุคคลคือ ผู้กู้ไม่จำเป็นต้องมีหลักประกัน ทำให้มีความเสี่ยงสูงแต่ก็แลกมาด้วยผลตอบแทนที่สูงด้วยเช่นกัน ความน่าสนใจของแพลตฟอร์ม peer to peer Lending คือความสะดวกรวดเร็ว ซึ่งเกิดจากการนำเทคโนโลยีมาใช้ลดขั้นตอนที่ยุ่งยากและใช้เวลานานทำให้มีต้นทุนถูกลง แต่ความเสี่ยงที่ผู้ใช้บริการ Peer to peer lending มีความกังวลมากที่สุด 3 อันดับแรก ได้แก่ อันดับ 1 ภัยคุกคาม ทางไซเบอร์ หรือความเสี่ยงที่จะถูกโจมตีทางอิเล็กทรอนิกส์ อันดับ 2 ธุรกิจแพลตฟอร์มล้ม (Platforms' collapse) จากการขาดความน่าเชื่อถือซึ่งอาจเกิดได้จากหลายประการเช่น ความไม่โปร่งใสของแพลตฟอร์มจากการหลีกเลี่ยง การเปิดเผยข้อมูล ทำให้นักลงทุนถอนเงินลงทุนออก ส่งผลให้แพลตฟอร์มขาดสภาพคล่องอย่างหนัก และนำไปสู่ การปิดตัวของแพลตฟอร์มในที่สุด อันดับ 3 การฉ้อโกงจากบุคคลที่สาม (Fraud) ด้วยวิธีการต่าง ๆ เช่น กู้ยืมเงิน ผ่านช่องทางออนไลน์ด้วยตัวตนที่ไม่มีจริง ผู้ใช้บริการแพลตฟอร์มก็มีความกังวลเรื่องการผิดนัดชำระหนี้ (Default) ซึ่งเป็นความเสี่ยงที่ผู้กู้ยืมสินเชื่อจะผิดนัดชาระหนี้ เป็นจำนวนมากโดยเฉพาะในช่วงที่ประสบกับวิกฤต เศรษฐกิจ และความเสี่ยงที่นักลงทุนรายใหญ่ อาทิ นักลงทุน สถาบัน จะเข้ามาแย่งการลงทุนของนักลงทุนรายย่อย (Crowding out of retail investors) ทำให้ผลตอบแทน โดยรวมลดลงได้ (เปมิกา ฐิติเพิ่มพงศ์, 2020)

1.1 ปัญหา

ทางบริษัท Lending club ยังคงประสบปัญหาการผิดนัดชำระหนี้ของผู้กู้อยู่สม่ำเสมอ ซึ่งเป็นสิ่งที่ไม่สามารถคาดการณ์ได้ ทำให้ไม่สามารถหาแนวทางการแก้ไขให้กับผู้กู้ได้ทันท่วงที ซึ่งส่งผลทำให้เกิดเป็นหนี้เสียให้กับทางบริษัทในอนาคตได้

1.2 วัตถุประสงค์ในการศึกษา

1.2.1. เพื่อสร้างตัวแบบในการวิเคราะห์และคาดการณ์ว่าผู้กู้คนไหนมีแนวโน้มที่จะผิดนัดชำระหนี้

1.2.2. เพื่อหาปัจจัยที่ส่งผลต่อการผิดนัดชำระหนี้ของผู้กู้

2. งานวิจัยที่เกี่ยวข้อง (Literature Review)

Author	Topic	Dataset	#Data	#Attribute	Method
Zhiqiang Li et al., (2021)	Application of XGBoost in P2P Default Prediction	Lending club	2,260,699	40	Xgboost
Anahita Namvar et al., (2018)	Credit risk prediction in an imbalanced social lending environment	Lending club	66,376	43	Logistic Regression, Linear Discriminate Analysis, Random Forest
Kim and Cho (2017)	Dempster-Shafer Fusion of Semi-supervised Learning Methods for Predicting Defaults in Social Lending	Lending club	332,844	17	Decision Tree
Fu (2017)	Combination of Random Forests and Neural Networks in Social Lending	Lending club	1,320,000	13	Combination of random forest and neural network
Zhang et al., (2017)	Determinants of loan funded successful in online P2P Lending	Paipai	193,614	21	Logistic regression

จากการศึกษางานวิจัยที่เกี่ยวข้อง การใช้เทคนิค XGBoost algorithm ทำนายหาผู้กู้ที่มีโอกาสผิดนัดชำระหนี้ โมเดลให้ค่า accuracy 97.71% (Zhiqiang Li., 2021) การใช้เทคนิค resampling มาช่วยแก้ปัญหา imbalance data ทำให้สามารถเพิ่มประสิทธิภาพของ model ได้ดี และวิธีที่ได้ผลดีที่สุดคือ การผสมกันระหว่าง Random forest และ Under sampling (Anahita Namvar., 2017) การนำทฤษฎี Dempster-Shafer มาช่วยในการ label data และใช้เทคนิค Support Vector Machine มาทำนายผล สามารถเพิ่มประสิทธิภาพได้ 6.15% (Kim and Cho., 2017) การใช้เทคนิค Ensemble Method ระหว่างโมเดล Random forest กับโมเดล Neural Network ได้ค่า accuracy 73.5% (Fu, 2017) การนำ data set ที่มีลักษณะคล้ายคลึงกันมาใช้เพื่อหาปัจจัยที่มีผลต่อการผิดนัดชำระหนี้โดยใช้ข้อมูลของผู้กู้ใน Paipai Platform ของประเทศจีน ซึ่งปัจจัยที่มีผลกระทบต่อการผิดนัดชำระหนี้ประกอบด้วย annual interest rate, repayment period, description, credit grade, successful loan number, failed loan number, gender, and borrowed credit score (Zhang., 2017)

3. ทฤษฎีที่เกี่ยวข้อง (Background)

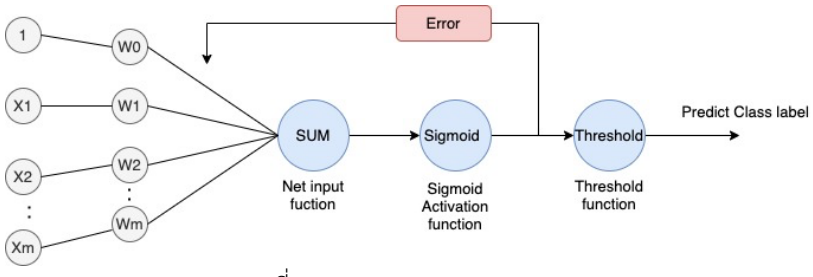
3.1 Credit Scoring

แบบจำลองที่ใช้กระบวนการทางสถิติในการจัดการข้อมูลเพื่อกำหนดเป็นค่าคะแนน Credit score ซึ่งใช้เป็นตัววัดความน่าจะเป็นในการชำระหนี้คืน Credit Scoring ถูกนำมาใช้อย่างแพร่หลายในการพิจารณาคำขอสินเชื่อ เนื่องจากคุณภาพของผู้กู้ส่งผลโดยตรงต่อความสามารถในการทำกำไรและความมั่นคงของสถาบันการเงิน การคัดกรองและพิจารณาคำขอของผู้กู้ยืมจึงเป็นขั้นตอนสำคัญในการป้องกันความเสี่ยงด้านเครดิต ชำระหนี้ โดยทางสถาบันการเงินจะใช้เป็นขั้นตอนการพิจารณาเพื่อลดความเสี่ยง และเพิ่มประสิทธิภาพในขั้นตอนการประเมินความ

เสี่ยงด้านเครดิต นอกจากนี้ Credit score ยังสามารถใช้เป็น Lending indicator เพื่อติดตามคุณภาพและความสามารถในการชำระหนี้ของลูกหนี้ที่จะเกิดจากการผิดนัดชำระหนี้ได้ (อโนทัย พุทธารี และคณะ, 2018)

3.2 Logistic Regression model

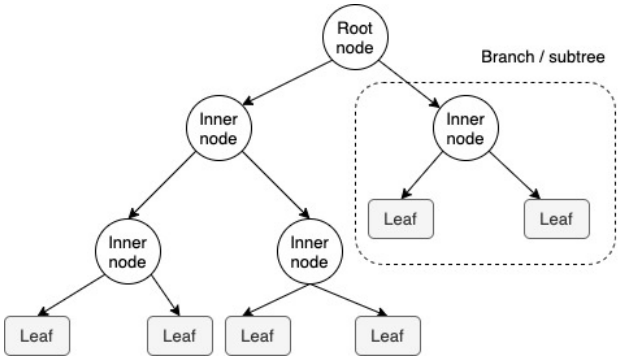
เป็นวิธีทาสถิติที่ใช้ในการหาความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม เพื่อทำนายโอกาสที่จะเกิดขึ้นและใช้ sigmoid function เพื่อทำนาย Class ที่เกิดขึ้นว่าเป็น 0, 1 ดังรูปภาพที่ 1



รูปที่ 1 Logistic Regression Model

3.3 Decision Tree model

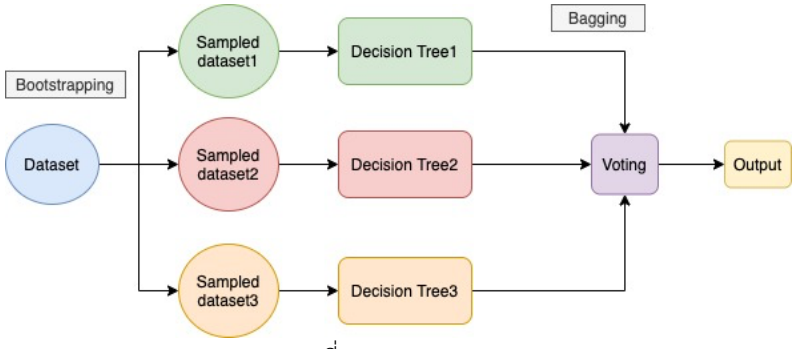
เป็นวิธีหนึ่งที่จะประมาณฟังก์ชันที่มีค่าไม่ต่อเนื่องด้วยแผนผังต้นไม้ ประกอบด้วยเซตของกฎต่าง ๆ เช่น ถ้า-แล้ว (if-then) เพื่อให้มนุษย์สามารถอ่านแล้วเข้าใจการตัดสินใจของต้นไม้ได้ดี โดยพิจารณาจากลักษณะของวัตถุ บัพภายใน (Inner node) ของต้นไม้จะแสดงตัวแปร ส่วนกิ่งของต้นไม้จะแสดงค่าที่เป็นไปได้ของตัวแปร ส่วนใบจะแทนประเภทของวัตถุ ต้นไม้ตัดสินใจ ถูกนำมาใช้งานอย่างแพร่หลาย ซึ่งนิยมใช้กันมากในการบริหารความเสี่ยง (Risk management) การทำงานนั้นต้นไม้การตัดสินใจจากบนลงล่างด้วยการถามว่าลักษณะใด ควรจะเป็นรากของต้นไม้การตัดสินใจนี้ ทำการถามซ้ำไปเรื่อย ๆ เพื่อหาต้นไม้ทั้งต้น โดยการเลือกว่าลักษณะใดดีที่สุดนั้นดูจากค่าของลักษณะ เรียกว่าเกนความรู้ (Information gain) และค่าที่ใช้บอกความไม่บริสุทธิ์ของข้อมูล เรียกว่า เอนโทรปี (Entropy) ดังรูปภาพที่ 2



รูปภาพที่ 2 Decision Tree model

3.4 Random Forest model

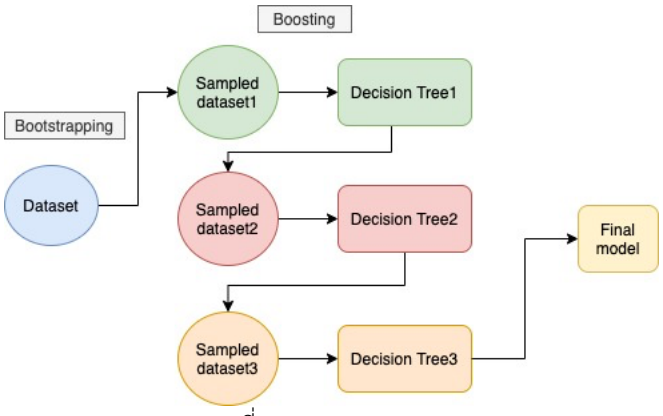
หลักการของ Random Forest คือการสร้างโมเดลจาก Decision Tree หลาย ๆ โมเดลโดยแต่ละโมเดลจะได้รับข้อมูลไม่เหมือนกัน แล้วทำการ Prediction ออกมาแต่ละโมเดล และทำการ Vote โมเดล Decision Tree ที่ให้ค่ามากที่สุด หรือเป็นการหาค่าเฉลี่ยของ Output ที่ออกมาทั้งหมด ดังรูปที่ 3



รูปภาพที่ 3 Bagging method

3.5 Gradient boosting model

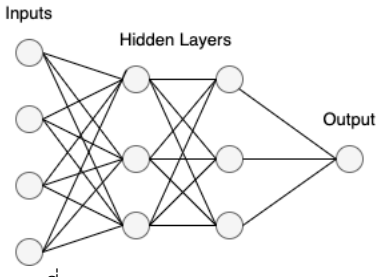
เป็นโมเดลที่นำเอา Decision Tree มาทำการ train ต่อกันหลาย ๆ tree โดยที่ Decision Tree จะเรียนรู้จากความผิดพลาดของต้นไม้ก่อนหน้า โดยให้น้ำหนักที่แตกต่างกันออกไป ถ้า classifier ไหนที่มีการทำนายผิดบ่อย ๆ จะให้น้ำหนักมากกว่าเพื่อให้มีการสุ่มขึ้นมาใหม่อีกครั้ง และทำการทำนายค่าไปเรื่อย ๆ ซึ่งจะทำให้มีความแม่นยำขึ้นเรื่อย ๆ ในแต่ละรอบและโมเดลจะทำการหยุดก็ต่อเมื่อ มีการ train ที่ลึกเกินไป และโมเดลไม่สามารถหารูปแบบของความผิดพลาดได้จากต้นไม้ก่อนหน้า หรือได้ผลรวมของ classifier ที่ดีที่สุด ซึ่งโมเดลที่มีลักษณะการทำงานคล้าย ๆ กันเช่น XGBoost, CatBoost, LightGBM ดังรูปที่ 4



รูปภาที่ 4 boosting method

3.6 Neural Network Model

เป็นชุดของ เซลล์ประสาท (neurons) ที่เชื่อมต่อ (connect) ระหว่างกัน หน้าที่ของมัน คือการรับข้อมูลทั้งหมด ดำเนินการผ่านฟังก์ชัน และส่งผลลัพธ์ไปยัง output แต่ละการเชื่อมต่อจะมีพารามิเตอร์ตัวหนึ่ง คือน้ำหนัก (Weight) ซึ่งบอกความน่าจะเป็นที่เกิดขึ้นถูกกำหนดโดยการสุ่ม และน้ำหนักจะถูกปรับในระหว่างการ Train โมเดล ด้วยวิธีคำนวณน้ำหนักย้อนกลับ (Back-propagation) โดยใช้ Gradient Descent Method เพื่อ Minimize ค่า SSE ดังรูปภาพที่ 5



รูปภาพที่ 5 Neural Network Architecture

3.7 เทคนิคการปรับเพิ่มและลดข้อมูลด้วยวิธีการสุ่ม

การจัดการความไม่สมดุลของข้อมูล (Imbalance) เกิดขึ้นเมื่อจำนวนข้อมูลของแต่ละคลาสแตกต่างกันมาก ทำให้ผลลัพธ์จากการจำแนกข้อมูลมีความโน้มเอียงไปทางข้อมูลที่มีจำนวนมาก ส่งผลให้ตัวแบบเรียนรู้ข้อมูลผิดพลาด ซึ่งเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีการสุ่ม โดยการสุ่มเพิ่มจำนวนของข้อมูลกลุ่มน้อย ๆ ให้มีปริมาณใกล้เคียงกับข้อมูลกลุ่มมาก และเทคนิคการปรับลดข้อมูลด้วยวิธีการสุ่มโดยการสุ่มลดจำนวนของข้อมูลกลุ่มมาก ให้มีปริมาณใกล้เคียงกับกลุ่มที่น้อย

3.7.1. SMOTE (synthetic minority over sampling technique)

เป็นการปรับเพิ่มข้อมูลกลุ่มน้อย ให้มีจำนวนเพิ่มขึ้น โดยทำการสุ่มค่าข้อมูลที่อยู่ในกลุ่มน้อย ขึ้นมา 1 ค่าด้วยวิธี K-nearest neighbor แล้วทำการคำนวณหาระยะห่างระหว่างจุดด้วยวิธี Euclidean distance ระหว่างค่าที่สุ่มกับข้อมูลใกล้เคียง และทำการเลือกข้อมูลที่สุ่มขึ้นมาใหม่จากระยะห่างระหว่างจุดที่ใกล้กันมากที่สุด ยกตัวอย่างงานวิจัยที่ใช้ SMOTE เทคนิคมาช่วยแก้ปัญหา Imbalance ในการจำแนกรายได้ของผู้ประกอบการร้านขายยาด้วยวิธีการสุ่มตัวอย่างเพิ่มด้วยการสังเคราะห์ซึ่งทำให้โมเดลทำนายผลออกมาดีที่สุด (นพมาศ อัครจันทโชต และคณะ, 2019)

3.7.2. ADASYN (ADAptive SYNthetic)

เป็นวิธีการสุ่มข้อมูลที่ถูกพัฒนามาจาก SMOTE เทคนิค ซึ่งขั้นตอนการสร้างข้อมูลขึ้นมาใหม่ ไม่จำเป็นต้องพิจารณา เฉพาะชุดข้อมูลของกลุ่มน้อยเท่านั้น โดยเทคนิคนี้จะใช้วิธีการ แจกแจงแบบถ่วงน้ำหนัก (Weight distribution) ของชุดข้อมูลในกลุ่มน้อย โดยจะพิจารณาจากความสำคัญของข้อมูลนั้น ถ้าข้อมูลตัวใดสามารถแบ่งกลุ่มยากก็จะให้น้ำหนักมากกว่า ถ้าข้อมูลตัวใดที่สามารถแบ่งกลุ่มได้ง่ายก็จะให้น้ำหนักน้อย ซึ่งวิธีนี้จะทำให้การสร้างข้อมูลชุดใหม่ มีการตัดสินใจแบ่งกลุ่มดีขึ้น ยกตัวอย่างงานวิจัยที่ใช้เทคนิค ADASYN ในการปรับข้อมูล

Imbalance ช่วย 2 เรื่องหลักคือ 1. ลดความไม่สมดุลของข้อมูล 2. ช่วยปรับของเขตการตัดสินใจให้ง่ายขึ้น (Haibo He et al., 2008)

3.7.3. IHT (Instant Hardness Threshold)

เป็นเทคนิคที่ใช้วิธีการสุ่มตัวอย่างหาข้อมูลในชุดข้อมูลที่มีคุณสมบัติที่เรียกว่า hardness ซึ่งบ่งชี้ความน่าจะเป็นที่ข้อมูลจะถูกจัดประเภทผิดโดยจะคำนวณค่า Instant Hardness ซึ่งจะได้ค่า Probability ของแต่ละ Data point ออกมาแล้วทำการลบ Data point ที่มีค่ามากกว่า threshold ที่กำหนด ยกตัวอย่างงานวิจัยที่ได้ทำการแก้ปัญหา Imbalance ในงาน hate speech classification ด้วยวิธี IHT สามารถทำงานได้เร็วขึ้น และเพิ่มค่า accuracy ให้กับโมเดลได้ (Naufal Azmi Verdikha., 2018)

4. ความสำคัญของปัญหา (Problem Statement)

4.1 แหล่งที่มาของข้อมูล

ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลจากเว็ปไซต์ Lending Club ซึ่งมีการเผยแพร่ข้อมูลไว้ในเว็ปไซต์ Kaggle (https://www.kaggle.com/ethon0426/lending-club-20072020q1) มีการเก็บรวมรวมข้อมูลของผู้กู้ที่เข้ามาใช้บริการเป็นระยะเวลาตั้งแต่ ปี 2007-2020Q3 มีจำนวนข้อมูลผู้กู้ทั้งหมด 2,925,493 และจำนวนคุณลักษณะของผู้กู้ทั้งหมด 141 คอลัมน์

4.2 ผลลัพธ์

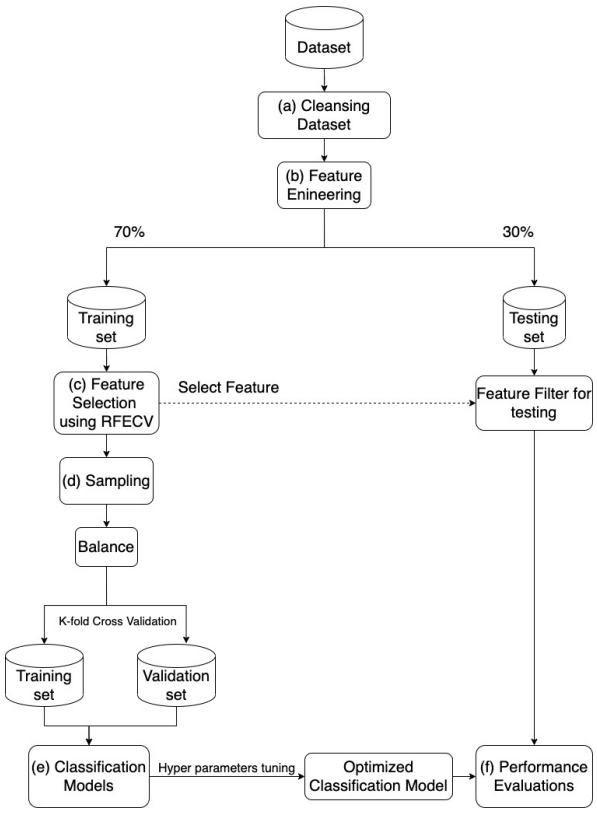
โมเดลจะทำนายผลออกมาเป็น Binary Classification โดยกำหนดให้ผู้กู้ที่มีแนวโน้มผิดนัดชำระหนี้เป็น 0 และผู้กู้ที่ไม่มีแนวโน้มผิดนัดชำระหนี้เป็น 1

4.3 วัตถุประสงค์

4.3.1. เพื่อสร้างตัวแบบในการหาผู้กู้ที่มีแนวโน้มในการผิดนัดชำระหนี้ได้

4.3.2. เพื่อศึกษาปัจจัยที่ส่งผลต่อการผิดนัดชำระหนี้ของผู้กู้

4.4 วิธีดำเนินการวิจัย (Methodology)



รูปที่ 6 กระบวนการทำงาน

4.4.1.การรวบรวมข้อมูล

ทำการเลือกข้อมูลของผู้กู้และผู้กู้ร่วมที่ใช้ในการพิจารณาปล่อยกู้ ทางผู้วิจัยได้รวบรวมตัวแปรที่มีความสำคัญและใช้ในการพิจารณามาทั้งหมด 31 ตัวแปร ดังตารางที่ 1

ตารางที่ 1 รายละเอียดตัวแปร

Variable	Description
annual_inc	Balance to credit limit on all trades
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
Emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

Emp_title	The job title supplied by the Borrower when applying for the loan.
Fico_range_low	The lower boundary ranges the borrower’s FICO at loan origination belongs to.
Grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are RENT, OWN, MORTGAGE, OTHER
installment	The monthly payment owed by the borrower if the loan originates.
Int_rate	Interest Rate on the loan.
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
Loan_status	Current status of the loan.
mort_acc	Number of mortgage accounts.
Pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
Revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
Sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
Verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
last_pymnt_d	Last month payment was received
open_acc	Number of open trades at time of application for the secondary applicant
total_acc	The total number of credit lines currently in the borrower's credit file
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
dti_joint	A ratio calculated using the co-borrowers’ total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers’ combined self-reported monthly income
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances
sec_app_fico_range_low	FICO range (high) for the secondary applicant
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant
sec_app_open_acc	Number of open trades at time of application for the secondary applicant

4.4.2.การพัฒนาตัวแบบการพยากรณ์

(a) ทำความสะอาดข้อมูล (Cleansing Data)

ทางผู้วิจัยเห็นว่าจำนวนข้อมูลที่นำมาใช้มีจำนวนมาก ทำให้การทำงานของโปรแกรมช้าลง และเห็นว่าสินเชื่อที่ทาง Lending Club ปล่อยให้กับผู้กู้มีระยะเวลาเพียง 3-5 ปีเท่านั้นในที่นี้ผู้วิจัยจึงเลือกช่วงเวลาที่นำมาวิเคราะห์เพียง 5 ปีล่าสุด ตั้งแต่ปี 2016-2020Q3 ซึ่งทำให้เหลือข้อมูลผู้กู้จำนวน 2,038,052 ผู้กู้

เนื่องจากในชุดข้อมูลที่ได้มีสถานะของผู้กู้ที่แตกต่างกันไป ในงานวิจัยนี้ต้องการที่จะหาว่าผู้กู้คนไหนที่มีแนวโน้มในการผิดนัดชำระหนี้ จึงทำการเลือกกลุ่มเป้าหมายคือ Fully Paid และ Charged Off โดยแทนค่าเป็น 0 และ 1 ตามลำดับ ทำให้ได้ข้อมูลผู้กู้ออกมาจำนวน 993,929 ผู้กู้ประกอบด้วย Fully Paid 788,161 จำนวนและ Charged Off 205,768 จำนวน

ทำการลบข้อมูลที่มี Missing values มากกว่า 50% ของทั้งหมดใน columns นั้นและทำการลบ values ที่เป็นค่า Null เนื่องจากข้อมูลมีจำนวนมาก และทางผู้วิจัยไม่ทราบถึงค่าที่ null อย่างแท้จริงการนำมาเติมค่าอาจจะทำให้การวิเคราะห์ผิดเพี้ยนไปได้

(b) Feature Engineering

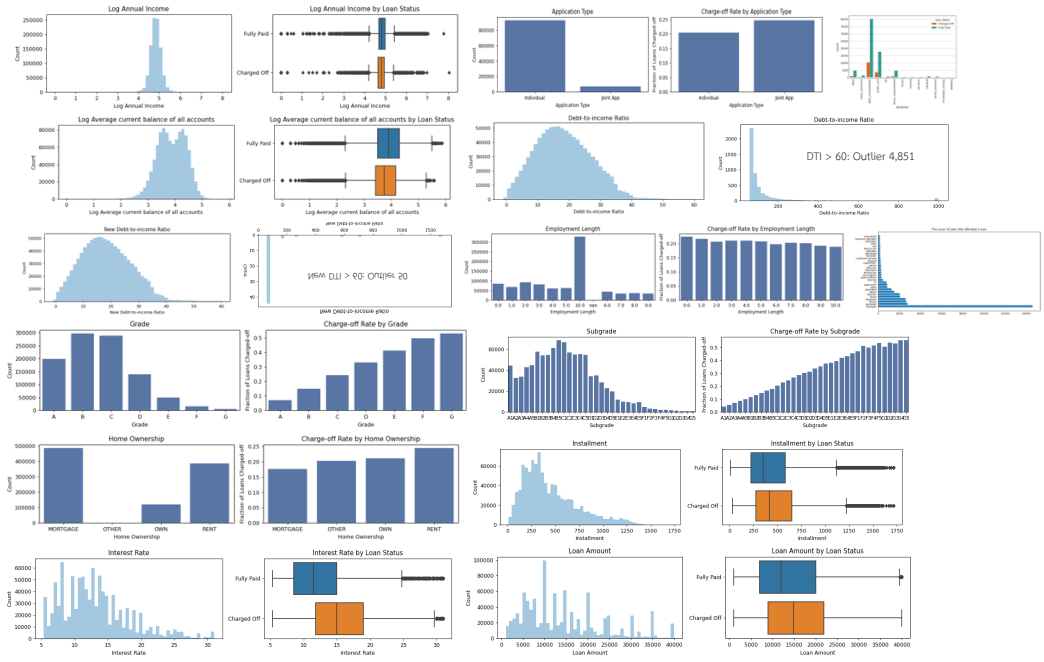
ทำการคำนวณตัวแปรใหม่ที่ใช้เป็นปัจจัยในการผิดนัดชำระหนี้ของผู้กู้ได้คือ New DTI โดยการนำ DTI ที่คำนวณได้มารวมกับหนี้ก้อนใหม่ที่ผู้กู้จะต้องชำระหนี้ โดยคำนวณจากสมการที่ 1

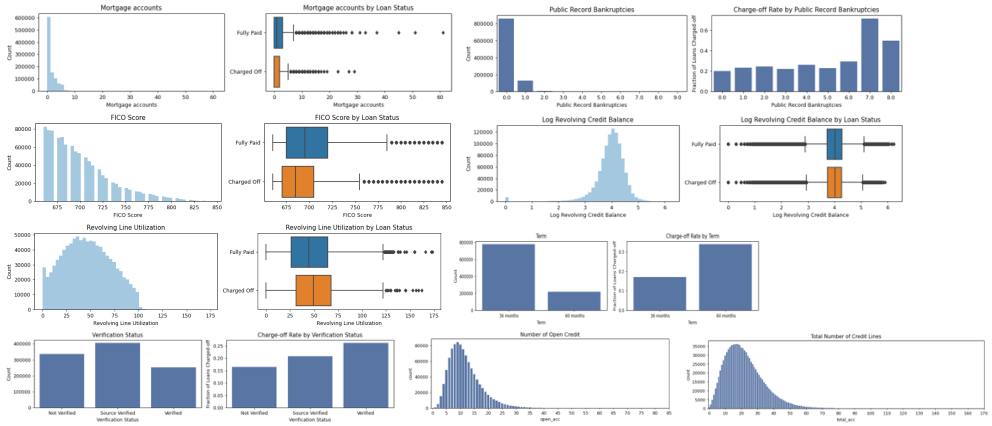
$$New\ DTI = \frac{\text{New monthly repayment amount}}{\text{monthly income}} \qquad (1)$$

เมื่อ $New\ monthly\ repayment\ amount = (DTI * Annual\ income) + installment$

เมื่อ $DTI = \frac{installment}{Annual\ income}$

หลังจากนั้นทำการ Exploratory data analysis เพื่อทำความเข้าใจของข้อมูลแต่ละตัวแปร ดูการกระจายตัวของข้อมูล และปรับรูปแบบข้อมูลบางตัวแปร โดยใช้เทคนิค Log transform ให้มีการกระจายตัวแบบปกติมากขึ้น ดังรูปที่ 7



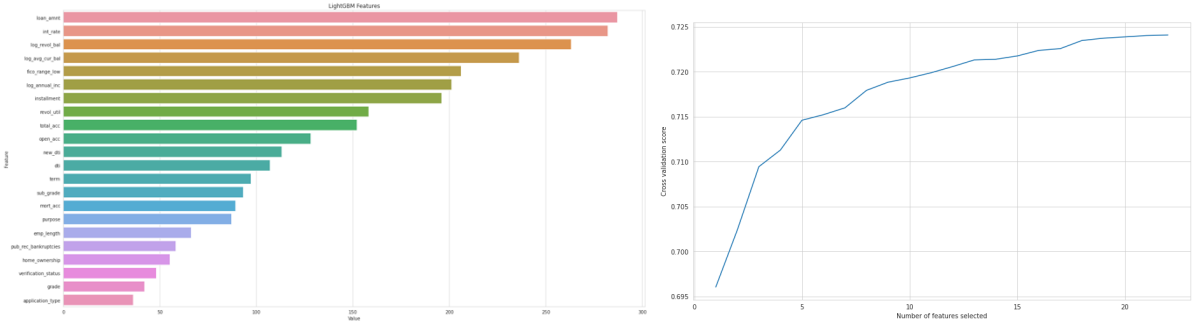


รูปภาพที่ 7 Exploratory Data Analysis

หลังจากทำ Feature Engineering เสร็จแล้วจะทำการ Transform Data โดยข้อมูลที่เป็น Numerical จะทำใช้วิธี Standardization และข้อมูลที่เป็น Categorical จะทำใช้วิธี Lebel Encoder

(a) Feature Selection Using REFCV (Recursive Feature elimination)

ใช้เพื่อหาคัดเลือก feature ให้มีจำนวนน้อยที่สุดแล้วยังได้ประสิทธิภาพมากที่สุด เพื่อช่วยให้ขั้นตอนการทำงานเร็วขึ้นและซับซ้อนน้อยลง จากภาพที่ 8 จะเห็นว่าเมื่อใช้วิธี RFE ในการเลือก feature เพื่อเข้าโมเดล LightGBM ประสิทธิภาพการทำนายผลเริ่มลดลงเรื่อย ๆ เมื่อตัด feature ออกไปต่ำกว่า 20 ตัวแปร ดังนั้น feature ที่เลือกไว้เข้าโมเดลจะเหลือจำนวนเพียง 20 ตัวแปรที่มีผลมากที่สุดเท่านั้น ดังรูปภาพที่ 8



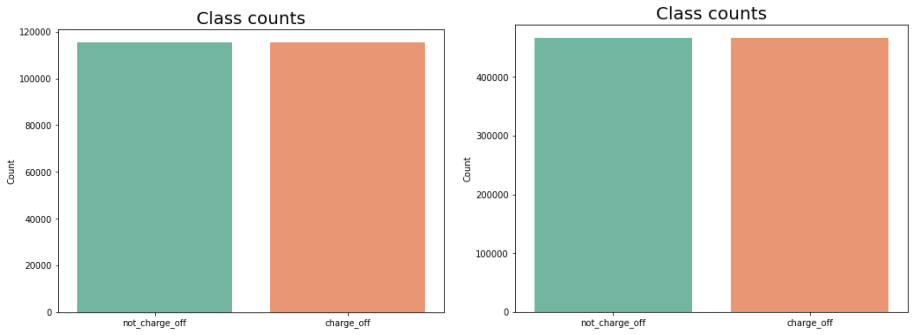
รูปภาพที่ 8 Feature Selection using REFCV

(b) Sampling method

ทำการสุ่มข้อมูลให้มีขนาดเท่า ๆ กันโดยการทดลองได้ใช้ทั้งหมด 2 วิธี

1. Under sampling คือการลดข้อมูลกลุ่มที่มาก ให้มีขนาดเท่ากับกลุ่มที่น้อย โดยใช้ RUS และ IHT จะได้ข้อมูลทั้ง 2 class เท่ากับ 115,329

2. Oversampling โดยการเพิ่มกลุ่มข้อมูลที่น้อยให้เท่ากับข้อมูลกลุ่มมากโดยทำการสุ่มตัวแปรขึ้นมาใหม่ โดยใช้ SMOTE และ ADASYN จะได้ข้อมูลทั้ง 2 class เท่ากับ 465,946 ดังรูปภาพที่ 9



รูปภาพที่ 9 Resampling Method

(c) Classification Model & Hyper parameter tuning

ในขั้นตอนการสร้างโมเดลผู้วิจัยได้ทดลองปรับค่าตัวแปรของโมเดลเพื่อให้ได้ค่าดีที่สุดโดยใช้ RamdomSearchCV ในการทดลองทีละตัวแปรโดยการตั้งค่าตัวแปรตามตารางที่ 2

ตารางที่ 2 Hyperparameter setup

Classifiers	Parameters
Logistic Regression	C = np.logspace(-3,3,7), penalty = ["l1","l2"]
DecisionTree	max_depth = [2, 4, 6, 8, 10], min_samples_leaf = [20, 40, 60 ,100], criterion = ['gini', 'entropy']
Random Forest	n_estimators = [100, 200, 300], max_features = ['auto', 'sqrt', 'log2'], max_depth = [2, 4, 6, 8, 10], criterion = ['gini', 'entropy']
XGBoost	min_child_weight = [1, 5, 10], gamma = [0.5, 1, 1.5, 2, 5], subsample = [0.6, 0.8, 1.0], colsample_bytree = [0.6, 0.8, 1.0], 'max_depth': [3, 4, 5]
Catboost	Depth = [2, 4, 6, 8, 10], learning_rate = [0.01, 0.02, 0.03, 0.04], iterations = [50, 100, 200, 300]
LightGBM	num_leaves = [31, 127], reg_alpha = [0.1, 0.5], min_data_in_leaf = [20, 40, 60, 100], lambda_l1 = [0, 1, 1.5], lambda_l2 = [0, 1]
Neural Network (MLP)	hidden_layer_sizes = [(50,50,50), (50,100,50), (100,)], activation = ['tanh', 'relu'], solver = ['sgd', 'adam'], alpha = [0.0001, 0.05], learning_rate = ['constant','adaptive']

(d) Evaluation

วิธีการวิเคราะห์ความถูกต้อง ในงานวิจัยนี้พิจารณาจากค่า Accuracy, Precision, Recall, F-measure, ROC Curve โดยการคำนวณค่าจากตาราง Confusion Matrix ดังตารางที่ 3

ตารางที่ 3 Confusion Matrix

		Predicted Class	
Actual		Class = 0	Class = 1
	Class = 0	True Negative (TN)	False Positive (FP)
	Class = 1	False Negative (FN)	True Positive (FP)

4.4.3 Experimental setup

ในการทดลองผู้วิจัยได้ใช้ Jupyter Notebook ในการเตรียมข้อมูล System specs: Macbook pro processor: 2.2 GHz Quad-Core Intel Core i7, Memory: 16 GB 1600 MHz DDR3, Graphics: Intel Iris Pro 1536 MB และ Google Colab Pro System specs: GPUs = [K80, P100, T4], CPUs: 2 x vCPU และ RAM: 24GB ในการรันโมเดล

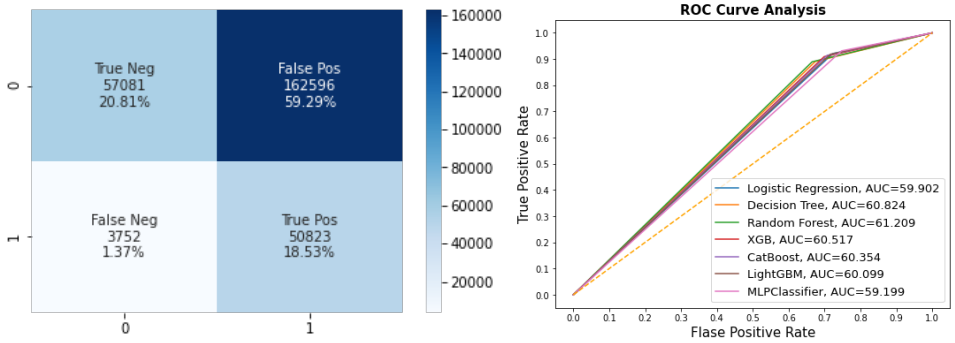
5. ผลการวิจัย (Experimental Result)

ตารางที่ 3 ผลที่ได้จากการทดลองแต่ละโมเดล

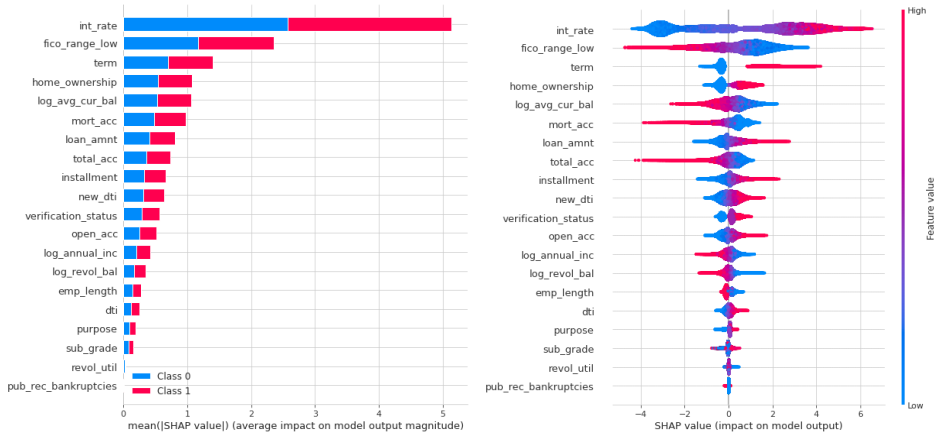
	Classifiers	Accuracy	Precision	Recall	F1-score	ROC
Non sampling	Logistic Regression	80.39	52.03	7.69	13.39	52.97
	Decision Tree	80.38	52.40	5.83	10.49	52.26
	Random Forest	80.44	56.60	3.47	6.54	51.41
	Xgboots	80.57	56.06	6.93	12.34	52.80
	Catboost	80.58	55.77	7.37	13.01	52.97
	LightGBM	80.60	55.50	8.31	14.45	53.33
	ANN	80.47	52.43	10.61	17.65	54.12
RUS	Logistic Regression	65.83	32.1	64.27	42.81	65.13
	Decision Tree	61.81	30.16	69.86	42.13	64.83
	Random Forest	63.81	31.3	68.49	42.97	65.57
	Xgboots	64.85	31.97	67.98	43.49	66.02
	Catboost	64.98	32.11	68.19	43.66	66.18
	LightGBM	65.06	32.11	67.84	43.59	66.1
	ANN	67.83	33.39	61.96	43.39	65.63
IHT	Logistic Regression	40.48	23.97	92.1	38.04	59.9
	Decision Tree	43.78	24.64	89.08	38.6	60.82
	Random Forest	44.48	24.86	88.95	38.86	61.21
	Xgboots	42.18	24.36	90.91	38.42	60.52

	Catboost	41.63	24.24	91.4	38.32	60.35
	LightGBM	40.95	24.09	91.85	38.16	60.1
	ANN	38.73	23.57	93.13	37.62	59.2
SMOTE	Logistic Regression	65.17	31.6	64.92	42.51	65.07
	Decision Tree	72.99	34.72	41.02	37.61	60.97
	Random Forest	69.47	33.56	55	41.69	64.03
	Xgboots	79.65	46.54	17.24	25.16	56.17
	Catboost	80.33	52.11	10.45	17.41	54.04
	LightGBM	80.35	52.35	10.82	17.93	54.19
	ANN	74.74	35.79	33.92	34.83	59.4
ADASYN	Logistic Regression	63.89	30.98	66.78	42.32	64.98
	Decision Tree	73.6	34.16	35.61	34.87	59.31
	Random Forest	71.12	34.25	49.54	40.5	63
	Xgboots	79.53	45.71	16.91	24.68	55.97
	Catboost	80.37	52.81	9.78	16.51	53.81
	LightGBM	80.4	53.19	10.08	16.95	53.94
	ANN	74.01	35.58	38.22	36.85	60.55

จากการทดลองทั้งหมด 7 โมเดล และใช้เทคนิค Resampling ทั้ง 4 แบบจะเห็นว่าโมเดลที่ให้ผลได้ดีที่สุดคือโมเดล Neural Network รวมกับเทคนิค IHT Under sampling ได้ค่า Recall สูงถึง 93.13% และค่า ROC Curve 59.2% ซึ่งผลการทดลองเมื่อเทียบกับตาราง Confusion matrix จะเห็นว่า % ในการสูญเสีย False Negative เพียง 1.37% หรือจำนวนผู้กู้ที่ทำนายว่า ไม่มีแนวโน้มผิดนัดชำระหนี้ แต่จริง ๆ ผิดนัดชำระหนี้ จำนวน 3,752 คน ดังรูปที่ 10 และปัจจัยที่ส่งผลในการทำนายว่าผู้กู้มีแนวโน้มในการผิดนัดชำระหนี้หรือไม่ โดยใช้ LightGBM ในการดู Feature Importance จะเห็นว่าปัจจัยที่ส่งผลต่อการทำนายมี 5 อันดับแรกคือ อัตราดอกเบี้ย, คะแนน credit scoring, ระยะเวลาในการกู้, สถานะบ้านที่อยู่อาศัยของผู้กู้ และ จำนวนเงินเฉลี่ยในบัญชีทั้งหมด ดังรูปที่ 11



รูปที่ 10 ตาราง Confusion matrix และ ROC Curve



รูปที่ 11 Feature Importance จาก LightGBM โมเดล

6. วิจารณ์และสรุปผล (Conclusion)

6.1 สรุปผลการทดลอง

6.1.1. จากการทดลองการใช้เทคนิค resampling สามารถเพิ่มประสิทธิภาพของ model ได้ดีเมื่อเทียบกับค่า recall และ model ที่ให้ผลได้ดีคือโมเดล Neural network รวมกับเทคนิค IHT Under sampling

6.1.2. ปัจจัยที่มีผลกระทบต่อการผิดนัดชำระหนี้ของผู้กู้ที่มากที่สุด 5 อันดับแรกคือ อัตราดอกเบี้ย, คะแนน credit scoring, ระยะเวลาในการกู้, สถานะบ้านที่อยู่อาศัยของผู้กู้ และ จำนวนเงินเฉลี่ยในบัญชีทั้งหมด

6.2 ประโยชน์ที่ได้รับ

6.2.1. ทาง Lending club สามารถนำเทคนิคนี้ไปใช้คาดการณ์หาแนวโน้มการผิดนัดชำระหนี้ของผู้กู้ เพื่อใช้เป็นกลยุทธ์ในการวางแผนจัดการหนี้เสียในอนาคตได้ เช่น การแจ้งเตือนการจ่ายเงินก่อนถึงวันครบกำหนด, การจัด campaign ให้ผู้กู้ที่มีแนวโน้มผิดนัดชำระหนี้ มีการกระตุ้นการจ่ายเงินมากขึ้น

6.2.2. Lending club สามารถทราบปัจจัยที่ต้องระวังของผู้กู้ที่มีแนวโน้มในการผิดนัดชำระหนี้ของผู้กู้ได้

6.2.3. สามารถนำเทคนิคไปประยุกต์ใช้กับรูปแบบการปล่อยสินเชื่อประเภทอื่น ๆ ได้

6.3 ข้อเสนอแนะ

6.3.1. งานวิจัยชิ้นนี้เป็นการวิเคราะห์เพียงแค่สินเชื่อส่วนบุคคล และช่วงระยะเวลาหนึ่งเท่านั้น ซึ่งสามารถนำไปวิเคราะห์สินเชื่อที่มีลักษณะคล้าย ๆ กันได้เช่น สินเชื่อรถ สินเชื่อบ้าน สินเชื่อการเกษตร

6.3.2. การติดตามประสิทธิภาพโดยรวม และปรับปรุงแบบจำลองอย่างต่อเนื่องจะทำให้แบบจำลองสะท้องความเสี่ยงของลูกหนี้ได้แม่นยำมากขึ้น หรือการวิเคราะห์แบบ real-time โดยใช้ Concept Diff กับข้อมูล Imbalance เพื่อใช้หา Pattern ของคนที่มีแนวโน้มผิดนัดชำระหนี้ได้รวดเร็วขึ้น

6.3.3. ใช้เทคนิค Clustering ในการจัดกลุ่มของผู้กู้ เพื่อให้ง่ายต่อการจัดแคมเปนให้ตรงกลุ่มเป้าหมายมากขึ้น

6.3.4. ใช้เทคนิค Social Network ในการหาความสัมพันธ์ระหว่างนักลงทุนกับผู้กู้ เพื่อใช้ในการหารูปแบบที่น่าสนใจ เช่น หาพฤติกรรมของผู้ลงทุน, หาพฤติกรรมของผู้กู้

7. เอกสารอ้างอิง Reference

นพมาศ อัครจันทโชติ และ นพมาศ อัครจันทโชติ (2019). **การเปรียบเทียบวิธีการแก้ปัญหาข้อมูลไม่สมดุลสำหรับการจำแนกกลุ่มรายได้ของผู้ประกอบการร้านยาประเภท ข.ย.1.** การประชุมเสนอผลงานวิจัยระดับชาติ มหาวิทยาลัยสุโขทัยธรรมาธิราช ครั้งที่ 9

พรชนก เทพขาม (2019). **เศรษฐกิจแบ่งปัน: บทเรียนการกู้ยืมผ่านช่องทางอิเล็กทรอนิกส์ (Peer-to-peer Lending)**. ธนาคารแห่งประเทศไทย

อโนทัย พุทธารี จิตรภณ หรูเจริญพรพานิช เพ็ญสิริ บำรุงเชาว์เกษม และชินวัฒน์ เทพหัสดิน ณ อยุธยา. (2018). CREDIT SCORING MODEL: **เครื่องมือในการประเมินคุณภาพสินเชื่อ**. ธนาคารแห่งประเทศไทย

เปมิกา ฐิติเพิ่มพงศ์ (2020, 17 กุมภาพันธ์). **มาทำความรู้จักกับ Peer-to-Peer Lending กันเถอะ | ตอนที่ 3** สืบค้นจาก https://www.nestifly.com/post/what-is-p2p-ep3

Aleum Kim and Sung-Bae Cho. (2017). **Dempster-Shafer Fusion of Semi-supervised Learning Methods for Predicting Defaults in Social Lending**. Seoul: Yonsei University.

Anahita Namvar, Mohammad Siami, Fethi Rabhi and Mohsen Naderpour. (2018). **Credit risk prediction in an imbalanced social lending environment**. Sydney: University of New South Wales.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. (2008) **ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning**. 2008 International Joint Conference on Neural Networks

Naufal Azmi Verdikha, Teguh Bharata Adji and Adhistya Erna Permanasari. (2018). **Study of Undersampling Method: Instance Hardness Threshold with Various Estimators for Hate Speech Classification**. IJITEE, Vol. 2, No. 2, June 2018

Yijie Fu (2017). **Combination of Random Forests and Neural Networks in Social Lending**. Shanghai: Shanghai Foreign Language School.

Yuejin Zhanga, Haifeng Lia, Mo Hai a, Jiaxuan Lia and Aihua Lib (2017). **Determinants of loan funded successful in online P2P Lending.** Beijing: Central University of Finance and Economics

Zhiqiang Li, Shouyan Li, Zhilong Li, Yixiang Hu and Hanlin Gao. (2021). **Application of XGBoost in P2P Default Prediction**. College of Business Jiangxi University of Science and Technology Nanchang