การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้ของเครื่อง

ANALYSIS OF CREDIT CARD DEBT DEFAULT RISK ANALYSIS BY USING MACHINE LEARNING ALGORITHM

เครือวัลย์ เนตรพนา

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2565

การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้ของเครื่อง

เครือวัลย์ เนตรพนา

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2565

ANALYSIS OF CREDIT CARD DEBT DEFAULT RISK ANALYSIS BY USING MACHINE LEARNING ALGORITHM

KRUEWAN NETPHANA

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2022

สารนิพนธ์

เรื่อง

การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้ของเครื่อง

ของ

เครือวัลย์ เนตรพนา

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

.....

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... ประธาน

(รองศาสตราจารย์ ดร.ดวงดาว วิชาดากุล)

..... กรรมการ

(อาจารย์ ดร.โสภณ มงคลลักษมี)

ชื่อเรื่อง	การวิเคราะห์ความเสี่ยงในการผิดนัดชำระของลูกหนี้บัตรเครดิต โดยการใช้อัลกอริทึมการเรียนรู้ของเครื่อง
ผู้วิจัย	เครือวัลย์ เนตรพนา
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2565
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

งานวิจัยนี้มีวัตถุประสงค์เพื่อการศึกษาการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการทดลองกับ ชุดข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิตซึ่งประกอบด้วย ข้อมูลจำนวนทั้งหมด 307,511 แถว และคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะเว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มใหญ่ๆ คือ กลุ่มลูกหนี้ปกติ คือกลุ่มลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และกลุ่มลูกหนี้ที่ไม่ปกติ คือกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร เครื่องมือหลักที่นักวิจัยใช้ ได้แก่ Machine Learning Algorithms เช่น Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting เป็นต้น โดยอาศัยการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งเป็นเครื่องมือสำหรับการพัฒนาแบบจำลอง ในการเรียนรู้แบบผู้สอน (Supervised Learning) โดยมีการทำงานแบบการแบ่งแยกประเภท (Classification) ซึ่งการเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนรู้ของเครื่องในการเรียนรู้ข้อมูล โดยอาศัยชุดข้อมูลที่ใช้ในการฝึกฝนเพื่อทำการพัฒนาแบบจำลองและชุดข้อมูลที่ใช้ในการทดสอบสำหรับใช้ในการการทดสอบแบบจำลอง โดยเราสามารถนำผลลัพธ์ที่ได้ ไปตรวจสอบกับชุดข้อมูลที่ใช้ในการทดสอบที่เรามีอยู่แล้ว ว่าแบบจำลองที่ถูกพัฒนาขึ้นนั้น มีประสิทธิภาพและความถูกต้อง (Accuracy) มากน้อยเพียงใด แต่จากชุดข้อมูลที่นำมาใช้ในการวิเคราะห์ข้อมูล พบว่าข้อมูลมีความไม่สมดุลกันของชุดข้อมูล (Imbalance data) สูงมาก ซึ่งทำให้ค่าความถูกต้อง (Accuracy) ที่ได้อาจมีค่าที่สูงมาก แต่มีประสิทธิภาพที่ไม่เพียงพอ เพราะค่า precision, recall และ F1-Score ที่ได้มีค่าที่ต่ำมาก โดยเราต้องอาศัยเทคนิคต่างๆ มาช่วยในการแก้ปัญหาความไม่สมดุลของชุดข้อมูล เช่น Oversampling, Under sampling และ Synthetic Minority Oversampling Technique (SMOTE) เพื่อทำให้แบบจำลองที่ได้มีประสิทธิภาพที่ดี ผลการศึกษาพบว่า การพัฒนาแบบจำลองโดยการใช้เทคนิควิธี Gradient Boosting ให้ค่าความไว (Recall) ที่มากที่สุด ซึ่งมีค่าเท่ากับ 0.65 มีค่าความถูกต้อง (Accuracy) เท่ากับ 0.62 และมีค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองเท่ากับ 0.54 แต่เทคนิควิธี K-Nearest Neighbors (KNN) ให้ค่าความไว (Recall) ที่น้อยที่สุด ซึ่งมีค่าเท่ากับ 0.58 มีค่าความถูกต้อง (Accuracy) เท่ากับ 0.55 และมีค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองเท่ากับ 0.47 ซึ่งมีค่าที่น้อยที่สุด

คำสำคัญ : การเรียนรู้ของเครื่อง, การเรียนรู้แบบผู้สอน, การแบ่งแยกประเภท, ความไม่สมดุลกันของชุดข้อมูล

Title	ANALYSIS OF CREDIT CARD DEBT DEFAULT RISK ANALYSIS BY USING MACHINE LEARNING ALGORITHM
Author	KRUEWAN NETPHANA
Degree	MASTER OF SCIENCE
Academic Year	2022
Thesis Advisor	Assistant Professor Dr. Sirisup Laohakiat

This research aims to study the prediction of debtors who are likely to default on their payments to the bank, using a dataset of credit card transactions. The dataset consists of 307,511 rows and 122 columns, sourced from a public data site. The data is divided into two main groups, normal debtors who comply with payments, and abnormal debtors who default on payments. The primary tool used by the researchers were Machine Learning Algorithms, such as Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest, Support Vector Classifier (SVC), and Gradient Boosting. Machine Learning is a tool used to develop models, with supervised learning, involving classification. The researchers used a training set to develop the model and a testing set to evaluate the performance of the model. However, they found that the data was significantly imbalanced, which affected the accuracy of the model, causing the precision of the model, recall and F1-Score values to be low. To overcome this problem, they employed techniques such as oversampling, under-sampling, and Synthetic Minority Oversampling Technique (SMOTE), to improve model performance. The study found that developing a model using Gradient Boosting technique provides the highest value of recall, equal to 0.65. However, the accuracy value was only 0.62 and the F1-Score is 0.54, which was used to measure the effectiveness of the model. On the other hand, K-Nearest Neighbors (KNN) technique provided the lowest value of recall, which was 0.58, which had an Accuracy value is 0.55 and the F1-Score is 0.47, which were the lowest values.

Keyword : Machine Learning Supervised Learning Classification Clustering Imbalance data

กิตติกรรมประกาศ

การจัดทำวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากการสนับสนุน ความรู้ความช่วยเหลือ คำแนะนำ ตลอดจนแนวทางในการทำวิจัยและจัดทำสารนิพนธ์ของ ผศ.ดร.ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษาและคณาจารย์ทุกท่านในภาควิชาวิทยาการข้อมูล คณะวิทยาศาสตร์มหาวิทยาลัยศรีนครินทรวิโรฒ การสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เครือวัลย์ เนตรพนา

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของการวิจัย

ปัจจุบันประชาชนส่วนใหญ่มีการใช้งานบัตรเครดิตที่เพิ่มมากขึ้น และมีการนำบัตรเครดิตมาใช้เป็นเครื่องมือหลักๆ ในการใช้จ่ายในชีวิตประจำวัน จึงทำให้ผู้คนส่วนใหญ่มีความสะดวกรวดเร็วในการใช้จ่ายมากยิ่งขึ้น และยังมีความปลอดภัยในการใช้จ่ายมากกว่าการพกเงินสดติดตัวเป็นจำนวนมาก ซึ่งจะทำให้เสี่ยงต่อการสูญหายหรือโจรกรรมมากยิ่งขึ้น โดยสมัยนี้ร้านค้า ห้างสรรพสินค้า และศูนย์การค้าต่างๆ ที่ให้บริการในหลายๆ แห่งทั่วประเทศ ได้มีการรับชำระเงินผ่านบัตรเครดิตที่เพิ่มมากขึ้น ทำให้ผู้คนส่วนใหญ่สามารถใช้บัตรเครดิตในการชำระค่าสินค้าและค่าบริการ แทนการชำระเงินด้วยเงินสด

บัตรเครดิต คือ ผลิตภัณฑ์ทางการเงินรูปแบบหนึ่ง ที่เป็นการกู้ยืมเงินจากทางธนาคารหรือจากสถาบันการเงินต่างๆ มาใช้จ่ายล่วงหน้า โดยใช้ในการชำระค่าสินค้าและค่าบริการแทนการชำระเงินด้วยเงินสด โดยวงเงินในการใช้จ่ายนั้นจะต้องไม่เกินยอดวงเงินที่สถาบันการเงินแต่ละสถาบันการเงินอนุมัติ ซึ่งจะต้องทำการชำระคืนในภายหลัง ซึ่งมีให้เลือกทั้งในรูปแบบของการชำระคืนแบบเต็มจำนวน ชำระคืนแบบจ่ายขั้นต่ำ หรือการผ่อนชำระผ่านบัตรเครดิต ซึ่งบัตรเครดิตในปัจจุบันมีให้เลือกหลากหลายรูปแบบและหลากหลายประเภท

ด้วยการสมัครบัตรเครดิตสมัยนี้นั้นเป็นเรื่องที่ง่ายมากยิ่งขึ้น จึงทำให้ผู้คนส่วนใหญ่หันมาใช้จ่ายผ่านบัตรเครดิตที่เพิ่มมากยิ่งขึ้น ซึ่งเอกสารในการสมัครบัตรเครดิต ใช้เพียงแค่สลิปเงินเดือนหรือหนังสือรับรองเงินเดือนภายใน 3 เดือน ก็สามารถสมัครบัตรเครดิตได้แล้ว และยังรองรับกับบุคคลที่ทำงานในหลากหลายอาชีพ ซึ่งไม่ว่าจะทำงานอาชีพไหนก็สามารถสมัครบัตรเครดิตได้ และด้วยเงื่อนไขการอนุมัติวงเงินที่สามารถทำได้สะดวกและรวดเร็วมากยิ่งขึ้น โดยไม่จำเป็นต้องมีหลักทรัพย์ค้ำประกัน และบัตรเครดิตบางประเภทยังมีเงื่อนไขในการยกเว้นค่าธรรมเนียมแรกเข้าและค่าธรรมเนียมรายปี การแลกคะแนนสะสมเพื่อแลกรับสิทธิ์ ของรางวัลหรือบัตรกำนัลต่างๆ นั้นจึงเป็นสาเหตุหลัก ที่ทำให้ประชาชนส่วนใหญ่หันมาใช้จ่ายผ่านบัตรเครดิตแทนการใช้จ่ายด้วยเงินสดเป็นจำนวนมากยิ่งขึ้น

แต่จากการเติบโตของการใช้จ่ายเงินสดผ่านบัตรเครดิต ซึ่งเติบโตขึ้นอย่างรวดเร็ว จึงทำให้ผู้คนส่วนใหญ่หันมาใช้จ่ายผ่านบัตรเครดิตแทนการชำระเงินด้วยเงินสด ถึงแม้ว่าจะช่วยทำให้ผู้คนส่วนใหญ่มีความสะดวกและรวดเร็วในการใช้จ่ายมากยิ่งขึ้น แต่ยังก่อให้เกิดปัญหาต่าง ๆ อีก

มากมาย เช่น ปัญหาหนี้ที่ไม่ก่อให้เกิดรายได้หรือที่เรียกว่าหนี้เสีย อันเนื่องมาจากการใช้จ่ายที่ฟุ่มเฟือยจนเกินความสามารถในการที่จะชำระเงินคืนให้กับทางธนาคาร

หนี้เสีย (Non-Performing Loan) เป็นสินเชื่อที่ไม่ก่อให้เกิดรายได้กับทางธนาคาร โดยธนาคารจะไม่สามารถนำเงินส่วนนี้ไปใช้จ่ายในด้านอื่นๆได้ และเมื่อเกิดหนี้เสียแล้ว ทางธนาคารจะพยายามติดตามหรือติดต่อลูกหนี้ เพื่อให้ลูกหนี้ได้มีการมาปรับโครงสร้างหนี้กับทางธนาคาร หนี้เสียจะเป็นเงินกู้ที่ลูกหนี้ได้มีการผิดนัดชำระกับทางธนาคาร โดยที่ไม่ได้ชำระเงินต้นและดอกเบี้ยรายเดือนตามระยะเวลาที่กำหนด โดยส่วนมากมักจะค้างชำระติดต่อกันนานเกินกว่า 90 วัน หนี้เสียของทางธนาคารมีโอกาสที่จะเกิดขึ้นได้เรื่อย ๆ เนื่องจากในสถานการณ์ปัจจุบัน ได้มีการระบาดของโรคที่สามารถติดต่อกันได้ อย่างเช่น โรค covid-19 ทำให้หลายบริษัทสู้สถานการณ์ทางเศรษฐกิจไม่ไหว จนทำให้หลายบริษัทได้มีการปิดตัวลง ประชาชนส่วนใหญ่จึงตกงานหรือขาดรายได้กันเป็นจำนวนมาก ทำให้สภาวะเศรษฐกิจของประเทศไทยในช่วงนี้เข้าสู่สภาวะเศรษฐกิจตกต่ำ ดังนั้นผู้คนส่วนใหญ่ที่มีสินเชื่อกับทางธนาคาร จึงไม่มีรายได้มากพอที่จะชำระหนี้สินเชื่อคืนให้กับทางธนาคาร และเมื่อไม่ได้ชำระหนี้สินเชื่อคืนเกินกว่าระยะเวลาที่กำหนด ก็จะนำไปสู่การเป็นสภาวะลูกหนี้ที่มีประวัติการชำระหนี้ที่ไม่ดี และข้อมูลพฤติกรรมเหล่านี้จะถูกส่งไปยังเครดิตบูโร ซึ่งเป็นสถาบันการเงินที่เก็บรวบรวมข้อมูลสินเชื่อที่เคยได้รับการอนุมัติจากสถาบันการเงินต่างๆ ประวัติการชำระหนี้สินเชื่อคืนให้กับทางธนาคาร ซึ่งถ้าหากลูกหนี้รายนั้นจะกู้ยืมสินเชื่อในครั้งถัดไป ทางธนาคารก็จะตรวจสอบเครดิตบูโรของบุคคลเหล่านั้นก่อน หากพบว่าลูกหนี้มีประวัติการชำระหนี้ที่ไม่ดี การกู้ยืมสินเชื่อกับทางธนาคารในครั้งถัดไป ทางธนาคารก็จะปล่อยกู้สินเชื่อให้ลูกหนี้รายนั้นยากมากยิ่งขึ้น และดอกเบี้ยที่ได้รับ ก็อาจเป็นดอกเบี้ยที่มีอัตราสูงกว่าบุคคลทั่วไป เนื่องจากถ้าหากทางธนาคารได้มีการอนุมัติในการปล่อยสินเชื่อส่วนใหญ่ให้กับบุคลลทั่วไป โดยไม่ได้มีการวิเคราะห์ถึงพฤติกรรมและปัจจัยต่างๆ ของลูกหนี้แต่ละราย อาจทำให้ทางธนาคารมีลูกหนี้ที่เป็นลูกหนี้เสียเป็นจำนวนมาก ซึ่งจะทำให้เกิดการขาดรายได้ของทางธนาคาร และธนาคารก็จะไม่สามารถนำเงินส่วนนี้ ไปหมุนเวียนเพื่อใช้จ่ายในการลงทุนธุรกิจทางด้านอื่นๆได้

งานวิจัยนี้ เน้นการศึกษาการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการทดลองในครั้งนี้ เราทดลองกับ ข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิตซึ่งประกอบด้วยข้อมูลจำนวนทั้งหมด 307,511 แถว และคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะ เว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มใหญ่ๆ คือ กลุ่มลูกหนี้ปกติ คือกลุ่มลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และกลุ่มลูกหนี้ที่ไม่ปกติ คือกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร

เครื่องมือหลักที่นักวิจัยใช้ ได้แก่ Machine Learning Algorithms เช่น Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting เป็นต้น โดยอาศัยการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งเป็นเครื่องมือสำหรับการพัฒนาแบบจำลอง ในการเรียนรู้แบบผู้สอน (Supervised Learning) โดยมีการทำงานแบบการแบ่งแยกประเภท (Classification) ซึ่งการเรียนรู้แบบมีผู้สอนเป็นการเรียนรู้ของเครื่องมือในการเรียนรู้ข้อมูล โดยอาศัยชุดข้อมูลที่ใช้ในการฝึกฝนเพื่อทำการพัฒนาแบบจำลองและชุดข้อมูลที่ใช้ในการทดสอบสำหรับใช้ในการการทดสอบแบบจำลอง โดยผลลัพธ์จากการเรียนรู้ของเครื่องมือในการเรียนรู้ข้อมูลสำหรับการพัฒนาแบบจำลอง คือ การคาดคะเนผลลัพธ์ที่อาจจะเกิดขึ้นจากข้อมูลที่ได้รับ โดยเราสามารถนำผลลัพธ์ที่ได้ ไปตรวจสอบกับชุดข้อมูลที่ใช้ในการทดสอบที่เรามีอยู่แล้ว ว่าแบบจำลองที่ถูกพัฒนาขึ้นนั้น มีประสิทธิภาพและความถูกต้อง (Accuracy) มากน้อยเพียงใด แต่จากชุดข้อมูลที่นำมาใช้ในการวิเคราะห์ข้อมูล พบว่าข้อมูลมีความไม่สมดุลกันของชุดข้อมูล (Imbalance data) สูงมาก ซึ่งทำให้ค่า Accuracy ที่ได้อาจมีค่าที่สูงมาก แต่อาจมีประสิทธิภาพที่ไม่เพียงพอ เพราะค่า precision และ recall ที่ได้มีค่าที่ต่ำมาก โดยเราต้องอาศัยเทคนิคต่างๆ มาช่วยในการแก้ปัญหาความไม่สมดุลของชุดข้อมูล เช่น Oversampling, Under sampling และ Synthetic Minority Oversampling Technique (SMOTE) เพื่อทำให้แบบจำลองที่ได้มีประสิทธิภาพที่ดียิ่งขึ้น

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาปัจจัย (Feature) ที่ส่งผลต่อการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าปัจจัยไหนที่ส่งผลต่อการพัฒนาแบบจำลอง ที่เมื่อเลือกปัจจัยเหล่านั้น มาทำการพัฒนาแบบจำลองแล้วทำให้แบบจำลองที่ได้ ทำนายโอกาสของการเกิดลูกหนี้ที่มีการผิดนัดชำระได้อย่างแม่นยำมากที่สุด

2. เพื่อศึกษาเทคนิคการเรียนรู้ของเครื่อง โดยการนำเทคนิคและอัลกอริทึมต่างๆ มาเรียนรู้ข้อมูลเพื่อพัฒนาเป็นแบบจำลอง

3. เพื่อนำเทคนิคต่างๆ มาประยุกต์ใช้ในการแก้ปัญหาของการพัฒนาแบบจำลอง เนื่องจากชุดข้อมูลมีข้อมูลที่ขาดหายไป (Missing Value) เป็นจำนวนมาก ทั้งในแนวระดับแถวและในแนวระดับคอลัมน์ จึงจำเป็นที่จะต้องมีการใช้เทคนิคต่างๆ มาเพื่อช่วยในการจัดการกับข้อมูลที่ขาดหายไป ก่อนการนำข้อมูลเหล่านี้ไปใช้ในการพัฒนาแบบจำลอง

4. เพื่อนำเทคนิคต่างๆ มาประยุกต์ใช้ในการแก้ปัญหาของการพัฒนาแบบจำลอง เนื่องจากชุดข้อมูลที่นำมาใช้ในการวิเคราะห์ มีความไม่สมดุลกันของชุดข้อมูล (Imbalance data)

สูงมาก จึงจำเป็นต้องมีการปรับความไม่สมดุลกันของชุดข้อมูล โดยการใช้เทคนิค Oversampling, Under sampling และ Synthetic Minority Oversampling Technique (SMOTE) มาเพื่อช่วยในการปรับความไม่สมดุลของชุดข้อมูล ก่อนการนำข้อมูลเหล่านี้ไปใช้ในการพัฒนาแบบจำลอง

5. เพื่อการเปรียบเทียบประสิทธิภาพของการพัฒนาแบบจำลอง ว่าแบบจำลองไหน ที่ให้ประสิทธิภาพในการทำนายที่ให้ผลลัพธ์ โอกาสของการเกิดลูกหนี้ที่มีการผิดนัดชำระได้อย่างแม่นยำมากที่สุด

1.3 ความสำคัญของการวิจัย

1. เพื่อนำแบบจำลองที่ได้ไปใช้ในการตรวจจับลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร เมื่อเราพัฒนาแบบจำลองขึ้นมา แล้วให้ผลลัพธ์ในการทำนายที่มีประสิทธิภาพและมีความอย่างแม่นยำสูง เราสามารถนำแบบจำลองที่ได้ ไปใช้ในการตรวจจับลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคารได้

2. เพื่อใช้ในการประเมินสภาวะของลูกหนี้แต่ละรายเบื้องต้น โดยการอาศัยชุดข้อมูลพฤติกรรมและปัจจัยพื้นฐานต่างๆ ของลูกหนี้ ซึ่งจะนำข้อมูลเหล่านี้มาใช้ในการทำนายเพื่อประเมินสภาวะของลูกหนี้แต่ละรายเบื้องต้นได้ ซึ่งจะช่วยลดภาระให้กับทางธนาคาร

3. เพื่อใช้ในการประเมินสถานะในการขอสินเชื่อของลูกหนี้แต่ละราย โดยการอาศัยชุดข้อมูลพฤติกรรมและปัจจัยพื้นฐานต่างๆ ของลูกหนี้ ซึ่งจะนำข้อมูลเหล่านี้มาใช้เป็นข้อมูลที่ใช้ในการวิเคราะห์สินเชื่อ เพื่อช่วยลดโอกาสที่จะเกิดหนี้เสียที่ไม่ก่อให้เกิดรายได้กับทางธนาคาร

1.4 ขอบเขตของการวิจัย

1.4.1 กลุ่มตัวอย่างประชากรที่ใช้ในการวิจัย

ข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิตซึ่งประกอบด้วย ข้อมูลจำนวนทั้งหมด 307,511 แถว และคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะ Kaggle.com เว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download

ตาราง 1 ตัวแปรของชุดข้อมูลที่ใช้ในการพัฒนาแบบจำลอง

ชื่อตัวแปรของข้อมูล	ชนิดของข้อมูล	คำอธิบายตัวแปรของข้อมูล
SK_ID_CURR	int64	ID ของเงินกู้
TARGET	int64	ตัวแปรเป้าหมาย (1 - ลูกค้าที่มีปัญหาในการชำระเงิน, 0 - กรณีอื่นๆ ทั้งหมด
NAME_CONTRACT_TYPE	int64	เงินสดหรือเงินหมุนเวียน
CODE_GENDER	int64	เพศ (M = ชาย, F = หญิง)
FLAG_OWN_CAR	int64	เจ้าของรถยนต์
FLAG_OWN_REALTY	int64	เจ้าของบ้าน
CNT_CHILDREN	int64	จำนวนบุตร
AMT_INCOME_TOTAL	float64	รายได้
AMT_CREDIT	float64	วงเงินกู้
AMT_ANNUITY	float64	จำนวนเงินที่ต้องจ่ายในแต่ละงวด
AMT_GOODS_PRICE	float64	ราคาของสินค้าที่ให้สินเชื่อ
NAME_TYPE_SUITE	int64	คนที่มากับลูกหนี้ เมื่อเขามายื่นขอสินเชื่อ
NAME_INCOME_TYPE	int64	ประเภทรายได้
NAME_EDUCATION_TYPE	int64	ระดับการศึกษาสูงสุด
NAME_FAMILY_STATUS	int64	สถานภาพทางครอบครัว
NAME_HOUSING_TYPE	int64	ที่อยู่อาศัย
REGION_POPULATION_RELATIVE	float64	คะแนนปกติ
DAYS_BIRTH	int64	วันเกิด
DAYS_EMPLOYED	int64	วันที่ทำงานจนถึงวันที่ปัจจุบัน
DAYS_REGISTRATION	int64	จำนวนวันที่ลูกค้าเปลี่ยนการลงทะเบียนก่อนการสมัคร
DAYS_ID_PUBLISH	int64	จำนวนวันที่ลูกค้าเปลี่ยนเอกสาร

ชื่อตัวแปรของข้อมูล	ชนิดของข้อมูล	คำอธิบายตัวแปรของข้อมูล
		ยืนยันตัวตนก่อนการสมัคร
OWN_CAR_AGE	int64	อายุรถยนต์
FLAG_MOBIL	int64	ให้โทรศัพท์มือถือ (1=YES, 0=NO)
FLAG_EMP_PHONE	int64	ให้โทรศัพท์ที่ทำงาน (1=YES, 0=NO)
FLAG_WORK_PHONE	int64	ให้โทรศัพท์บ้าน (1=YES, 0=NO)
FLAG_CONT_MOBILE	int64	โทรศัพท์มือถือสามารถติดต่อได้ (1=YES, 0=NO)
FLAG_PHONE	int64	ให้โทรศัพท์บ้าน (1=YES, 0=NO)
FLAG_EMAIL	int64	ให้อีเมล (1=YES, 0=NO)
OCCUPATION_TYPE	int64	อาชีพ
CNT_FAM_MEMBERS	float64	จำนวนสมาชิกในครอบครัว
REGION_RATING_CLIENT	int64	คะแนนของภูมิภาคที่อาศัยอยู่ (1,2,3)
REGION_RATING_CLIENT_W_CITY	int64	คะแนนของภูมิภาคที่อาศัยอยู่โดยคำนึงถึงเมือง (1,2,3)
WEEKDAY_APPR_PROCESS_START	int64	สมัครสินเชื่อวันไหนของสัปดาห์
HOUR_APPR_PROCESS_START	int64	สมัครสินเชื่อเวลากี่โมง
REG_REGION_NOT_LIVE_REGION	int64	ที่อยู่ไม่ตรงกับที่อยู่ติดต่อ (1= แตกต่าง, 0=เหมือนกัน)
REG_REGION_NOT_WORK_REGION	int64	ที่อยู่ไม่ตรงกับที่อยู่ที่ทำงาน (1= แตกต่าง, 0=เหมือนกัน)
LIVE_REGION_NOT_WORK_REGION	int64	ที่อยู่ติดต่อไม่ตรงกับที่อยู่ที่ทำงาน (1=แตกต่าง, 0=เหมือนกัน)
REG_CITY_NOT_LIVE_CITY	int64	ที่อยู่ถาวรไม่ตรงกับที่อยู่ติดต่อ (1= แตกต่าง, 0=เหมือนกัน)
REG_CITY_NOT_WORK_CITY	int64	ที่อยู่ถาวรไม่ตรงกับที่อยู่ที่ทำงาน

ชื่อตัวแปรของข้อมูล	ชนิดของข้อมูล	คำอธิบายตัวแปรของข้อมูล
		(1=แตกต่าง, 0=เหมือนกัน)
LIVE_CITY_NOT_WORK_CITY	int64	ที่อยู่ติดต่อไม่ตรงกับที่อยู่ที่ทำงาน (1=แตกต่าง, 0=เหมือนกัน)
ORGANIZATION_TYPE	int64	ประเภทขององค์กรที่ทำงาน
EXT_SOURCE_1	float64	คะแนนปกติ
EXT_SOURCE_2	float64	คะแนนปกติ
EXT_SOURCE_3	float64	คะแนนปกติ
APARTMENTS_AVG	float64	ค่าเฉลี่ยขนาดอพาร์ตเมนต์
BASEMENTAREA_AVG	float64	ค่าเฉลี่ยพื้นที่นั่งเล่น
YEARS_BEGINEXPLUATATION_AVG	float64	ค่าเฉลี่ยอายุอาคาร
YEARS_BUILD_AVG	float64	ค่าเฉลี่ยอายุอาคาร
COMMONAREA_AVG	float64	ค่าเฉลี่ยพื้นที่ส่วนกลาง
ELEVATORS_AVG	float64	ค่าเฉลี่ยจำนวนลิฟต์
ENTRANCES_AVG	float64	ค่าเฉลี่ยจำนวนทางเข้า
FLOORSMAX_AVG	float64	ค่าเฉลี่ยจำนวนชั้นมากที่สุด
FLOORSMIN_AVG	float64	ค่าเฉลี่ยจำนวนชั้นน้อยที่สุด
LANDAREA_AVG	float64	ค่าเฉลี่ยพื้นที่ที่ดิน
LIVINGAPARTMENTS_AVG	float64	ค่าเฉลี่ยพื้นที่พักอาศัย
LIVINGAREA_AVG	float64	ค่าเฉลี่ยพื้นที่พักอาศัย
NONLIVINGAPARTMENTS_AVG	float64	ค่าเฉลี่ยพื้นที่ส่วนที่ไม่พักอาศัย
NONLIVINGAREA_AVG	float64	ค่าเฉลี่ยพื้นที่ส่วนที่ไม่พักอาศัย
APARTMENTS_MODE	float64	ค่ากลางขนาดอพาร์ตเมนต์
BASEMENTAREA_MODE	float64	ค่ากลางพื้นที่นั่งเล่น
YEARS_BEGINEXPLUATATION_MODE	float64	ค่ากลางอายุอาคาร
YEARS_BUILD_MODE	float64	ค่ากลางอายุอาคาร
COMMONAREA_MODE	float64	ค่ากลางพื้นที่ส่วนกลาง
ELEVATORS_MODE	float64	ค่ากลางจำนวนลิฟต์

ชื่อตัวแปรของข้อมูล	ชนิดของข้อมูล	คำอธิบายตัวแปรของข้อมูล
ENTRANCES_MODE	float64	ค่ากลางจำนวนทางเข้า
FLOORSMAX_MODE	float64	ค่ากลางจำนวนชั้นมากที่สุด
FLOORSMIN_MODE	float64	ค่ากลางจำนวนชั้นน้อยที่สุด
LANDAREA_MODE	float64	ค่ากลางพื้นที่ที่ดิน
LIVINGAPARTMENTS_MODE	float64	ค่ากลางพื้นที่พักอาศัย
LIVINGAREA_MODE	float64	ค่ากลางพื้นที่พักอาศัย
NONLIVINGAPARTMENTS_MODE	float64	ค่ากลางพื้นที่ส่วนที่ไม่พักอาศัย
NONLIVINGAREA_MODE	float64	ค่ากลางพื้นที่ส่วนที่ไม่พักอาศัย
APARTMENTS_MEDI	float64	ค่ามัธยฐานขนาดอพาร์ตเมนต์
BASEMENTAREA_MEDI	float64	ค่ามัธยฐานพื้นที่นั่งเล่น
YEARS_BEGINEXPLUATATION_MEDI	float64	ค่ามัธยฐานอายุอาคาร
YEARS_BUILD_MEDI	float64	ค่ามัธยฐานอายุอาคาร
COMMONAREA_MEDI	float64	ค่ามัธยฐานพื้นที่ส่วนกลาง
ELEVATORS_MEDI	float64	ค่ามัธยฐานจำนวนลิฟต์
ENTRANCES_MEDI	float64	ค่ามัธยฐานจำนวนทางเข้า
FLOORSMAX_MEDI	float64	ค่ามัธยฐานจำนวนชั้นมากที่สุด
FLOORSMIN_MEDI	float64	ค่ามัธยฐานจำนวนชั้นน้อยที่สุด
LANDAREA_MEDI	float64	ค่ามัธยฐานพื้นที่ที่ดิน
LIVINGAPARTMENTS_MEDI	float64	ค่ามัธยฐานพื้นที่พักอาศัย
LIVINGAREA_MEDI	float64	ค่ามัธยฐานพื้นที่พักอาศัย
NONLIVINGAPARTMENTS_MEDI	float64	ค่ามัธยฐานพื้นที่ส่วนที่ไม่พักอาศัย
NONLIVINGAREA_MEDI	float64	ค่ามัธยฐานพื้นที่ส่วนที่ไม่พักอาศัย
FONDKAPREMONT_MODE	float64	ค่ากลาง
HOUSETYPE_MODE	float64	ค่ากลางแบบบ้าน
TOTALAREA_MODE	float64	ค่ากลางพื้นที่ทั้งหมด
WALLSMATERIAL_MODE	float64	ค่ากลางผนังวัสดุ
EMERGENCYSTATE_MODE	float64	ค่ากลางภาวะฉุกเฉิน

ชื่อตัวแปรของข้อมูล	ชนิดของข้อมูล	คำอธิบายตัวแปรของข้อมูล
OBS_30_CNT_SOCIAL_CIRCLE	float64	วันที่เลยกำหนดผิดนัดชำระ
DEF_30_CNT_SOCIAL_CIRCLE	float64	ผิดนัดชำระภายใน 30 วัน (วันที่เลยกำหนดผิดนัดชำระ)
OBS_60_CNT_SOCIAL_CIRCLE	float64	วันที่เกินกำหนดผิดนัดชำระ
DEF_60_CNT_SOCIAL_CIRCLE	float64	ผิดนัดชำระภายใน 60 วัน (วันที่เกินกำหนดผิดนัดชำระ)
DAYS_LAST_PHONE_CHANGE	float64	เปลี่ยนเบอร์โทรศัพท์มือถือ ก่อนการสมัครกี่วัน
FLAG_DOCUMENT_2	int64	ประเภทเอกสาร
FLAG_DOCUMENT_3	int64	ประเภทเอกสาร
FLAG_DOCUMENT_4	int64	ประเภทเอกสาร
FLAG_DOCUMENT_5	int64	ประเภทเอกสาร
FLAG_DOCUMENT_6	int64	ประเภทเอกสาร
FLAG_DOCUMENT_7	int64	ประเภทเอกสาร
FLAG_DOCUMENT_8	int64	ประเภทเอกสาร
FLAG_DOCUMENT_9	int64	ประเภทเอกสาร
FLAG_DOCUMENT_10	int64	ประเภทเอกสาร
FLAG_DOCUMENT_11	int64	ประเภทเอกสาร
FLAG_DOCUMENT_12	int64	ประเภทเอกสาร
FLAG_DOCUMENT_13	int64	ประเภทเอกสาร
FLAG_DOCUMENT_14	int64	ประเภทเอกสาร
FLAG_DOCUMENT_15	int64	ประเภทเอกสาร
FLAG_DOCUMENT_16	int64	ประเภทเอกสาร
FLAG_DOCUMENT_17	int64	ประเภทเอกสาร
FLAG_DOCUMENT_18	int64	ประเภทเอกสาร
FLAG_DOCUMENT_19	int64	ประเภทเอกสาร
FLAG_DOCUMENT_20	int64	ประเภทเอกสาร

ชื่อตัวแปรของข้อมูล	ชนิดของข้อมูล	คำอธิบายตัวแปรของข้อมูล
FLAG_DOCUMENT_21	int64	ประเภทเอกสาร
AMT_REQ_CREDIT_BUREAU_HOUR	float64	จำนวนคำถามของเครดิตบูโร หนึ่งชั่วโมงก่อนการสมัคร
AMT_REQ_CREDIT_BUREAU_DAY	float64	จำนวนคำถามของเครดิตบูโร หนึ่งวันก่อนการสมัคร (ไม่รวมก่อนการสมัครหนึ่งชั่วโมง)
AMT_REQ_CREDIT_BUREAU_WEEK	float64	จำนวนคำถามของเครดิตบูโร หนึ่งสัปดาห์ก่อนการสมัคร (ไม่รวมหนึ่งวันก่อนการสมัคร)
AMT_REQ_CREDIT_BUREAU_MON	float64	จำนวนคำถามของเครดิตบูโร หนึ่งเดือนก่อนการสมัคร (ไม่รวมหนึ่งสัปดาห์ก่อนการสมัคร)
AMT_REQ_CREDIT_BUREAU_QRT	float64	จำนวนคำถามของเครดิตบูโร 3 เดือนก่อนการสมัคร (ไม่รวมก่อนการสมัคร 1 เดือน)
AMT_REQ_CREDIT_BUREAU_YEAR	float64	จำนวนคำถามของเครดิตบูโร 1 วันต่อปี (ไม่รวม 3 เดือนล่าสุดก่อนการสมัคร)
YEARS_BIRTH	int64	อายุของลูกหนี้

1.4.2 กรอบแนวคิดในงานวิจัย

1. การสำรวจข้อมูล Exploratory Data Analysis (EDA) โดยการวิเคราะห์ระดับความสัมพันธ์ของแต่ละตัวแปร ค้นหาว่าตัวแปรแต่ละตัวมีความสัมพันธ์กันมากน้อยเพียงใด วัตถุประสงค์เพื่อลดจำนวนตัวแปรที่ใช้ในการพัฒนาแบบจำลอง เพื่อเพิ่มความรวดเร็วในการพัฒนาแบบจำลอง และยังทำให้แบบจำลองที่ได้มีประสิทธิภาพในการทำนายที่มีความแม่นยำมากยิ่งขึ้น

2. การเตรียมข้อมูล (Preparing Data) ซึ่งจะทำการจัดการกับข้อมูลที่ขาดหายไป ซึ่งจากชุดข้อมูลที่เรานำมาใช้ในการวิเคราะห์ข้อมูล มีข้อมูลที่ขาดหายไปเป็นจำนวนมาก ซึ่งจะจัดการกับข้อมูลที่ขาดหายไป โดยจะจัดการกับข้อมูลทั้งในแนวระดับแถวและในแนวระดับคอลัมน์ และมีการปรับความไม่สมดุลของชุดข้อมูล โดยการนำเทคนิค Oversampling, Under sampling และ Synthetic Minority Oversampling Technique (SMOTE) มาใช้ เพื่อช่วยเพิ่มประสิทธิภาพในการพัฒนาแบบจำลอง ทำให้ชุดข้อมูลก่อนการนำไปใช้ในการพัฒนาแบบจำลอง เป็นชุดข้อมูลที่มีความสมดุลกันของชุดข้อมูล ซึ่งจะช่วยส่งผลให้แบบจำลองที่ได้ มีประสิทธิภาพในการทำนายที่มีความแม่นยำมากยิ่งขึ้น

3. การพัฒนาแบบจำลอง ซึ่งเราจะทำการพัฒนาแบบจำลอง โดยการใช้ Machine Learning Algorithms โดยอาศัยการเรียนรู้ของเครื่อง ซึ่งเป็นเครื่องมือสำหรับการพัฒนาแบบจำลอง ในการเรียนรู้แบบผู้สอน โดยมีการทำงานแบบการแบ่งแยกประเภท ซึ่งในงานวิจัยนี้อัลกอริทึมที่ถูกนำมาใช้ในการพัฒนาแบบจำลอง ได้แก่ Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting

4. การประเมินประสิทธิภาพของแบบจำลอง เพื่อประเมินประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้ Confusion matrix โดยดูที่ค่า Accuracy, Precision, Recall และ F1-Score

1.5 **สมมุติฐานในการวิจัย**

1. การจัดการกับข้อมูลที่ขาดหายไปเป็นจำนวนมาก ก่อนนำข้อมูลเหล่านี้ไปใช้ในการพัฒนาแบบจำลอง ทำให้ข้อมูลก่อนการนำไปพัฒนาแบบจำลองเป็นชุดข้อมูลที่มีประสิทธิภาพ

2. การปรับความไม่สมดุลของชุดข้อมูล ด้วยเทคนิค Synthetic Minority Over-sampling (SMOTE) จะมีส่วนช่วยทำให้แบบจำลองที่ได้มีประสิทธิภาพในการทำนายที่มีความแม่นยำมากยิ่งขึ้น

3. อัลกอริทึมต่างๆ ที่ใช้ในการพัฒนาแบบจำลอง อัลกอริทึม XGBoostClassifier ให้ผลลัพธ์ในการทำนายโอกาสของการเกิดลูกหนี้ที่มีการผิดนัดชำระได้อย่างแม่นยำมากที่สุด

4. เพื่อประเมินประสิทธิภาพของแบบจำลอง โดยการใช้ Confusion matrix โดยค่า Accuracy, Precision, Recall และ F1-Score ที่ได้จะมีค่าสูง และมีความแม่นยำในการทำนายโอกาสของการเกิดลูกหนี้ที่มีการผิดนัดชำระได้อย่างแม่นยำและมีประสิทธิภาพ

1.6 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. เพื่อช่วยให้ธนาคารสามารถประเมินสถานะในการขอสินเชื่อของลูกหนี้แต่ละรายในการวิเคราะห์สินเชื่อได้

2. เพื่อช่วยให้ธนาคารสามารถประเมินสภาวะของลูกหนี้แต่ละรายเบื้องต้นได้ ว่ามีโอกาสของการเกิดลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคารไหม

3. ทำให้ทราบถึงปัจจัยที่อาจส่งผลต่อการพัฒนาแบบจำลอง ที่ใช้ในการทำนายโอกาสของการเกิดลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร

4. สามารถนำแบบจำลองที่ได้ ไปประยุกต์ใช้ในสถาบันทางการเงินต่างๆได้

5. สามารถนำแบบจำลองที่ได้ไปต่อยอด เพื่อพัฒนาเป็นแบบจำลองที่ดี และมีประสิทธิภาพที่ดียิ่งขึ้นได้

โดยโครงสร้างของงานวิจัยเล่มนี้จะประกอบไปด้วยเนื้อหาต่างๆ ดังนี้ บทที่ 2 ทบทวนวรรณกรรม บทที่ 3 วิธีการดำเนินการวิจัย บทที่ 4 ผลการดำเนินการวิจัย และบทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

บทที่ 2
ทบทวนวรรณกรรม

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

1. งานวิจัยที่เกี่ยวข้องกับการทำ Credit card fraud detection
2. ทฤษฎีเกี่ยวกับหลักการทำงานของแบบจำลองที่ใช้ในงานวิจัยนี้

ผู้วิจัยได้ศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับการพัฒนาแบบจำลองในการตรวจจับการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระ ซึ่งงานวิจัยส่วนใหญ่ที่พบจะเกี่ยวข้องกับการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต

2.1 งานวิจัยที่เกี่ยวข้องกับการทำ Credit card fraud detection

2.1.1 Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis (1)

งานวิจัยนี้ได้กล่าวถึง การตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต โดยการใช้เทคนิค Machine Learning บทความนี้ได้ทำการตรวจสอบประสิทธิภาพของการพัฒนาแบบจำลองโดยการใช้อัลกอริทึม naïve bayes, k-nearest neighbor และ logistic regression กับชุดข้อมูลธุรกรรมการใช้งานบัตรเครดิตที่ได้มาจากผู้ถือบัตรเครดิตในยุโรป ซึ่งมีธุรกรรมการทำรายการทั้งหมด 284,807 รายการ โดยธุรกรรม 492 รายการเป็นธุรกรรมที่เป็นการฉ้อโกง ซึ่งจะเห็นได้ว่าชุดข้อมูลนี้มีความไม่สมดุลอย่างมาก โดยมีเพียง 0.173% ของธุรกรรมที่ถูกระบุว่าเป็นการฉ้อโกง โดยจะใช้เทคนิค under-sampling กับชุดข้อมูลประวัติพฤติกรรมของคนที่ใช้งานบัตรเครดิตที่ถูกกฎหมาย และ over-sampling กับชุดข้อมูลประวัติพฤติกรรมของคนที่ใช้งานบัตรเครดิตในการฉ้อโกง เพื่อแก้ไขปัญหาความไม่สมดุลของชุดข้อมูล และมีการใช้เทคนิค Feature Selection เพื่อคัดเลือก Feature ที่สำคัญที่จะนำมาใช้ในการพัฒนาแบบจำลอง ซึ่งในงานวิจัยนี้มีการทดลองกับชุดข้อมูล 3 รูปแบบ คือ วิธีการที่ไม่ได้สุ่มชุดข้อมูล วิธีการสุ่มชุดข้อมูลแบบ hybrid ด้วยอัตราส่วน 10:90 และวิธีการสุ่มชุดข้อมูลแบบ hybrid ด้วยอัตราส่วน 34:66 และแสดงการเปรียบเทียบประสิทธิภาพของการพัฒนาแบบจำลองกับงานวิจัยอื่นๆที่เกี่ยวข้อง ซึ่งผลลัพธ์ประสิทธิภาพของการพัฒนาแบบจำลองจะดูจากค่า accuracy, sensitivity, specificity,

precision, Matthews correlation coefficient และ balanced classification rate ซึ่งแบบจำลองที่ให้ผลลัพธ์ค่าความแม่นยำที่เหมาะสมที่ดีที่สุด คือ แบบจำลองที่พัฒนาโดยอัลกอริทึม k-nearest neighbor ซึ่งให้ค่าความแม่นยำเท่ากับ 97.92% และรองลงมาคือแบบจำลองที่พัฒนาโดยอัลกอริทึม naïve bayes ซึ่งให้ค่าความแม่นยำเท่ากับ 97.69% และสุดท้ายแบบจำลองที่พัฒนาโดยอัลกอริทึม logistic regression ซึ่งให้ค่าความแม่นยำเท่ากับ 54.86% โดยสรุปจากผลลัพธ์ค่าความแม่นยำที่เกิดขึ้น แสดงให้เห็นว่า k-nearest neighbor มีประสิทธิภาพดีที่สุด

2.1.2 Credit Card Fraud Detection - Machine Learning methods (2)

งานวิจัยนี้ได้กล่าวถึง การตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต โดยการใช้เทคนิค Machine Learning บทความนี้ จะแสดง Machine Learning หลายๆอัลกอริทึมที่นำมาใช้ในการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต มีธุรกรรมการทำรายการทั้งหมด 284,807 รายการ โดยธุรกรรม 492 รายการเป็นธุรกรรมที่เป็นการฉ้อโกง ซึ่งจะเห็นได้ว่าชุดข้อมูลนี้มีความไม่สมดุลอย่างมาก โดยมีเพียง 0.173% ของธุรกรรมที่ถูกระบุว่าเป็นการฉ้อโกง มีการใช้เทคนิค Feature Selection เพื่อลดการเกิด overfitting ลดเวลาในการพัฒนาแบบจำลอง และเพื่อเพิ่มค่าความแม่นยำของการพัฒนาแบบจำลอง ทำให้แบบจำลองที่ได้มีประสิทธิภาพที่ดียิ่งขึ้น ซึ่งได้มีการลด Feature ทำให้เหลือ 27 Feature ที่ถูกนำมาใช้ในการพัฒนาแบบจำลอง ซึ่งจากชุดข้อมูลที่นำมาใช้ในการพัฒนาแบบจำลอง ชุดข้อมูลมีความไม่สมดุลกันของชุดข้อมูลสูงมาก จึงได้มีการใช้เทคนิค SMOTE (Synthetic Minority Over-sampling) เพื่อนำมาช่วยในการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล โดยจะทำการสุ่มตัวอย่างของชุดข้อมูลขึ้นมา เพื่อให้ชุดของข้อมูลเกิดความสมดุลกัน ซึ่งจะทำให้ชุดข้อมูลเกิดความสมดุลกันอยู่ในอัตราส่วน 50:50 และมีการทำ scaling เนื่องจากข้อมูลเวลาและข้อมูลจำนวนเงินมีความแตกต่างกันอย่างมาก จึงต้องมีการปรับ scale ของข้อมูลเพื่อให้ข้อมูลอยู่ในระดับเดียวกัน ในการพัฒนาแบบจำลองได้มีการใช้โปรแกรม Spyder ในการเขียนภาษา Python เพื่อการพัฒนาแบบจำลอง ซึ่งเป็นส่วนหนึ่งของแพลตฟอร์ม Anaconda โดยมีการใช้ libraries ได้แก่ numpy, pandas, matplotlib, sklearn และ imblearn โดยอัลกอริทึมที่ใช้ในการพัฒนาแบบจำลอง ได้แก่ Logistic Regression, Random Forest, Naive Bayes และ Multilayer Perceptron ในการพิจารณาว่าอัลกอริทึมใดเหมาะสมที่สุดสำหรับการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต ซึ่งเกณฑ์ที่ใช้ในการพิจารณาผลลัพธ์ของอัลกอริทึมการเรียนรู้ของเครื่องคือค่า Accuracy, Recall และ Precision ซึ่งแบบจำลองที่ให้ผลลัพธ์ค่าความแม่นยำที่เหมาะสมที่ดีที่สุด คือ แบบจำลองที่พัฒนาโดยอัลกอริทึม Random Forest

ซึ่งให้ค่าความแม่นยำเท่ากับ 99.96% และอันดับที่สองคือแบบจำลองที่พัฒนาโดยอัลกอริทึม Multilayer Perceptron ซึ่งให้ค่าความแม่นยำเท่ากับ 99.93% และอันดับที่สามคือแบบจำลองที่พัฒนาโดยอัลกอริทึม Naive Bayes ซึ่งให้ค่าความแม่นยำเท่ากับ 99.23% และสุดท้ายแบบจำลองที่พัฒนาโดยอัลกอริทึม Logistic Regression ซึ่งให้ค่าความแม่นยำเท่ากับ 97.46% โดยสรุปจากผลลัพธ์ค่าความแม่นยำที่เกิดขึ้น แสดงให้เห็นว่า Random Forest มีประสิทธิภาพดีที่สุด

2.2 ทฤษฎีเกี่ยวกับหลักการทำงานของแบบจำลองที่ใช้ในงานวิจัยนี้

Random Forest for Credit Card Fraud Detection (3)

2.2.1 Random Forest

Random Forest เป็นหนึ่งในวิธีในการนำมาใช้ในการตรวจจับการฉ้อโกงของการใช้งานบัตรเครดิต โดย Random Forest เป็นอัลกอริทึมที่นิยมในการทำต้นไม้ตัดสินใจ เนื่องจากมีความยืดหยุ่นในการจัดการคุณลักษณะข้อมูลประเภทต่างๆ อย่างไรก็ตามแบบจำลองต้นไม้ตัดสินใจต้นเดียว อาจมีประสิทธิภาพที่ไม่เพียงพอและทำให้เกิดการ Overfit ซึ่งจะใช้เทคนิค Ensemble ในการแก้ปัญหาเหล่านี้ โดยการรวมกลุ่มกันของต้นไม้ตัดสินใจหลายๆต้น จะช่วยเพิ่มความแม่นยำมากกว่าการทำต้นไม้ตัดสินใจต้นเดียว

Random Forest คือหนึ่งในเทคนิคการทำ Ensemble ซึ่งหลักการทำงานของ Random Forest คือการรวมกันของต้นไม้ตัดสินใจหลายๆต้น เพื่อช่วยกันในการทำนายให้มีประสิทธิภาพที่ดียิ่งขึ้น โดยที่ต้นไม้ตัดสินใจแต่ละต้นจะได้รับคุณลักษณะและข้อมูลที่ไม่เหมือนกัน เพื่อให้ต้นไม้ตัดสินใจแต่ละต้นมีความแตกต่างและเป็นอิสระต่อกันมากยิ่งขึ้น ความสามารถของ Random Forest ไม่เพียงแต่ขึ้นอยู่กับความแข็งแกร่งของต้นไม้ตัดสินใจแต่ละต้นเท่านั้น แต่ยังขึ้นอยู่กับความสัมพันธ์ระหว่างต้นไม้ตัดสินใจแต่ละต้น ยิ่งต้นไม้ตัดสินใจแต่ละต้นมีความสัมพันธ์ที่แตกต่างกันมากเท่าไหร่ ประสิทธิภาพของ Random Forest ก็จะยิ่งดีมากขึ้นเท่านั้น ความผันแปรของต้นไม้ตัดสินใจ มาจากเทคนิคการทำ Bootstrap และการสุ่มเลือกชุดย่อยของข้อมูลที่แตกต่างกัน โดยค่าที่ได้จากการทำนาย จะเป็นค่าการทำนายที่ได้ของต้นไม้ตัดสินใจแต่ละต้น จากนั้นค่าการทำนายสุดท้าย ในกรณีปัญหาแบบการจำแนกประเภท (Classification) จะใช้ผลโหวตที่มากที่สุด (Majority vote) โดยค่าการทำนายของต้นไม้ตัดสินใจต้นไหนที่ได้รับผลโหวตมากที่สุด จะถูกเลือกให้เป็นค่าการทำนายของปัญหานั้นๆ ซึ่งข้อดีของ Random Forest คือมีความแข็งแกร่งต่อสัญญาณรบกวน (noise) และมีความแข็งแกร่งต่อค่าที่ผิดปกติ (outlier)

Credit card fraud detection using Naïve Bayes model based and KNN classifier (4)

2.2.2 Naïve Bayes

Bayesian network classifiers ได้รับความนิยมอย่างมากในด้านการเรียนรู้ของเครื่อง และเป็นตัวในการจำแนกประเภทแบบการเรียนรู้แบบมีผู้สอน (Supervised Learning) Naïve Bayes เป็นเทคนิคที่ใช้ทฤษฎีความน่าจะเป็นตามกฎของเบย์ โดยอาศัยหลักการของความน่าจะเป็นเข้ามาช่วยในการหาคำตอบของเหตุการณ์หนึ่งๆที่สนใจ โดยจะพยายามทำนายคลาสที่เรียกว่าคลาสผลลัพธ์ ซึ่งจะพิจารณาจากความน่าจะเป็นแบบมีเงื่อนไขว่าคลาสนั้นเกิดขึ้นกี่ครั้งจากข้อมูลการฝึกอบรม เทคนิค Naïve Bayes มีข้อดีคือมีประสิทธิภาพ มีความรวดเร็ว มีความแม่นยำสูง ง่ายต่อการนำไปใช้ และมีสมมติฐานที่ชัดเจน ซึ่ง Naïve Bayes จะให้ผลลัพธ์ที่ค่อนข้างแม่นยำ และยังได้รับการพิสูจน์มาแล้วหลายครั้งว่า Naïve Bayes ทำงานได้อย่างมีประสิทธิภาพในงานทางด้านต่างๆ ที่เกี่ยวข้องกับการเรียนรู้ของเครื่อง

(5) Naïve Bayes จะประกอบด้วย nodes และ edges โดยที่ nodes เป็นตัวแทนของตัวแปรสุ่ม และ edges แสดงถึงความสัมพันธ์ระหว่างตัวแปรสุ่ม โดยวิธีการคำนวณ จะมีการกำหนดค่าความน่าจะเป็นขั้นต่ำและค่าความน่าจะเป็นสูงสุดไว้ล่วงหน้าของธุรกรรมที่เป็นการฉ้อโกงหรือธุรกรรมที่เป็นธุรกรรมที่ถูกกฎหมาย จากนั้นสำหรับธุรกรรมที่เข้ามาใหม่ เราจะคำนวณค่าความน่าจะเป็นของธุรกรรมนั้น หากค่าความน่าจะเป็นที่ถูกกฎหมายน้อยกว่าค่าต่ำสุดที่กำหนดไว้สำหรับธุรกรรมที่ถูกกฎหมาย และมากกว่าค่าสูงสุดที่กำหนดไว้สำหรับธุรกรรมที่เป็นการฉ้อโกง หากเป็นจริงธุรกรรมที่เข้ามาใหม่จะถูกจัดประเภทเป็นการฉ้อโกง

2.2.3 K-Nearest Neighbor

K-Nearest Neighbor เป็นหนึ่งในอัลกอริทึมที่ใช้มากที่สุดสำหรับทั้งปัญหาการจำแนกประเภท (classification) และการทำนายการถดถอย (regression) ประสิทธิภาพของ K-Nearest Neighbor จะขึ้นอยู่กับปัจจัย 3 อย่าง ได้แก่ ตัววัดระยะทาง กฎระยะทาง และค่าของตัววัดระยะทาง (k) ซึ่งตัววัดระยะทางใช้ในการวัดเพื่อค้นหาเพื่อนบ้านที่ใกล้ที่สุดของจุดข้อมูลใหม่ที่เข้ามา กฎระยะทางใช้ในการจัดประเภทจุดข้อมูลใหม่โดยเปรียบเทียบคุณลักษณะของจุดข้อมูลใหม่ที่เข้ามากับจุดข้อมูลเดิมที่อยู่ในพื้นที่ใกล้เคียงกัน และค่าของ k จะกำหนดจำนวนเพื่อนบ้านที่จะนำมาใช้ในการเปรียบเทียบ (5)

เนื่องจากอัลกอริทึมนี้มีการเรียนรู้ที่ช้ามาก ถ้าเปรียบเทียบกับอัลกอริทึมอื่นๆ เนื่องจาก K-Nearest Neighbor ต้องมีการคำนวณระยะห่างระหว่างข้อมูลที่ต้องการใช้ในการพิจารณากับชุดข้อมูลตัวอย่าง จำนวน k ชุด หลักการทำงานของ K-Nearest Neighbor คือมีการแบ่งข้อมูลออกเป็นกลุ่มต่างๆ โดยจะทำการวิเคราะห์ข้อมูลใหม่จากข้อมูลเดิม ที่อยู่ในบริเวณใกล้เคียงกัน โดยจะกำหนดค่า k ซึ่งค่า k คือค่าที่ใช้ในการกำหนดว่าจะวิเคราะห์ข้อมูลที่อยู่ใกล้กับข้อมูลที่ต้องการจำแนกที่สุดกี่ข้อมูล ซึ่งการกำหนดค่า k ที่แตกต่างกันนั้น จะทำให้ได้ค่าความแม่นยำที่ไม่เท่ากัน ซึ่งนั้นก็เป็นส่วนหนึ่งในการหาค่า k ที่ดีที่สุด และทำการพิจารณาตัวแปรคลาสที่มีคะแนนโหวตสูงสุด และนำค่าตัวแปรคลาสนั้นมาเป็นคำตอบของปัญหานั้นๆ ข้อดีของ K-Nearest Neighbor คือเป็นเทคนิคที่เรียบง่าย สามารถใช้ในการแก้ปัญหาที่มีความซับซ้อนได้ และมีประสิทธิภาพสูง

A Comparative Analysis of Various Credit Card Fraud Detection Techniques (5)

2.2.4 Logistic Regression

Logistic Regression เป็นเทคนิคในการวิเคราะห์สถิติเชิงคุณภาพ เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสความน่าจะเป็นที่จะเกิดเหตุการณ์หนึ่งๆที่สนใจ หรือไม่เกิดเหตุการณ์หนึ่งๆที่สนใจ โดยอาศัยสมการ Logistic ที่สร้างขึ้นจากชุดตัวแปรทำนาย โดยที่ระหว่างตัวแปรทำนายจะต้องมีความสัมพันธ์กันต่ำ Logistic Regression แบ่งออกเป็น 2 ประเภท ได้แก่ การวิเคราะห์ Logistic ทวิ ซึ่งจะใช้กับตัวแปรคลาส ที่มีค่า 2 คลาส ได้แก่ ทำนายโอกาสความน่าจะเป็นที่จะเกิดเหตุการณ์ของเหตุการณ์หนึ่งๆที่สนใจมีค่าเป็น 1 หรือไม่เกิดเหตุการณ์ของเหตุการณ์หนึ่งๆที่สนใจมีค่าเป็น 0 และการวิเคราะห์การถดถอย Logistic พหุกลุ่ม ซึ่งจะใช้กับตัวแปรคลาส ที่มีค่ามากกว่า 2 คลาส โดย Logistic ทั้ง 2 ประเภท จะแตกต่างกันในด้านตัวแปรตาม

Logistic Regression มีเป้าหมายในการประมาณค่าของสัมประสิทธิ์ของพารามิเตอร์โดยการใช้ sigmoid function เมื่อธุรกรรมดำเนินไป ค่าของ attributes จะทำการตรวจสอบและบอกว่าธุรกรรมดังกล่าวควรดำเนินการต่อไปหรือไม่ ถ้าเป็นธุรกรรมที่เป็นธุรกรรมที่ถูกกฎหมายทั่วไปธุรกรรมจะยังคงดำเนินต่อไป แต่ถ้าเป็นธุรกรรมที่เป็นการฉ้อโกงก็จะหยุดการดำเนินการ

Imbalance-XGBoost: leveraging weighted and focal losses forbinary label-imbalanced classification with XGBoost (6)

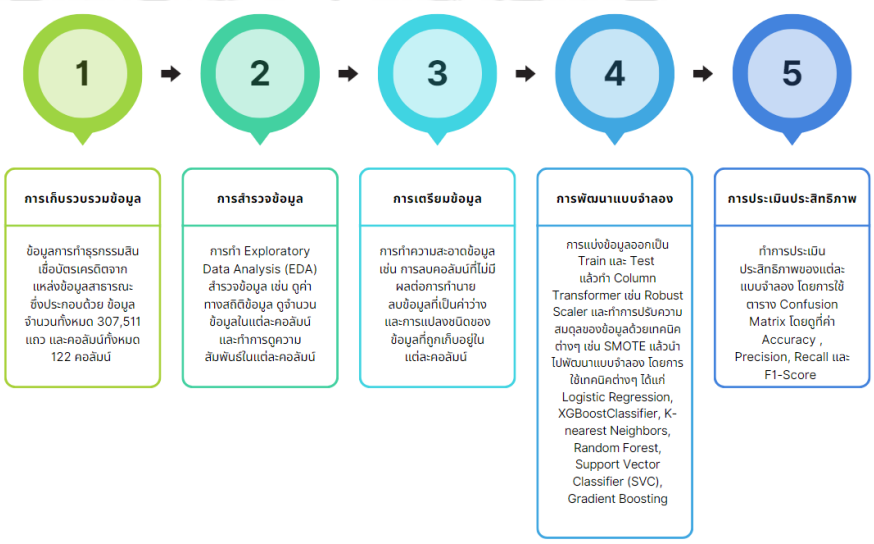
2.2.5 XGBoost

XGBoost เป็นอัลกอริทึมที่ถูกพัฒนาขึ้นมาจาก Gradient Tree Boosting ในแง่ของความเร็วและขนาดในการคำนวณ ซึ่งสามารถจัดการงานที่มีข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ XGBoost เป็นอัลกอริทึมที่มีประสิทธิภาพ และใช้เวลาในการพัฒนาแบบจำลองที่เหมาะสม ซึ่งมีการประยุกต์ใช้งานในด้านการวิจัยที่หลากหลายตั้งแต่การวินิจฉัยโรคมะเร็งไปจนถึงการประเมินความเสี่ยงด้านการใช้งานบัตรเครดิต โดยจะใช้ภาษา Python ในการพัฒนาแบบจำลอง ซึ่งในปัจจุบัน XGBoost ได้กลายเป็นวิธีการทางเลือกแรกสำหรับใช้ในการพัฒนาแบบจำลองของข้อมูลขนาดใหญ่ และยังเป็นวิธีการในการพัฒนาแบบจำลองที่ได้รับความนิยมมากที่สุด ถึงแม้ว่า XGBoost ประสบความสำเร็จอย่างมาก ทั้งในด้านปัญหาการทำนายการถดถอย (regression) และการจำแนกประเภท (classification) แต่ประสิทธิภาพของมันมักจะลดลง เมื่อชุดข้อมูลที่ถูกนำมาใช้ในการวิเคราะห์มีปัญหาความไม่สมดุลกันของชุดข้อมูล (Imbalance Data) แต่มีหลายงานวิจัยที่บอกว่า XGBoost สามารถใช้ในการจัดการกับปัญหาความไม่สมดุลกันของชุดข้อมูลได้ดี ซึ่งสามารถทำงานได้อย่างมีประสิทธิภาพเหนือกว่าวิธีการอื่นๆ ในการจัดการกับปัญหาความไม่สมดุลกันของชุดข้อมูล ซึ่งในงานวิจัยนี้ได้มีการแนะนำ imbalance-XGBoost ซึ่งเป็นแพ็คเกจ Python ที่ใช้ XGBoost ในการแก้ไขปัญหาความไม่สมดุลกันของชุดข้อมูล

บทที่ 3
วิธีการดำเนินงานวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้มีการวางแผนขั้นตอนในการดำเนินงานวิจัยซึ่งมีรายละเอียดดังต่อไปนี้

1. การเก็บรวบรวมข้อมูล
2. การจัดกระทำข้อมูลและการวิเคราะห์ข้อมูล
3. การเตรียมข้อมูล
4. การพัฒนาแบบจำลอง
5. การประเมินประสิทธิภาพของแบบจำลอง



ภาพประกอบ 1 Flow Chart วิธีดำเนินการพัฒนาแบบจำลอง

ในภาพประกอบที่ 1 จะแสดงขั้นตอนวิธีการดำเนินการพัฒนาแบบจำลอง ประกอบไปด้วย 5 ขั้นตอนหลักๆ คือ การเก็บรวบรวมข้อมูล การสำรวจข้อมูล การเตรียมข้อมูล การพัฒนาแบบจำลอง และการประเมินประสิทธิภาพของแบบจำลอง

3.1 การเก็บรวบรวมข้อมูล

ข้อมูลการทำธุรกรรมสินเชื่อบัตรเครดิต ประกอบด้วย ข้อมูลจำนวนทั้งหมด 307,511 แถว และประกอบด้วยคอลัมน์ทั้งหมด 122 คอลัมน์ จากแหล่งข้อมูลสาธารณะ Kaggle.com จากเว็บไซต์ https://www.kaggle.com/datasets/mishra5001/credit-card?resource=download โดยการแบ่งข้อมูลออกเป็น 2 กลุ่มใหญ่ๆ กลุ่มลูกหนี้ปกติ คือกลุ่มลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และกลุ่มลูกหนี้ที่ไม่ปกติ คือกลุ่มลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร

3.2 การจัดกระทำข้อมูลและการวิเคราะห์ข้อมูล

งานวิจัยนี้ มีการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยอาศัยการเรียนรู้ของเครื่องมือที่ช่วยในการตัดสินใจ โดยมีขั้นตอนการพัฒนาแบบจำลองและการเปรียบเทียบประสิทธิภาพของแบบจำลองดังต่อไปนี้

โดยการวิเคราะห์ระดับความสัมพันธ์ของตัวแปร ค้นหาว่าตัวแปรแต่ละตัวมีความสัมพันธ์กันมากน้อยเพียงใด วัตถุประสงค์เพื่อลดจำนวนตัวแปรที่ใช้ในการพัฒนาแบบจำลอง เพื่อความรวดเร็วในการพัฒนาแบบจำลอง และเพื่อเพิ่มประสิทธิภาพในการเรียนรู้ข้อมูลของแบบจำลอง และยังทำให้แบบจำลองที่ได้มีประสิทธิภาพในการทำนายที่มีความแม่นยำมากยิ่งขึ้น

โดยการสำรวจข้อมูล จะทำการสำรวจและตรวจสอบข้อมูลเบื้องต้น ว่าข้อมูลที่นำมาใช้ในการวิเคราะห์มีประสิทธิภาพมากน้อยเพียงใด โดยจะทำการวิเคราะห์ข้อมูลทั้งหมดที่นำมาใช้ในการพัฒนาแบบจำลอง และสำรวจดูว่าแต่ละคอลัมน์ของชุดข้อมูลเก็บข้อมูลเป็นชนิดอะไร

ซึ่งชุดข้อมูลมีจำนวนแถวทั้งหมด 307,511 แถว และจำนวนคอลัมน์ทั้งหมด 122 คอลัมน์ และมีการเก็บข้อมูลเป็นชนิดต่างๆ ดังต่อไปนี้

1. int64 จะเก็บข้อมูลที่มีลักษณะของข้อมูลเป็นตัวเลขจำนวนเต็ม (integer)
2. float64 จะเก็บข้อมูลที่มีลักษณะของข้อมูลเป็นตัวเลขที่เป็นจุดทศนิยม (float)
3. object จะเก็บข้อมูลที่มีลักษณะของข้อมูลเป็นตัวอักษร (object)

และจะสำรวจดูค่าทางสถิติของแต่ละคอลัมน์ ว่ามีค่าทางสถิติเป็นอย่างไรบ้าง

โดยค่าทางสถิติ จะเป็นค่าทางสถิติที่สรุปแนวโน้มรูปแบบของการกระจายตัว โดยจะไม่นับรวมข้อมูลที่ขาดหายไป โดยค่าทางสถิติที่แสดงจะเป็นค่าทางสถิติของคอลัมน์ที่เก็บข้อมูลอยู่ในรูปแบบของตัวเลข (numerical) ได้แก่ int64, float64 ซึ่งถ้าคอลัมน์ไหนที่เก็บข้อมูลอยู่ในรูปแบบของตัวอักษร (object) จะไม่ได้ถูกนำมาคำนวณด้วย

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	307499.000000	3.072330e+05
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	27108.573909	5.383962e+05
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	14493.737315	3.694465e+05
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	1615.500000	4.050000e+04
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16524.000000	2.385000e+05
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000	4.500000e+05
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06

ภาพประกอบ 2 ค่าทางสถิติของข้อมูลที่อยู่ในรูปแบบของตัวเลข

ในภาพประกอบที่ 2 จะแสดงค่าทางสถิติของข้อมูลที่อยู่ในรูปแบบของตัวเลข ซึ่งประกอบด้วยค่าต่างๆ ดังต่อไปนี้

1. count คือ จำนวนแถวของข้อมูลทั้งหมด ที่ถูกเก็บอยู่ในคอลัมน์นั้น
2. mean คือ ค่าเฉลี่ยของข้อมูล ผลรวมของข้อมูลทั้งหมด หารด้วยจำนวนของข้อมูลทั้งหมดที่ถูกเก็บอยู่ในคอลัมน์นั้น
3. min คือ ค่าที่น้อยที่สุดของข้อมูลทั้งหมด ที่ถูกเก็บอยู่ในคอลัมน์นั้น
4. max คือ ค่าที่มากที่สุดของข้อมูลทั้งหมด ที่ถูกเก็บอยู่ในคอลัมน์นั้น
5. standard deviation คือ ค่าเบี่ยงเบนมาตรฐานของข้อมูลทั้งหมด ที่ถูกเก็บอยู่ในคอลัมน์นั้น

และเมื่อต้องการ ดูค่าทางสถิติของคอลัมน์ที่เก็บข้อมูลอยู่ในรูปแบบของตัวอักษร (object) โดยจะไม่นับรวมข้อมูลที่ขาดหายไป

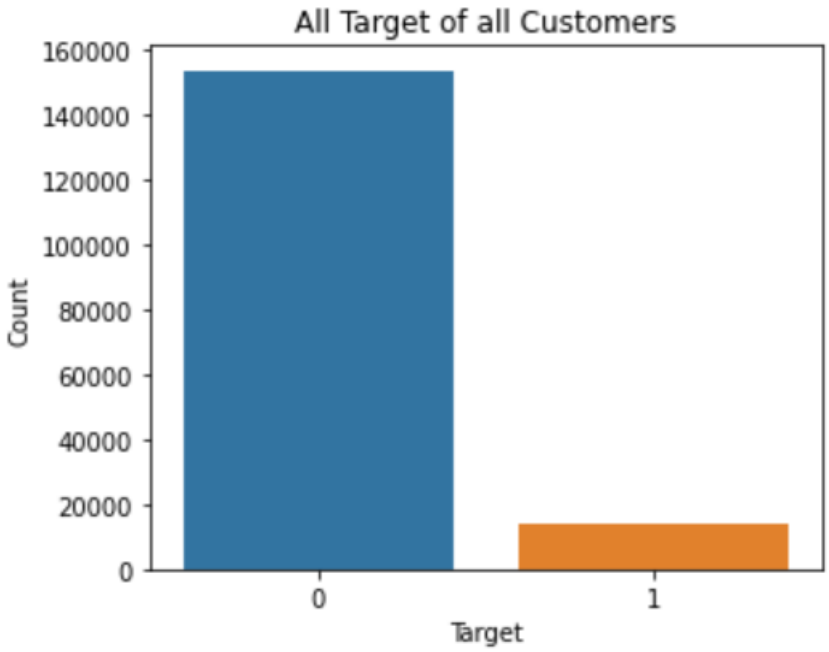
	count	unique	top	freq
NAME_CONTRACT_TYPE	307511	2	Cash loans	278232
CODE_GENDER	307511	3	F	202448
FLAG_OWN_CAR	307511	2	N	202924
FLAG_OWN_REALTY	307511	2	Y	213312
NAME_TYPE_SUITE	306219	7	Unaccompanied	248526
NAME_INCOME_TYPE	307511	8	Working	158774
NAME_EDUCATION_TYPE	307511	5	Secondary / secondary special	218391
NAME_FAMILY_STATUS	307511	6	Married	196432
NAME_HOUSING_TYPE	307511	6	House / apartment	272868
OCCUPATION_TYPE	211120	18	Laborers	55186
WEEKDAY_APPR_PROCESS_START	307511	7	TUESDAY	53901
ORGANIZATION_TYPE	307511	58	Business Entity Type 3	67992
FONDKAPREMONT_MODE	97216	4	reg oper account	73830
HOUSETYPE_MODE	153214	3	block of flats	150503
WALLSMATERIAL_MODE	151170	7	Panel	66040
EMERGENCYSTATE_MODE	161756	2	No	159428

ภาพประกอบ 3 ค่าทางสถิติของข้อมูลที่อยู่ในรูปแบบของตัวอักษร (object)

ในภาพประกอบที่ 3 จะแสดงค่าทางสถิติของข้อมูลที่อยู่ในรูปแบบของตัวอักษร (object) ซึ่งประกอบด้วยค่าต่างๆ ดังต่อไปนี้

1. count คือ จำนวนแถวของข้อมูลทั้งหมด ที่ถูกเก็บอยู่ในคอลัมน์นั้น
2. unique คือ คอลัมน์นั้นเก็บข้อมูลที่แตกต่างกันทั้งหมดกี่ค่า
3. top คือ ข้อมูลที่มีความถี่หรือจำนวนของข้อมูลมากที่สุด คือข้อมูลอะไร
4. freq คือ ข้อมูลที่มีความถี่หรือจำนวนของข้อมูลมากที่สุด มีทั้งหมดกี่ค่า

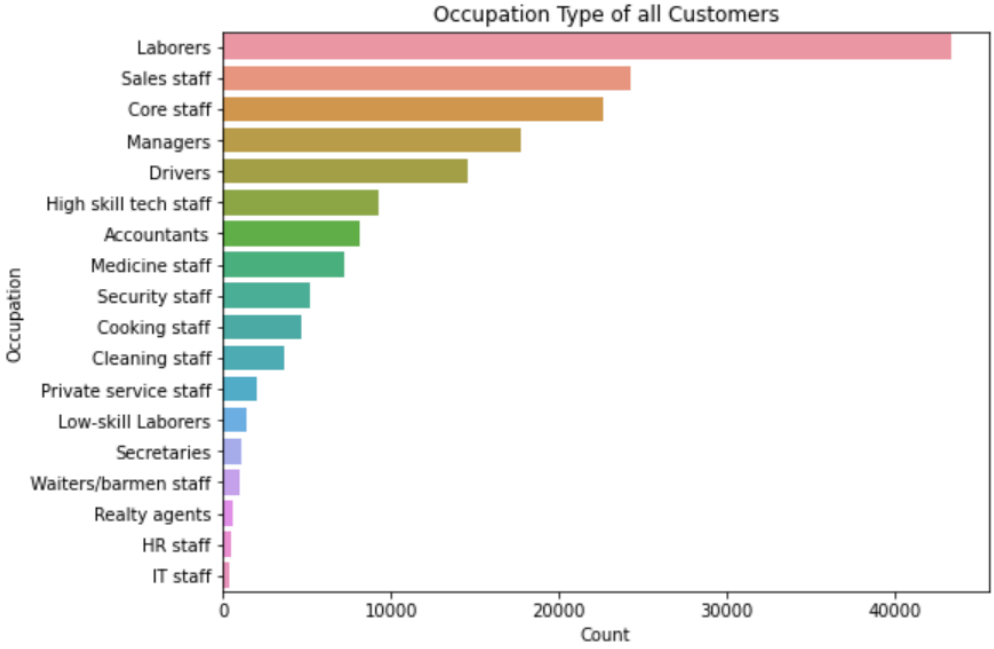
และจากการสำรวจพบว่าข้อมูลที่นำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง มีความไม่สมดุลกันของชุดข้อมูลสูงมาก



ภาพประกอบ 4 จำนวนข้อมูลของตัวแปรเป้าหมาย (Target class)

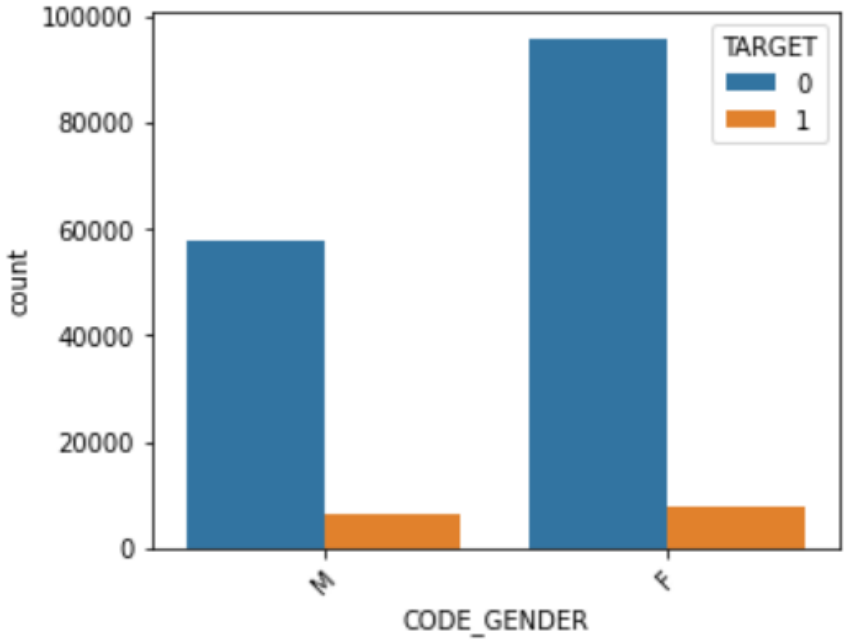
ในภาพประกอบ 4 จะแสดงว่าข้อมูลมีความไม่สมดุลกันของชุดข้อมูลสูงมาก ซึ่งมีจำนวนแถวชุดข้อมูลของลูกหนี้ปกติ ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคารอยู่จำนวน 153,525 แถว และมีจำนวนแถวชุดข้อมูลของลูกหนี้ผิดปกติ ที่มีการผิดนัดชำระกับทางธนาคารอยู่จำนวน 14,207 แถว

และจะทำการ plot เพื่อดูข้อมูลในแต่ละคอลัมน์ว่าเก็บข้อมูลอะไรบ้าง และข้อมูลที่จัดเก็บในแต่ละคอลัมน์มีมากน้อยเพียงใด



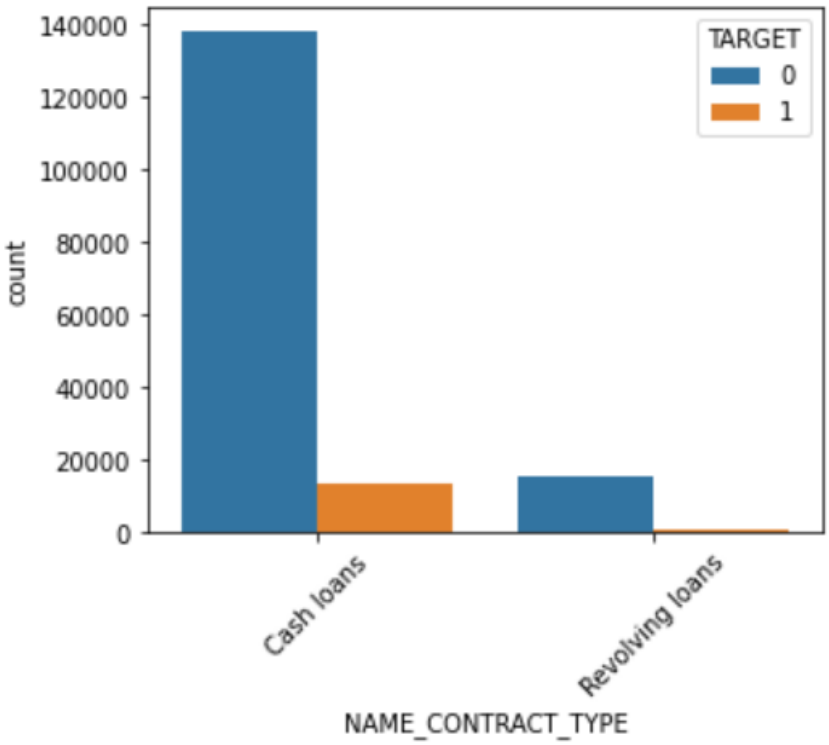
ภาพประกอบ 5 จำนวนข้อมูลอาชีพของลูกหนี้

ในภาพประกอบ 5 จะใช้ในการบอกอาชีพของลูกหนี้ โดยลูกหนี้ส่วนใหญ่จะมีอาชีพเป็นกรรมกร ซึ่งมีจำนวนข้อมูลทั้งหมด 43,437 แถว



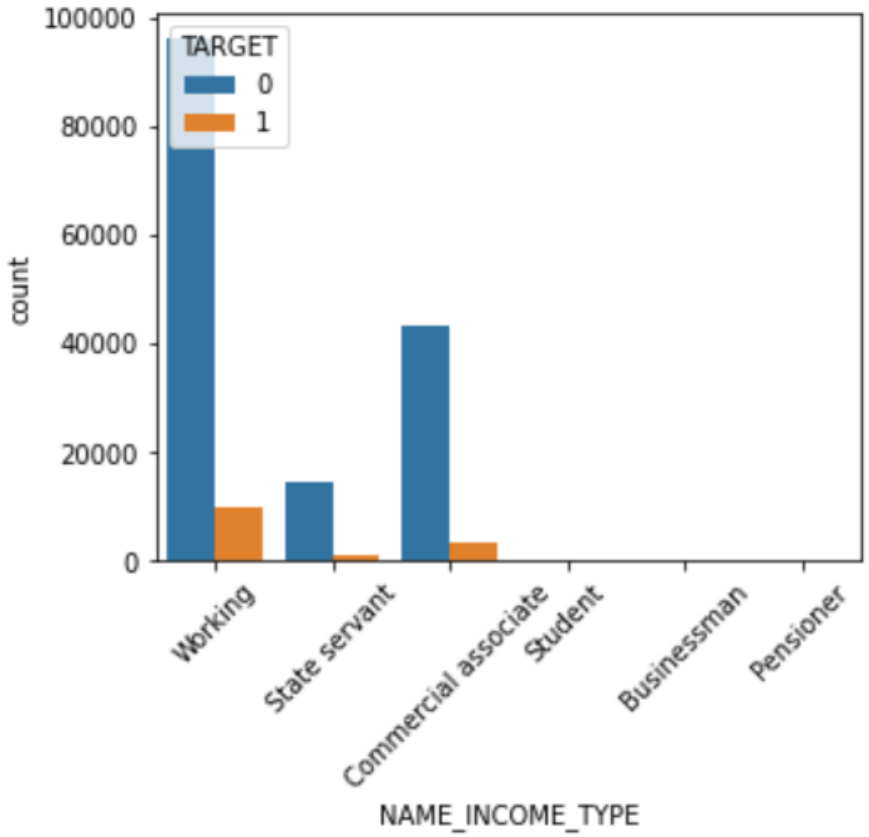
ภาพประกอบ 6 จำนวนข้อมูลเพศของลูกหนี้

ในภาพประกอบ 6 จะใช้ในการบอกเพศของลูกหนี้ จะพบว่าลูกหนี้ส่วนใหญ่เป็นเพศหญิงมากกว่าเพศชาย ซึ่งมีจำนวนข้อมูลเพศหญิงทั้งหมด 103,737 แถว และจำนวนข้อมูลเพศชายทั้งหมด 63,993 แถว



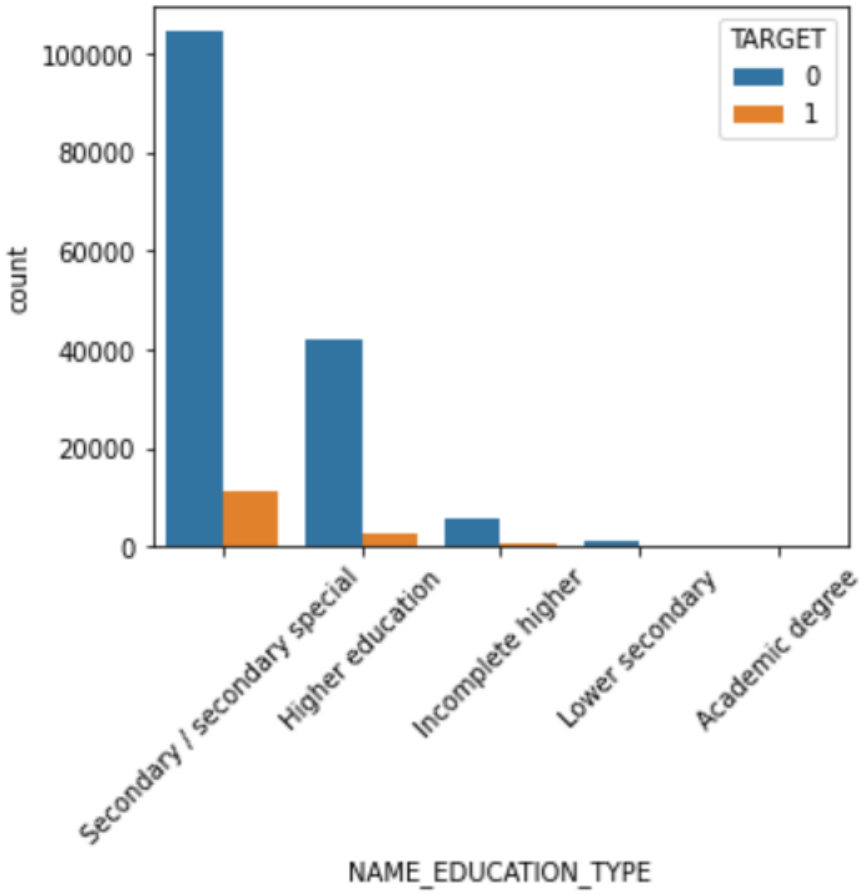
ภาพประกอบ 7 จำนวนข้อมูลประเภทในการกู้ยืมสินเชื่อของลูกหนี้

ในภาพประกอบ 7 จะใช้ในการบอกประเภทในการกู้ยืมสินเชื่อของลูกหนี้ ว่าเป็นเงินสดหรือเงินทุนหมุนเวียน โดยข้อมูลส่วนใหญ่ที่จัดเก็บจะเป็นข้อมูลประเภทเงินสด (Cash Loans) ซึ่งมีจำนวนข้อมูลมากถึง 151,480 แถว



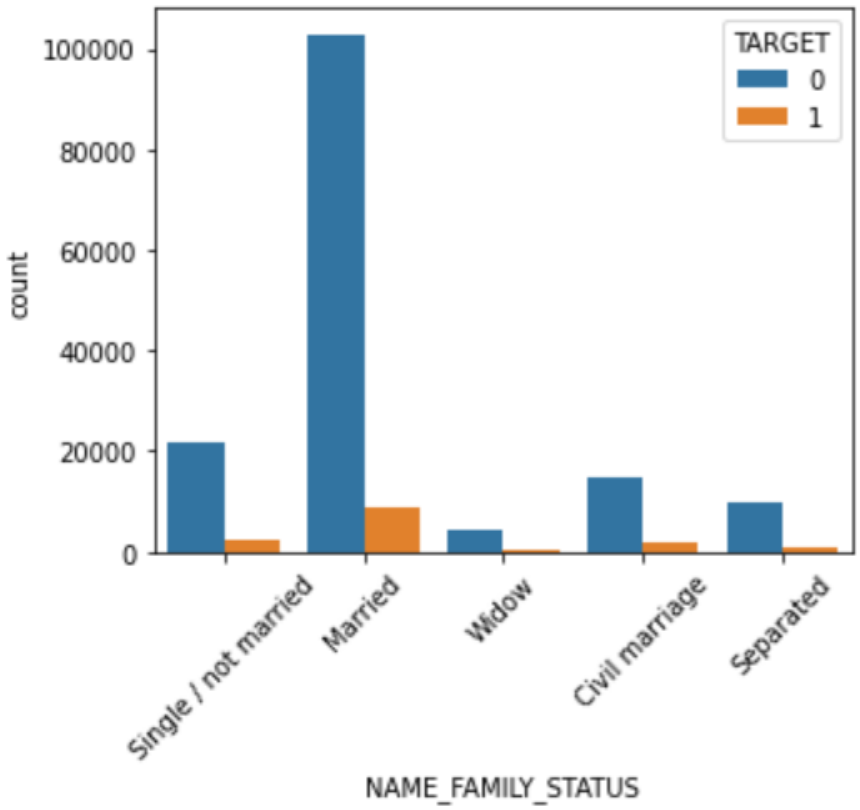
ภาพประกอบ 8 จำนวนข้อมูลประเภทรายได้ของลูกหนี้

ในภาพประกอบ 8 จะใช้ในการบอกประเภทรายได้ของลูกหนี้ โดยลูกหนี้ส่วนใหญ่มีรายได้มาจากการทำงาน ซึ่งมีจำนวนข้อมูลมากถึง 106,018 แถว



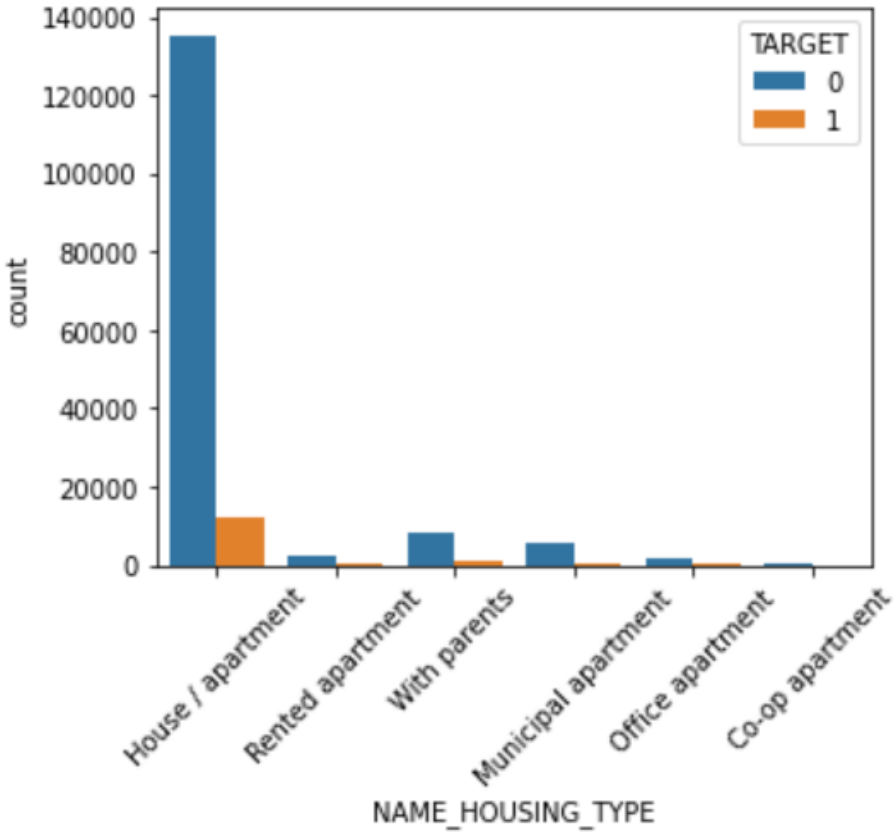
ภาพประกอบ 9 จำนวนข้อมูลระดับการศึกษาสูงสุดของลูกหนี้

ในภาพประกอบ 9 จะใช้ในการบอกระดับการศึกษาสูงสุดของลูกหนี้ โดยลูกหนี้ส่วนใหญ่จบการศึกษาระดับ มัธยมศึกษาตอนต้น (Secondary) ซึ่งมีจำนวนข้อมูลมากถึง 115,721 แถว



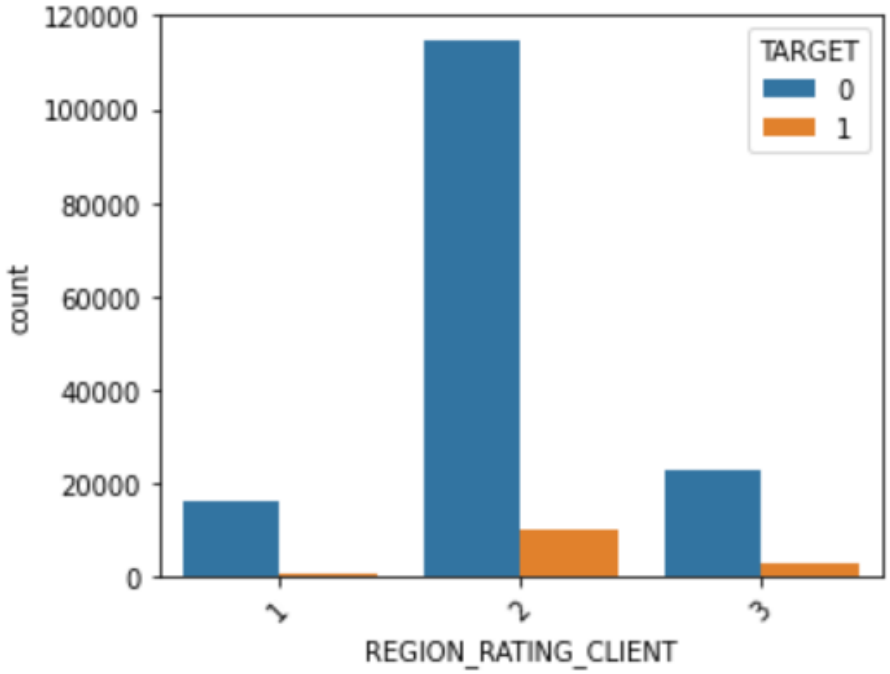
ภาพประกอบ 10 จำนวนข้อมูลสถานภาพทางครอบครัวของลูกหนี้

ในภาพประกอบ 10 จะใช้ในการบอกสถานภาพทางครอบครัวของลูกหนี้ โดยลูกหนี้ส่วนใหญ่มีสถานภาพทางครอบครัว คือ สมรมหรือแต่งงานแล้ว ซึ่งมีจำนวนข้อมูลมากถึง 111,842 แถว



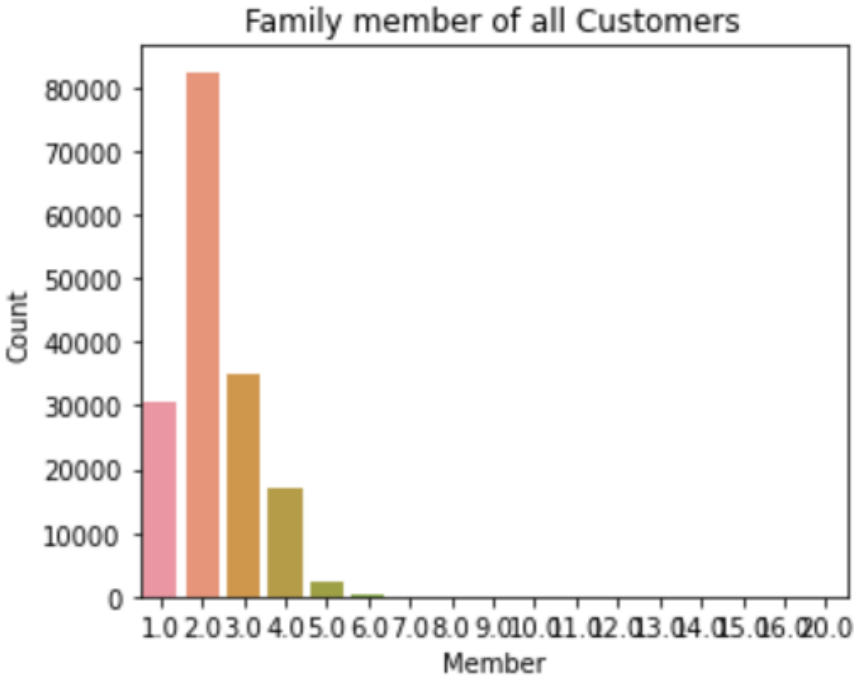
ภาพประกอบ 11 จำนวนข้อมูลที่อยู่อาศัยของลูกหนี้

ในภาพประกอบ 11 จะใช้ในการบอกที่อยู่อาศัยของลูกหนี้ โดยลูกหนี้ส่วนใหญ่พักอาศัยอยู่ที่บ้านหรือ Apartment ซึ่งมีจำนวนข้อมูลมากถึง 147,524 แถว



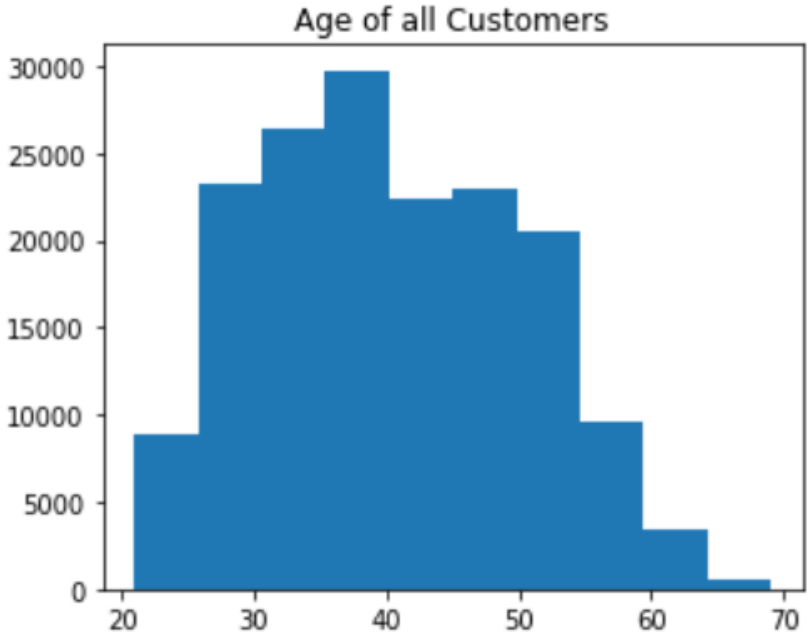
ภาพประกอบ 12 จำนวนข้อมูลระดับ Rating ในแต่ละภูมิภาคที่อยู่อาศัยของลูกหนี้

ในภาพประกอบ 12 จะใช้ในการบอกระดับ Rating ที่ลูกหนี้ให้ ที่เกี่ยวข้องกับคะแนนภูมิภาคที่อยู่อาศัยของลูกหนี้ โดยจะมีค่าให้เลือก คือ 1,2,3 ซึ่งลูกหนี้ส่วนใหญ่ จะให้ Rating เท่ากับ 2



ภาพประกอบ 13 จำนวนสมาชิกในครอบครัวของลูกหนี้

ในภาพประกอบ 13 จะใช้ในการบอกจำนวนสมาชิกในครอบครัวของลูกหนี้ โดยลูกหนี้ส่วนใหญ่จะมีจำนวนสมาชิกในครอบครัวทั้งหมด 2 คน ซึ่งมีจำนวนข้อมูลมากถึง 82,408 แถว

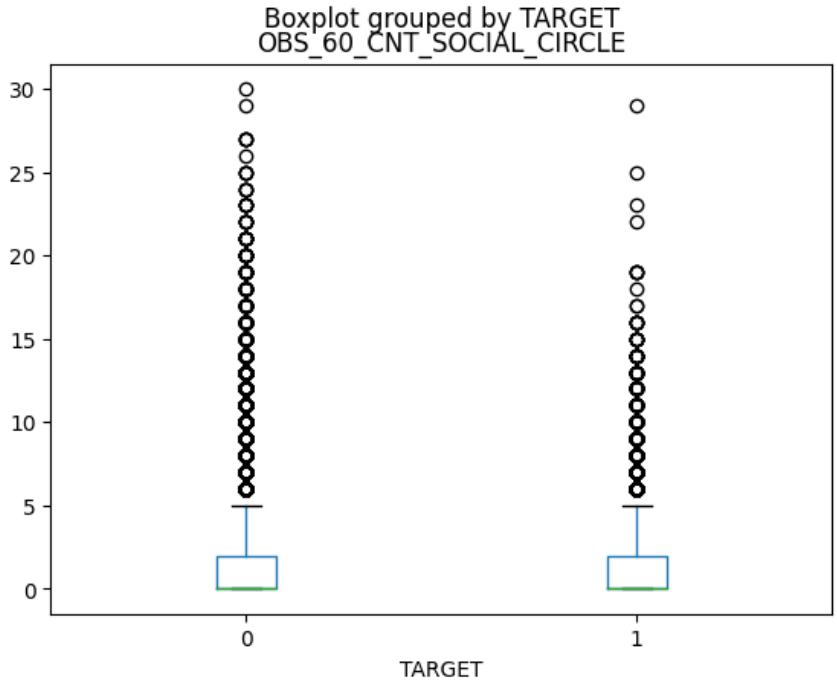


ภาพประกอบ 14 จำนวนข้อมูลอายุของลูกหนี้

ในภาพประกอบ 14 จะใช้ในการบอกอายุของลูกหนี้ โดยลูกหนี้ส่วนใหญ่จะมีอายุอยู่ระหว่าง 25-55 ปี นอกจากนั้นก็ยังมีลูกหนี้รายอื่นๆอีกมากมาย ซึ่งก็มีอายุที่แตกต่างกันออกไป

กราฟ Box Plot เป็นการแสดงผลกราฟที่อยู่ในรูปแบบของการกระจายของข้อมูลในคอลัมน์นั้นๆ โดยการแสดงผลจะแสดงผลของค่า median, Q1, Q3 และข้อมูลที่อยู่นอกเหนือจากขอบเขตของชุดข้อมูลหรือที่เรียกว่า outliers ซึ่ง box plot จะแสดงค่าทางสถิติ ดังต่อไปนี้

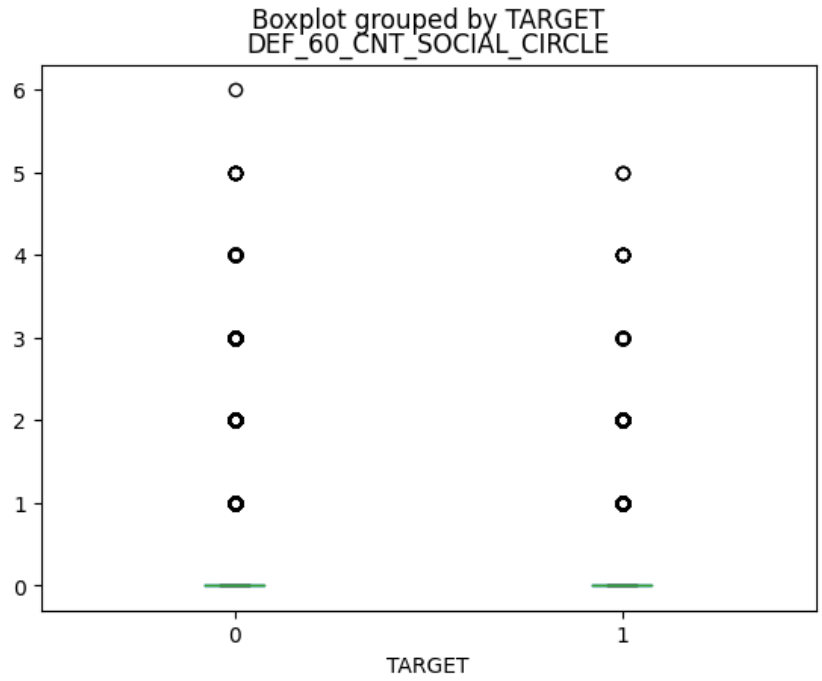
1. median คือค่ากลางที่แบ่งข้อมูลออกเป็นสองส่วนเท่า ๆ กัน อย่างละ 50%
2. Q1 คือค่าของข้อมูลที่แบ่งข้อมูลออกเป็นสัดส่วน 25-75 โดยมีจำนวนของข้อมูลที่น้อยกว่าค่า Q1 อยู่ 25%
3. Q3 คือค่าของข้อมูลที่แบ่งข้อมูลออกเป็นสัดส่วน 75-25 โดยมีจำนวนของข้อมูลที่น้อยกว่าค่า Q3 อยู่ 75%
4. IQR คือค่าของข้อมูลที่อยู่ระหว่าง Q1 และ Q3



ภาพประกอบ 15 แสดงภาพ Boxplot ของคอลัมน์ OBS_60_CNT_SOCIAL_CIRCLE

ในภาพประกอบ 15 จะแสดงภาพ Boxplot ของข้อมูล โดยจะแสดงข้อมูลวันที่ลูกค้าเลยกำหนดในการผิดนัดชำระภายใน 60 วัน ซึ่งข้อมูลส่วนใหญ่จะมีค่าอยู่ระหว่าง 0-3 วัน ซึ่งข้อมูลในคอลัมน์นี้มีการกระจายตัวที่ไม่ปกติ โดยมีการกระจายตัวแบบเบ้ขวาเพราะข้อมูลส่วนใหญ่กระจุกตัวอยู่ใกล้ 0 ซึ่งมีค่า median เท่ากับ 0 วัน และมีค่ามากที่สุดเท่ากับ 5 วัน ถึงแม้ว่าจะมีข้อมูลบางส่วนที่มากกว่า 5 วัน แต่ข้อมูลเหล่านั้นเป็นเพียงข้อมูลที่เป็นค่า outlier เนื่องจากข้อมูลมีความ

เหมือนกันทั้งสองคลาส ระหว่างคลาสที่เป็นลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และคลาสที่เป็นลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร จึงสรุปได้ว่าข้อมูลในคอลัมน์นี้ ไม่ค่อยดีเท่าไหร่ ที่จะนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้อย่างชัดเจน



ภาพประกอบ 16 แสดงภาพ Boxplot ของคอลัมน์ DEF_60_CNT_SOCIAL_CIRCLE

ในภาพประกอบ 16 จะแสดงภาพ Boxplot ของข้อมูล โดยจะแสดงข้อมูลจำนวนรายการที่ลูกค้าผิดนัดชำระภายใน 60 วัน ซึ่งข้อมูลส่วนใหญ่จะมีค่าเท่ากับ 0 วัน ถึงแม้ว่าจะมีข้อมูลบางส่วนที่มีค่ามากกว่า 1 วัน แต่ข้อมูลเหล่านั้นเป็นเพียงข้อมูลที่เป็นค่า outlier เนื่องจากข้อมูลมีความเหมือนกันทั้งสองคลาส ระหว่างคลาสที่เป็นลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และคลาสที่เป็นลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร จึงสรุปได้ว่าข้อมูลในคอลัมน์นี้ ไม่ค่อยดีเท่าไหร่ ที่จะนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้อย่างชัดเจน

ภาพประกอบ 17 ภาพรวมคอลัมน์ทั้งหมดที่จัดเก็บข้อมูลชนิด int64

ในภาพประกอบ 17 ทำการ plot ภาพรวมของคอลัมน์ทั้งหมดที่จัดเก็บข้อมูลชนิด int64 ว่าเมื่อ plot เป็น histogram แล้ว ข้อมูลแต่ละคอลัมน์มีลักษณะของการจัดเก็บข้อมูลเป็นอย่างไร

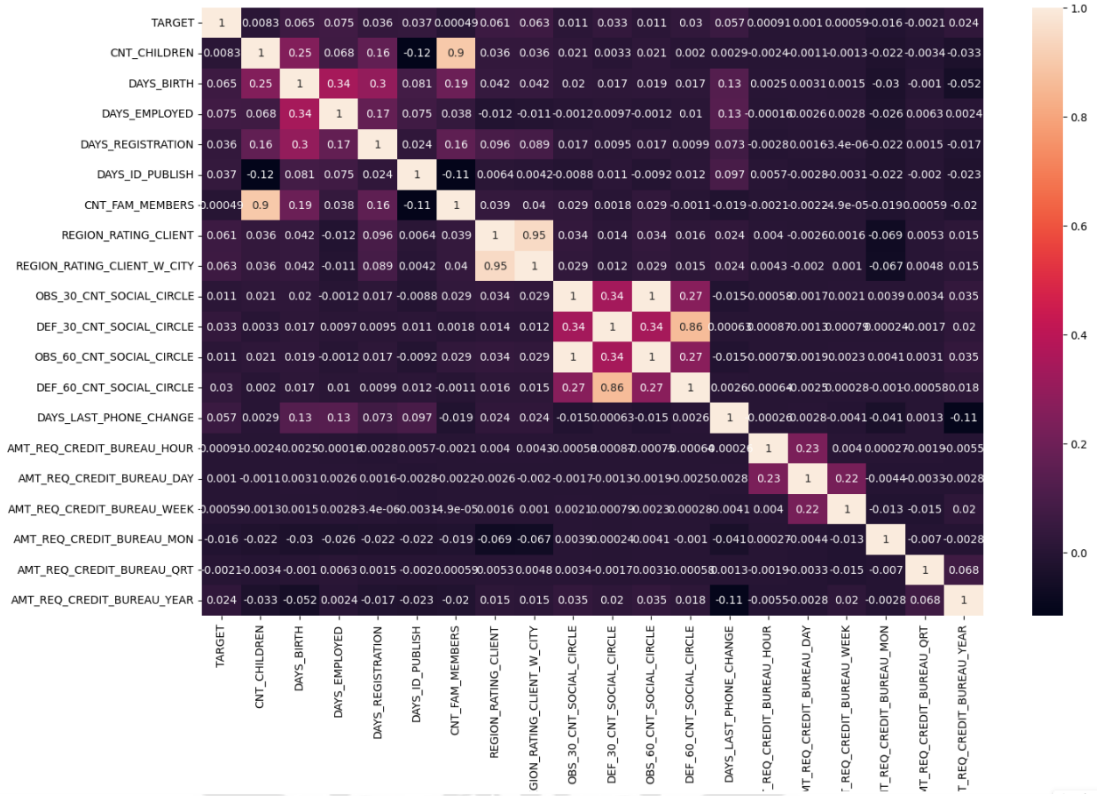
จากนั้นจะทำการสำรวจความสัมพันธ์ของแต่ละคอลัมน์ โดยการใช้เทคนิคที่เรียกว่า correlation ในการแสดงค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ โดยค่า Correlation จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยที่

-1 คือ ตัวแปรทั้งสองตัวมีความสัมพันธ์กัน ในทิศทางตรงกันข้าม

1 คือ ตัวแปรทั้งสองตัวมีความสัมพันธ์กัน ในทิศทางเดียวกัน

0 คือ ตัวแปรทั้งสองตัว ไม่มีความสัมพันธ์ต่อกันเลย

เนื่องจากชุดข้อมูลที่นำมาใช้ในการพิจารณามีจำนวนคอลัมน์ที่มากมหาศาล หากนำคอลัมน์ทั้งหมดมาดูค่า Correlation อาจจะทำให้ดูไม่รู้เรื่อง จึงทำการแสดงค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ โดยการแบ่งแยกตามชนิดของข้อมูล



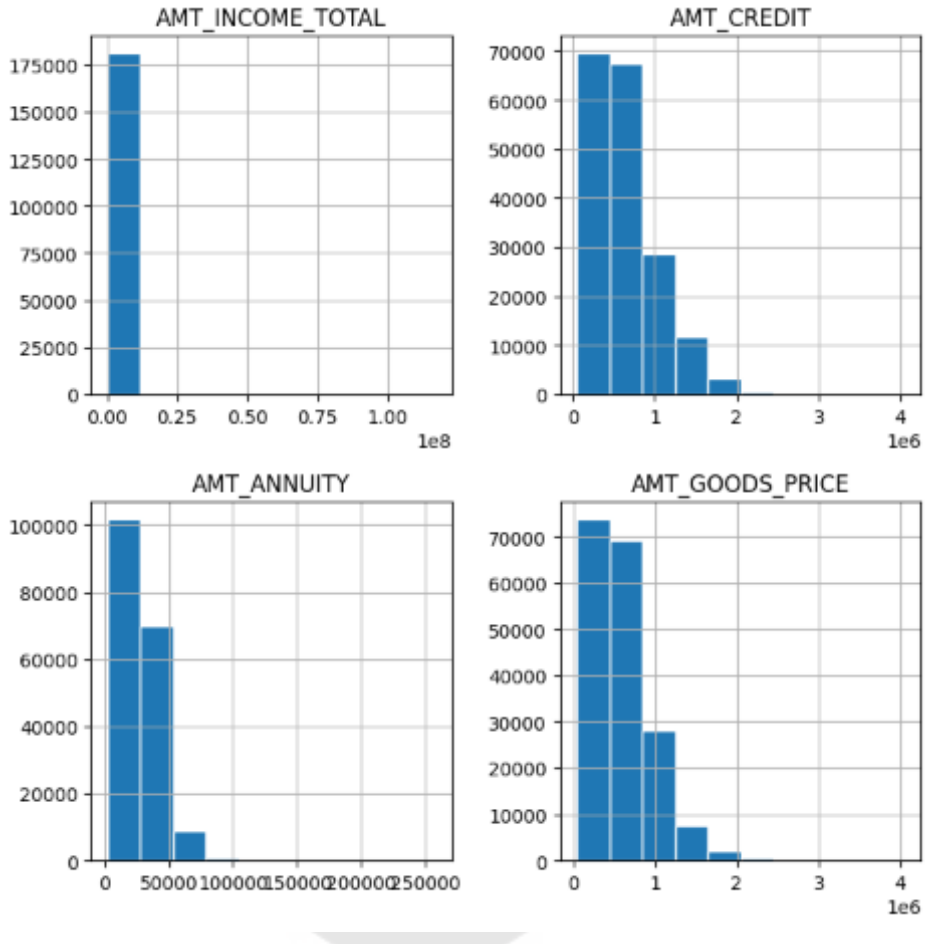
ภาพประกอบ 18 ค่าความสัมพันธ์ของข้อมูลชนิด int64

ในภาพประกอบ 18 จะแสดงค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ ที่จัดเก็บข้อมูลชนิด int64

ซึ่งจากการหาค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ ที่จัดเก็บข้อมูลชนิด int64 จะเห็นได้ว่ามีบางคอลัมน์ที่มีค่าความสัมพันธ์สูงมากๆ โดย

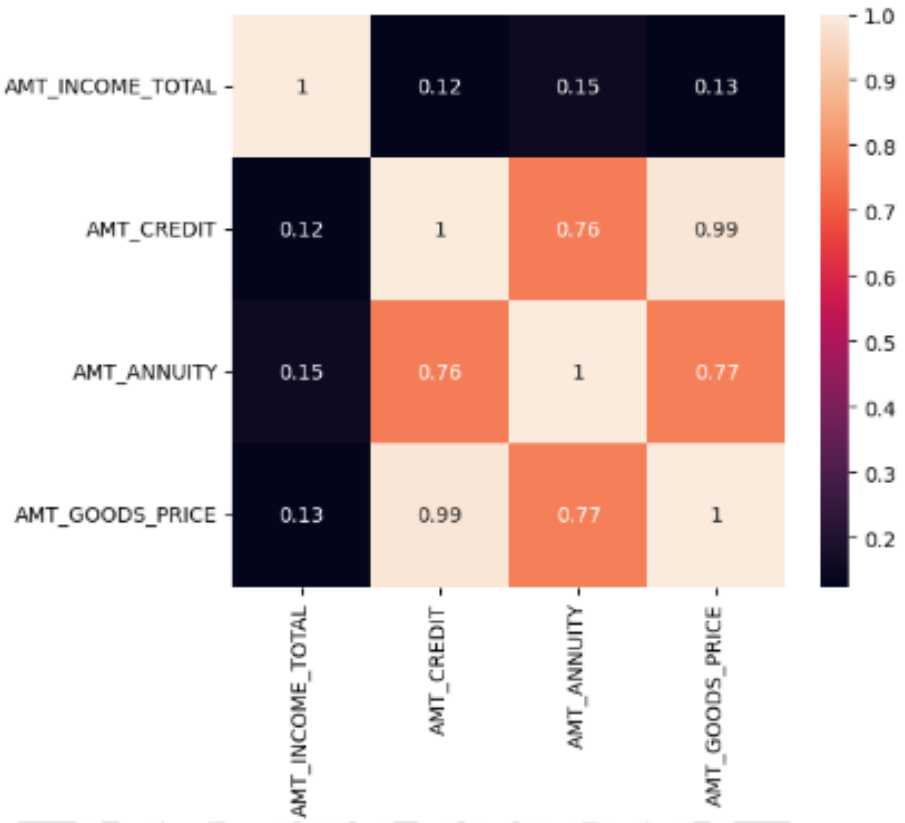
1. OBS_30_CNT_SOCIAL_CIRCLE และ OBS_60_CNT_SOCIAL_CIRCLE มีค่าความสัมพันธ์ เท่ากับ 1
2. DEF_30_CNT_SOCIAL_CIRCLE และ DEF_60_CNT_SOCIAL_CIRCLE มีค่าความสัมพันธ์ เท่ากับ 0.86

3. CNT_CHILDREN และ CNT_FAM_MEMBERS มีค่าความสัมพันธ์ เท่ากับ 0.9
4. REGION_RATING_CLIENT และ REGION_RATING_CLIENT_W_CITY มีค่าความสัมพันธ์ เท่ากับ 0.95



ภาพประกอบ 19 ภาพรวมคอลัมน์ทั้งหมดที่จัดเก็บข้อมูลชนิด float64

ในภาพประกอบ 19 ทำการ plot ภาพรวมของคอลัมน์ทั้งหมดที่จัดเก็บข้อมูลชนิด float64 ว่าเมื่อ plot เป็น histogram แล้ว ข้อมูลแต่ละคอลัมน์มีลักษณะของการจัดเก็บข้อมูลเป็นอย่างไร

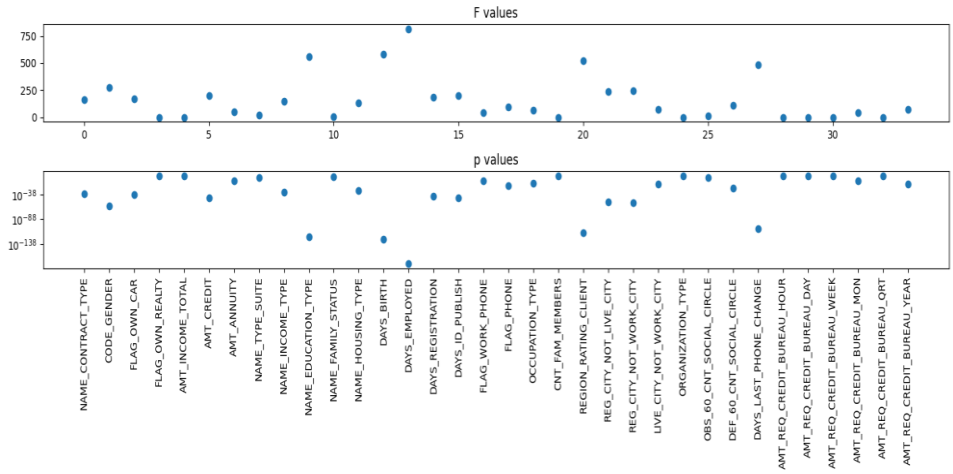


ภาพประกอบ 20 ค่าความสัมพันธ์ของข้อมูลชนิด float64

ในภาพประกอบ 20 จะแสดงค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ ที่จัดเก็บข้อมูลชนิด float64

ซึ่งจากการหาค่าความสัมพันธ์ของข้อมูลในแต่ละคอลัมน์ ที่จัดเก็บข้อมูลชนิด float64 จะเห็นได้ว่ามีบางคอลัมน์ที่มีค่าความสัมพันธ์สูงมากๆ โดย

1. AMT_CREDIT และ AMT_GOODS_PRICE มีค่าความสัมพันธ์ เท่ากับ 0.99



ภาพประกอบ 21 ค่าความสำคัญของแต่ละคุณลักษณะของข้อมูล

ในภาพประกอบ 21 จะแสดงค่า F-value และ P-value จะแสดงคุณลักษณะที่สำคัญ ที่จะนำใช้ในการพัฒนาแบบจำลอง โดยที่ค่า F-value จะแสดงถึงค่าความสัมพันธ์ระหว่างแต่ละคุณลักษณะของข้อมูลกับตัวแปรเป้าหมาย และค่า P-value จะแสดงถึงค่าความน่าจะเป็นของการสังเกตสถิติทดสอบ โดยทั้งค่า F-value และ P-value สามารถนำไปใช้ในการทำ Feature Selection เพื่อช่วยในการเลือกคุณลักษณะที่สำคัญ ที่จะนำไปใช้ในการพัฒนาแบบจำลอง และยังช่วยลดจำนวนคุณลักษณะของข้อมูลที่นำไปใช้ในการพัฒนาแบบจำลอง โดยเลือกเฉพาะคุณลักษณะที่มีความสำคัญต่อตัวแปรเป้าหมาย ซึ่งจะช่วยลดขนาดของแบบจำลอง ลดความซ้ำซ้อนของการพัฒนาแบบจำลอง และยังช่วยเพิ่มประสิทธิภาพของการพัฒนาแบบจำลองให้มีประสิทธิภาพที่ดียิ่งขึ้น

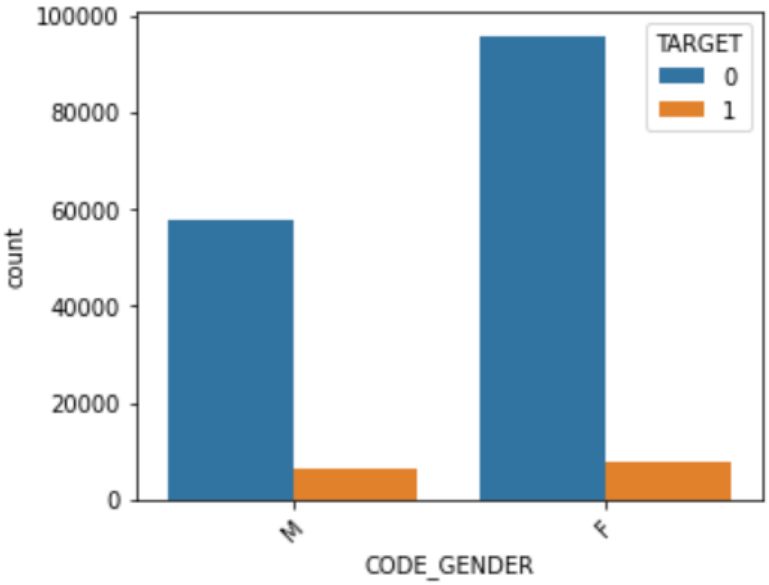
3.3 การเตรียมข้อมูล

1. จากชุดข้อมูลที่เรานำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง มีข้อมูลที่ขาดหายไปเป็นจำนวนมาก ซึ่งจะจัดการกับข้อมูลที่ขาดหายไปก่อนที่จะนำข้อมูลเหล่านี้ไปพัฒนาเป็นแบบจำลอง เพื่อให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น โดยจะจัดการกับข้อมูลที่ขาดหายไปทั้งในแนวระดับแถวและในแนวระดับคอลัมน์ โดยจะกำจัดคอลัมน์ที่มีข้อมูลที่ขาดหายไป ที่มากกว่า 45% ซึ่งเราจะไม่นำคอลัมน์เหล่านั้น มาใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลอง จากในตอนแรกที่มีคอลัมน์ที่นำมาใช้ในการวิเคราะห์ข้อมูลทั้งหมด 122 คอลัมน์ แล้วทำการลบคอลัมน์ที่มีจำนวนข้อมูลที่ขาดหายไป ที่มากกว่า 45% จะทำให้เหลือคอลัมน์ที่นำมาใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลองทั้งหมด 73 คอลัมน์

2. ลบคอลัมน์ที่เป็น FLAG_DOCUMENT ออกทั้งหมด เนื่องจากเป็นคอลัมน์ที่ใช้ในการเก็บข้อมูลที่เกี่ยวข้องกับการเก็บเอกสาร ซึ่งไม่น่าจะมีประโยชน์ที่จะนำคอลัมน์เหล่านี้ไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลอง โดยในขั้นตอนนี้ ได้มีการลบคอลัมน์ที่เกี่ยวข้องกับการเก็บเอกสาร จะทำให้เหลือคอลัมน์ที่จะนำไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลองทั้งหมด 52 คอลัมน์

3. ลบข้อมูลบางแถวในชุดข้อมูลออก เพราะข้อมูลในแถวนั้นส่วนใหญ่ จะเก็บข้อมูลที่เป็นค่าว่าง (NaN) เนื่องจากมีข้อมูลจำนวน 307,511 แถว จึงสามารถลบข้อมูลส่วนน้อย ที่เก็บข้อมูลที่เป็นค่าว่างออกได้ แล้วเมื่อทำการทำความสะอาดข้อมูลเรียบร้อยแล้ว จะพบว่าข้อมูลที่ถูกจัดเก็บอยู่ในแต่ละคอลัมน์ ไม่มีคอลัมน์ไหนที่มีการจัดเก็บข้อมูลที่เป็นค่าว่างอีกแล้ว

4. ลบข้อมูลบางแถวที่เก็บข้อมูลเพศเป็นค่า “XNA” ออก เนื่องจากเมื่อทำการนับจำนวนข้อมูลในคอลัมน์เพศแล้ว พบว่ามีข้อมูลเพศทั้งหมด 3 ค่า คือ M, F, XNA ซึ่งข้อมูลจริงๆที่ถูกต้องที่ควรจัดเก็บ ควรมีแค่ 2 ค่า คือ เพศชายและเพศหญิง (M, F)



ภาพประกอบ 22 จำนวนข้อมูลเพศของลูกหนี้

ในภาพประกอบ 22 จะแสดงข้อมูลเพศของลูกหนี้ จะเหลือเฉพาะข้อมูลที่เป็นเพศชายและเพศหญิงเท่านั้น

5. ลบบางคอลัมน์ที่เก็บข้อมูลส่วนใหญ่ที่เป็นค่า “1” ออก เพราะคอลัมน์เหล่านี้ เมื่อนำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง คอลัมน์เหล่านี้จะไม่มีผลต่อการทำนาย เนื่องจากไม่มีความแตกต่าง ที่จะสามารถนำไปใช้ในการแบ่งแยกคลาสได้

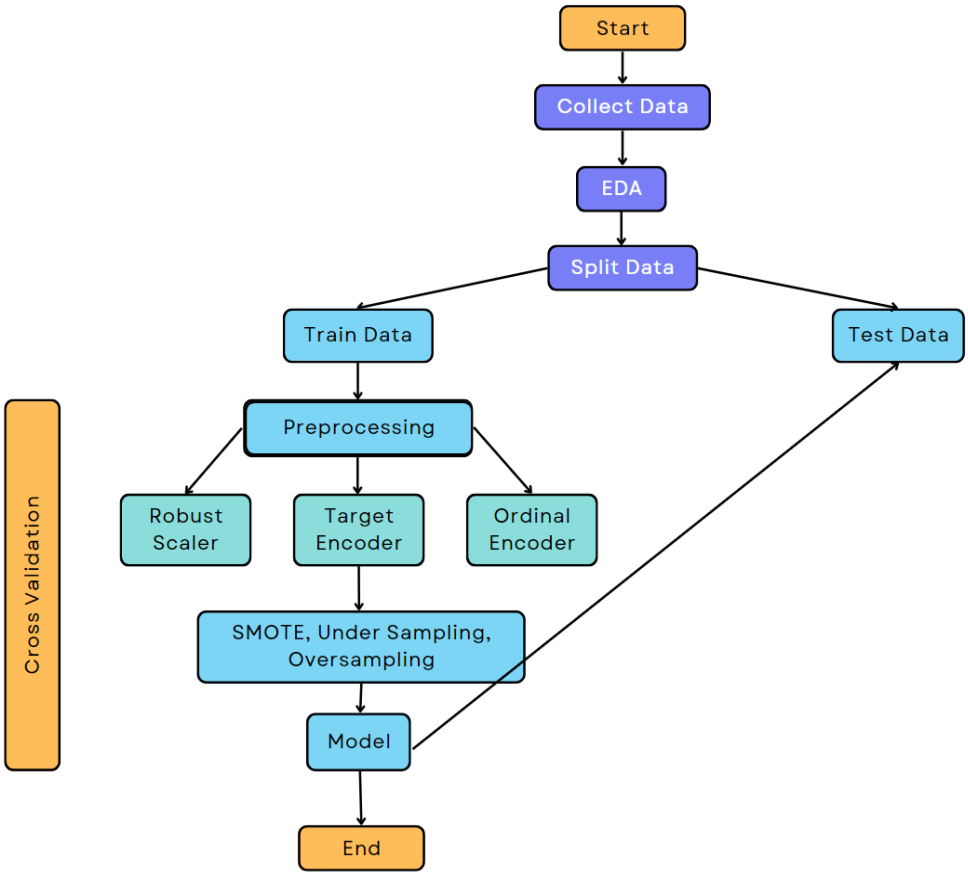
6. ทำการแปลงชนิดของข้อมูล เนื่องจากเมื่อเราทำการดูชนิดของข้อมูลแล้ว ยังมีบางคอลัมน์ ที่เก็บชนิดของข้อมูลไม่ตรงกับลักษณะของข้อมูลที่ถูกจัดเก็บอยู่จริงๆ จึงได้มีการแปลงชนิดของข้อมูล ให้มีชนิดของข้อมูลที่ถูกต้อง ก่อนนำไปพัฒนาแบบจำลอง เพื่อทำให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น

7. ลบบางคอลัมน์ ที่มีค่าความสัมพันธ์สูงๆออก และใช้เพียงคอลัมน์เดียวในการเรียนรู้ข้อมูลเพื่อพัฒนาแบบจำลอง เพื่อลดความซับซ้อนของการเรียนรู้ข้อมูลเพื่อพัฒนาแบบจำลอง และลดเวลาในการเรียนรู้ข้อมูลเพื่อพัฒนาแบบจำลอง

```
Data columns (total 34 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   NAME_CONTRACT_TYPE           36234 non-null  object
 1   CODE_GENDER                  36234 non-null  object
 2   FLAG_OWN_CAR                 36234 non-null  object
 3   FLAG_OWN_REALTY              36234 non-null  object
 4   AMT_INCOME_TOTAL             36234 non-null  float64
 5   AMT_CREDIT                   36234 non-null  float64
 6   AMT_ANNUITY                  36234 non-null  float64
 7   NAME_TYPE_SUITE              36234 non-null  object
 8   NAME_INCOME_TYPE             36234 non-null  object
 9   NAME_EDUCATION_TYPE          36234 non-null  object
 10  NAME_FAMILY_STATUS           36234 non-null  object
 11  NAME_HOUSING_TYPE            36234 non-null  object
 12  DAYS_BIRTH                   36234 non-null  int64
 13  DAYS_EMPLOYED                36234 non-null  int64
 14  DAYS_REGISTRATION            36234 non-null  int64
 15  DAYS_ID_PUBLISH              36234 non-null  int64
 16  FLAG_WORK_PHONE              36234 non-null  object
 17  FLAG_PHONE                   36234 non-null  object
 18  OCCUPATION_TYPE              36234 non-null  object
 19  CNT_FAM_MEMBERS              36234 non-null  int64
 20  REGION_RATING_CLIENT         36234 non-null  int64
 21  REG_CITY_NOT_LIVE_CITY       36234 non-null  object
 22  REG_CITY_NOT_WORK_CITY       36234 non-null  object
 23  LIVE_CITY_NOT_WORK_CITY      36234 non-null  object
 24  ORGANIZATION_TYPE            36234 non-null  object
 25  OBS_60_CNT_SOCIAL_CIRCLE     36234 non-null  int64
 26  DEF_60_CNT_SOCIAL_CIRCLE     36234 non-null  int64
 27  DAYS_LAST_PHONE_CHANGE       36234 non-null  int64
 28  AMT_REQ_CREDIT_BUREAU_HOUR   36234 non-null  int64
 29  AMT_REQ_CREDIT_BUREAU_DAY    36234 non-null  int64
 30  AMT_REQ_CREDIT_BUREAU_WEEK   36234 non-null  int64
 31  AMT_REQ_CREDIT_BUREAU_MON    36234 non-null  int64
 32  AMT_REQ_CREDIT_BUREAU_QRT    36234 non-null  int64
 33  AMT_REQ_CREDIT_BUREAU_YEAR   36234 non-null  int64
dtypes: float64(3), int64(15), object(16)
```

ภาพประกอบ 23 คอลัมน์ทั้งหมดที่นำไปใช้ในการพัฒนาแบบจำลอง

3.4 การพัฒนาแบบจำลอง



ภาพประกอบ 24 กระบวนการของการพัฒนาแบบจำลอง

อธิบายถึงกระบวนการของการพัฒนาแบบจำลอง โดยมีการเก็บข้อมูลจากแหล่งข้อมูลสาธารณะ และทำการสำรวจข้อมูล เช่น ดูค่าทางสถิติของข้อมูล ดูจำนวนของข้อมูลที่ถูกเก็บอยู่ในแต่ละคอลัมน์ และดูความสัมพันธ์ระหว่างข้อมูลที่ถูกเก็บอยู่ในแต่ละคอลัมน์ แล้วทำการแบ่งข้อมูลออกเป็น Train 80% และ Test 20% และทำการเตรียมข้อมูล (Preprocessing) โดยมีการแปลงข้อมูลให้อยู่ใน 3 รูปแบบ คือ Robust Scaler, Target Encoder และ Ordinal Encoder แล้วนำไปปรับความไม่สมดุลของข้อมูลด้วยเทคนิคต่างๆ เช่น SMOTE, Under Sampling และ Oversampling หลังจากนั้นนำข้อมูล Train ที่ได้ไปทำการ Fit เพื่อพัฒนาแบบจำลอง โดยทำการตรวจสอบกับข้อมูล Train ในทุกๆส่วน โดยการทำ Cross validation เพื่อปรับปรุงประสิทธิภาพของแบบจำลองให้มีความถูกต้องแม่นยำมากยิ่งขึ้น หลังจากนั้นนำข้อมูลที่ถูกแบ่งแยกไว้ในส่วนของ Test ไปทำการทดสอบกับแบบจำลองที่ถูกสร้างขึ้น โดยทำการ Predict กับแบบจำลอง แล้วนำผลลัพธ์ที่ได้มาทำการประเมินประสิทธิภาพของแบบจำลอง (Evaluation model)

การพัฒนาแบบจำลอง ประกอบด้วยขั้นตอนหลักๆ ดังต่อไปนี้

3.4.1 การแบ่ง training data และ test data (Data splitting)

โดยจะมีการแบ่งชุดข้อมูลออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลที่ใช้ในการฝึกฝน (Train Data) เพื่อใช้ในการพัฒนาแบบจำลอง และชุดข้อมูลที่ใช้ในการทดสอบ (Test Data) เพื่อใช้ในการทดสอบแบบจำลอง โดยจะแบ่งข้อมูลออกเป็น 2 ส่วน ด้วยอัตราส่วนละ 80 : 20 ซึ่งจะได้ข้อมูลเท่ากับ 36,234 : 9,059

3.4.2 การทำ Column Transformer

1. RobustScaler

การปรับขนาดของข้อมูลตัวเลขก่อนนำไปพัฒนาแบบจำลอง ทำให้ขนาดของข้อมูลตัวเลข อยู่ในรูปแบบของค่าที่เป็นมาตรฐาน ซึ่งจะมีค่าอยู่ระหว่าง 0-1

2. OrdinalEncoder

การแปลงข้อมูลที่เป็นตัวเลขที่ไม่มีลำดับ (Nominal number) ให้อยู่ในรูปแบบของตัวเลขที่มีลำดับ (Ordinal number)

3. TargetEncoder

การแปลงข้อมูลที่ใช้ในการเข้ารหัสตัวแปรประเภทหมวดหมู่ในชุดข้อมูลโดยการแทนที่แต่ละหมวดหมู่ด้วยค่าเฉลี่ยของตัวแปรเป้าหมายสำหรับหมวดหมู่นั้นๆ

3.4.3 การปรับความไม่สมดุลของชุดข้อมูล (Imbalance Data)

มีการปรับชุดข้อมูลเพื่อให้ชุดของข้อมูลเกิดความความสมดุลกัน เนื่องจากชุดข้อมูลมีความไม่สมดุลกันของชุดข้อมูลสูงมาก คือมีจำนวนชุดข้อมูลของลูกหนี้ปกติคือลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคารมากกว่าจำนวนชุดข้อมูลของลูกหนี้ที่ไม่ปกติคือลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร จึงต้องมีการปรับความสมดุลของชุดข้อมูล ด้วยวิธีการเพิ่มจำนวนของชุดข้อมูลประเภทของลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร ให้มีความใกล้เคียงกันหรือเท่ากันกับชุดข้อมูลของลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร

1. Synthetic Minority Over-Sampling Technique (SMOTE)

การเพิ่มจำนวนชุดข้อมูลของกลุ่มน้อย (minor class) ให้มีจำนวนใกล้เคียงกับชุดข้อมูลกลุ่มหลัก (major class) โดยการเพิ่มจำนวนข้อมูลจะไม่ได้เพิ่มจากการสุ่ม แต่เกิดจากการสร้างข้อมูลขึ้นมาใหม่โดยใช้ข้อมูลเดิมที่มีอยู่แล้ว

2. Oversampling

การสุ่มเพิ่มจำนวนชุดข้อมูลของกลุ่มน้อย (minor class) ให้มีจำนวนมากขึ้นจนใกล้เคียงกับชุดข้อมูลของกลุ่มหลัก (major class)

3. Under Sampling

การสุ่มลดจำนวนชุดข้อมูลของกลุ่มหลัก (Major class) ให้ลดน้อยลงพอๆ กันกับชุดข้อมูลของกลุ่มน้อย (Minor class)

3.4.4 การพัฒนาแบบจำลอง

มีการใช้ Cross Validation ซึ่งเป็นเทคนิคที่ใช้ในการค้นหาค่า Hyperparameter ด้วยการลองใช้ค่าพารามิเตอร์ต่างๆ ที่ได้มีการกำหนดไว้ พารามิเตอร์แต่ละตัวจะถูกนำมาใช้ในการพัฒนาแบบจำลอง และประเมินประสิทธิภาพหรือหาค่าความแม่นยำของแต่ละแบบจำลอง แบบจำลองไหนที่ให้ค่าความแม่นยำสูงสุดจะถือว่าแบบจำลองนั้นดีที่สุด โดยการใช้เทคนิคที่เรียกว่า “GridSearchCV” เพื่อนำมาใช้ในการ Tunning Parameter เพื่อหา Hyperparameter ที่ดีที่สุด เพื่อนำไปใช้ในการพัฒนาแบบจำลองทำให้ได้แบบจำลองที่ดีที่สุด

ซึ่งในงานวิจัยนี้ อัลกอริทึมที่จะถูกนำมาใช้ในการพัฒนาแบบจำลอง ได้แก่

1. Logistic Regression
2. XGBoostClassifier
3. K-nearest Neighbors
4. Random Forest
5. Support Vector Classifier (SVC)
6. Gradient Boosting

3.5 การประเมินประสิทธิภาพของแบบจำลอง

การวัดประสิทธิภาพของการพัฒนาแบบจำลอง เพื่อประเมินประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้ Confusion matrix โดยดูที่ค่า Accuracy , Precision, Recall และ F1-Score ซึ่ง Confusion matrix จะมีลักษณะเป็นตาราง หากข้อมูลที่ต้องการจำแนกมี 2 ประเภท คือ ทายถูก (Positive) และ ทายผิด (Negative) ตาราง Confusion Matrix จะมีลักษณะตามตาราง

ตาราง 2 Confusion matrix

	ค่าจริง (Actual)	
ค่าการทำนาย (Predict)	True positive (TP)	False positive (FP)
	False negative (FN)	True negative (TN)

จากรูป เมื่อมีการทำนาย 2 ประเภท ผลการทำนายทั้งหมดที่เป็นไปได้ จะมีทั้งหมด 4 ค่า ดังต่อไปนี้

1. True positive (TP) คือ การทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระ ***ถูก***
2. True negative (TN) คือ การทำนายลูกหนี้ปกติที่ไม่ได้มีการผิดนัดชำระ ***ถูก***
3. False positive (FP) คือ การทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระ ***ผิด***
4. False negative (FN) คือ การทำนายลูกหนี้ปกติที่ไม่ได้มีการผิดนัดชำระ ***ผิด***

ซึ่งสามารถนำค่าเหล่านี้ มาคำนวณหาประสิทธิภาพได้ ดังต่อไปนี้

1. ค่าความไว (Recall)
   คือ ค่าความถูกต้องของการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าจริง เทียบกับ จำนวนครั้งของเหตุการณ์ ทั้งการทำนายและการเกิดขึ้นจริง ว่าเป็นจริง
   สูตรในการคำนวณ คือ TP / (TP+FN)
2. ค่าความถูกต้อง (Accuracy)
   คือ ค่าความถูกต้องและความแม่นยำของแบบจำลอง
   สูตรในการคำนวณ คือ (TP+TN) / (TP+FP+TN+FN)
3. ค่าความแม่นยำ (Precision)

คือ การเปรียบเทียบการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าจริง แล้วเกิดขึ้นจริง (TP) เทียบกับการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระว่าจริง แต่สิ่งที่เกิดขึ้นไม่จริง (FP)

สูตรในการคำนวณ คือ TP / (TP+FP)

4. F1-Score

เป็นค่าเฉลี่ยระหว่างค่า precision และ recall เพื่อใช้ในการวัดความสามารถของแบบจำลอง

สูตรในการคำนวณ คือ 2 x (Precision x Recall) / (Precision + Recall)

บทที่ 4
ผลการดำเนินงานวิจัย

การวิจัยนี้ เป็นงานวิจัยในการศึกษาการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยอาศัยการเรียนรู้ของเครื่องมือที่ช่วยในการตัดสินใจ ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามขบวนการและขั้นตอนต่างๆ จนกระทั่งประเมินประสิทธิภาพของแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยมีขั้นตอนต่างๆ ดังนี้

1. ผลลัพธ์ของการเตรียมข้อมูล
2. ผลลัพธ์ของการพัฒนาแบบจำลอง

4.1 ผลลัพธ์ของการเตรียมข้อมูล

เมื่อทำการทำความสะอาดข้อมูลเรียบร้อยแล้ว จะเหลือชุดข้อมูลที่นำไปใช้ในการวิเคราะห์ข้อมูลเพื่อพัฒนาแบบจำลองทั้งหมด 181,391 แถว และคอลัมน์ทั้งหมด 34 คอลัมน์

ซึ่งจะทำการพัฒนาแบบจำลอง โดยการใช้ Machine Learning Algorithms โดยอาศัยการเรียนรู้ของเครื่อง ซึ่งเป็นเครื่องมือสำหรับการพัฒนาแบบจำลอง ในการเรียนรู้แบบผู้สอน ซึ่งจะทำงานแบบการแบ่งแยกประเภท เพื่อนำไปใช้ในการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร

4.2 ผลลัพธ์ของการพัฒนาแบบจำลอง

ผลลัพธ์ของการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการใช้เทคนิคต่างๆ เมื่อมีการเตรียมข้อมูลเรียบร้อยแล้ว เราจะนำข้อมูลเหล่านั้นมาใช้ในการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร การทดลองของงานวิจัยนี้ ได้มีการพัฒนาแบบจำลอง โดยการใช้อัลกอริทึม Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting ซึ่งได้มีการเปรียบเทียบประสิทธิภาพของแบบจำลอง โดยดูที่ค่า Accuracy, Precision, Recall และ F1-Score โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก เนื่องจากเป็นตัววัดความสามารถของแบบจำลอง และทำการแสดงผลลัพธ์ประสิทธิภาพของการพัฒนาแบบจำลองต่างๆ ดังนี้

4.2.1 Logistic Regression

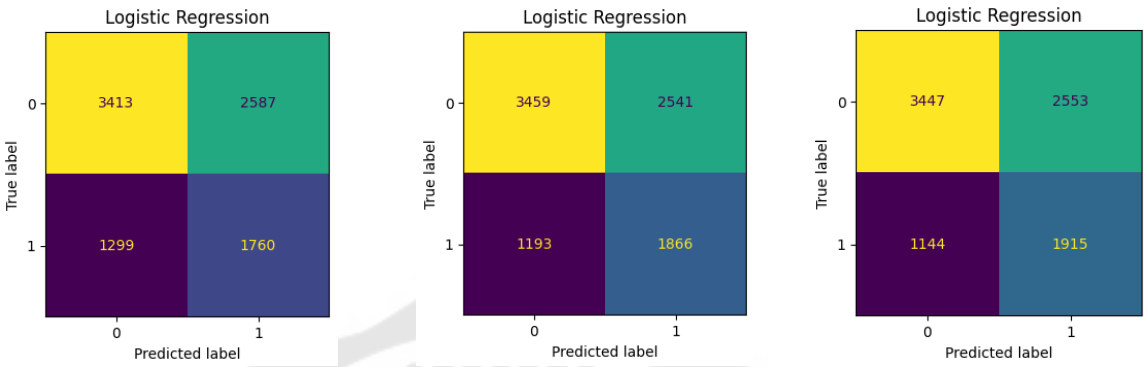
ผู้วิจัยทำการพัฒนาแบบจำลอง โดยการใช้เทคนิคการวิเคราะห์การถดถอย มาใช้ในการพัฒนาแบบจำลอง ซึ่งเป็นเทคนิคการวิเคราะห์ตัวแปรเพื่อประมาณค่าหรือทำนายเหตุการณ์ที่สนใจว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร หรือไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร

โดยได้มีการกำหนดค่าตัวแปร C เพื่อช่วยปรับความซับซ้อนของการพัฒนาแบบจำลอง เนื่องจากถ้าค่า C มีค่าน้อยจะทำให้เกิด regularization ที่มากขึ้นและอาจทำให้แบบจำลองเกิดการ underfitting ได้ แต่ถ้าค่า C มีค่ามากจะทำให้เกิด regularization ที่น้อยลงและอาจทำให้แบบจำลองเกิดการ overfitting ได้ โดยได้มีการกำหนดค่า Hyperparameter C ให้มีค่าเท่ากับ [0.01, 0.1, 1, 10, 100, 1000] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 10

ตาราง 3 ผลลัพธ์การเปรียบเทียบประสิทธิภาพค่า Hyperparameter C ของการพัฒนาแบบจำลอง Logistic Regression

Hyperparameter C	Accuracy
0.01	0.58
0.1	0.58
1	0.57
10	0.59
100	0.57
1000	0.56

ทำการเปรียบเทียบกันระหว่างการเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้เทคนิค Oversampling, Under sampling และ SMOTE



ภาพประกอบ 25 Confusion Matrix ของการพัฒนาแบบจำลอง Logistic Regression

ตาราง 4 ผลลัพธ์ของการพัฒนาแบบจำลอง Logistic Regression

Logistic Regression	Oversampling	Under sampling	SMOTE
Accuracy	0.57	0.59	0.59
Precision	0.40	0.42	0.43
Recall	0.58	0.61	0.63
F1-Score	0.48	0.50	0.51

จากการเปรียบเทียบประสิทธิภาพของทั้ง 3 แบบจำลอง จากวิธีการปรับความไม่สมดุลของข้อมูล ด้วยวิธีการ Oversampling, Under sampling และ SMOTE โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก ซึ่งจะสรุปได้ว่า วิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Synthetic Minority Oversampling Technique (SMOTE) ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่า F1-Score เท่ากับ 0.51 โดยจะทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,915 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,447 คน

1. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,760 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,413 คน
2. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling

สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,866 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,459 คน

3. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ SMOTE สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,915 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,447 คน

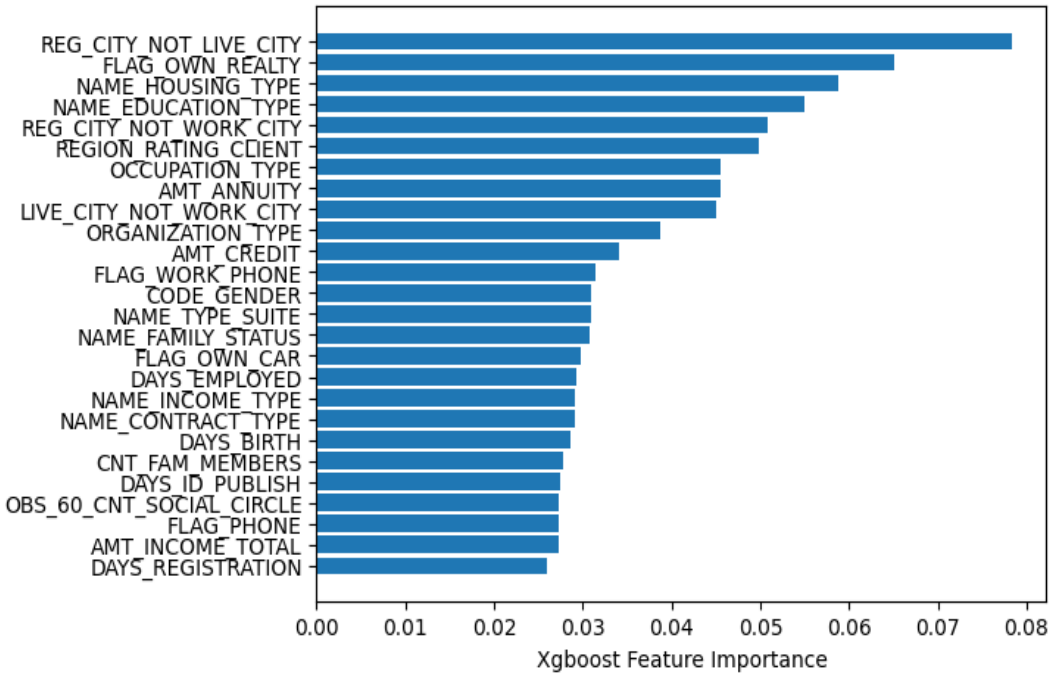
4.2.2 XGBoostClassifier

ผู้วิจัยทำการพัฒนาแบบจำลอง โดยการใช้เทคนิคต้นไม้ตัดสินใจหลายๆต้นมาช่วยกันในการตัดสินใจ ซึ่งการทำงานของแบบจำลอง XGBoost คือการสร้างต้นไม้ตัดสินใจหลายๆต้น โดยที่ต้นไม้ตัดสินใจแต่ละต้นจะถูกสร้างขึ้นมาจากการปรับปรุงประสิทธิภาพของแบบจำลองที่ถูกสร้างขึ้นก่อนหน้า แล้วจะพยายามแก้ไขความผิดพลาด (error) ของแบบจำลองที่ถูกสร้างขึ้นก่อนหน้า ให้แบบจำลองที่ถูกสร้างขึ้นในครั้งถัดๆไป มีความถูกต้องแม่นยำในการทำนายมากยิ่งขึ้นเรื่อย ๆ เมื่อมีการเรียนรู้ของต้นไม้ตัดสินใจต่อเนื่องกันจนมีความลึกมากพอแบบจำลองจะหยุดการเรียนรู้ก็ต่อเมื่อไม่เหลือค่าความผิดพลาดจากต้นไม้ตัดสินใจก่อนหน้าให้เรียนรู้แล้ว ซึ่งผลลัพธ์สุดท้ายของการพัฒนาแบบจำลอง จะมีการนำทุกแบบจำลองมารวมกัน เป็น 1 classifier แต่มีการกำหนดให้น้ำหนักของแต่ละแบบจำลองไม่เท่ากัน ซึ่งจะมีการกำหนดให้น้ำหนักของแบบจำลองตัวแรก มีน้ำหนักที่มากที่สุด และแบบจำลองตัวถัดๆไป มีน้ำหนักที่มีค่าน้อยลง เพื่อลดการเกิด Overfitting ของการพัฒนาแบบจำลอง ซึ่งการทำงานของ XGBoost จะมีลักษณะการทำงานในลักษณะที่เป็นลำดับแบบ sequential คือต้องรอให้แบบจำลองต้นแรกสร้างเสร็จเรียบร้อยก่อน จึงจะนำไปพัฒนาเป็นแบบจำลองต้นที่สองได้ ซึ่งในตอนนี้ การพัฒนาแบบจำลองโดยการใช้เทคนิค XGBoost ถือว่าเป็นเทคนิคในการพัฒนาแบบจำลองที่ได้รับความนิยมมากที่สุด และเป็นเทคนิคในการพัฒนาแบบจำลองที่ดีที่สุด

XGBoost สามารถนำไปใช้ในการค้นหา feature importance ซึ่งใช้ในการบอกว่า feature ไหนเป็นตัวทำนายที่สำคัญที่สุดสำหรับการพัฒนาแบบจำลอง ซึ่งวิธีการคำนวณ feature importance ของ XGBoost จะประกอบไปด้วยค่าต่างๆ ดังต่อไปนี้

1. Gain คือการวัดการมีส่วนร่วมของ feature ในการเพิ่มค่าความแม่นยำของการพัฒนาแบบจำลอง
2. Weight คือการวัดจำนวนครั้งที่ feature นั้นถูกนำมาใช้ในการแบ่งข้อมูลของการพัฒนาแบบจำลอง

3. Coverage คือการวัดเปอร์เซ็นต์ของข้อมูลที่ผ่าน feature นั้น ๆ ในการแบ่งข้อมูลของการพัฒนาแบบจำลอง
4. Frequency คือการวัดเปอร์เซ็นต์ของการใช้งาน feature นั้น ทั้งในการแบ่งข้อมูลและในการเป็น node สุดท้ายของต้นไม้ตัดสินใจ (Leaf node)



ภาพประกอบ 26 จำนวน feature importance ของการพัฒนาแบบจำลอง XGBoostClassifier

โดยในการหา feature importance ของการพัฒนาแบบจำลอง XGBoostClassifier จะได้ค่า feature importance ซึ่งเรียงลำดับตามความสำคัญจากมากไปหาน้อย จะได้ค่าดังตารางที่ 5

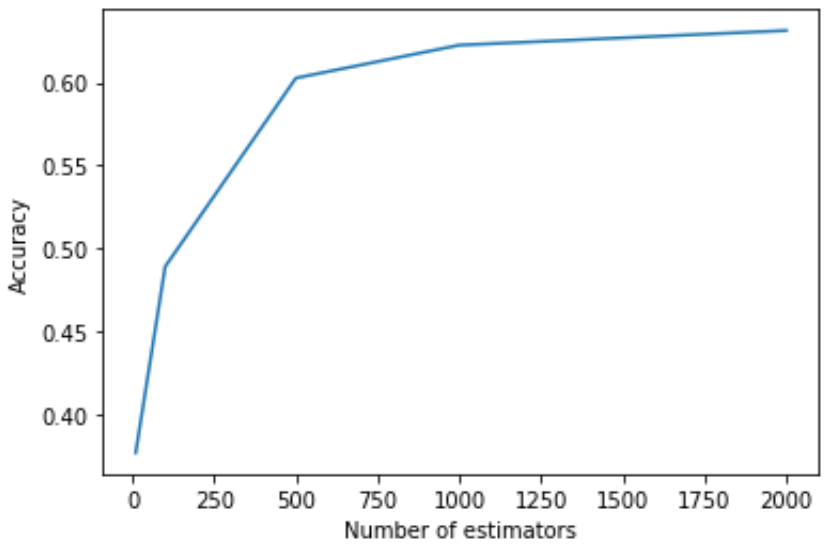
ตาราง 5 ผลลัพธ์ feature importance ของการพัฒนาแบบจำลอง XGBoostClassifier

Feature	Importance
REG_CITY_NOT_LIVE_CITY	0.08
FLAG_OWN_REALTY	0.065
NAME_HOUSING_TYPE	0.06
NAME_EDUCATION_TYPE	0.055
REG_CITY_NOT_WORK_CITY	0.05
REGION_RATING_CLIENT	0.05
OCCUPATION_TYPE	0.045
AMT_ANNUITY	0.045
LIVE_CITY_NOT_WORK_CITY	0.045
ORGANIZATION_TYPE	0.04
AMT_CREDIT	0.035
FLAG_WORK_PHONE	0.03
CODE_GENDER	0.03
NAME_TYPE_SUITE	0.03
NAME_FAMILY_STATUS	0.03
FLAG_OWN_CAR	0.03
DAYS_EMPLOYED	0.03
NAME_INCOME_TYPE	0.03
NAME_CONTRACT_TYPE	0.03
DAYS_BIRTH	0.03
CNT_FAM_MEMBERS	0.03
DAYS_ID_PUBLISH	0.03

OBS_60_CNT_SOCIAL_CIRCLE	0.03
FLAG_PHONE	0.03
AMT_INCOME_TOTAL	0.03
DAYS_REGISTRATION	0.025

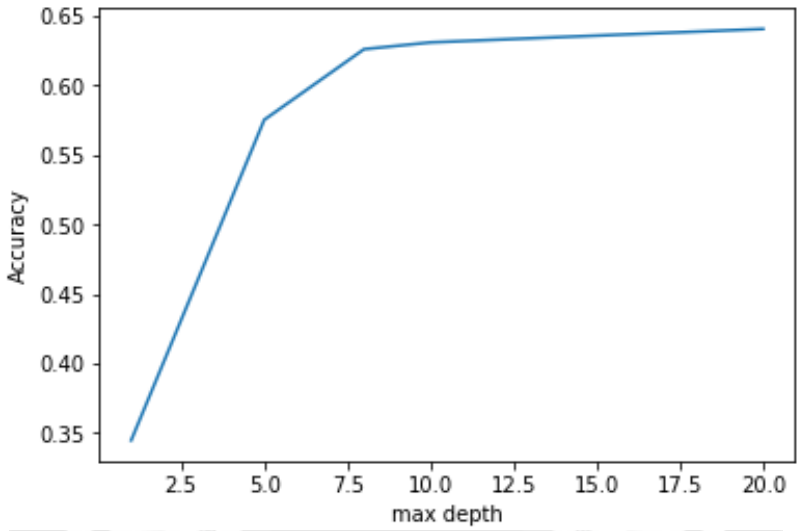
โดย REG_CITY_NOT_LIVE_CITY เป็น feature ที่มีความสำคัญสูงเท่ากับ 0.08 ซึ่งเป็น feature ที่มีความสำคัญสูงสุดในการพัฒนาแบบจำลอง XGBoostClassifier ที่สามารถนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้ รองลงมาจะเป็น FLAG_OWN_REALTY ซึ่งมีค่าความสำคัญเท่ากับ 0.065 และสุดท้าย NAME_HOUSING_TYPE ซึ่งมีค่าความสำคัญเท่ากับ 0.06 และ feature อื่นๆ ร่วมด้วย ซึ่ง feature เหล่านี้เป็น feature หลักๆที่สำคัญที่นำไปใช้ในการพัฒนาแบบจำลอง XGBoostClassifier เพื่อทำให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น

โดยได้มีการกำหนดค่าตัวแปร n_estimators เพื่อกำหนดจำนวนของต้นไม้ตัดสินใจที่จะนำมาช่วยกันในการตัดสินใจในการทำนายเพื่อทำให้ได้ผลลัพธ์ที่ดียิ่งขึ้น โดยได้มีการกำหนดค่า Hyperparameter n_estimators ให้มีค่าเท่ากับ [10, 100, 500, 1000, 1500, 2000] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 1000



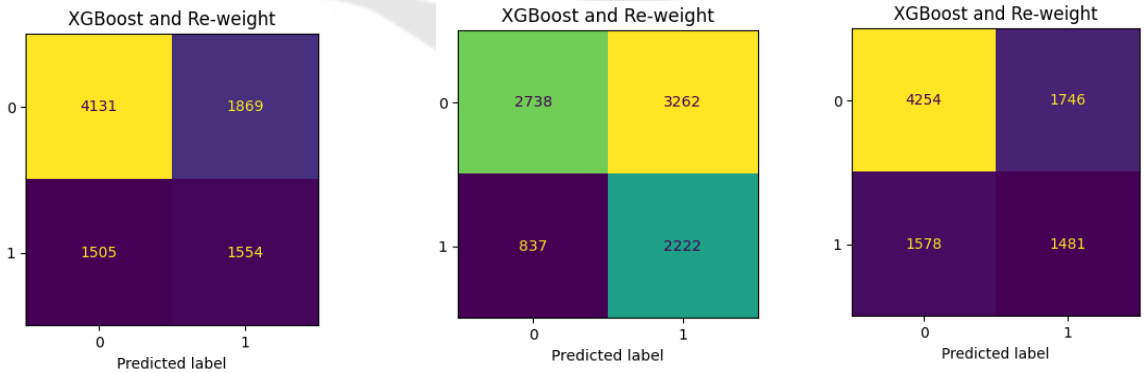
ภาพประกอบ 27 จำนวน n_estimators ของการพัฒนาแบบจำลอง XGBoostClassifier

และมีการกำหนดค่าตัวแปร max_depth เพื่อช่วยในการกำหนดความสูงของต้นไม้ตัดสินใจแต่ละต้น เพื่อลดการเกิด Overfitting ของการพัฒนาแบบจำลอง โดยได้มีการกำหนดค่า Hyperparameter max_depth ให้มีค่าเท่ากับ [1, 5, 8, 10, 15, 20] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 10



ภาพประกอบ 28 จำนวน max_depth ของการพัฒนาแบบจำลอง XGBoostClassifier

ทำการเปรียบเทียบกันระหว่างการเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้เทคนิค Oversampling, Under sampling และ SMOTE



ภาพประกอบ 29 Confusion Matrix ของการพัฒนาแบบจำลอง XGBoostClassifier

ตาราง 6 ผลลัพธ์ของการพัฒนาแบบจำลอง XGBoostClassifier

XGBoostClassifier	Oversampling	Under sampling	SMOTE
Accuracy	0.63	0.55	0.63
Precision	0.45	0.41	0.46
Recall	0.51	0.73	0.48
F1-Score	0.48	0.52	0.47

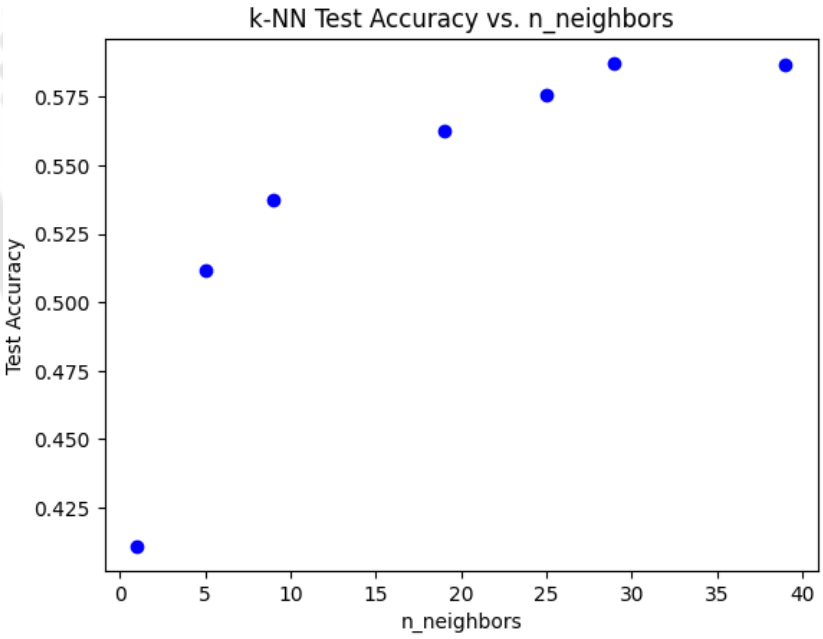
จากการเปรียบเทียบประสิทธิภาพของทั้ง 3 แบบจำลอง จากวิธีการปรับความไม่สมดุลของข้อมูล ด้วยวิธีการ Oversampling, Under sampling และ SMOTE โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก ซึ่งจะสรุปได้ว่า วิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่า F1-Score เท่ากับ 0.52 โดยจะทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,222 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,738 คน

1. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling. สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,554 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 4,131 คน
2. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,222 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,738 คน
3. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ SMOTE สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,481 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 4,254 คน

4.2.3 K-Nearest Neighbors (KNN)

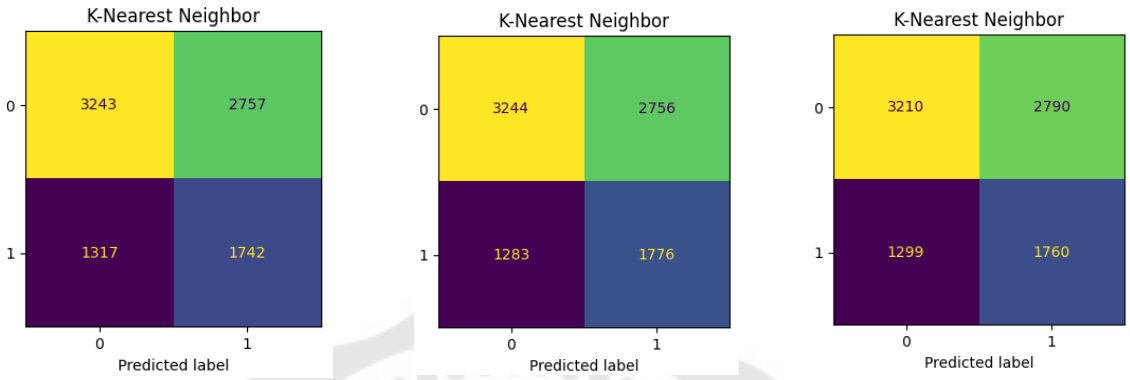
ผู้วิจัยทำการพัฒนาแบบจำลอง โดยการใช้เทคนิคเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุด โดยวิธีการทำงานของเทคนิคเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุด คือการค้นหาเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุด แล้วแบ่งกลุ่มข้อมูลและทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียงเป็นจำนวน K ตัว ซึ่งค่า K คือค่าที่แบบจำลองนำมาใช้ในการพิจารณา ว่าต้องการดูเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุดจำนวนกี่จุดข้อมูล แล้วผลลัพธ์สุดท้ายของการทำนาย จะนำผลลัพธ์ทุกค่าที่ได้จากการทำนายมาหาผลลัพธ์ โดยการทำ Majority vote หรือการใช้เสียงข้างมาก ซึ่งนั้นจะเป็นคำตอบสุดท้ายหรือผลลัพธ์ที่ได้จากการทำนาย

โดยได้มีการกำหนดค่าตัวแปร n_neighbors ซึ่งเป็นตัวกำหนดว่าเราต้องการดูเพื่อนบ้านที่อยู่ใกล้ฉันมากที่สุดทั้งหมดกี่จุดข้อมูล โดยได้มีการกำหนดค่า Hyperparameter n_neighbors ให้มีค่าเท่ากับ [1, 5, 9, 19, 25, 29, 39] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 29



ภาพประกอบ 30 จำนวน n_neighbors ของการพัฒนาแบบจำลอง K-Nearest Neighbors (KNN)

ทำการเปรียบเทียบกันระหว่างการเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้เทคนิค Oversampling, Under sampling และ SMOTE



ภาพประกอบ 31 Confusion Matrix ของการพัฒนาแบบจำลอง K-Nearest Neighbors (KNN)

ตาราง 7 ผลลัพธ์ของการพัฒนาแบบจำลอง K-Nearest Neighbors (KNN)

K-Nearest Neighbors	Oversampling	Under sampling	SMOTE
Accuracy	0.55	0.55	0.55
Precision	0.39	0.39	0.39
Recall	0.57	0.58	0.58
F1-Score	0.46	0.47	0.46

จากการเปรียบเทียบประสิทธิภาพของทั้ง 3 แบบจำลอง จากวิธีการปรับความไม่สมดุลของข้อมูล ด้วยวิธีการ Oversampling, Under sampling และ SMOTE โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก ซึ่งจะสรุปได้ว่า วิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่า F1-Score เท่ากับ 0.47 โดยจะทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,776 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,244 คน

1. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling. สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,742 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,243 คน

2. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,776 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,244 คน

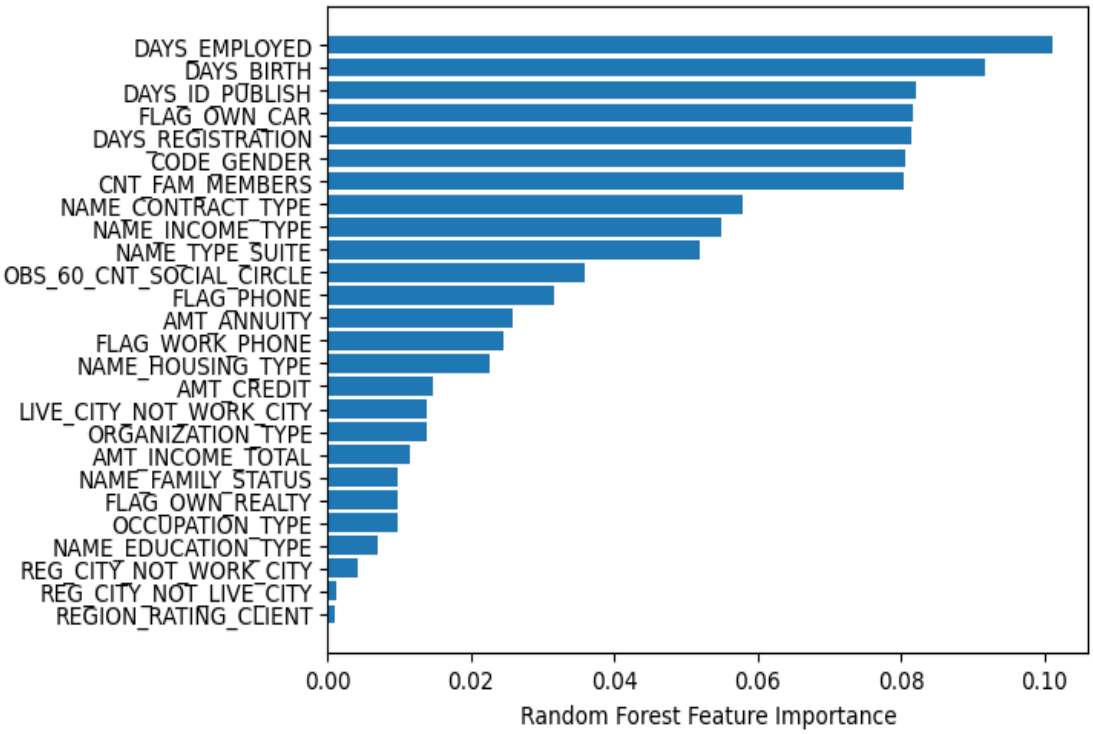
3. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ SMOTE สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,760 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,210 คน

4.2.4 Random Forest

ผู้วิจัยทำการพัฒนาแบบจำลอง โดยการใช้เทคนิคต้นไม้ตัดสินใจหลายๆต้นมาช่วยกันในการตัดสินใจ ซึ่งการทำงานของแบบจำลอง Random Forest คือการสร้างต้นไม้ตัดสินใจหลายๆต้น โดยที่ต้นไม้ตัดสินใจแต่ละต้น จะถูกสร้างขึ้นมาจาก 2 วิธี คือ การทำ Bagging และ random feature projection ซึ่งการทำงานของ Bagging คือการสร้าง dataset ใหม่ขึ้นมาหลายๆ dataset จาก dataset เดิมที่มีอยู่แล้ว โดยใช้วิธีการหยิบสุ่มและคืนข้อมูลลงใน dataset แต่ละตัว โดยในแต่ละ dataset ที่ได้นั้น จะมีจำนวนข้อมูลที่น้อยกว่าจำนวนข้อมูลที่อยู่ใน dataset เดิมตั้งต้นที่มีอยู่แล้ว และวิธีการทำงานของ random feature projection คือการสุ่ม feature ขึ้นมาในจำนวน $\sqrt{P}$ ซึ่งค่า P คือจำนวนของ Feature ที่เรานำมาใช้ในการพัฒนาแบบจำลอง โดยแต่ละครั้งของการ split ในแต่ละ node จะมีการสุ่ม Feature ตามจำนวนของ $\sqrt{P}$ ทุกครั้ง แล้วนำมาหาว่า Feature ไหน ที่ให้ค่า Gini ที่น้อยที่สุด ถือว่า Feature นั้นเป็น Feature ที่มีความสะอาดมากที่สุด ก็จะนำ Feature นั้น มาใช้ในการพัฒนาแบบจำลอง ซึ่งการทำงานของ Random Forest จะมีลักษณะการทำงานที่ขนานกันไปแบบ Parallelly คือสามารถทำงานไปพร้อมๆกันได้

Random Forest สามารถนำไปใช้ในการค้นหา feature importance ซึ่งใช้ในการบอกว่า feature ไหนเป็นตัวทำนายที่สำคัญที่สุดสำหรับการพัฒนาแบบจำลอง ซึ่ง feature importance ของ Random Forest ถูกกำหนดด้วยการวัดค่าการลดความสกปรกที่เกิดจากการนำ feature ไปใช้ในการแบ่งกลุ่มข้อมูล ค่าความสกปรกของ node ในต้นไม้ตัดสินใจจะถูกคำนวณจากค่า Gini impurity และการลดความสกปรกหลังจากการแบ่งกลุ่มจะใช้เพื่อคำนวณหาค่าความสำคัญของ feature ที่ใช้ในการแบ่งกลุ่มนั้น ซึ่ง feature ตัวไหนที่ลดความสกปรกได้มาก feature นั้นก็จะเป็น feature ที่มีความสำคัญมาก โดยการจัดอันดับความสำคัญของ feature จะใช้ค่าเฉลี่ยของการลด

ความสกปรก ซึ่งเป็นค่าเฉลี่ยของการลดความสกปรกที่เกิดจาก feature นั้น ในต้นไม้ตัดสินใจทั้งหมดใน Random Forest



ภาพประกอบ 32 จำนวน feature importance ของการพัฒนาแบบจำลอง Random Forest

โดยในการหา feature importance ของการพัฒนาแบบจำลอง Random Forest จะได้ค่า feature importance ซึ่งเรียงลำดับตามความสำคัญจากมากไปหาน้อย จะได้ค่าดังตารางที่ 8

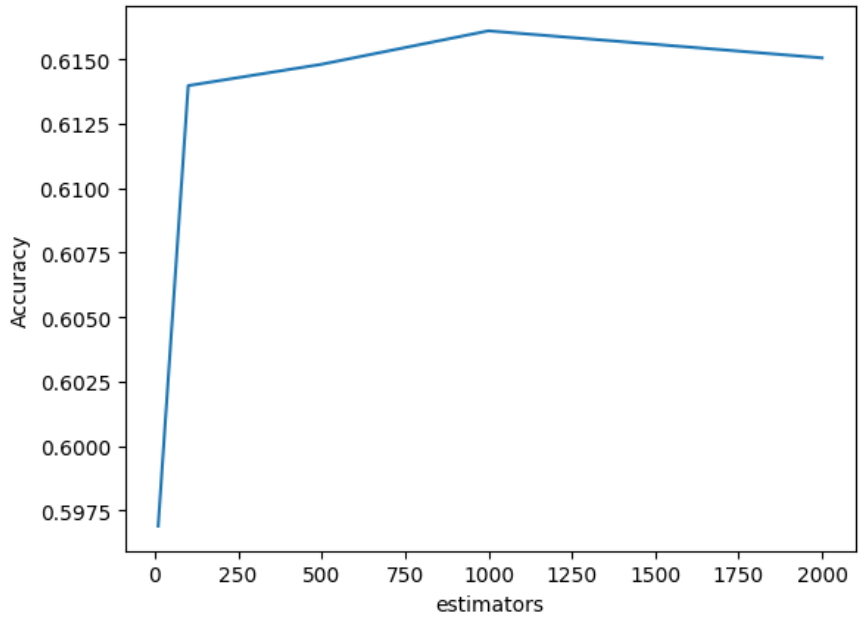
ตาราง 8 ผลลัพธ์ feature importance ของการพัฒนาแบบจำลอง Random Forest

Feature	Importance
DAYS_EMPLOYED	0.1
DAYS_BIRTH	0.09
DAYS_ID_PUBLISH	0.08
FLAG_OWN_CAR	0.08
DAYS_REGISTRATION	0.08
CODE_GENDER	0.08

CNT_FAM_MEMBERS	0.08
NAME_CONTRACT_TYPE	0.06
NAME_INCOME_TYPE	0.055
NAME_TYPE_SUITE	0.05
OBS_60_CNT_SOCIAL_CIRCLE	0.04
FLAG_PHONE	0.035
AMT_ANNUITY	0.03
FLAG_WORK_PHONE	0.03
NAME_HOUSING_TYPE	0.02
AMT_CREDIT	0.015
LIVE_CITY_NOT_WORK_CITY	0.015
ORGANIZATION_TYPE	0.015
AMT_INCOME_TOTAL	0.01
NAME_FAMILY_STATUS	0.01
FLAG_OWN_REALTY	0.01
OCCUPATION_TYPE	0.01
NAME_EDUCATION_TYPE	0.005
REG_CITY_NOT_WORK_CITY	0.003

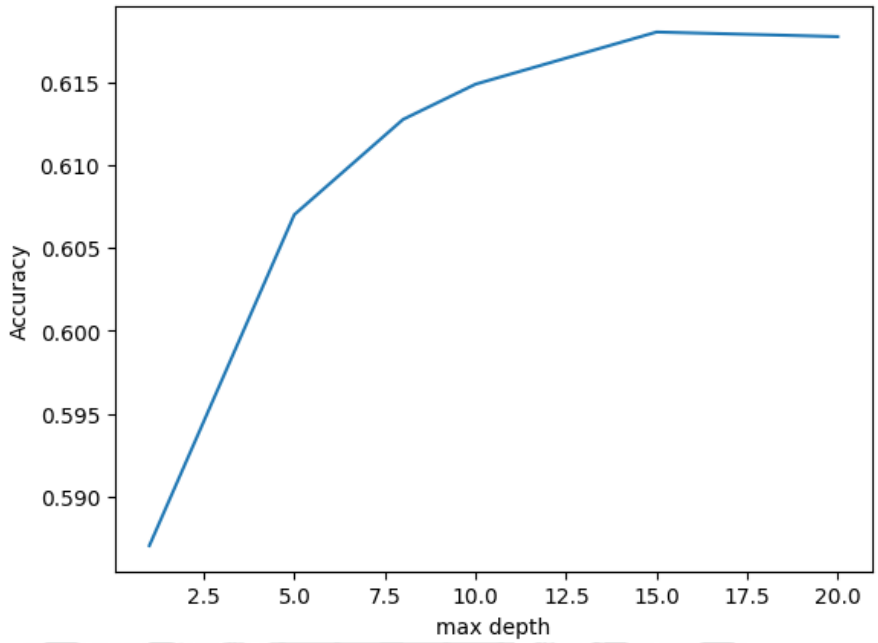
โดย DAYS_EMPLOYED เป็น feature ที่มีความสำคัญสูงเท่ากับ 0.1 ซึ่งเป็น feature ที่มีความสำคัญสูงสุดในการพัฒนาแบบจำลอง Random Forest ที่สามารถนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้ รองลงมาจะเป็น DAYS_BIRTH ซึ่งมีค่าความสำคัญเท่ากับ 0.09 และสุดท้าย DAYS_ID_PUBLISH,FLAG_OWN_CAR,DAYS_REGISTRATION,CODE_GENDER, CNT_FAM_MEMBERS ซึ่งมีค่าความสำคัญเท่ากับ 0.08 และ feature อื่นๆ ร่วมด้วย ซึ่ง feature เหล่านี้เป็น feature หลักๆที่สำคัญที่นำไปใช้ในการพัฒนาแบบจำลอง Random Forest เพื่อทำให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น

โดยได้มีการกำหนดค่าตัวแปร n_estimators เพื่อกำหนดจำนวนของต้นไม้ตัดสินใจที่จะนำมาช่วยกันในการตัดสินใจในการทำนายเพื่อทำให้ได้ผลลัพธ์ที่ดียิ่งขึ้น โดยได้มีการกำหนดค่า Hyperparameter n_estimators ให้มีค่าเท่ากับ [10, 100, 500, 1000, 1500, 2000] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 1000



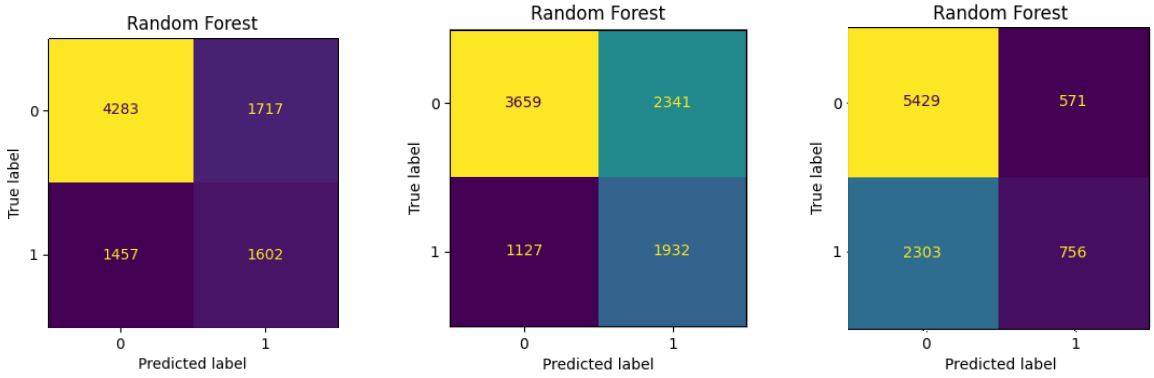
ภาพประกอบ 33 จำนวน n_estimators ของการพัฒนาแบบจำลอง Random Forest

และมีการกำหนดค่าตัวแปร max_depth เพื่อช่วยในการกำหนดความสูงของต้นไม้ตัดสินใจแต่ละต้น เพื่อลดการเกิด Overfitting ของการพัฒนาแบบจำลอง โดยได้มีการกำหนดค่า Hyperparameter max_depth ให้มีค่าเท่ากับ [1, 5, 8, 10, 15, 20] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 15



ภาพประกอบ 34 จำนวน max_depth ของการพัฒนาแบบจำลอง Random Forest

ทำการเปรียบเทียบกันระหว่างการเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้เทคนิค Oversampling, Under sampling และ SMOTE



ภาพประกอบ 35 Confusion Matrix ของการพัฒนาแบบจำลอง Random Forest

ตาราง 9 ผลลัพธ์ของการพัฒนาแบบจำลอง Random Forest

Random Forest	Oversampling	Under sampling	SMOTE
Accuracy	0.65	0.62	0.68
Precision	0.48	0.45	0.57
Recall	0.52	0.64	0.27
F1-Score	0.50	0.53	0.36

จากการเปรียบเทียบประสิทธิภาพของทั้ง 3 แบบจำลอง จากวิธีการปรับความไม่สมดุลของข้อมูล ด้วยวิธีการ Oversampling, Under sampling และ SMOTE โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก ซึ่งจะสรุปได้ว่า วิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่า F1-Score เท่ากับ 0.53 โดยจะทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,959 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,686 คน

1. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling. สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,602 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 4,283 คน
2. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,932 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,659 คน
3. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ SMOTE สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 756 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 5,429 คน

4.2.5 Support Vector Machine (SVC)

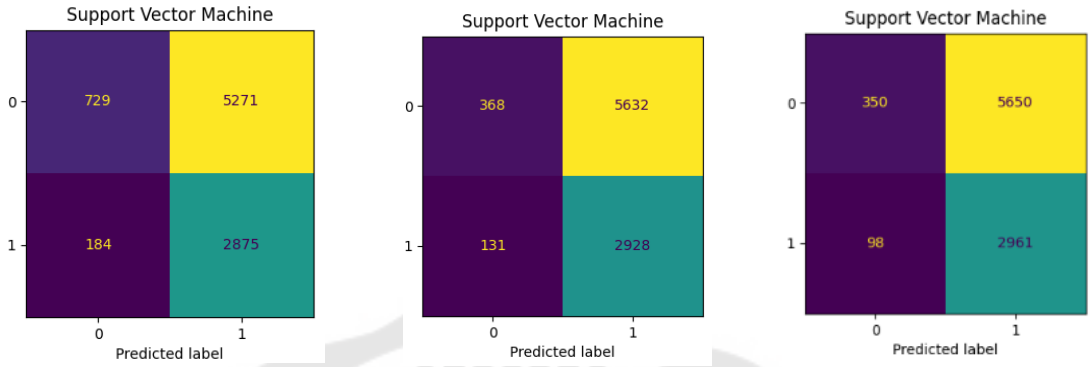
ผู้วิจัยทำการพัฒนาแบบจำลอง โดยการใช้เทคนิค Support Vector Machine (SVC) ซึ่งเป็นเทคนิคที่มีความยืดหยุ่นและทำงานได้ดี โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีความซับซ้อน และมีข้อมูลหลายๆ Feature โดยการทำงานของ Support Vector Machine (SVC) คือการพยายามหาเส้นแบ่งระหว่างคลาสต่างๆ ในข้อมูล ให้เส้นแบ่งที่ได้มีความกว้างมากที่สุด และยอมให้มีข้อมูลบางจุดอยู่ระหว่างเส้นแบ่ง เพื่อไม่ให้เส้นมันเพี้ยนมากจนเกินไป โดยมันพยายามที่จะแยกข้อมูลของทั้ง 2 คลาสออกจากกัน

โดยได้มีการกำหนดค่าตัวแปร C เพื่อช่วยปรับความซับซ้อนของการพัฒนาแบบจำลอง เนื่องจากถ้าค่า C มีค่าน้อยจะทำให้เกิด margin ที่กว้างขึ้นแยกแยะระหว่างคลาสต่างๆได้ดี แต่ก็อาจจะทำให้เกิด overfitting ได้ แต่ถ้าค่า C มีค่ามากก็จะทำให้เกิด margin ที่แคบลง และก็อาจจะทำให้เกิด overfitting ได้เช่นกัน โดยได้มีการกำหนดค่า Hyperparameter C ให้มีค่าเท่ากับ [0.01, 0.1, 1, 10, 100, 1000] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 1

ตาราง 10 ผลลัพธ์การเปรียบเทียบประสิทธิภาพค่า Hyperparameter C ของการพัฒนาแบบจำลอง Support Vector Machine (SVC)

Hyperparameter C	Accuracy
0.01	0.54
0.1	0.55
1	0.57
10	0.56
100	0.56
1000	0.56

ทำการเปรียบเทียบกันระหว่างการเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้เทคนิค Oversampling, Under sampling และ SMOTE



ภาพประกอบ 36 Confusion Matrix ของการพัฒนาแบบจำลอง Support Vector Machine (SVC)

ตาราง 11 ผลลัพธ์ของการพัฒนาแบบจำลอง Support Vector Machine (SVC)

SVC	Oversampling	Under sampling	SMOTE
Accuracy	0.40	0.36	0.37
Precision	0.35	0.34	0.34
Recall	0.94	0.96	0.97
F1-Score	0.51	0.50	0.51

จากการเปรียบเทียบประสิทธิภาพของทั้ง 3 แบบจำลอง จากวิธีการปรับความไม่สมดุลของข้อมูล ด้วยวิธีการ Oversampling, Under sampling และ SMOTE โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก ซึ่งจะสรุปได้ว่า วิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Synthetic Minority Oversampling Technique (SMOTE) ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่า F1-Score เท่ากับ 0.51 โดยจะทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,961 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 350 คน

1. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling. สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,875 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 729 คน

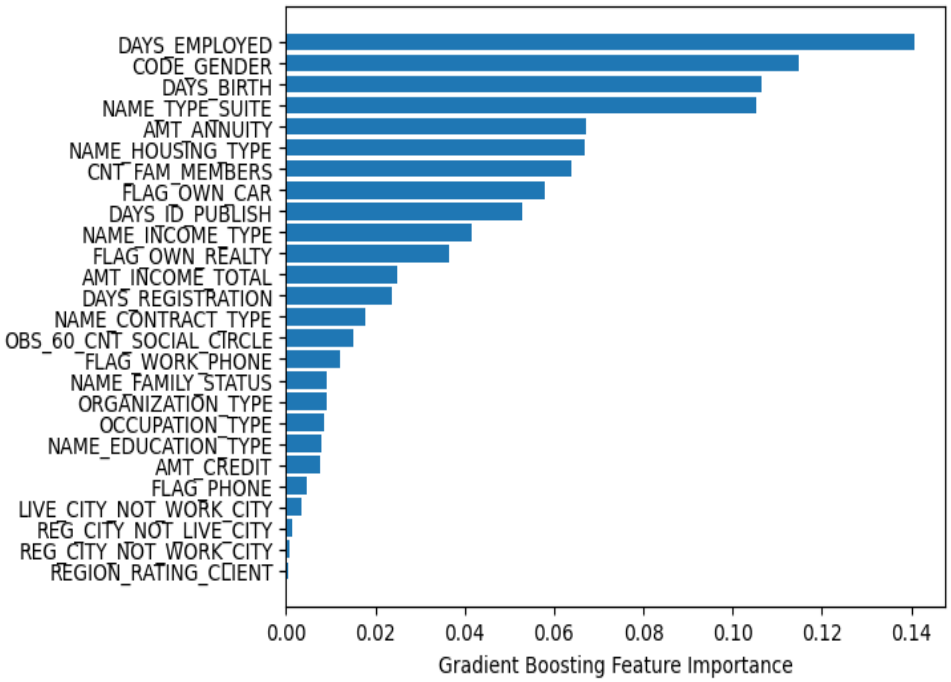
2. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,928 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 368 คน

3. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ SMOTE สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 2,961 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 350 คน

4.2.6 Gradient Boosting

ผู้วิจัยทำการพัฒนาแบบจำลอง โดยการใช้เทคนิคต้นไม้ตัดสินใจหลายๆต้นมาช่วยกันในการตัดสินใจ ซึ่งการทำงานของแบบจำลอง Gradient Boosting คือเทคนิคการเรียนรู้ของเครื่องมือสำหรับการแก้ปัญหา โดยจะใช้เทคนิคการเพิ่มการรวมจำนวนต้นไม้ตัดสินใจที่มีความแม่นยำต่ำ เพื่อสร้างเป็นต้นไม้ตัดสินใจต้นใหม่ โดยต้นไม้ตัดสินใจต้นใหม่จะถูกสร้างขึ้นจากความผิดพลาด (error) จากการคำนวณของต้นไม้ตัดสินใจก่อนหน้า ซึ่งหลักการทำงานของ Gradient Boosting จะมีลักษณะการทำงานในลักษณะที่เป็นลำดับแบบ sequential คือต้องรอให้แบบจำลองต้นแรกสร้างเสร็จเรียบร้อยก่อน จึงจะนำไปพัฒนาเป็นแบบจำลองต้นที่สองได้

Gradient Boosting สามารถนำไปใช้ในการค้นหา feature importance ซึ่งใช้ในการบอกว่า feature ไหนเป็นตัวทำนายที่สำคัญที่สุดสำหรับการพัฒนาแบบจำลอง ซึ่งหลักการทำงานคล้ายกับ XGBoostClassifier และ Random Forest โดยจะคำนวณจากผลรวมของการลดค่าความผิดพลาด (loss) จากการใช้งานแต่ละ feature นั้น โดยวิธีการคำนวณในการส่งผ่านต้นไม้ตัดสินใจแต่ละต้น โดยจะมีการใช้งาน feature ที่มีความสำคัญมากขึ้นในการลดความผิดพลาด ซึ่งสามารถนำมาใช้ในการหาค่าความสำคัญของ feature ได้ การคำนวณค่าความสำคัญของ feature ใน Gradient Boosting จะช่วยให้เราเข้าใจว่า feature ใดเป็นปัจจัยที่สำคัญที่สุดในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร ซึ่งสามารถนำไปช่วยในการปรับการพัฒนาแบบจำลอง เพื่อทำให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น และสามารถนำไปใช้ในการเลือก feature ที่สำคัญสำหรับการพัฒนาแบบจำลอง



ภาพประกอบ 37 จำนวน feature importance ของการพัฒนาแบบจำลอง Gradient Boosting

โดยในการหา feature importance ของการพัฒนาแบบจำลอง Gradient Boosting จะได้ค่า feature importance ซึ่งเรียงลำดับตามความสำคัญจากมากไปหาน้อย จะได้ค่าดังตารางที่ 12

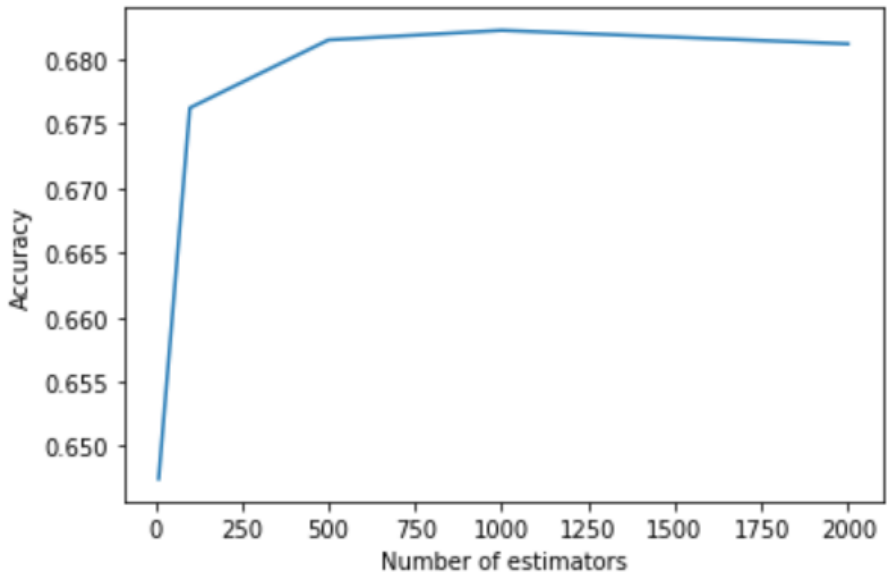
ตาราง 12 ผลลัพธ์ feature importance ของการพัฒนาแบบจำลอง Gradient Boosting

Feature	Importance
DAYS_EMPLOYED	0.14
CODE_GENDER	0.12
DAYS_BIRTH	0.1
NAME_TYPE_SUITE	0.1
AMT_ANNUITY	0.07
NAME_HOUSING_TYPE	0.07
CNT_FAM_MEMBERS	0.065
FLAG_OWN_CAR	0.06

DAYS_ID_PUBLISH	0.05
NAME_INCOME_TYPE	0.04
FLAG_OWN_REALTY	0.04
AMT_INCOME_TOTAL	0.03
DAYS_REGISTRATION	0.03
NAME_CONTRACT_TYPE	0.02
OBS_60_CNT_SOCIAL_CIRCLE	0.015
FLAG_WORK_PHONE	0.01
NAME_FAMILY_STATUS	0.01
ORGANIZATION_TYPE	0.01
OCCUPATION_TYPE	0.01
NAME_EDUCATION_TYPE	0.01
AMT_CREDIT	0.01
FLAG_PHONE	0.005
LIVE_CITY_NOT_WORK_CITY	0.005

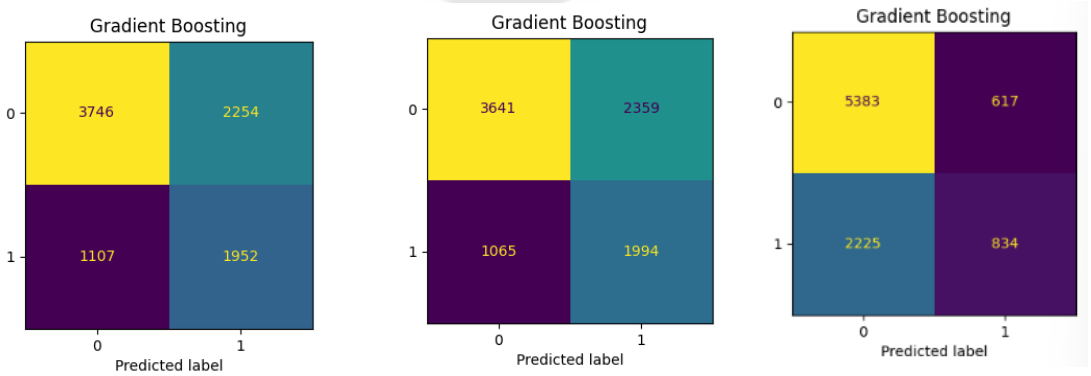
โดย DAYS_EMPLOYED เป็น feature ที่มีความสำคัญสูงเท่ากับ 0.14 ซึ่งเป็น feature ที่มีความสำคัญสูงสุดในการพัฒนาแบบจำลอง Gradient Boosting ที่สามารถนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้ รองลงมาจะเป็น CODE_GENDER ซึ่งมีค่าความสำคัญเท่ากับ 0.12 และสุดท้าย DAYS_BIRTH,NAME_TYPE_SUITE ซึ่งมีค่าความสำคัญเท่ากับ 0.1 และ feature อื่นๆ ร่วมด้วย ซึ่ง feature เหล่านี้เป็น feature หลักๆที่สำคัญที่นำไปใช้ในการพัฒนาแบบจำลอง Gradient Boosting เพื่อทำให้ได้แบบจำลองที่มีประสิทธิภาพที่ดียิ่งขึ้น

โดยได้มีการกำหนดค่าตัวแปร n_estimators เพื่อกำหนดจำนวนของต้นไม้ตัดสินใจที่จะนำมาช่วยกันในการตัดสินใจในการทำนายเพื่อทำให้ได้ผลลัพธ์ที่ดียิ่งขึ้น โดยได้มีการกำหนดค่า Hyperparameter n_estimators ให้มีค่าเท่ากับ [10, 100, 500, 1000, 1500, 2000] ซึ่งค่า Hyperparameter ตัวที่ดีที่สุด ที่เรานำไปใช้ในการพัฒนาแบบจำลอง มีค่าเท่ากับ 500



ภาพประกอบ 38 จำนวน n_estimators ของการพัฒนาแบบจำลอง Gradient Boosting

ทำการเปรียบเทียบกันระหว่างการเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง โดยการใช้เทคนิค Oversampling, Under sampling และ SMOTE



ภาพประกอบ 39 Confusion Matrix ของการพัฒนาแบบจำลอง Gradient Boosting

ตาราง 13 ผลลัพธ์ของการพัฒนาแบบจำลอง Gradient Boosting

Gradient Boosting	Oversampling	Under sampling	SMOTE
Accuracy	0.63	0.62	0.69
Precision	0.46	0.46	0.57
Recall	0.64	0.65	0.27
F1-Score	0.54	0.54	0.37

จากการเปรียบเทียบประสิทธิภาพของทั้ง 3 แบบจำลอง จากวิธีการปรับความไม่สมดุลของข้อมูล ด้วยวิธีการ Oversampling, Under sampling และ SMOTE โดยจะสนใจที่ค่า F1-Score ของ Positive class เป็นหลัก ซึ่งจะสรุปได้ว่า วิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่า F1-Score เท่ากับ 0.54 โดยจะทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,994 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,641 คน

1. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Oversampling. สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,952 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,746 คน
2. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 1,994 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 3,641 คน
3. แบบจำลองของการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ SMOTE สามารถทำนายว่าจะเกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 834 คน และทำนายว่าไม่เกิดเหตุการณ์ลูกหนี้ผิดนัดชำระกับทางธนาคาร 5,383 คน

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการทำวิจัยเรื่องการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร ผู้วิจัยได้ทำการประเมินประสิทธิภาพของการพัฒนาแบบจำลอง เพื่อนำมาใช้ในการเปรียบเทียบและสรุปผล โดยแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

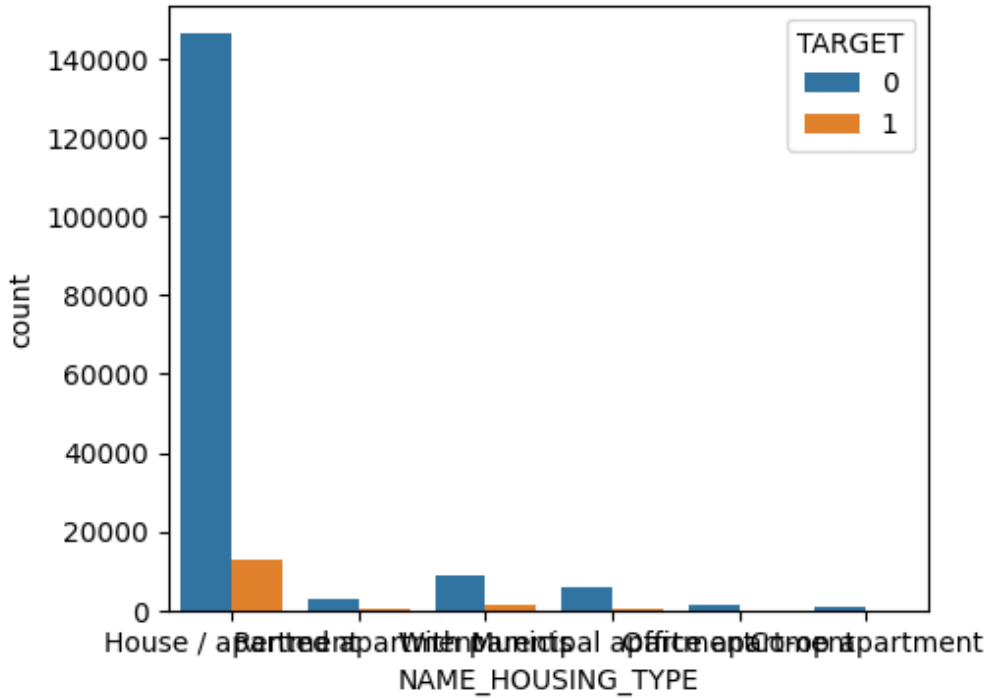
1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในการเปรียบเทียบค่า Feature importance ของการพัฒนาแบบจำลอง ระหว่างแบบจำลอง XGBoostClassifier , Random Forest , Gradient Boosting ซึ่งจะทำการเปรียบเทียบ 15 feature ที่มีความสำคัญมากที่สุด ที่นำมาใช้ในการพัฒนาแบบจำลอง

ซึ่งจากผลลัพธ์ของการเปรียบเทียบค่า feature importance ของการพัฒนาแบบจำลองระหว่างแบบจำลอง XGBoostClassifier , Random Forest , Gradient Boosting ของทั้ง 15 feature จะพบว่าแบบจำลองทั้ง 3 แบบจำลองให้ความสำคัญกับ NAME_HOUSING_TYPE, AMT_ANNUITY, CODE_GENDER, NAME_TYPE_SUITE

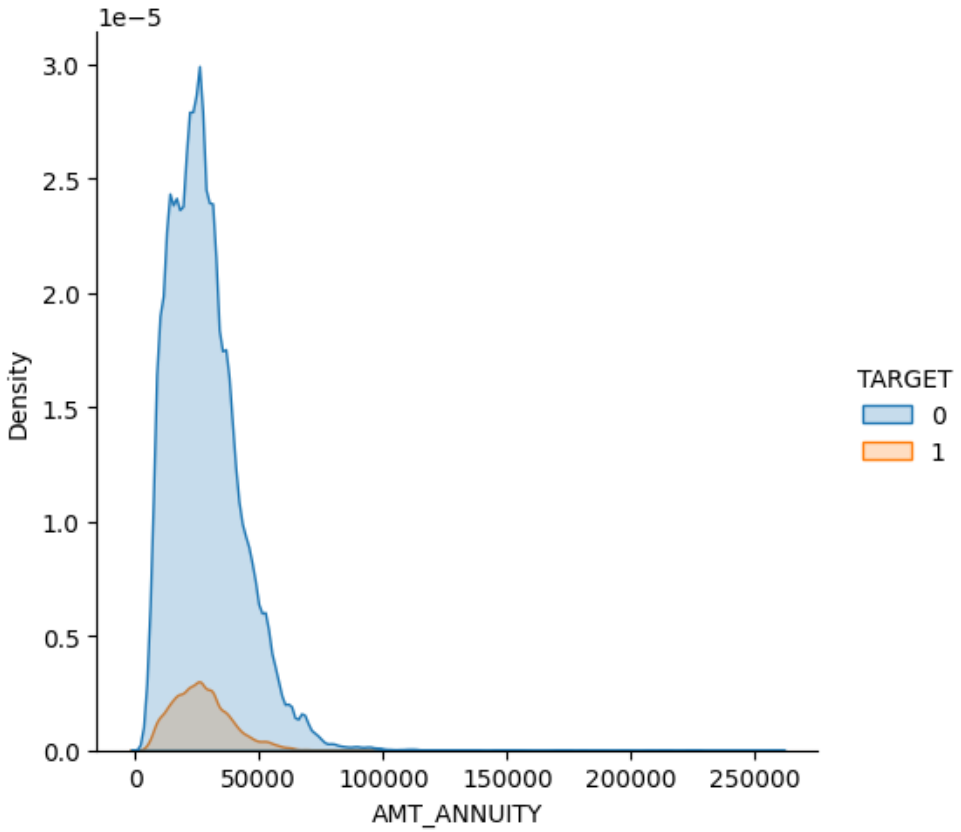
ข้อมูล NAME_HOUSING_TYPE ใช้ในการบอกที่อยู่อาศัยของลูกหนี้ ว่าอาศัยอยู่บ้านของตนเอง อาศัยอยู่กับพ่อแม่ และอื่นๆ



ภาพประกอบ 40 จำนวนข้อมูลที่ใช้ในการบอกที่อยู่อาศัยของลูกหนี้

เมื่อทำการ plot เพื่อดูข้อมูลของ NAME_HOUSING_TYPE จะพบว่าข้อมูลส่วนใหญ่ลูกหนี้จะอาศัยอยู่ที่บ้านของตนเอง ซึ่งถ้านำข้อมูลของ NAME_HOUSING_TYPE มาใช้ในการตรวจจับการผิดนัดชำระของลูกหนี้ จะไม่สามารถใช้ในการตรวจจับการผิดนัดชำระของลูกหนี้ได้ เนื่องจากทั้งข้อมูลลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคารและข้อมูลลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร มีข้อมูลส่วนใหญ่ที่เหมือนกัน คือลูกหนี้อาศัยอยู่ที่บ้านของตนเอง ซึ่งไม่ใช่ตัวที่ใช้ในการจำแนกลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร

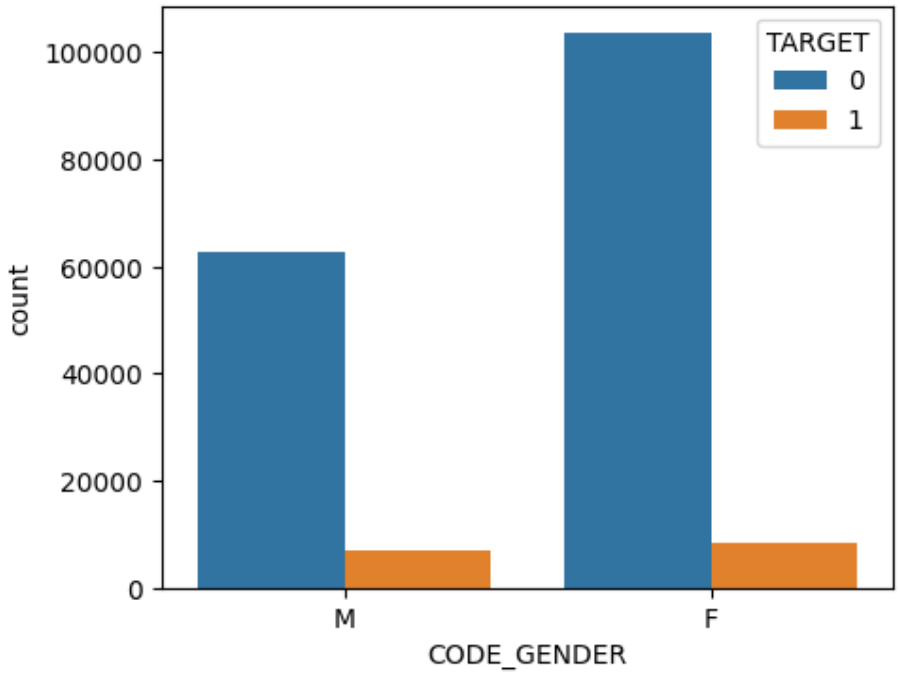
ข้อมูล AMT_ANNUITY ใช้ในการบอกจำนวนเงินที่ลูกหนี้จะต้องชำระในแต่ละงวด



ภาพประกอบ 41 จำนวนข้อมูลของจำนวนเงินที่ลูกหนี้จะต้องชำระในแต่ละงวด

เมื่อทำการ plot เพื่อดูข้อมูลของ AMT_ANNUITY จะพบว่าข้อมูลลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร มีค่างวดที่ต้องชำระอยู่ระหว่าง 0-100,000 ในขณะที่ลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร มีค่างวดที่ต้องชำระอยู่ระหว่าง 0-60,000 ซึ่งไม่น่าจะสอดคล้องกับความเป็นจริง เนื่องจากถ้าหากลูกหนี้ที่มีค่างวดที่ต้องชำระสูงก็อาจจะมีโอกาสที่จะเป็นลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคารเยอะกว่า ซึ่งเมื่อได้ทำการ plot ข้อมูลออกมาดู พบว่าข้อมูลที่ได้ ไม่ค่อยสอดคล้องกับความเป็นจริงที่เกิดขึ้น แต่ในความเป็นจริงข้อมูลของจำนวนเงินที่ลูกหนี้จะต้องชำระในแต่ละงวด อาจเป็นสาเหตุที่ทำให้เกิดโอกาสลูกหนี้ผิดนัดชำระกับทางธนาคารได้ ถ้าข้อมูลของจำนวนเงินที่ลูกหนี้จะต้องชำระในแต่ละงวดมีค่าสูงๆ

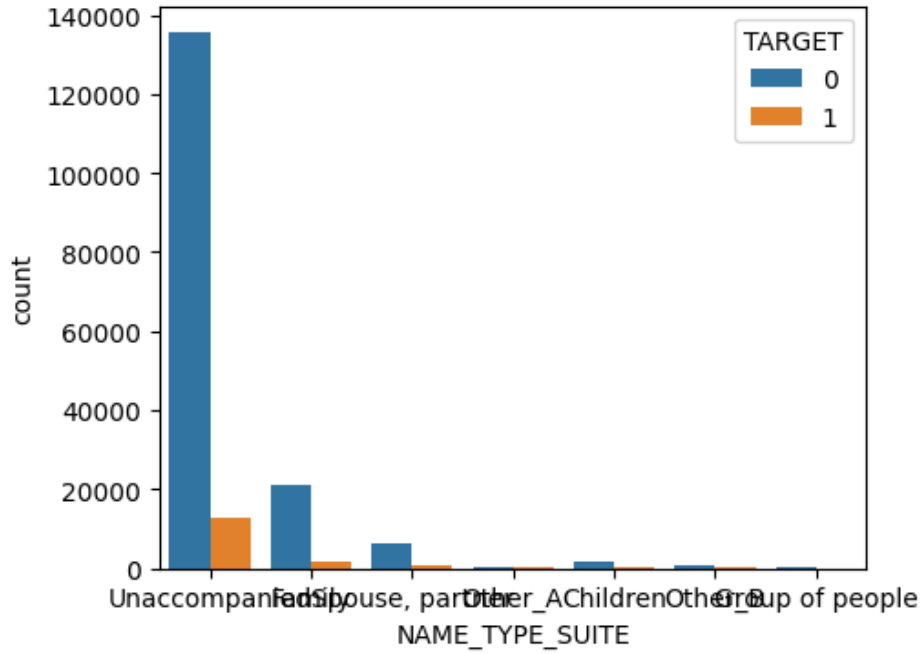
ข้อมูล CODE_GENDER ใช้ในการบอกเพศของลูกหนี้ ว่าเป็นเพศชายหรือเพศหญิง



ภาพประกอบ 42 จำนวนข้อมูลเพศของลูกหนี้

เมื่อทำการ plot เพื่อดูข้อมูลของ CODE_GENDER จะพบว่าข้อมูลลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคารมีความเท่าเทียมกันระหว่างเพศชายและเพศหญิง เพศจึงไม่สามารถเป็นตัวกำหนดได้ว่าจะเป็นลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคารไหม และในความจริงที่เกิดขึ้นพบว่าทั้งเพศชายและเพศหญิงมีโอกาสในการเป็นลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร

ข้อมูล NAME_TYPE_SUITE ใช้ในการบอกคนที่มากับลูกหนี้ เมื่อเวลาที่ลูกหนี้มายื่นขอสมัครสินเชื่อ



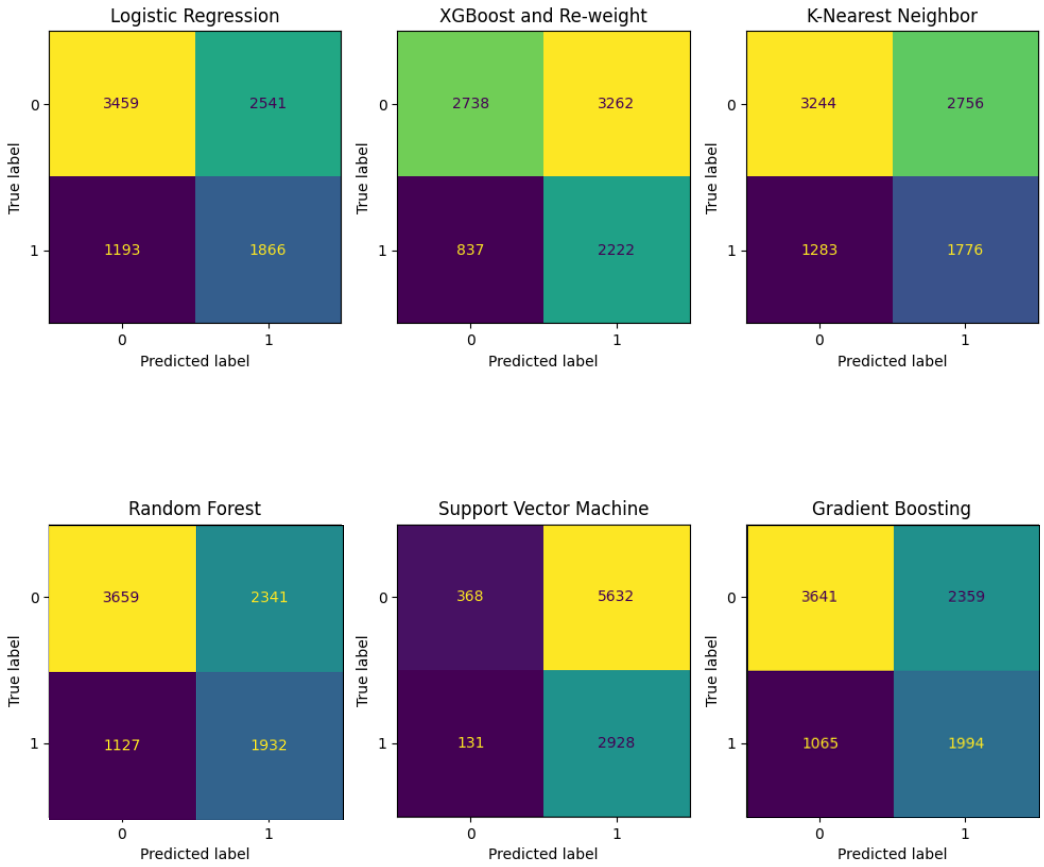
ภาพประกอบ 43 จำนวนข้อมูลของคนที่มากับลูกหนี้เวลามายื่นขอสมัครสินเชื่อ

เมื่อทำการ plot เพื่อดูข้อมูลของ NAME_TYPE_SUITE จะพบว่าข้อมูลส่วนใหญ่ ไม่ว่าจะเป็นข้อมูลลูกหนี้ที่ไม่ได้มีการผิดนัดชำระกับทางธนาคาร และข้อมูลลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร ไม่ได้มีคนมาด้วย ซึ่งมาคนเดียว นั้นจึงทำให้ข้อมูลของ NAME_TYPE_SUITE ไม่มีผลต่อการทำนาย และไม่สอดคล้องกับความเป็นจริงที่เกิดขึ้น เนื่องจากลูกหนี้ที่จะมีโอกาสในการเป็นลูกหนี้ที่มีการผิดนัดชำระกับทางธนาคาร ต้องขึ้นอยู่กับสภาวะทางการเงินของลูกหนี้คนนั้น ไม่เกี่ยวข้องกับบุคคลที่มากับลูกหนี้เวลาที่ลูกหนี้มายื่นขอสมัครสินเชื่อ

ในการเปรียบเทียบค่า Accuracy, Precision, Recall และ F1-Score ระหว่างแบบจำลอง Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting จะได้ค่าดังตารางที่ 15

ตาราง 14 ผลลัพธ์ของการพัฒนาแบบจำลอง โดยการใช้อัลกอริทึม Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting

Matrix	Logistic Regression	XGBoost	K-Nearest Neighbor	Random Forest	Support Vector Machine	Gradient Boosting
Accuracy	0.59	0.55	0.55	0.62	0.36	0.62
Precision	0.42	0.41	0.39	0.45	0.34	0.46
Recall	0.61	0.73	0.58	0.64	0.96	0.65
F1-Score	0.50	0.52	0.47	0.53	0.50	0.54

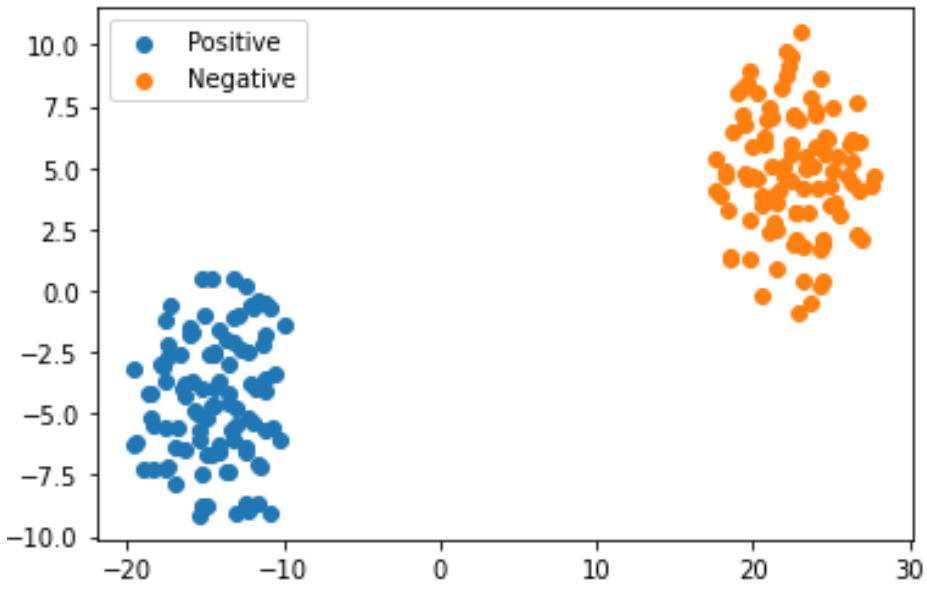


ภาพประกอบ 44 Confusion Matrix ของการพัฒนาแบบจำลอง โดยการใช้อัลกอริทึม Logistic Regression, XGBoostClassifier, K-nearest Neighbors, Random Forest , Support Vector Classifier (SVC), Gradient Boosting

จากผลการศึกษาการพัฒนาแบบจำลองเพื่อใช้ในการทำนายลูกหนี้ที่มีโอกาสในการผิดนัดชำระกับทางธนาคาร โดยการใช้เทคนิคต่าง ๆ เมื่อเปรียบเทียบประสิทธิภาพของการพัฒนาแบบจำลองของหลายๆอัลกอริทึม พบว่าเทคนิควิธีการปรับความไม่สมดุลของข้อมูลด้วยวิธีการ Under sampling เมื่อนำมาใช้ในการปรับความไม่สมดุลของการพัฒนาแบบจำลองจะทำให้แบบจำลองที่ได้มีประสิทธิภาพที่ดียิ่งขึ้น และจะเห็นได้ว่าการพัฒนาแบบจำลองโดยการใช้เทคนิควิธี Gradient Boosting ให้ค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองที่มากที่สุด ซึ่งมีค่าเท่ากับ 0.54 ค่าความไวมีค่าเท่ากับ 0.65 และมีค่าความถูกต้องเท่ากับ 0.62 แต่เทคนิควิธี K-Nearest Neighbor ให้ค่า F1-Score ที่ใช้ในการวัดความสามารถของแบบจำลองที่น้อยที่สุด ซึ่งมีค่าเท่ากับ 0.47 ค่าความไวมีค่าเท่ากับ 0.58 และมีค่าความถูกต้องเท่ากับ 0.55

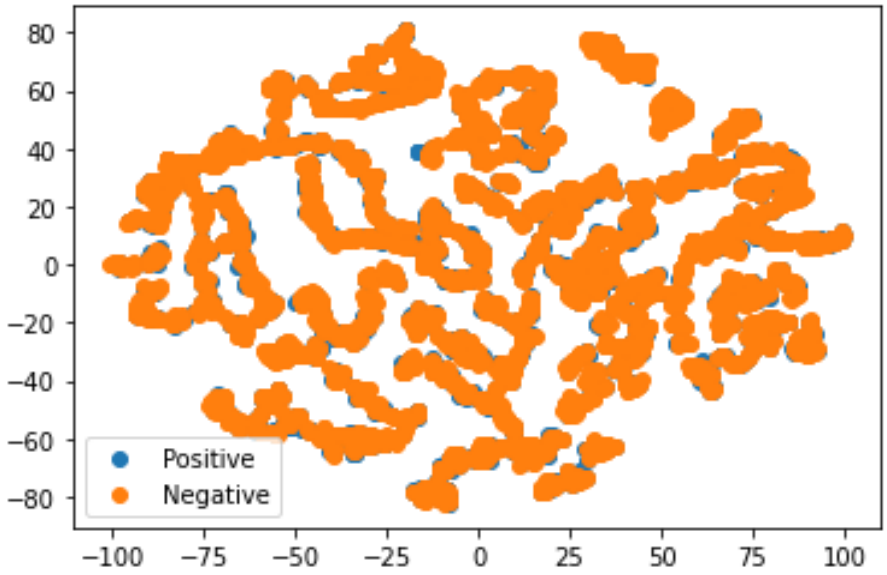
5.2 อภิปรายผลการวิจัย

จากผลลัพธ์ของการพัฒนาแบบจำลอง โดยการใช้วิธีการเทคนิคอัลกอริทึมต่างๆ จะเห็นได้ว่าคุณลักษณะของข้อมูลต่างๆ มีคุณลักษณะของข้อมูลที่ไม่เพียงพอในการแบ่งแยกความแตกต่างระหว่างแต่ละคลาสได้ ซึ่งค่าของคุณลักษณะข้อมูลเหล่านั้นไม่สัมพันธ์กับข้อมูลตัวแปรเป้าหมาย และค่าของคุณลักษณะข้อมูลเหล่านั้นไม่แตกต่างกันมากนัก ระหว่าง Positive class และ Negative class ซึ่งค่าของคุณลักษณะข้อมูลไม่สามารถตรวจจับความแตกต่างระหว่าง Positive class และ Negative class ได้ ทำให้ค่าคุณลักษณะของข้อมูลเหล่านั้นไม่สามารถใช้ในการทำนายผลลัพธ์ได้อย่างมีประสิทธิภาพ โดยจะทำการวิเคราะห์ Error Analysis หรือข้อผิดพลาดของการพัฒนาแบบจำลอง โดยการใช้เทคนิค T-distributed Stochastic Neighbor Embedding (t-SNE) ซึ่งเป็นเทคนิคที่ใช้สำหรับแสดงผลข้อมูลให้อยู่ในรูปแบบสองมิติหรือสามมิติ โดยมีวัตถุประสงค์เพื่อช่วยให้เราเข้าใจและวิเคราะห์ข้อมูลที่มีมิติสูงๆ ทำให้อยู่ในรูปแบบของมิติต่ำ ๆ ทำให้เราเข้าใจข้อมูลได้ดีมากยิ่งขึ้น โดยการใช้เทคนิคนี้จะช่วยให้เราเห็นภาพรวมของข้อมูลได้ง่ายมากยิ่งขึ้น และมีประสิทธิภาพในการวิเคราะห์ข้อมูลที่มีมิติสูงได้ดีมากยิ่งขึ้น



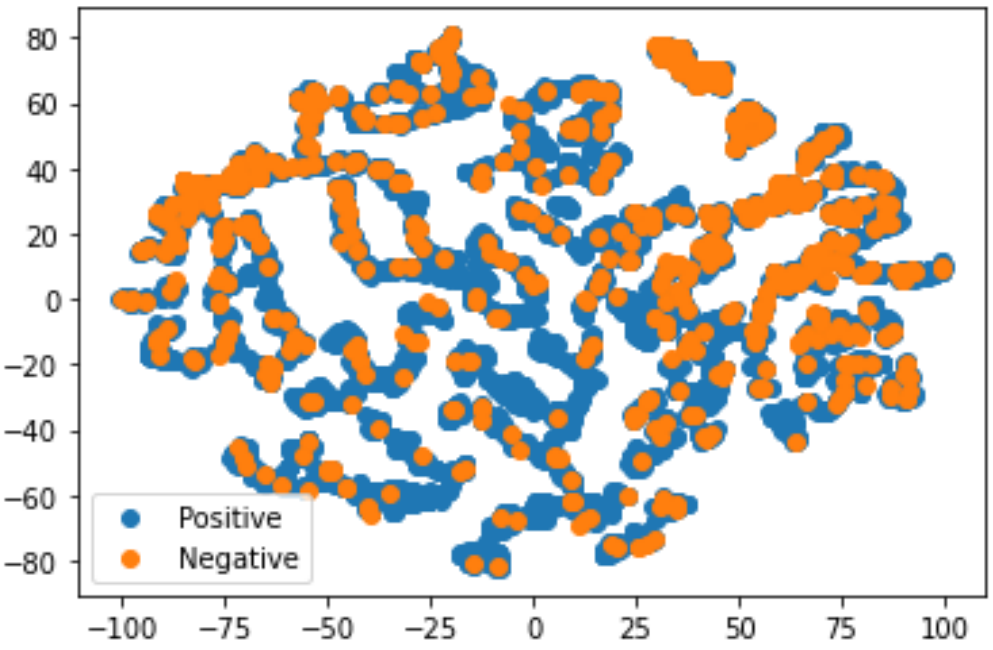
ภาพประกอบ 45 ภาพสองมิติจากการสุ่มตัวอย่างข้อมูลระหว่าง Positive class และ Negative class

เมื่อเราทำการแสดงผลของข้อมูลให้อยู่ในรูปแบบของสองมิติ สิ่งที่ควรจะได้คือ ตัวแปรเป้าหมาย ควรถูกแบ่งแยกออกจากกันอย่างชัดเจน ระหว่างคลาสที่เป็น Positive Class และ Negative Class ภาพที่ได้เกิดจากการสุ่มข้อมูลขึ้นมาเพื่อมาทำการแสดงผลว่า ถ้าข้อมูลมีคุณลักษณะของข้อมูลที่แตกต่างกัน ที่สามารถนำไปใช้ในการแบ่งแยกคลาสได้ เมื่อเราทำการแสดงผลของข้อมูลให้อยู่ในรูปแบบของสองมิติ คลาสที่ได้ระหว่าง Positive Class และ Negative Class จะถูกแบ่งแยกออกจากกันอย่างชัดเจน ตามภาพประกอบที่ 45



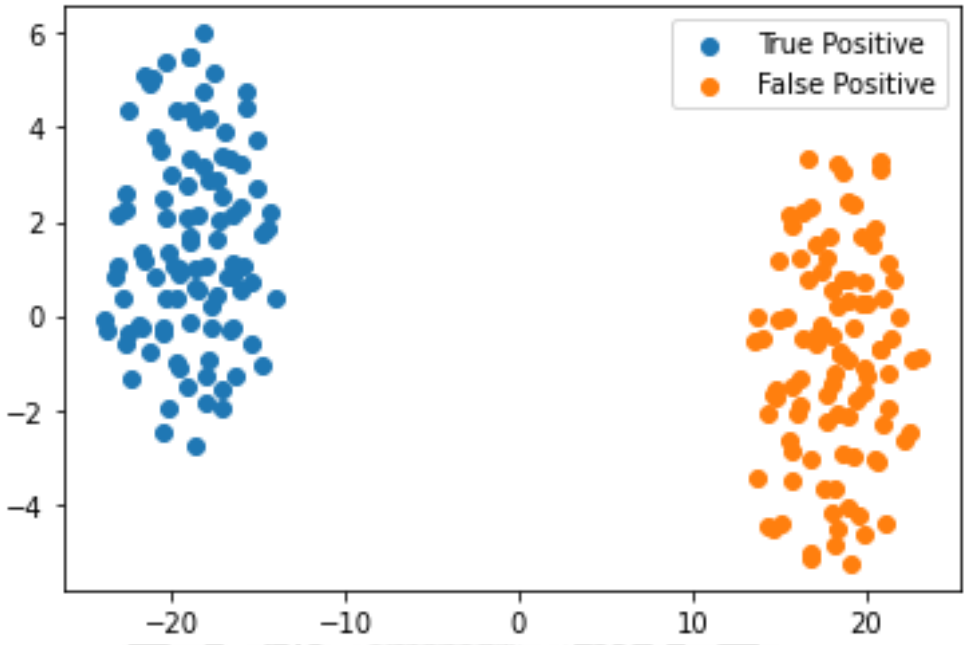
ภาพประกอบ 46 ภาพสองมิติจากการทำนายคลาสตัวแปรเป้าหมายระหว่าง Positive class และ Negative class

เมื่อทำการแสดงผลของข้อมูลของเราให้อยู่ในรูปแบบของสองมิติ จะเห็นได้ว่าข้อมูลของเรา ไม่ได้มีการกระจายตัวของข้อมูลที่ดี ข้อมูลตัวแปรเป้าหมายของ Positive class ปะปนอยู่กับข้อมูลตัวแปรเป้าหมายที่เป็น Negative class ทำให้เราไม่สามารถแบ่งแยก Positive class และ Negative class ออกจากกันได้อย่างชัดเจน เมื่อเรานำข้อมูลเหล่านั้นมาใช้ในการพัฒนาแบบจำลอง ทำให้แบบจำลองที่ได้จึงมีประสิทธิภาพที่ไม่ค่อยดี และไม่เพียงพอต่อการนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้อย่างชัดเจน



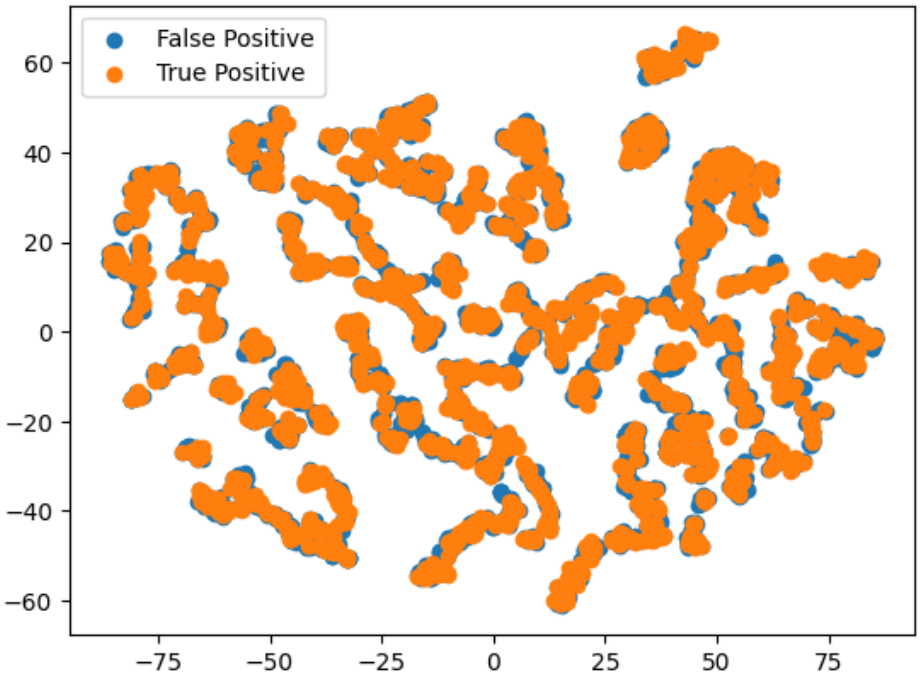
ภาพประกอบ 47 ภาพสองมิติจากคลาสตัวแปรเป้าหมายที่เป็นค่าจริงระหว่าง Positive class และ Negative class

การนำชุดข้อมูลที่ใช้ในการทดสอบกับข้อมูลตัวแปรเป้าหมายที่เป็นค่าจริง นำมาแสดงผลของข้อมูลให้อยู่ในรูปแบบของสองมิติ จะเห็นได้ว่าจากข้อมูลจริง ๆ ของทั้ง Positive class และ Negative class ข้อมูลไม่ได้ถูกแบ่งแยกออกจากกันอย่างชัดเจน ซึ่งเมื่อเรานำแบบจำลองที่ได้ ไปทดสอบกับชุดข้อมูลที่ใช้ในการทดสอบ ผลลัพธ์ที่ได้คือ แบบจำลองที่ได้ ไม่สามารถที่จะแบ่งแยกคลาสที่เป็น Positive class และ Negative class ออกจากกันได้อย่างชัดเจน ทำให้แบบจำลองที่ได้ไม่ค่อยมีประสิทธิภาพ และไม่เพียงพอต่อการนำไปใช้ในการจำแนกแต่ละคลาสออกจากกันได้อย่างชัดเจน



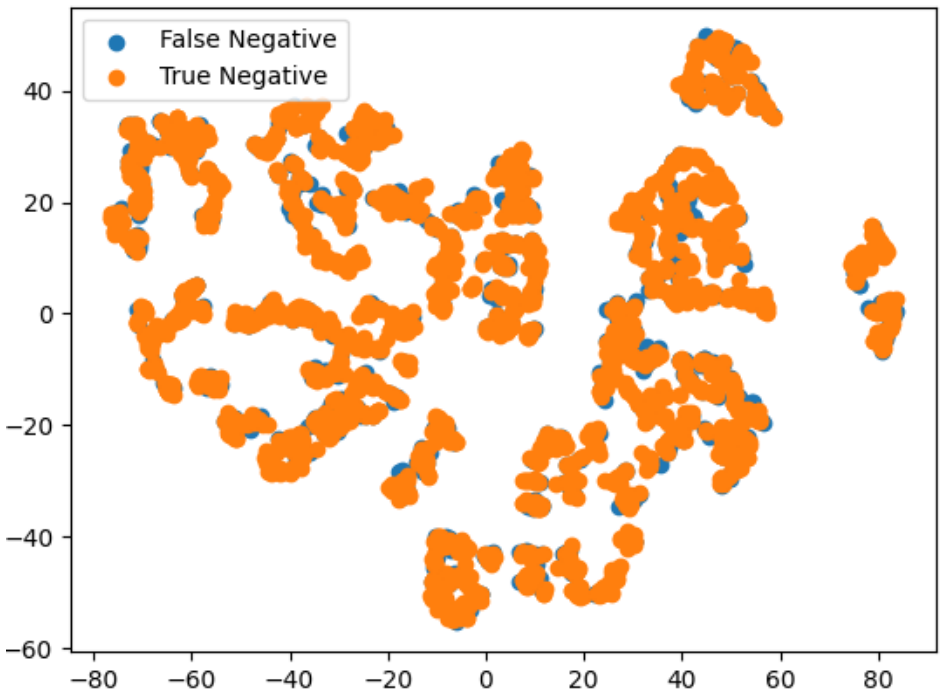
ภาพประกอบ 48 ภาพสองมิติจากการสุ่มตัวอย่างข้อมูลระหว่าง True Positive และ False Positive

เมื่อเราลองทำการวัดประสิทธิภาพของการพัฒนาแบบจำลอง โดยการแสดงผลประสิทธิภาพของแบบจำลองที่ทำนาย Positive Class ได้ ถ้าข้อมูลมีคุณลักษณะข้อมูลที่ดี มีประสิทธิภาพ เมื่อเรานำข้อมูลเหล่านั้น มาทำการแสดงผลของข้อมูลให้อยู่ในรูปแบบของสองมิติ ค่าที่ได้จากการทำนายคลาสของ True Positive และ False Positive ควรจะถูกแบ่งแยกออกจากกันอย่างชัดเจน ตามภาพประกอบที่ 48



ภาพประกอบ 49 ภาพสองมิติจากการทำนายคลาสตัวแปรเป้าหมายระหว่าง True Positive และ False Positive

ข้อมูลที่เรานำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลอง แล้วนำแบบจำลองที่ได้ ไปใช้ในการทำนาย ซึ่งจากผลลัพธ์ที่ได้จะเห็นได้ว่าค่าที่ได้จากการทำนาย Positive Class จะเห็นได้ว่าคลาสที่เป็น True Positive และ False Positive อยู่ในตำแหน่งเดียวกัน ปะปนกันอยู่ ไม่สามารถแบ่งแยกคลาสหรือแยกความแตกต่างออกจากกันได้อย่างชัดเจน



ภาพประกอบ 50 ภาพสองมิติจากการทำนายคลาสตัวแปรเป้าหมายระหว่าง True Negative และ False Negative

ค่าที่ได้จากการทำนาย Negative Class จะเห็นได้ว่าคลาสที่เป็น True Negative และ False Negative อยู่ในตำแหน่งเดียวกันเช่นกัน ปะปนกันอยู่ ไม่สามารถแบ่งแยกคลาสหรือแยกความแตกต่างออกจากกันได้อย่างชัดเจน

5.3 ข้อเสนอแนะ

ถ้าข้อมูลที่นำมาใช้ในการวิเคราะห์เพื่อพัฒนาแบบจำลองไม่สามารถแบ่งแยกความแตกต่างระหว่าง Positive class และ Negative class ได้ อาจทำให้ประสิทธิภาพของการพัฒนาแบบจำลองลดลง เนื่องจากแบบจำลองที่ได้จะไม่สามารถแยกแยะความแตกต่างระหว่างคลาสได้ ซึ่งสามารถแก้ไขปัญหาได้หลายวิธี

1. การเพิ่มจำนวนข้อมูล การเพิ่มจำนวนข้อมูลอาจช่วยทำให้แบบจำลองเข้าใจและเรียนรู้ Insight ของข้อมูลได้ดีมากยิ่งขึ้น และช่วยปรับปรุงประสิทธิภาพในการแยกแยะระหว่าง Positive class และ Negative class ได้ดีมากยิ่งขึ้น

2. เทคนิค Resampling เช่น oversampling หรือ under sampling อาจช่วยปรับสมดุลของการกระจายระหว่าง Positive class และ Negative class ในข้อมูลและช่วยเพิ่มประสิทธิภาพของการพัฒนาแบบจำลอง

3. Feature engineering การออกแบบ Feature ใหม่ๆหรือการแปลง Feature ที่มีอยู่แล้วอาจช่วยเพิ่มประสิทธิภาพในการแยกแยะความแตกต่างระหว่างแต่ละคลาสได้

4. การเลือก Algorithm ที่นอกเหนือจากนี้ อาจเหมาะสมกับการจัดการข้อมูลที่ไม่มีความแตกต่างระหว่างคลาส และอาจช่วยเพิ่มประสิทธิภาพของแบบจำลอง

5. การทำ Ensemble การรวมการทำนายของแบบจำลองหลายๆ ตัวเข้าด้วยกัน อาจช่วยปรับปรุงประสิทธิภาพโดยรวมของแบบจำลอง

บรรณานุกรม

1. Awoyemi JO, Adetunmbi AO, Oluwadare SA, editors. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 international conference on computing networking and informatics (ICCNI); 2017: IEEE.

2. Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A, editors. Credit card fraud detection-machine learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH); 2019: IEEE.

3. Xuan S, Liu G, Li Z, Zheng L, Wang S, Jiang C, editors. Random forest for credit card fraud detection. 2018 IEEE 15th international conference on networking, sensing and control (ICNSC); 2018: IEEE.

4. Kiran S, Guru J, Kumar R, Kumar N, Katariya D, Sharma M. Credit card fraud detection using Naïve Bayes model based and KNN classifier. International Journal of Advance Research, Ideas and Innovations in Technoloy. 2018;4(3):44.

5. Jain Y, Tiwari N, Dubey S, Jain S. A comparative analysis of various credit card fraud detection techniques. Int J Recent Technol Eng. 2019;7(5S2):402-7.

6. Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters. 2020;136:190-7.

ประวัติผู้เขียน