การพัฒนาแบบจำลองจำแนกลูกหนี้ที่มีความน่าจะเป็นในการชำระคืนหนี้กับลูกหนี้ที่มีความน่าจะเป็นในการไม่ชำระคืนหนี้ โดยใช้เทคนิคเหมืองข้อมูล

1. บทคัดย่อ

งานวิจัยฉบับนี้ได้นำเสนอแบบจำลองการจำแนกลูกหนี้สินเชื่อที่ดีกับไม่ดี โดยใช้เทคนิคเหมืองข้อมูล 4 เทคนิค คือ ต้นไม้ตัดสินใจ ตัวจำแนกของเบย์อย่างง่าย เพื่อนบ้านใกล้สุด และโครงข่ายประสาทเทียม โดยทดสอบกับชุดข้อมูลสินเชื่อและข้อมูลส่วนตัวของคนในประเทศเยอรมัน ซึ่งเป็นข้อมูลในวันที่ 17 พฤศจิกายน ค.ศ 1994

งานวิจัยฉบับนี้ได้พัฒนาแบบจำลองมาจากบทความวิจัย “การพัฒนาแบบจำลองการพิจารณาให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูล” ของคุณ ชลลดา ม่วงธนัง, คุณสุรศักดิ์ มังสิงห์ และคุณ นิเวศ จิระวิชิตชัย ซึ่งได้เผยแพร่บทความทางอินเทอร์เน็ต เมื่อวันที่ 25 มิถุนายน 2564
โดยได้มีการปรับสมดุลของข้อมูลให้จำนวนตัวอย่างของลูกหนี้ที่ดีกับลูกหนี้ที่ไม่ดีมีจำนวนอย่างละ 300 เท่ากัน เพื่อแก้ปัญหาความโน้มเอียงของตัวแบบที่มีโอกาสจะจำแนกตัวอย่างไปยังประเภทที่มีตัวอย่างมากกว่า จากการใช้เสียงส่วนใหญ่ หรือ ความน่าจะเป็นในการจำแนก

ผลการทดสอบพบว่าความแม่นยำของแบบจำลองโดยใช้เทคนิค โครงข่ายประสาทเทียมให้ความแม่นยำเฉลี่ยสูงที่สุดที่ 71.83% และตัวแปรต้น 5 อันดับที่มีความสัมพันธ์กับตัวแปรตามมากที่สุด คือ 1.สถานะของบัญชี/เงินเดือน (status of existing checking account) 2.ระยะเวลาการกู้ (duration in month) 3. ประวัติของเครดิต (credit history) 4.วงเงิน (credit amount) 5.ที่อยู่อาศัย (housing)

ที่มาและความสำคัญ

การสร้างแบบจำลองที่ใช้ในการประเมินลูกค้าที่จะมาขอสินเชื่อว่าควรจะอนุมัติสินเชื่อหรือไม่ และควรอนุมัติที่วงเงินเท่าไหร่ ควรคิดอัตราดอกเบี้ยเท่าไหร่ ต้องใช้หลักทรัพย์อะไรบ้างในการค้ำประกัน หรือแม้แต่การติดตามสินเชื่อที่ปล่อยไปแล้วว่ามีความเสี่ยงที่จะผิดนัดชำระหนี้หรือไม่ สิ่งเหล่านี้เป็นเรื่องสำคัญสำหรับธนาคารพาณิชย์ บริษัทให้สินเชื่อ และบริษัทอื่นๆที่เกี่ยวข้อง เพราะ ธนาคารพาณิชย์ บริษัทให้สินเชื่อมีรายได้หลักมาจากดอกเบี้ยให้สินเชื่อ ถ้าหากมีระบบการกรองลูกค้าที่จะปล่อยสินเชื่อไม่ดีก็จะทำบริษัทรับลูกค้ากลุ่มเสี่ยงเข้ามาและมีโอกาสที่จะเกิดหนี้เสียหรือหนี้ NPL เป็นจำนวนมาก ซึ่งหนี้เหล่านี้จะเป็นภาระทางเงินที่บริษัทต้องแบกรับต่อไป ดังนั้นการมีเครื่องมือหรือตัวชี้วัดการกรองลูกค้าที่ดี จึงเป็นเรื่องสำคัญ

ปัจจุบัน หนึ่งในเครื่องมือที่ธนาคารพาณิชย์และสถาบันการเงินต่างๆ ใช้ในการประเมินสิ่งเหล่านี้ คือ Credit Score หรือ คะแนนเครดิต ซึ่งจัดทำและเก็บข้อมูลโดย บริษัท ข้อมูลเครดิตแห่งชาติ จำกัด โดยมีระดับคะแนน 8 ระดับ คือ AA, BB, CC , DD, EE, FF, GG, HH เริ่มตั้งแต่ 300 จนถึง 900 ซึ่งระดับ AA หมายความว่ามีโอกาสในการผิดนัดชำระหนี้ต่ำสุด ส่วนระดับ HH หมายความว่ามีโอกาสในการผิดนัดชำระหนี้สูงสุด โดย Credit Score มาจาก Credit Scoring model หรือแบบจำลองที่ใช้กระบวนการทางสถิติในการสร้างแบบจำลองโดยใช้ข้อมูลประวัติการก่อหนี้ พฤติกรรมการชำระหนี้ และข้อมูลอื่นๆในอดีตของลูกค้า มาประเมินและคำนวณออกมาเป็นค่าคะแนน Credit Score ซึ่งเป็นวิธีการสร้างแบบจำลองแบบ Traditional statistical method คือใช้เทคนิค discriminant analysis, probit analysis และ logistic regression แต่ในงานวิจัยนี้จะใช้วิธีการสร้างแบบจำลองแบบ Advanced statistical method โดยใช้เทคนิคการทำเหมืองข้อมูล 4 เทคนิค คือ decision tree, k-nearest neighbors, naïve bayes classifier และ artificial neural networks ในการประเมินและจำแนกลูกค้าที่มาขอสินเชื่อเป็นลูกค้าที่มีความน่าจะเป็นในการชำระคืนหนี้สามารถปล่อยสินเชื่อได้ กับ ลูกค้าที่มีความน่าจะเป็นในการไม่ชำระคืนหนี้ ไม่ควรปล่อยสินเชื่อให้

2. งานที่เกี่ยวข้อง

ในการทำวิจัยครั้งนี้ผู้วิจัยได้ทำการพัฒนาแบบจำลอง โดยต่อยอดมาจากงานวิจัยที่เคยศึกษาแบบจำลองการพิจารณาการให้คะแนนสินเชื่อ หรือ credit scoring โดยใช้เทคนิคต่างๆ ซึ่งเป็นข้อมูลโดยย่อ ดังนี้

1.การพัฒนาแบบจำลองการพิจารณาให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูล

งานวิจัยได้นำเสนอแบบจำลองการให้คะแนนสินเชื่อ และทดสอบประสิทธิภาพของแบบจำลองการให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูล โดยทดสอบกับชุดข้อมูล 2 กลุ่ม คือ คะแนนสินเชื่อประเทศออสเตรเลียและประเทศเยอรมัน สร้างแบบจำลองการให้คะแนนสินเชื่อด้วยอัลกอริทึม 5 วิธี ได้แก่ ต้นไม้การตัดสินใจ ตัวจำแนกของเบย์อย่างง่าย โครงข่ายประสาท การถดถอยแบบโลจิสติก และซับพอร์ตเวคเตอร์แมชชีน ทำการทดสอบประสิทธิภาพของแบบจำลองจากค่าความแม่นยำ ผลการทดสอบพบว่าแบบจำลองที่ดีที่สุดคือโครงข่ายประสาทเทียมทั้ง 2 กลุ่มข้อมูล โดยแบบจำลองมีประสิทธิภาพสูงสุดได้ประสิทธิภาพ 90.14% กับชุดข้อมูลจากประเทศออสเตรเลีย และ 90.08% กับชุดข้อมูลจากประเทศเยอรมันตามลำดับ

2. Logistic regression and its application in credit scoring

Credit scoring is a mechanism used to quantify the risk factors relevant for an obligor's ability and willingness to pay. Credit scoring has become the norm in modern banking, due to the large number of applications received on a daily basis and the increased regulatory requirements for banks. In this study, the concept and application of credit scoring in a South African banking environment is explained, with reference to the International Bank of Settlement's regulations and requirements. The steps necessary to develop a credit scoring model is looked at with focus on the credit risk context, but not restricted to it. Applications of the concept for the whole life cycle of a product are mentioned. The statistics behind credit scoring is also explained, with particular emphasis on logistic regression. Linear regression and its assumptions are first shown, to demonstrate why it cannot be used for a credit scoring model. Simple logistic regression is first shown before it is expanded to a multivariate view. Due to the large number of variables available for credit scoring models provided by credit bureaus, techniques for reducing the number of variables included for modeling purposes is shown, with reference to specific credit scoring notions. Stepwise and best subset logistic regression methodologies are also discussed with mention to a study on determining the best significance level for forward stepwise logistic regression. Multinomial and ordinal logistic regression is briefly looked at to illustrate how binary logistic regression can be expanded to model scenarios with more than two possible outcomes, whether on a nominal or ordinal scale. As logistic regression is not the only method used in credit scoring, other methods will also be noted, but not in extensive detail. The study ends with a practical application of logistic regression for a credit scoring model on data from a South African bank.

3. A Comparative Study of Data Mining Techniques for Credit Scoring in Banking

Credit is becoming one of the most important incomes of banking. Past studies indicate that the credit risk scoring model has been better for Logistic Regression and Neural Network. The purpose of this paper is to conduct a comparative study on the accuracy of

classification models and reduce the credit risk. In this paper, we use data mining of enterprise software to construct four classification models, namely, decision tree, logistic regression, neural network and support vector machine, for credit scoring in banking. We conduct a systematic comparison and analysis on the accuracy of 17 classification models for credit scoring in banking. The contribution of this paper is that we use different classification methods to construct classification models and compare classification models accuracy, and the evidence demonstrates that the support vector machine models have higher accuracy rates and therefore outperform past classification methods in the context of credit scoring in banking

3. ขั้นตอนวิธีการวิจัย

ข้อมูลที่ใช้ในการวิจัย

ในการพัฒนาแบบจำลองจำแนกลูกค้าที่มาขอสินเชื่อโดยเทคนิคเหมืองข้อมูลนั้น ทางผู้วิจัยได้ที่ใช้ข้อมูลเครดิตพลเมืองของประเทศเยอรมัน จำนวน 1000 ตัวอย่าง ซึ่งมีข้อมูลทั้งหมด 21 ตัวแปร ประกอบด้วย ตัวแปรต้น 20 ตัวแปร คือ

1) สถานะของบัญชี/เงินเดือน (status of existing checking account)
2) ระยะเวลาการกู้ (duration in month)
3) ประวัติของเครดิต (credit history)
4) วัตถุประสงค์(purpose)
5) วงเงิน (credit amount)
6) บัญชีออมทรัพย์/พันธบัตร (savings account/bonds)
7) ระยะเวลาของการทำงานปัจจุบัน (present employment since)
8) อัตราการผ่อนชำระเป็นเปอร์เซ็นต์ของรายได้ (installment rate in percentage of disposable income)
9) สถานะและเพศ (personal status and sex)
10) ลูกหนี้/ผู้ค้ำประกัน (other debtors/guarantors)
11) ระยะเวลาการพักอาศัยของที่อยู่ปัจจุบัน (present residence since)
12) ทรัพย์สิน (property)
13) อายุ (age in years)
14) แผนการผ่อนชำระ (other installment plans)
15) ที่อยู่อาศัย (housing)
16) จำนวนเครดิตที่มีอยู่ของธนาคารนี้ (number of existing credits at this bank)
17) งาน (job)
18) จำนวนผู้รับผิดชอบในการดูแล (number of people being liable to provide maintenance for)
19) โทรศัพท์ (telephone)
20) แรงงานต่างด้าว (Foreign worker)

และตัวแปรตาม 1 ตัวแปร คือ ผลการให้คะแนนสินเชื่อ โดย 1 แปลว่า มีความน่าจะเป็นในการชำระคืนหนี้ สามารถปล่อยสินเชื่อได้ 2 แปลว่า มีความน่าจะเป็นในการไม่ชำระคืนหนี้ ไม่ควรปล่อยสินเชื่อให้

การเตรียมข้อมูลก่อนประมวลผล(Data Preproccessing)

ในข้อมูลตัวอย่าง 1000 ตัวอย่าง มีสินเชื่อที่ชำระหนี้คนให้มีค่าเป็น 1 แปลว่ามีตัวอย่างที่ชำระหนี้คืน 700 ตัวอย่าง และมีสินเชื่อไม่ที่ถูกชำระหนี้คนให้มีค่าเป็น 2 แปลว่ามีตัวอย่างที่ไม่ชำระหนี้คืน 300 ตัวอย่าง ดังนั้น เพื่อทำให้ตัวแบบไม่มีความเอนเอียงในการให้คะแนนสินเชื่อ class ที่มีจำนวนตัวอย่างมากกว่า อันเนื่องมาจากการใช้ majority vote หรือ probability ในจำแนก class ทางผู้วิจัยจึงกำหนดให้ตัวอย่างของทั้ง 2 class มีจำนวนตัวอย่างเท่ากัน class ละ 300 ตัวอย่าง โดยใช้ operator ที่ชื่อว่า sample ใน RapidMiner ในการสุ่มตัวอย่างให้ทั้ง 2 class เท่ากัน

จากนั้นใช้ operator ที่ชื่อว่า weight by correlation ในการเรียงตัวแปรต้นที่มีความสัมพันธ์กับตัวแปรตาม จากมากไปน้อย จากนั้นใช้ operator ที่ชื่อว่า select by weights กำหนด k=5 ในการเลือกตัวแปรต้นที่มีความสัมพันธ์กับตัวแปรตามมากที่สุด 5 อันดับแรก หลังจากนั้นจะมี 2 เทคนิคที่ต้องมีการปรับสเกลตัวแปรต้น คือ เพื่อนบ้านใกล้สุด และโครงข่ายประสาทเทียม โดยใช้ operator ที่ชื่อว่า nominal to numerical แปลงตัวแปรต้นเชิงคุณภาพให้กลายเป็นตัวแปรหุ่น และใช้ operator normalize โดยใช้วิธี range transformation หรือ min-max normalization ในการแปลงข้อมูลให้อยู่ในช่วง 0-1 ส่วนเทคนิค ต้นไม้ตัดสินใจ และ ตัวจำแนกของเบย์อย่างง่าย ไม่ต้องแปลงข้อมูลตัวแปรให้อยู่ในช่วง 0-1

แบบจำลองการจำแนกต้นไม้ตัดสินใจ (Decision Trees)

ใช้ gain ratio ในการแบ่งบัพ และหยุดแบ่งบัพเมื่อ ต้นไม้ตัดสินใจมีความลึกเท่ากับ 10 หรือมีจำนวนตัวอย่างเท่ากับ 4

แบบจำลองการจำแนกเพื่อนบ้านใกล้สุด (k-Nearest Neighbors : k-NN)

เนื่องจาก k หรือ ตัวอย่างที่ใกล้ที่สุดไม่ควรมากหรือน้อยเกินไปและควรน้อยกว่ารากที่สองของชุดข้อมูลฝึกฝน (ในที่นี้คือ 402) ทำให้ K ที่มากที่สุดที่เป็นไปได้คือ 19 ดังนั้นทางผู้วิจัยจึงได้ทำการทดความแม่นยำของตัวแบบที่ k=3,7,11,15,19

แบบจำลองตัวจำแนกของเบย์อย่างง่าย (Naïve Bayes Classifier)

ใช้ laplace correlation ในการจำแนก

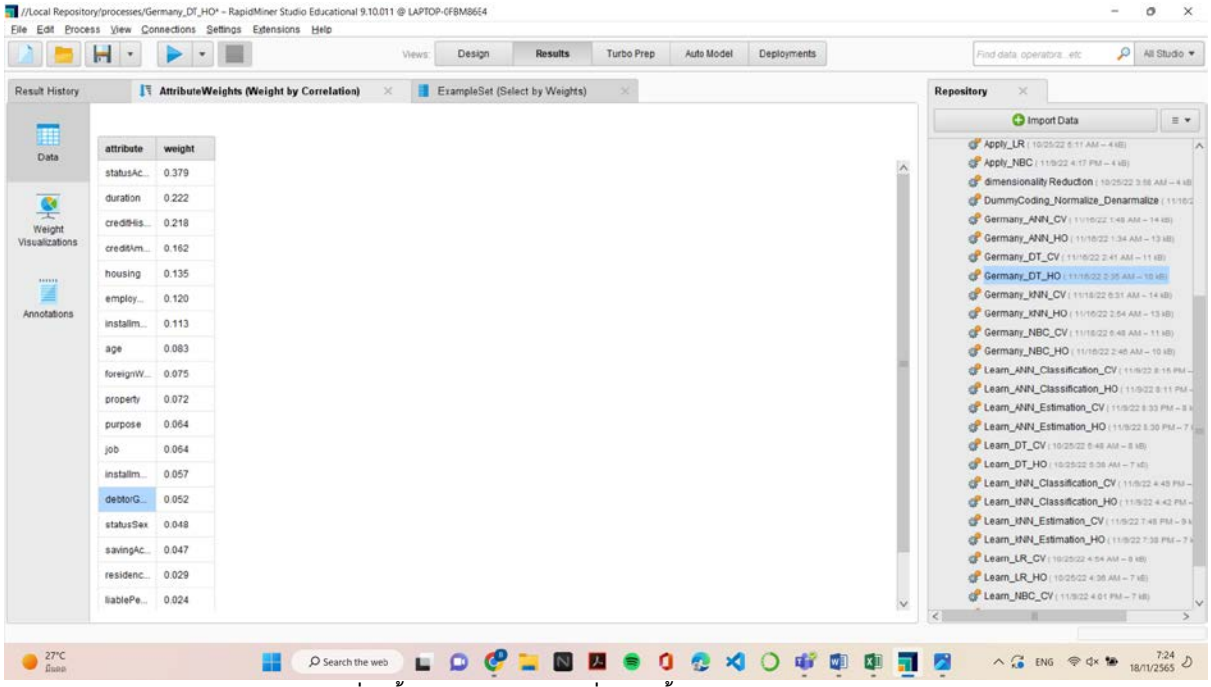
แบบจำลองการจำแนกโครงข่ายประสาทเทียม (Artificial Neural Networks : ANNs)

ใช้ค่า default ในการจำแนก คือ มี 1 hidden layer มีค่า epoch=200 และค่า learning rate=0.01

การทดสอบความถูกต้องของตัวแบบ (Model Validation)

ใช้การตรวจสอบไขว้ (Cross-Validation) ที่ k-fold cross validation ที่ 5 และ 10 ในการทดสอบ เนื่องจากมีชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบที่มีขนาดเล็ก เพื่อให้ตัวแบบได้เรียนรู้อย่างเพียงพอ ตัวแบบมีสมรรถนะในการจำแนกที่ดี และเพื่อให้ผลการตรวจสอบตัวแบบมีความน่าเชื่อถือ จึงเหมาะกับวิธีการตรวจสอบแบบไขว้มากกว่าการใช้การตรวจสอบแบบแบ่งข้อมูลบางส่วนเพื่อทดสอบภายหลัง

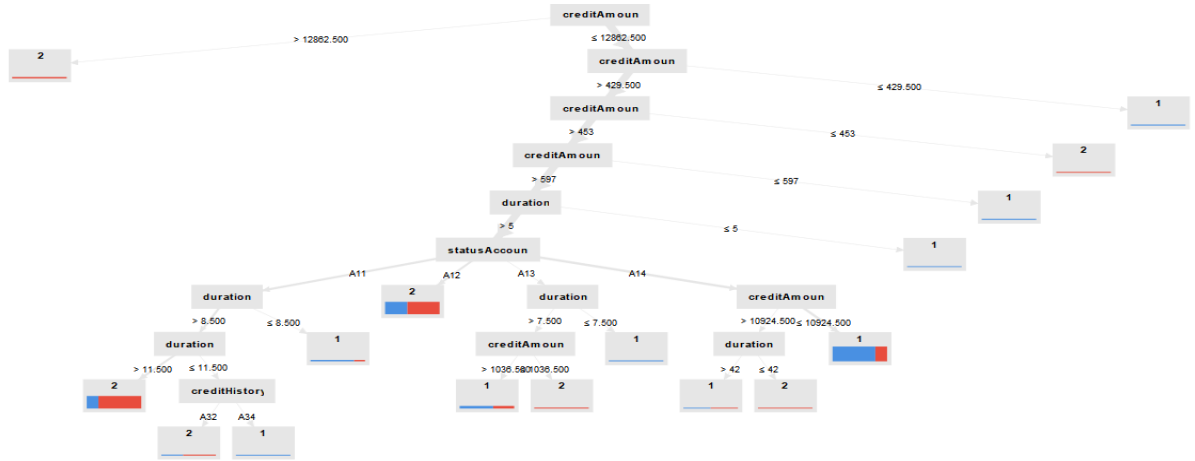
4. ผลการทดลอง



ภาพที่ 0 น้ำหนักของตัวแปรต้นเมื่อให้ค่าน้ำหนักด้วยค่า correlation

จากการหาความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรต้น โดยใช้ operator ที่ชื่อว่า weight by correlation และ select by weights ที่ k=5 พบว่าตัวแปรต้น 5 อันดับที่มีความสัมพันธ์กับตัวแปรตามมากที่สุด คือ 1.สถานะของบัญชี/เงินเดือน (status of existing checking account) 2. ระยะเวลาการกู้ (duration in month) 3. ประวัติของเครดิต (credit history) 4.วงเงิน (credit amount) 5.ที่อยู่อาศัย (housing)

แบบจำลองการจำแนกต้นไม้ตัดสินใจ



ภาพที่ 1 แบบจำลองต้นไม้ตัดสินใจ ที่ cross validation fold=10

accuracy: 69.00% +/- 4.92% (micro average: 69.00%)

	true 1	true 2	class precision
pred. 1	196	82	70.50%
pred. 2	104	218	67.70%
class recall	65.33%	72.67%	

ตารางที่ 1 ค่าความถูกต้องแม่นยำของแบบจำลองต้นไม้ตัดสินใจ ที่ cross validation fold=10

จากตารางที่ 1 จะได้ว่าแบบจำลองต้นไม้ตัดสินใจมีความถูกต้องแม่นยำเฉลี่ย 69% และมีความน่าจะเป็นในการชำระคืนหนี้ 65.33% และมีความน่าจะเป็นในการไม่ชำระคืนหนี้ 76.67%

แบบจำลองการจำแนกเพื่อนบ้านใกล้สุด

accuracy: 70.50% +/- 3.85% (micro average: 70.50%)

	true 1	true 2	class precision
pred. 1	219	96	69.52%
pred. 2	81	204	71.58%
class recall	73.00%	68.00%	

ตารางที่ 2 ค่าความถูกต้องแม่นยำของแบบจำลองเพื่อนบ้านใกล้สุด ที่ cross validation fold=5

จากการทดลองสุ่มค่า K=3,7,11,15,19 ที่ cross validation fold=5,10 พบว่าที่ค่า k=15 และ cross validation fold=5 ให้ค่าความแม่นยำที่สูงที่สุด 70.50% และมีความน่าจะเป็นในการชำระคืนหนี้ 65.33% กับมีความน่าจะเป็นในการไม่ชำระหนี้คืน 76.67% ดังตารางที่ 2

แบบจำลองตัวจำแนกของเบย์อย่างง่าย

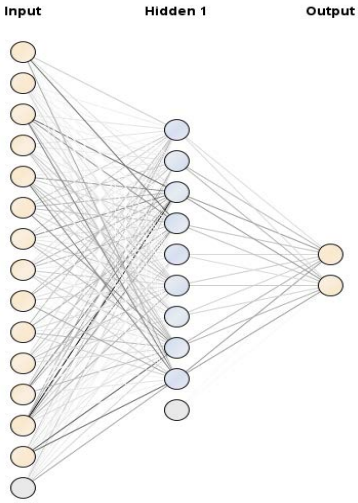
accuracy: 70.00% +/- 0.83% (micro average: 70.00%)

	true 1	true 2	class precision
pred. 1	229	109	67.75%
pred. 2	71	191	72.90%
class recall	76.33%	63.67%	

ตารางที่ 3 ค่าความถูกต้องแม่นยำของแบบจำลองตัวจำแนกของเบย์อย่างง่าย ที่ cross validation fold=5

จากตารางที่ 3 จะได้ว่าแบบจำลองตัวจำแนกของเบย์อย่างง่ายมีความถูกต้องแม่นยำเฉลี่ย 70% และมีความน่าจะเป็นในการชำระคืนหนี้ 76.33% และมีความน่าจะเป็นในการไม่ชำระคืนหนี้ 63.67%

แบบจำลองการจำแนกโครงข่ายประสาทเทียม



ภาพที่ 2 แบบจำลองโครงข่ายประสาทเทียม ที่ cross validation fold=5

accuracy: 71.83% +/- 2.60% (micro average: 71.83%)

	true 1	true 2	class precision
pred. 1	217	86	71.62%
pred. 2	83	214	72.05%
class recall	72.33%	71.33%	

ตารางที่ 4 ค่าความถูกต้องแม่นยำของแบบจำลองการจำแนกโครงข่ายประสาทเทียม ที่ cross validation fold=5

จากตารางที่ 4 จะได้ว่าแบบจำลองการจำแนกโครงข่ายประสาทเทียม มีความถูกต้องแม่นยำเฉลี่ย 71.83% และมีความน่าจะเป็นในการชำระคืนหนี้ 72.33% และมีความน่าจะเป็นในการไม่ชำระคืนหนี้ 71.33%

5. การอภิปรายและสรุปผล

จากการทดสอบแบบจำลองการจำแนกลูกหนี้ที่มีความน่าจะเป็นในการชำระคืนหนี้กับลูกหนี้ที่มีความน่าจะเป็นในการไม่ชำระคืนหนี้ โดยใช้เทคนิคทางเหมืองข้อมูล 4 เทคนิค พบว่า เทคนิคโครงข่ายประสาทเทียมให้ความแม่นยำสูงที่สุดที่ 71.83% ทั้งนี้การจะทำให้ตัวแบบจำลองมีค่าความแม่นยำที่สูงขึ้นนั้นขึ้นอยู่กับคุณภาพของตัวแปรตามด้วยซึ่งถ้าหากสามารถเก็บข้อมูลเชิงลึกของกลุ่มตัวอย่าง เช่น ข้อมูลอัตราส่วนทางการเงิน อย่างสภาพคล่องทางการเงินในกลุ่มที่ทำธุรกิจ หรือข้อมูลค่าใช้จ่ายต่อเดือนของกลุ่มตัวอย่างบุคคลทั่วไป เป็นต้น ก็จะทำให้ความแม่นยำของตัวแบบดีขึ้น และถ้าให้เพิ่มจำนวนตัวอย่างของข้อมูลในแต่ละ class ให้ได้อีกผลการวิจัยก็จะทำให้แบบจำลองนี้ดียิ่งขึ้นต่อไป

6. เอกสารอ้างอิง

[1] ชลลดำ ม่วงธนัง, สุรศักดิ์ มังสิงห์ และ นิเวศ จิระวิชิตชัย. (2564). การพัฒนาแบบจำลองการพิจารณให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูล. สืบค้นเมื่อ 19 พฤศจิกายน 2565, จาก https://ph02.tci-thaijo.org/index.php/scihcu/article/view/244011/165786

[2] อโนทัย พุทธารี, จิตรภณ หรูเจริญพรพานิช, เพ็ญสิริ บารุงเชาว์เกษม และชินวัฒน์ เทพหัสดิน ณ อยุธยา. (2561). Credit Scoring model : เครื่องมือในการประเมินคุณภาพสินเชื่อ. 19 พฤศจิกายน 2565, จาก https://www.bot.or.th/Thai/MonetaryPolicy/ArticleAndResearch/FAQ/FAQ_132.pdf

[3] Christine Bolton. (2552). Logistic regression and its application in credit scoring. สืบค้นเมื่อ 2552, จาก https://repository.up.ac.za/bitstream/handle/2263/27333/dissertation.pdf; sequence=1

[4] Shin-Chen Huang, Min-Yuh Day. (2556). A Comparative Study of Data Mining Techniques for Credit Scoring in Banking. สืบค้นเมื่อ 19 พฤศจิกายน 2565, จาก https://core.ac.uk/download/pdf/225228868.pdf

[5] Mr.Rushi Longadge, Ms.Snehlata S.Dongre, Dr.Latesh Malik. (2556). Class Imbalance Problem in Data Mining: Review. สืบค้นเมื่อ 19 พฤศจิกายน 2565, จาก https://arxiv.org/ftp/arxiv/papers/1305/1305.1707.pdf

[6] Sunil Kumar Cheruku. (2562). What is imbalanced dataset and its impacts on machine learning models?. สืบค้นเมื่อ 19 พฤศจิกายน 2565, จาก https://www.linkedin.com/pulse/what-imbalanced-dataset-its-impacts-machine-learning-models-cheruku