การพัฒนาแบบจำลองการพิจารณาให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูล

Development of a credit scoring model using data mining techniques

ชลลดา ม่วงธนัง[1*] สุรศักดิ์ มังสิงห์[1] และ นิเวศ จิระวิชิตชัย[2]

[1] สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีปทุม กรุงเทพมหานคร 10900

[2] สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์และเทคโนโลยี
สถาบันการจัดการปัญญาภิวัฒน์ นนทบุรี 11120

Chonlada Muangthanang[1*], Surasak Mungsing[1], Nivet Chirawichitchai[2]

[1] Program in Information and CommunicationTechnology, School of Information Technology, Sripatum University, Bangkok 10900

[2] Program in Computer Engineering, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi 11120

Received: 20 April 2021/ Revised: 20 June 2021/ Accepted: 25 June 2021

บทคัดย่อ

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

คำสำคัญ: ต้นไม้การตัดสินใจ การจำแนกกลุ่มนาอีฟเบย์ โครงข่ายประสาท การถดถอยแบบโลจิสติก ซับพอร์ตเวคเตอร์แมชชีน

Corresponding author: chonlada.mua@spulive.net

Abstract

The research presented a credit scoring model and tested the performance of the credit scoring model using data mining techniques. The tests were conducted on two datasets, including Australia and Germany credit score data sets. The credit scoring model using 5 algorithms: decision tree, Naïve Bayes classification, neural network, logistic regression and support vector machine. Perform a model performance test based on accuracy values. The experimental results showed that the best model was the neural network in both data groups, with the highest performing model performing 90.14% with the Australian dataset and 90.08% with the German dataset, respectively.

Keywords: Decision tree, Naïve Bayes classification, Neural network, Logistic regression, Support vector machine

บทนำ

การพัฒนาและยกมาตรฐานระดับสินเชื่อจะไม่สามารถดำเนินการได้อย่างมีประสิทธิภาพหากไม่ทราบสถานะทางสินเชื่อ (credit standing) ที่เกิดขึ้น การประสบปัญหาขาดข้อมูลสถานะทางสินเชื่อเพื่อการจัดการ จึงทำให้ส่งผลต่อปัญหาด้านการเข้าถึงเงินทุนและด้านการจัดการภายใน อันดับความน่าเชื่อถือ (credit rating) [1] เป็นการประเมินความน่าเชื่อถือของผู้ออกตราสารหนี้โดย “สถาบันจัดอันดับความน่าเชื่อถือ (credit rating agencies)” ที่ได้รับการรับรองจากสำนักงานกำกับหลักทรัพย์และตลาดหลักทรัพย์ (กลต.) [2] ในประเทศไทยมี 2 แห่งคือ บริษัท ทริสเรทติ้ง จำกัด และ บริษัท ฟิทช์ เรทติ้งส์ (ประเทศไทย) จำกัด โดยสถาบันจัดอันดับความน่าเชื่อถือจะวิเคราะห์จากผลการดำเนินงานและความเสี่ยงต่าง ๆ ที่มีผลกระทบต่อบริษัท การจัดอันดับเครดิตจึงเป็นข้อมูลหนึ่งที่สำคัญในการพิจารณาว่า ผู้ออกตราสาร ซึ่งมีสถานะเป็นลูกหนี้จะมีความสามารถในการจ่ายคืนมากน้อยแค่ไหน

บริษัท ข้อมูลเครดิตแห่งชาติ จำกัด (National Credit Bureau) [3] ที่มีหน้าที่เป็นตัวกลางในการจัดเก็บรวบรวมข้อมูลเครดิตซึ่งมีทั้งประวัติการชำระหนี้ที่ดีและไม่ดีของลูกค้าตามที่สถาบันการเงินหรือบริษัทที่เป็นสมาชิกจัดส่งให้ จึงได้พัฒนาเครื่องมือที่เรียกว่า แบบจำลองคะแนนเครดิต (credit scoring) [4] คือ เครื่องมือที่ใช้กระบวนการทางสถิติทำขึ้น เพื่อกำหนดตัวชี้วัดความน่าจะเป็นในการชำระคืนหนี้ โดยใช้ข้อมูลเฉพาะในส่วนที่ไม่สามารถระบุตัวของเจ้าของข้อมูลมาเป็นปัจจัยหนึ่งในการจัดทำข้อมูลเครดิตที่สร้างขึ้นมาดังกล่าว เพื่อช่วยลดปัญหาในการปล่อยกู้ให้แก่ผู้ที่ไม่พร้อมจะชำระหนี้คืน โดยจะบอกได้ว่าผู้กู้รายนั้นมีความน่าเชื่อถือขนาดไหน โดยที่ข้อมูลเครดิตถูกพัฒนาขึ้นจากข้อมูลธุรกรรมทางการเงินทั้งหมดในทุก ๆ สถาบันการเงินของลูกค้า ในอดีตตลอดช่วงระยะเวลา 3 ปี ว่ามีการใช้จ่ายเป็นอย่างไร มียอดการกู้ยืมอะไรบ้าง จำนวนเท่าไร มีการชำระเงินตรงเวลาไหม ผิดนัดชำระหรือไม่ หรือกระทั่งมีการขอวงเงินที่พร้อมจะกู้ยืมไว้มากน้อยขนาดไหน เป็นต้น [5] จากนั้นจึงนำข้อมูลมาสร้างแบบจำลองเพื่อคำนวณเป็นคะแนนออกมาใช้ในการตัดสินใจวิเคราะห์ลูกค้า ซึ่งเรียกว่าคะแนนเครดิต (credit score) เป็นตัวชี้วัดความน่าจะเป็นในการชำระคืนหนี้โดยใช้วิธีทางสถิติในการประมวลผลข้อมูล

วิธีดำเนินการวิจัย

ข้อมูลที่ใช้ในการวิจัย กลุ่มข้อมูลประกอบด้วย 2 กลุ่มข้อมูล ดังนี้ 1) กลุ่มข้อมูลเครดิตของพลเมืองประเทศออสเตรเลีย (Australian credit) [6] จำนวนกลุ่มตัวอย่าง 690 ตัวอย่าง ซึ่งมีข้อมูลทั้งหมด 14 attribute โดยมีความเกี่ยวข้องกับการสมัครบัตรเครดิต ชื่อและ attribute ทั้งหมดถูกเปลี่ยนเป็นสัญลักษณ์ที่ไม่มีความหมายเพื่อ

ป้องกันความลับของข้อมูล ประกอบด้วย ตัวแปรต้น 13 attribute และตัวแปรตาม 1 attribute คือ ผลการให้คะแนนสินเชื่อ โดย 0 แปลว่า มีความน่าจะเป็นในการชำระคืนหนี้ 1 แปลว่า มีความน่าจะเป็นในการไม่ชำระคืนหนี้ และ 2) กลุ่มข้อมูลเครดิตของพลเมืองประเทศเยอรมัน (German credit) [7] จำนวนกลุ่มตัวอย่าง 988 ตัวอย่าง ซึ่งมีข้อมูลทั้งหมด 21 attribute ประกอบด้วย ตัวแปรต้น 20 attribute คือ

1) สถานะของบัญชี (status of existing checking account)
2) ระยะเวลา (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)

โดยตัวแปรตาม 1 attribute คือ ผลการให้คะแนนสินเชื่อ โดย good แปลว่า มีความน่าจะเป็นในการชำระคืนหนี้ bad แปลว่า มีความน่าจะเป็นในการไม่ชำระคืนหนี้

การพัฒนาแบบจำลอง ข้อมูลการพัฒนาตัวแบบจำลอง มี 5 ตัวแบบ ดังนี้

1. ตัวแบบจำลองต้นไม้การตัดสินใจ (decision tree) เป็นการรวบรวมของโหนดการตัดสินใจ (decision node) แล้วเชื่อมต่อกันด้วยกิ่งก้านต่าง ๆ (branches) ขยายออกจากโหนดราก (root node) ไปจนถึงจุดสิ้นสุดที่โหนดใบ (leaf node) โดยจุดเริ่มต้นอยู่ที่โหนดรากซึ่งอยู่ตำแหน่งบนสุดของต้นไม้ตัดสินใจ ในแต่ละโหนดการตัดสินใจแสดงการทดสอบคุณลักษณะ (attribute) ของข้อมูล โดยแต่ละกิ่งก้านแสดงผลลัพธ์ที่เป็นไปได้จากการทดสอบ ในแต่ละกิ่งก้านจะนำไปสู่โหนดการตัดสินใจอีกโหนดหนึ่ง หรือสิ้นสุดอยู่ที่โหนดใบซึ่งแสดงคำตอบหรือกลุ่ม (class) ที่กำหนดไว้ [8] ต้นไม้การตัดสินใจสร้างโดยอัลกอริทึมต้นไม้การจำแนกและการถดถอย (Classification and Regression Trees Algorithm : CART) ซึ่งเป็นไบนารี่ ประกอบด้วยกิ่งก้านสำหรับแต่ละโหนดการติดสินใจ 2 กิ่ง CART แบ่งแยกระเบียนในชุดข้อมูลออกเป็นระเบียนที่มีค่าเหมือนกันสำหรับคุณลักษณะที่เป็นค่าเป้าหมาย CART ทำการขยายต้นไม้โดยดำเนินงานสำหรับแต่ละโหนดการตัดสินใจ ค้นหาตัวแปรที่เป็นประโยชน์และค่าในการแบ่งแยกที่เป็นไปได้ทั้งหมด แล้วเลือกการแบ่งแยกที่เหมาะสมที่สุดตามเกณฑ์ดังสมการต่อไปนี้ [9]

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#classes} |P(j|t_L) - P(j|t_R)| \quad (1)$$

t_L แทน โหนดทางซ้ายของโหนด t

t_R แทน โหนดทางขวาของโหนด t

$$P_L = \frac{\text{จำนวนของระเบียนที่โหนด } t_L}{\text{จำนวนของระเบียนในชุดข้อมูลทั้งหมด}}$$

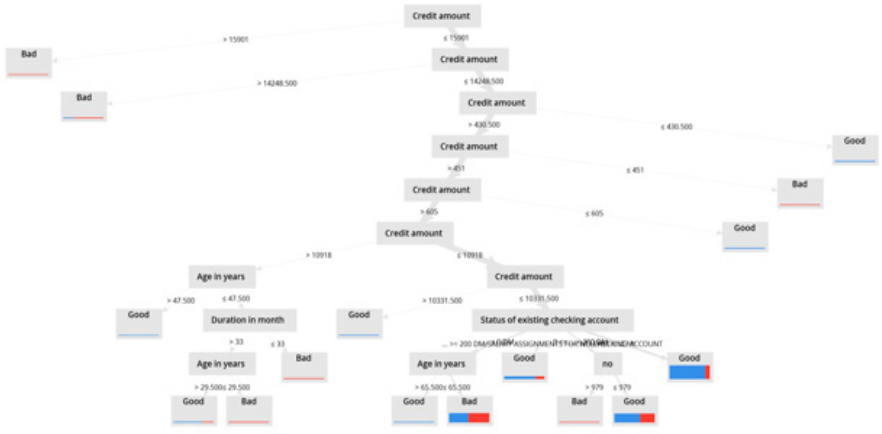
$$P_R = \frac{\text{จำนวนของระเบียนที่โหนด } t_R}{\text{จำนวนของระเบียนในชุดข้อมูลทั้งหมด}}$$

$$P(j|t_L) = \frac{\text{จำนวนของคำตอบ j และระเบียนที่โหนด } t_L}{\text{จำนวนของระเบียนที่โหนด } t_L}$$

$$P(j|t_R) = \frac{\text{จำนวนของคำตอบ j และระเบียนที่โหนด } t_R}{\text{จำนวนของระเบียนที่โหนด } t_R}$$

ดังนั้นการแบ่งแยกที่เหมาะสมที่สุดคือการแบ่งแยกที่ทำให้การวัด นี้มีค่ามากที่สุดสำหรับการแบ่งแยกที่เป็นไปได้ทั้งหมดที่โหนด t แสดงดังภาพที่ 1 และ 2

ภาพที่ 1 แบบจำลอง decision tree จากกลุ่มข้อมูลของ Australian credit



ภาพที่ 2 แบบจำลอง decision tree จากกลุ่มข้อมูลของ German credit

2. ตัวแบบจำลองการจำแนกกลุ่มนาอีฟเบย์ (Naïve Bayes classification) โดยทั่วไปค่าของความน่าจะเป็นที่จำเป็นในการคำนวณเพื่อหาการจำแนกกลุ่มภายหลังสูงสุดมีจำนวน ค่า เมื่อ k เป็นจำนวนของคำตอบสำหรับตัวแปรเป้าหมาย และ m เป็นจำนวนของตัวแปรทำนาย [10] สำหรับตัวจำแนกกลุ่มภายหลังสูงสุดกล่าวอีกอย่างหนึ่งคือให้นำจำนวนตัวแปรทำนายคูณกับจำนวนค่าที่แตกต่างกันของตัวแปรเป้าหมาย [11] ดังสมการต่อไปนี้

$$\theta_{NB} = \arg \max_{\theta} \prod_{i=1}^{m} p(X_i=x_i|\theta)p(\theta) \qquad (2)$$

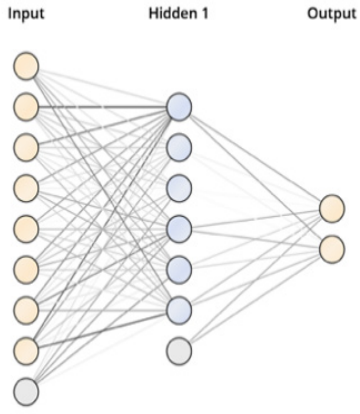
3. ตัวแบบจำลองโครงข่ายประสาท (neural network) ประเภท Multi - Layer Perceptron (MLP) ลักษณะของโครงข่ายประสาท คือ การจดจำระบบการเรียนรู้ที่ซับซ้อนในสมองของสัตว์ได้ ซึ่งประกอบด้วยชุดของเซลล์ประสาท (neuron) ที่เชื่อมต่อกัน แม้ว่าเซลล์ประสาทเซลล์หนึ่งอาจจะมีโครงสร้างอย่างง่าย โครงข่ายของเซลล์ประสาทที่เชื่อมต่อกันสามารถทำการเรียนรู้งานที่ซับซ้อน [8] ดังสมการต่อไปนี้

$$net_j = \sum_{i} W_{ij}x_{ij} = W_{0j}x_{0j} + W_{1j}x_{1j} + \ldots + W_{Ij}x_{Ij} \qquad (3)$$

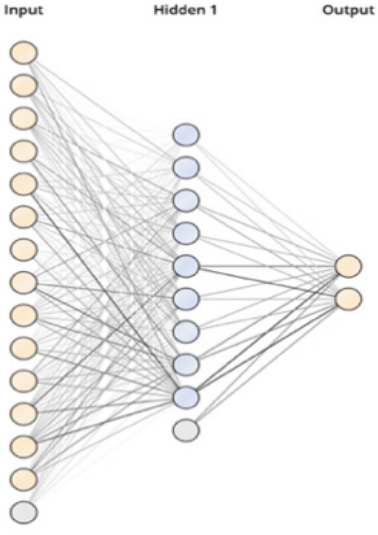
x_{ij} แทน ข้อมูลเข้าที่ i ไปยังโหนด j

W_{ij} แทน น้ำหนักถ่วงที่สัมพันธ์กับข้อมูลเข้าที่ i ไปยังโหนด j และมีข้อมูลเข้าจำนวน I+1 ไปยังโหลด j

จากสมการข้างต้นจึงได้แก่ การจำแนกและการจดจำรูปแบบได้ จะทำการเลียนแบบการเรียนรู้ที่ไม่ใช่เชิงเส้นตรงซึ่งเกิดขึ้นในโครงข่ายของเซลล์ประสาทที่พบในธรรมชาติ แสดงดังภาพที่ 3 และ 4



ภาพที่ 3 แบบจำลอง neural network จากกลุ่มข้อมูลของ Australian credit



ภาพที่ 4 แบบจำลอง neural network จากกลุ่มข้อมูลของ German credit

4. ตัวแบบจำลองการถดถอยแบบโลจิสติก (logistic regression model) ในการวิเคราะห์การถดถอยโลจิสติกเป็นวิธีการสำหรับอธิบายความสัมพันธ์ระหว่างตัวแปรผลตอบสนองเชิงกลุ่มตัวแปรเดียวและตัวแปรทำนายหลายตัว [11] ดังสมการต่อไปนี้

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad (4)$$

5. ตัวแบบจำลองซับพอร์ตเวคเตอร์แมชชีน (support vector machine) เป็นตัวจำแนกเชิงเส้น (linear classifier) แบบ 2 คลาส ซึ่งเป็นที่ยอมรับถึงประสิทธิภาพของการจำแนกที่เหนือว่าวิธีการจำแนกอื่นๆ ข้อได้เปรียบของ SVM คือ มีประสิทธิภาพในการจำแนกข้อมูลที่มีมิติจำนวนมากได้ นอกจากนี้การใช้ฟังก์ชันเคอร์เนล (Kernel function) เพื่อแปลงข้อมูลไปยังมิติที่สูงขึ้นในปริภูมิคุณลักษณะ (feature space) สามารถจำแนกข้อมูลที่มีความคลุมเครือได้อย่างมีประสิทธิภาพ [12] ดังสมการต่อไปนี้ [13]

$$+1 \text{ ถ้า } w \cdot x + b >= +1 \quad (5)$$

$$-1 \text{ ถ้า } w \cdot x + b <= -1 \quad (6)$$

$$\text{ถ้า } -1 < w \cdot x + b < +1 \quad (7)$$

การวิเคราะห์ข้อมูลจากการวิจัย วิธีการวิเคราะห์การเกาะกลุ่ม และวิธีการจำแนกข้อมูลจากแบบจำลองเพื่อวัดค่าความถูกต้องแม่นยำ (accuracy) ของข้อมูลได้เลือกใช้ operators ของแต่ละวิธีจากโปรแกรม RapidMiner Studio [14-16] มีจำนวน 5 วิธีการ ดังนี้

1. วิธีการวิเคราะห์จากแบบจำลองต้นไม้การตัดสินใจ (decision tree) (ตารางที่ 1 และ 2)

ตารางที่ 1 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง decision tree จากกลุ่มข้อมูลของ Australian credit

	true 0	true 1	class precision
pred. 0	324	15	95.58%
pred. 1	59	292	83.19%
class recall	84.60%	95.11%	

accuracy: 89.28%

ตารางที่ 2 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง decision tree จากกลุ่มข้อมูลของ German credit

	true Good	true Bad	class precision
pred. Good	689	285	70.74%
pred. Bad	0	14	100.00%
class recall	100.00%	4.68%	

accuracy: 71.15%

2. วิธีการวิเคราะห์จากแบบจำลองการจำแนกแบบเบย์อย่างง่าย (Naïve Bayes) (ตารางที่ 3 และ 4)

ตารางที่ 3 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง Naïve Bayes Classification จากกลุ่มข้อมูลของ Australian credit

	true 0	**true 1**	**class precision**
pred. 0	349	99	77.90%
pred. 1	34	208	85.95%
class recall	91.12%	67.75%	

accuracy: 80.72%

ตารางที่ 4 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง Naïve Bayes classification จากกลุ่มข้อมูลของ German credit

	true Good	**true Bad**	**class precision**
pred. Good	513	100	84.15%
pred. Bad	158	199	55.74%
class recall	77.07%	66.56%	

accuracy: 73.89%

3. วิธีการวิเคราะห์จากแบบจำลองข่ายงานประสาท (neural net) ประเภท Multi - Layer Perceptron (MLP) (ตารางที่ 5 และ 6)

ตารางที่ 5 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง neural net ประเภท Multi - Layer Perceptron (MLP) จากกลุ่มข้อมูลของ Australian credit

	true 0	**true 1**	**class precision**
pred. 0	354	39	90.08%
pred. 1	29	268	90.24%
class recall	92.43%	87.30%	

accuracy: 90.14%

ตารางที่ 6 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง neural net ประเภท Multi - Layer Perceptron (MLP) จากกลุ่มข้อมูลของ German credit

	true Good	**true Bad**	**class precision**
pred. Good	669	78	89.56%
pred. Bad	20	221	91.70%
class recall	97.10%	73.91%	

accuracy: 90.08%

4. วิธีการวิเคราะห์จากแบบจำลองสมการถดถอยแบบโลจิสติก (logistic regression) (ตารางที่ 7 และ 8)

ตารางที่ 7 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง logistic regression จากกลุ่มข้อมูลของ Australian credit

	true 0	true 1	class precision
pred. 0	364	75	82.92%
pred. 1	19	232	92.43%
class recall	95.04%	75.57%	

accuracy: 86.38%

ตารางที่ 8 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง logistic regression จากกลุ่มข้อมูลของ German credit

	true Good	true Bad	class precision
pred. Good	654	200	76.58%
pred. Bad	35	99	73.88%
class recall	94.92%	33.11%	

accuracy: 76.21%

5. วิธีการวิเคราะห์จากแบบจำลองซับพอร์ตเวคเตอร์แมชชีน (support vector machine) (ตารางที่ 9 และ 10)

ตารางที่ 9 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง support vector machine จากกลุ่มข้อมูลของ Australian credit

	true 0	true 1	class precision
pred. 0	306	23	93.01%
pred. 1	77	284	78.67%
class recall	79.90%	92.51%	

accuracy: 85.51%

ตารางที่ 10 ค่าความถูกต้องแม่นยำของข้อมูลแบบจำลอง support vector machine จากกลุ่มข้อมูลของ German credit

	true Good	true Bad	class precision
pred. Good	620	158	79.69%
pred. Bad	69	141	67.14%
class recall	89.99%	47.16%	

accuracy: 77.02%

ผลการวิจัย

ในการพัฒนาแบบจำลองการศึกษาที่ได้มีข้อมูลทำให้เห็นว่า ข้อมูลจาก Australian แบ่งออกเป็นมีความน่าจะเป็นในการชำระคืนหนี้ (0) และมีความน่าจะเป็นในการไม่ชำระคืนหนี้ และข้อมูลจาก German แบ่งออกเป็นมีความน่าจะเป็นในการชำระคืนหนี้ มีความน่าจะเป็นในการไม่ชำระคืนหนี้ ได้ผลการศึกษากลุ่มข้อมูลดังตารางที่ 11-12

ตารางที่ 11 เปรียบเทียบค่าความน่าจะเป็นในการชำระและไม่ชำระคืนหนี้ กลุ่มข้อมูล Australian credit จากการจำแนกข้อมูลของแบบจำลอง

Model	ความน่าจะเป็นในการคืนหนี้(%)	
	ชำระ	ไม่ชำระ
Decision Tree	84.60	95.11
Naïve Bayes Classification	91.12	67.75
Neural Net (MLP)	92.43	87.30
Logistic Regression	95.04	75.57
Support Vector Machine	79.90	92.51

จากข้อมูลของกลุ่ม Australian credit มีความน่าจะเป็นในการชำระคืนนี้สูงที่สุดจากการจำแนกข้อมูลของแบบจำลอง logistic regression คือ 95.04% คิดเป็น 656 ตัวอย่างจากจำนวนกลุ่มตัวอย่างทั้งหมด 690 ตัวอย่าง และมีความน่าจะเป็นในการไม่ชำระคืนนี้สูงที่สุดจากการจำแนกข้อมูลของแบบจำลอง Decision Tree คือ 95.11% คิดเป็น 657 ตัวอย่างจากจำนวนกลุ่มตัวอย่างทั้งหมด 690 ตัวอย่าง

ตารางที่ 12 เปรียบเทียบค่าความน่าจะเป็นในการชำระและไม่ชำระคืนหนี้ กลุ่มข้อมูล German Credit จากการจำแนกข้อมูลของแบบจำลอง

Model	ความน่าจะเป็นในการคืนหนี้ (%)	
	ชำระ	ไม่ชำระ
Decision Tree	100.00	4.68
Naïve Bayes Classification	77.07	66.56
Neural Net (MLP)	97.10	73.91
Logistic Regression	94.92	33.11
Support Vector Machine	89.99	47.16

จากข้อมูลของกลุ่ม German credit มีความน่าจะเป็นในการชำระคืนนี้สูงที่สุดจากการจำแนกข้อมูลของแบบจำลอง decision tree คือ 100.00% คิดเป็น 988 ตัวอย่างจากจำนวนกลุ่มตัวอย่างทั้งหมด 988 ตัวอย่าง และมีความน่าจะเป็นในการไม่ชำระคืนนี้สูงที่สุดจากการจำแนกข้อมูลของแบบจำลอง neural net คือ 73.91% คิดเป็น 731 ตัวอย่างจากจำนวนกลุ่มตัวอย่างทั้งหมด 988 ตัวอย่าง

อภิปรายและสรุปผลการวิจัย

งานวิจัยได้นำเสนอแบบจำลองการให้คะแนนสินเชื่อ และทดสอบประสิทธิภาพของแบบจำลองการให้คะแนนสินเชื่อโดยใช้เทคนิคเหมืองข้อมูล โดยทดสอบกับชุดข้อมูล 2 กลุ่มคือ คะแนนสินเชื่อประเทศออสเตรเลียและประเทศเยอรมัน สร้างแบบจำลองการให้คะแนนสินเชื่อด้วยอัลกอริทึม 5 วิธี ได้แก่ ต้นไม้การตัดสินใจ การจำแนกกลุ่มนาอีฟเบย์ โครงข่ายประสาท การถดถอยแบบโลจิสติก และซับพอร์ตเวคเตอร์แมชชีน ทำการทดสอบประสิทธิภาพของแบบจำลองจากค่าความแม่นยำ ผลการทดสอบพบว่าแบบจำลองที่ดีที่สุดคือโครงข่ายประสาทเทียมทั้ง 2 กลุ่มข้อมูล โดยแบบจำลองมีประสิทธิภาพสูงสุดได้ประสิทธิภาพ 90.14% กับชุดข้อมูลจากประเทศออสเตรเลีย และ 90.08% กับชุดข้อมูลจากประเทศเยอรมันตามลำดับ ประสิทธิภาพความแม่นยำในการจำแนกข้อมูลของแบบจำลองที่ได้ศึกษาได้ผลการศึกษาจากกลุ่มข้อมูล Australian credit และ German credit (ตารางที่ 13)

ตารางที่ 13 เปรียบเทียบค่าความแม่นยำในการจำแนกข้อมูลของแบบจำลอง

Model	Accuracy (%)	
	Australia	German
Decision Tree	89.28	71.15
Naïve Bayes Classification	80.72	73.89
Neural Net (MLP)	90.14	90.08
Logistic Regression	86.38	76.21
Support Vector Machine	85.51	77.02

จากข้อมูลของกลุ่มมีความแม่นยำในการจำแนกข้อมูลสูงที่สุดจากแบบจำลอง neural net ทั้ง 2 กลุ่มข้อมูล คือ Australian credit คิดเป็น 90.14% และ German credit คิดเป็น 90.08%

เอกสารอ้างอิง

1. สมาคมตลาดตราสารหนี้ไทย (ThaiBMA). อันดับเครดิต (credit rating). [อินเทอร์เน็ต]. 2560 [เข้าถึงเมื่อ 11 มิ.ย. 2562]. เข้าถึงได้จาก: http://www.thaibma.or.th/EN/Investors/Individual/Blog/CreditRating.aspx
2. ตลาดหลักทรัพย์แห่งประเทศไทย. การเข้าจดทะเบียนตราสารหนี้. [อินเทอร์เน็ต]. 2562. [เข้าถึงเมื่อ 12 มิ.ย. 2562]. เข้าถึงได้จาก: https://www.set.or.th/th/products/listing/bond_listing_p2.html
3. บริษัท ข้อมูลเครดิตแห่งชาติ จำกัด (National Credit Bureau). สารพันน่ารู้เรื่องเครดิตบูโร. [อินเทอร์เน็ต]. 2559 [เข้าถึงเมื่อ 12 มิ.ย. 2562]. เข้าถึงได้จาก: https://www.ncb.co.th/all-about-credit-bureau
4. ธนาคารแห่งประเทศไทย (Bank of Thailand). ความรู้เบื้องต้นเกี่ยวกับ credit scoring. [อินเทอร์เน็ต]. 2559 [เข้าถึงเมื่อ 12 มิ.ย. 2562]. เข้าถึงได้จาก: https://www.creditinfocommittee.or.th/Thai/Pages/News.aspx
5. บริษัท CMSK จำกัด (CMSK Academy). Credit scoring เครื่องมือเช็คความน่าเชื่อถือในการชำระหนี้. [อินเทอร์เน็ต]. 2562 [เข้าถึงเมื่อ 12 มิ.ย. 2562]. เข้าถึงได้จาก: https://cmsk-academy.com/article/360/credit-scoring

6. Statlog (Australian credit approval) data set. California: The UCI machine learning repository. [อินเทอร์เน็ต]. [เข้าถึงเมื่อ 11 ม.ค. 2563]. เข้าถึงได้จาก: http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)
7. Statlog (German credit data) data set. California: The UCI machine learning repository. [อินเทอร์เน็ต]. 2537 [เข้าถึงเมื่อ 11 ม.ค. 2563]. เข้าถึงได้จาก: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
8. สายชล สินสมบูรณ์ทอง. การทำเหมืองข้อมูล เล่ม 1: การค้นหาความรู้จากข้อมูล data mining 1: discovering knowledge in data. พิมพ์ครั้งที่ 2. กรุงเทพฯ: จามจุรีโปรดักส์; 2560.
9. Kennedy RL, Lee Y, Roy BV, Reed CD, Lippman RP. Solving data mining problems through pattern recognition. New Jersey: Pearson Education, Inc.; 1995.
10. Hand D, Mannila H, Smyth P. Principles of data mining. Cambridge: MIT Press; 2001.
11. สายชล สินสมบูรณ์ทอง. การทำเหมืองข้อมูล เล่ม 2: วิธีการและตัวแบบ data mining 2: methods and models. กรุงเทพฯ: จามจุรีโปรดักส์; 2559.
12. ณิชกุล วงศ์เกษกิต. ซัพพอร์ตเวกเตอร์แมชชิน (Support Vector Machine: SVM). [อินเทอร์เน็ต]. 2563 [เข้าถึงเมื่อ 1 ต.ค. 2563]. เข้าถึงได้จาก: https://knowledge.snru.ac.th/ซัพพอร์ตเวกเตอร์แมชชิน/
13. ศาสตรา วงศ์ธนวสุ. ปัญญาประดิษฐ์ ทฤษฎี โปรแกรม และการประยุกต์. ขอนแก่น: มหาวิทยาลัยขอนแก่น; 2560.
14. นิเวศ จิระวิชิตชัย. แบบจำลองการตรวจสอบการทุจริตสำหรับข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการลดมิติข้อมูลร่วมกับการเรียนรู้ของเครื่อง. วารสารวิจัย มทร.ธัญบุรี 2563;10(1);215-25.
15. เสถียร จันทร์ปลา, ปรัชญนันท์ นิลสุข. มหาวิทยาลัยอัจฉริยะ: แนวทางการดำเนินงานของมหาวิทยาลัยอัจฉริยะ. วารสารวิชาการ การจัดการเทคโนโลยี มหาวิทยาลัยราชภัฏมหาสารคาม 2562;6(1):147-58.
16. RapidMiner Studio. USA: RapidMiner. [อินเทอร์เน็ต]. 2563 [เข้าถึงเมื่อ 11 ม.ค. 2563]. เข้าถึงได้จาก: https://rapidminer.c