เปรียบเทียบผลลัพธ์ของการอนุมัติสินเชื่อด้วย 3 แบบจำลองของระเบียบวิธี Machine Learning โดยใช้โปรแกรมอาร์

Comparative result of credit approval with 3 models of machine learning algorithm using program R

สุเมธ จุฑาจันทร์[1] และ สมพร ปั่นโภชา[2]

[1] สาขาวิศวกรรมการเงิน คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยหอการค้าไทย, sumet.jutha@gmail.com

[2] คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยหอการค้าไทย, sompon_punpocha@yahoo.com

---

บทคัดย่อ

นวัตกรรมการเงินในปัจจุบันนักวิทยาศาสตร์ข้อมูล (Data Scientist) ได้มีนำเอาระบบปัญญาประดิษฐ์ (Artificial Intelligence: AI) ช่วยวิเคราะห์ข้อมูลขนาดใหญ่โดยผ่านการเรียนรู้ของเครื่อง (Machine Learning: ML) เพื่อให้การบริการลูกค้ามีความสะดวกและรวดเร็วมากขึ้น เช่น การนำเอไอ (AI) มาใช้ในการพิจารณาวงเงินกู้แทนเจ้าหน้าที่หรือการอนุมัติวงเงินฉุกเฉินของบัตรเครดิตภายใน เป็นต้น เนื่องจากข้อจำกัดในการเข้าถึงและใช้ข้อมูลในประเทศไทย ดังนั้นในการศึกษาวิจัยผู้วิจัยได้ใช้ข้อมูลของเจอมันเครดิต (German Credit ที่มาจาก UC Irvine Machine Learning Reposity) โดยใช้โปรแกรมอาร์เป็นเครื่องมือในการสร้างระเบียบวิธีแมชชีนเลิร์นนิง

ผลการศึกษาพบว่า แบบจำลองKNN (k-Nearest Neighbors)มีประสิทธิภาพสูงที่สุด โดยมี ค่าความถูกต้อง (Accuracy Score), ค่าความแม่นยำ(Precision Score), ค่ารีคอล (Recall) และค่าเอฟ-1สกอร์(F1-Score) คือ 99.33%, 99.03%, 100% และ 99.51% ตามลำดับ รองลงมาคือแบบจำลองโลจิสติค รีเกรสชัน(Logistic Regression)โดยมีค่าความถูกต้อง (Accuracy Score), ค่าความแม่นยำ (Precision Score), ค่ารีคอล (Recall) และ ค่าเอฟ-1 สกอร์ (F1-Score) คือ 73.00%, 75.00%, 90.73% และ 82.12% ตามลำดับ และแบบจำลองคาร์ท โมเดล (CART Model) โดยมีค่าความถูกต้อง (Accuracy Score), ค่าความแม่นยำ (Precision Score), ค่ารีคอล(Recall) และค่าเอฟ-1 สกอร์ (F1-Score) คือ 69.00%, 87.32%, 72.76% และ 79.38% ตามลำดับ

คำสำคัญ: Credit Approval Analysis, Machine Learning, KNN (k-Nearest Neighbors), Accuracy Score, Precision Score, Recall, F1-Score

ABSTRACT

At present of financial innovations, data scientists have used Artificial Intelligence (AI) to analyze big data with Machine Learning (ML) for serve customers to more convenient and faster such as using AI to analyze credit approval that acts on behalf of the financial officer or to analyze credit card emergency loan approval etc. Due to limited access to and use of the data in Thailand. So the objective of this study was to study machine learning in financial data and measure efficiency of 3 ML algorithms of the Credit Approval Analysis by using German Credit data, from the UC Irvine Machine Learning Reposity, and using R Programming as a tool to create Machine Learning.

The results showed that the KNN model had the most efficient, with Accuracy Score, Precision Score, Recall and F1-Score as 99.33%, 99.03%, 100% and 99.51%, respectively, followed by Logistic Regression simulated with Accuracy Score, Precision Score, Recall and F1-Score as 73.00%, 75.00%, 90.73% and 82.12% respectively, whereas CART Model with Accuracy Score 69.00%, Precision Score 87.32%, Recall 72.76% and F1-Score 79.38%.

Keywords: Credit Approval Analysis, Machine Learning, KNN (k-Nearest Neighbors), Accuracy Score, Precision Score, Recall, F1-Score

1. บทนำ

ปัจจุบันการบริหารจัดการข้อมูลขนาดใหญ่เป็นสิ่งจำเป็นและมีความสำคัญมากขึ้นเรื่อยๆ จึงทำให้สายอาชีพ “Data Scientist” หรือนักวิทยาศาสตร์ข้อมูล ซึ่งเป็นผู้ที่ต้องมีความสามารถในการประยุกต์ใช้ความรู้ทางคณิตศาสตร์และสถิติช่วยในการวิเคราะห์ และจัดการฐานข้อมูลขนาดใหญ่ เป็นที่รู้จักและได้รับความสนใจและเป็นที่ต้องการเป็นอย่างมากในสายอาชีพทางการเงิน ซึ่งผู้วิจัยมีความสนใจในศาสตร์และการประกอบอาชีพสายงาน การบริหารจัดการข้อมูล (Data) จึงมุ่งเน้นศึกษาถึงความเชื่อมโยงข้อมูลทางการเงินและความรู้ทางคณิตศาสตร์และสถิติที่มี โดยจากการค้นคว้าข้อมูลและพบว่ามีโครงงานที่หลากหลายที่มีการใช้ปัญญาประดิษฐ์เข้ามาประยุกต์ใช้ทางการเงิน หนึ่งในนั้นคือการใช้ในการพิจารณาการอนุมัติสินเชื่อ และเป็นการพัฒนาหนึ่งในนวัตกรรมทางการเงินที่ใช้ความรู้ความเข้าใจและศาสตร์ของวิทยาศาสตร์ข้อมูลหรือ Data Scientist มาใช้ทางการเงินซึ่งในปัจจุบันคงหนีไม่พ้นการใช้ระบบ AI / ML มาอำนวยความสะดวกและเร่งในการประมวลผลสำหรับให้บริการลูกค้าด้านต่างๆ แทนมนุษย์ โดยเฉพาะอย่างยิ่ง การนำ AI มาใช้ในการพิจารณาวงเงินกู้แทนเจ้าหน้าที่ ตั้งแต่การอนุมัติวงเงินฉุกเฉินของบัตรเครดิตภายในไม่กี่นาที

การพิจารณาการให้สินเชื่อบุคคลนั้นขึ้นอยู่กับ ข้อมูลการเงินในอดีตของแต่ละบุคคลนั้นๆ เพื่อให้การพิจารณานั้นมีความแม่นยำและถูกต้องมากขึ้นจำเป็นต้องมีข้อมูลพื้นฐานของผู้ขอสินเชื่ออาทิเช่น อายุ, เพศ,รายได้ การวิเคราะห์ด้านเครดิตเกี่ยวข้องกับการใช้ศาสตร์ของสถิติ การวิเคราะห์เชิงปริมาณ การวัดข้อมูลเชิงคุณภาพ เพื่อตรวจสอบหาความเป็นไปได้ในการพิจารณาการอนุมัติการกู้ยืม แต่เดิมภาระการสร้างหลักเกณฑ์การพิจารณาเป็นของเจ้าหน้าที่การพิจารณาสินเชื่อ แต่ด้วยเทคโนโลยีในปัจจุบันความพร้อมในการประมวลผลข้อมูลที่ใหญ่ขึ้น (Big Data) ความรู้ทางด้านการเรียนรู้ของเครื่องจักรหรือที่เรียกกันว่า Machine Leaning (ML) สามารถสร้างความเป็นไป

ได้ที่สมเหตุสมผล ดังที่กล่าวไปแล้วข้างต้นคือข้อมูลถึงข้อมูลหรือ Feature ที่จำเป็นที่ใช้ในการสร้าง ML โดยการใช้การเรียนรู้แบบมีผู้สอน หรือ Supervised Machine Learning สร้างแบบจำลองโดยมีเบื้องหลังเป็น ระเบียบวิธีทางคณิตศาสตร์เพื่อมาจำแนกหาความน่าเชื่อถือของผู้มาขอสินเชื่อโดยจัดในกลุ่มต่อไปนี้คือ 1 (Creditability = 1 : น่าเชื่อถือ อนุมัติให้กู้ผ่าน) , และ (Creditability = 0 : ไม่มีความน่าเชื่อถือเพียงพอ อนุมัติให้กู้ไม่ผ่าน)

1.1 **KNN (k-Nearest Neighbors)**

KNN เป็นหนึ่งใน ML ที่สามารถเข้าใจได้ง่ายและทำงานได้ดีในเชิงของการปฏิบัติงานจริงช่วยแก้ปัญหาได้ทั้ง classification และ regression เป็นระเบียบวิธีที่ไม่มีพารามิเตอร์(non-parametric algorithm) หมายความว่าไม่มีการสร้างสมมติฐานเกี่ยวกับการกระจายตัวของความน่าจะเป็นของข้อมูลตัวอย่าง ข้อมูลในโลกแห่งความเป็นจริงอาจไม่ตรงกันในทางทฤษฎี KNN จึงเป็นอีกหนึ่งแบบจำลองที่มีประโยชน์ในการใช้งาน

การวัดความใกล้ชิดด้วยการวัดระยะทาง KNN มีการคำนวณระยะระหว่างจุดที่ต้องการแยกประเภทหรือจัดกลุ่ม การพยากรณ์นั้นจะถูกกำหนดโดยการโหวตด้วยเสียงส่วนใหญ่ โดยส่วนใหญ่กำหนดให้เป็นเลขคี่ระยะห่างระหว่างการสังเกตของ i และ j สามารถวัดได้หลายทิศทาง ใช้ Euclidean distance ในการคำนวณหาระยะทาง โดยมีสูตรการคำนวณดังสมการที่ 1.1 นี้คือ (Dr. N.D Lewis, 2017)

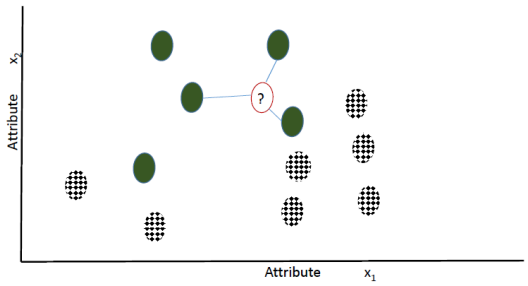
$$\mathcal{D}(x_i, x_j) = \sqrt{\sum_{m=1}^{n} (x_{im} - x_{jm})^2} \quad (1.1)$$

เมื่อ n เป็นจำนวนของ feature

สูตรการหาระยะเมตริกซ์ที่เป็นที่นิยมอีกอันนั่นก็คือ **Manhattan distance ตามสมการ 1.2** (Dr. N.D Lewis, 2017)

$$\mathcal{D}(x_i, x_j) = \sum_{m=1}^{n} |x_{im} - x_{jm}|; \quad (1.2)$$

จากการศึกษาการกระจายของข้อมูลนั้นคล้ายคลึงกับรูปที่ 1.1 ดังตัวอย่างที่ยกมา (Dr. N.D Lewis, 2017)



(Dr. N.D Lewis. 2017. Machine Learning Made Easy With R)

รูปที่ 1.1 การกระจายตัวของความน่าจะเป็นของข้อมูลตัวอย่าง

ข้อดีของ KNN ทำได้ง่าย โดยไม่มีแบบจำลองทางคณิตศาสตร์ภายใต้ ระเบียบวิธี ที่ซับซ้อน ง่ายสำหรับการทำงานและการทำความเข้าใจ เหมาะกับข้อมูลที่ไม่มีความซับซ้อนมากมายนัก เหมาะสำหรับการเริ่มต้นเรียนรู้การทำ ML (Dr. N.D Lewis. 2017. Machine Learning Made Easy With R) (Dr. N.D Lewis, 2017)

ข้อเสียของ KNN ยังไม่ใช่ statistic model ที่แท้จริง เพราะตอน train ข้อมูลทำได้แค่การนำข้อมูลเป็น storage ไว้ไม่ได้มีการเรียนรู้ใดๆที่แน่ชัดจริงๆ ยกตัวอย่างเช่น ถ้าพิจารณาโมเดลรูปแมวแล้วแมวทุกตัวหันไปทางซ้าย หากมีข้อมูลใหม่เข้ามาโดยมีแมวหันไปทางขวาแบบจำลองจะแยกไม่ออกทันทีว่าเป็นแมว และมีความล่าช้ากว่าแบบจำลองอื่นเช่น Logistic Regression หรือ CART (Dr. N.D Lewis, 2017)

1.2 Logistic Regression เป็นการถดถอยชนิดพิเศษที่ตัวแปรเป้าหมายคือ Binary เช่น “Yes” หรือ “No” ลักษณะของ Feature เป็นแบบ discrete หรือ continuous ก็ได้ เช่นเดียวกับการถดถอยเชิงเส้นก็สามารถทำได้เช่น การพยากรณ์ การศึกษาเชิงพรรณนาหรือการทดสอบสมมติฐาน แต่จำคำนึงถึงความแม่นยำของการพยากรณ์เป็นหลัก การวัดผลของโมเดลที่มีลักษณะ Binary จึงนิยมใช้ Confusion matrix ในการวัดผลดังกล่าว (Dr. N.D Lewis, 2017)

Logistic regression จะรับค่า 0 หรือ 1 หรือที่เป็นไบนารี(binary) สามารถทำได้โดยใช้แบบจำลองการถดถอยโลจิสติก(Logistic Regression Model) **ตามสมการ 1.3** (Dr. N.D Lewis, 2017)

$$\ln\left(\frac{p(y)}{1-p(y)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{1.3}$$

เป็นการเปลี่ยนแปลงจากสมการด้านบนดังกล่าวของ linear regression ไปเป็น target variable ที่เป็นไบนารี

Odds Ratio เมื่อมีการแบ่งชั้น(class)แบบไบนารีแล้ว มักสนใจในส่วนของความน่าจะเป็นที่ทำการสังเกตในแบบเฉพาะเจาะจงของชั้นนั้นๆ **การคำนวณ odd ratio ตามสมการ 1.4** (Dr. N.D Lewis, 2017)

$$Odds = \frac{p(y)}{1-p(y)} \tag{1.4}$$

Log odds ratio Natural logarithm ของ odds ratio เรียกว่า log odds ratio หรือ “logit” ทีนี้มาตรวจสอบถึง odd ratio ของ 1 **ตามสมการ 1.5** (Dr. N.D Lewis, 2017)

$$\begin{aligned} &\ln\left(\frac{p(y)}{1-p(y)}\right) = 1 \\ &\Rightarrow p(y) = [1-p(y)] \\ &\Rightarrow p(y) + p(y) = 1 \\ &\Rightarrow p(y) = 0.5 \end{aligned} \tag{1.5}$$

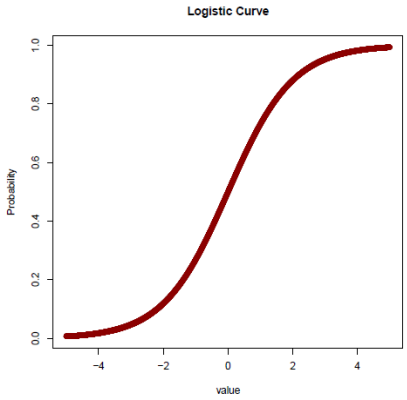
odd ratio เท่ากับ 1 แสดงถึงความน่าจะเป็นว่า y =1 คือ 0.5 logistic regression models logit คือ ฟังก์ชันเชิงเส้น ของตัว feature นั่นเอง พยายามจำลอง p(y) เป็นการถดถอยเชิงเส้น อย่างไรก็ตามสิ่งนี้ทำให้เกิดความน่าจะเป็นนอกช่วง [0,1] odd function เองรับค่าระหว่าง 0 และ ∞ เมื่อ take natural log เข้าไปยัง odds function จะได้รับช่วงของค่าตั้งแต่ -1 ถึง 1 ได้ (Dr. N.D Lewis, 2017)

- ถ้า p > 0.5 , odds ratio > 1, และ log odds ratio จะเป็นบวก
- ถ้า p < 0.5 , odds ratio < 1, แบะ log odd ratio จะเป็นลบ

p คือค่าความน่าจะเป็น จะเห็นว่า การถดถอยโลจิสติกส์ ถูกสร้างเพื่อรับค่า p ในระหว่าง $0 \leqslant p \leqslant 1$

The Logistic Curve หรือ Sigmoid function เป็นฟังก์ชั่นที่จับความสัมพันธ์ของตัวแปร binary และ features จากข้อมูลที่ใช้ในการพิจารณานั้นๆ สามารถคำนวณได้จาก **ตามสมการ 1.6** (Dr. N.D Lewis, 2017)

$$p(y) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad (1.6)$$



(Dr. N.D Lewis. 2017. Machine Learning Made Easy With R)

รูปที่ 1.2 Logistic Curve

ค่าของ $\propto$ และ β เป็นตัวกำหนดตำแหน่งและการกระจายของเส้นโค้ง Logistic

Relationship to Logistic Regression

ถ้าใช้ Logistic Regression กับ feature x เขียนสมการได้**ตามสมการที่ 1.7**

$$\log\left(\frac{p(y)}{1 - p(y)}\right) = \alpha + \beta x \quad (1.7)$$

เขียนให้อยู่ในพจน์ของ odds ได้**ตามสมการ 1.8**

$$\frac{p(y)}{1 - p(y)} = \exp(\alpha + \beta x) \quad (1.8)$$

จัดรูปโดยสนใจใน term ของ P(y) ได้**ตามสมการ 1.9**

$$p(y) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}, \quad (1.9)$$

ข้อดีของ Logistic Regression เนื่องจากลักษณะที่เรียบง่ายและมีประสิทธิภาพไม่ต้องการพลังในการคำนวณใช้งานง่ายตีความได้ง่ายซึ่งนักวิเคราะห์ข้อมูลและนักวิทยาศาสตร์ใช้กันอย่างแพร่หลาย ยังไม่ต้องการการปรับขนาดของคุณสมบัติ การถดถอยโลจิสติกให้คะแนนความน่าจะเป็นสำหรับการสังเกต (Dr. N.D Lewis, 2017)

ข้อเสียของ Logistic Regression การถดถอยโลจิสติกไม่สามารถจัดการกับคุณลักษณะ / ตัวแปรเชิงหมวดหมู่จำนวนมากได้ นอกจากนี้ไม่สามารถแก้ปัญหาที่ไม่เป็นเชิงเส้นด้วยการถดถอยโลจิสติกส์ได้นั่นคือเหตุผลว่าทำไมจึงต้องมีการเปลี่ยนแปลงคุณสมบัติที่ไม่ใช่เชิงเส้น การถดถอยโลจิสติกจะทำงานได้ไม่ดีกับตัวแปรอิสระที่ไม่มีความสัมพันธ์กับตัวแปรเป้าหมายและมีความคล้ายคลึงหรือสัมพันธ์กันมา (Dr. N.D Lewis, 2017)

1.3 CART Model เป็นโมเดลที่สามารถใช้ได้ทั้งการจำแนกประเภทและการถดถอย ความแตกต่างนั้นจะอยู่ใน Target Variable การจำแนกประเภท เช่นการ หาค่า Yes หรือ No เช่น หาว่าฝนจะตกในวันพรุ่งนี้หรือไม่ การถดถอยเช่น การทำนายราคาบ้าน (Dr. N.D Lewis, 2017)

Classification Tree หลักการของ Classification Tree เหมือนกับ Regression Tree แตกต่างกันแค่ เปลี่ยน cost function จาก RSS เป็น Gini impurity หรือ Entropy เพื่อความเหมาะสมกับปัญหา classification (Dr. N.D Lewis, 2017)

Gini Impurity เป็นการวัดความไม่บริสุทธิ์ หรือความไม่เพียวของ class ในแต่ละกลุ่มข้อมูลที่แบ่งตามแต่ละ split point… สำหรับปัญหา classification แบบ binary ที่มี target variable เป็น 0 หรือ 1 การ split ที่ดี ควรจะได้กลุ่มข้อมูลออกมา 2 กลุ่มที่สามารถแยก class 0 กับ class 1 ออกมาได้ชัดเจนในแต่ละกลุ่ม ยิ่งสามารถแบ่งแยกชั้นของ target variable ออกมาได้ดี ค่า Gini impurity : G ก็จะยิ่งต่ำ **ตามสมการ 1.10** (Dr. N.D Lewis, 2017)

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \qquad (1.10)$$

Entropy เป็นการวัดความไม่แน่นอน (randomness) ของข้อมูล เช่น การโยนเหรียญด้วย fair coin ที่มีโอกาสเกิดหัว/ก้อย ที่ 50% และ 50% ก็จะมีค่า Entropy โดยคำนวณจาก Eq-3 เท่ากับ -(1/2*log2(1/2) + 1/2*log2(1/2)) = 1 **ตามสมการ 1.11** ซึ่งถือว่าเป็นค่า entropy ที่สูงที่สุด เพราะไม่สามารถคาดเดาเหตุการณ์ที่ไม่มี bias แบบนี้ได้ (ถ้าเกิดว่าเหรียญไม่ใช่ fair coin และมีการเอนเอียง(bias) เช่น โอกาสออกหัว 90% แปลว่าสามารถคาดเดาเหตุการณ์ได้ง่ายว่าจะออกหัวมากกว่าออกก้อย) (Dr. N.D Lewis, 2017)

การจำแนกประเภทของแบบจำลอง ต้องการทำนายชั้นของ target variable ให้แม่นยำ หมายความว่าต้องการลดความไม่แน่นอนให้น้อยที่สุด นั่นก็คือการพยายามแยกชั้นของ target variable ให้ได้สัดส่วนของ ชั้นใด ชั้นหนึ่งมากที่สุด เพื่อเพิ่มความแน่นอนในการทำนาย (เพิ่ม certainty ลด randomness เหมือนกรณีโยนเหรียญที่มี bias) (Dr. N.D Lewis, 2017)

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \qquad (1.11)$$

กำหนดให้ ชั้น ของ target variable มีทั้งหมด K class (กรณี binary classification คือ K=2), p_hat_mk = สัดส่วนหรือ % ของ class k ภายในกลุ่ม => คล้ายๆกับ Gini impurity คือ ถ้าในกลุ่ม หรือใน node ที่แบ่งออกมาได้สามารถแยกชั้นของ target variable ออกมาได้ 1 ชั้นแบบเพียวๆ ทำให้ค่า entropy = 0 เนื่องจาก ค่า p_hat_mk ของ ชั้นนั้นมีค่าเท่ากับ 1 (log 1 = 0) ส่วน ชั้นอื่นๆ มีค่าเท่ากับ 0 (ทำให้ค่านอกวงเล็บ = 0) (Dr. N.D Lewis, 2017)

cost function ทั้ง 2 ตัวมีจุดประสงค์เหมือนกัน คือ พยายามทำให้การ split node แต่ละครั้ง ได้กลุ่ม observation ออกมาให้มีความ pure ของ class ใด class หนึ่งใน target variable มากที่สุด (Dr. N.D Lewis, 2017)

ข้อดีของ Cart Model ถูกออกแบบมาเพื่อให้รองรับข้อมูลแบบไม่เป็นเชิงเส้นให้ผล Trained Model ที่ถูกตีความออกมาให้ง่าย,สามารถทำได้ทั้ง Regression และ Classification (Dr. N.D Lewis, 2017)

ข้อเสียของ Cart Model ข้อมูลที่มีความแปรปรวนสูง ทำให้เกิดการแตกกิ่งได้ง่าย ทำให้มีความแม่ยำที่ต่ำเมื่อทำการทดสอบแบบจำลอง **และ**ไม่เหมาะกับข้อมูลที่เป็นลักษณะการกระจายแบบเชิงเส้น ธรรมชาติของระเบียบวิธีของ CART คือการค้นหา พารามิเตอร์ที่ดีที่สุดจากบนลงล่างโดยตัดสินใจจากเงื่อนไขที่พบในลำดับชั้นปัจจุบันเท่านั้น โดยไม่ได้ตรวจสอบว่าการตัดสินใจในชั้นนั้นส่งผลให้ลำดับชั้นล่างๆ ลงมามีค่าความไม่บริสุทธิ์น้อยที่สุดหรือไม่ เรียกพฤติกรรมแบบนี้ว่า Greedy algorithm ซึ่งพฤติกรรมแบบนี้ทำให้ต้องใช้เวลานานจนแทบเป็นอนันต์จึงจะหาต้นไม้ที่ดีที่สุดได้ ในทางคณิตศาสตร์เรียกปัญหานี้ว่า ปัญหา NP-Complete ดังนั้นเมื่อใช้ การตัดสินใจแบบกิ่งไม้

(Decision tree) จึงต้องยอมรับผลลัพธ์ที่อาจจะไม่สมบูรณ์แบบ แต่ก็สามารถใช้ได้กับงานส่วนใหญ่ (Dr. N.D Lewis, 2017)

จากงานวิจัย “Credit Approval Analysis” โดย Abiola Smith, Brendan Maher และ Deepesh Khaneja โดยวิจัยดังกล่าวมีการเปรียบเทียบร้อยละของความแม่นยำของ ML โดยพบว่า MLที่มีความแม่นยำ ที่มากที่สุด 3 ลำดับ คือ Classification and Regresstion Tree (CART) 87.2% , Logistic Regression 85.6%, K-Nearest Neighbors 85.6% ดังนั้นผู้วิจัยจึงเลือก แบบจำลอง ML 3 โมเดลข้างต้นมาทำการศึกษา และเนื่องจากมีข้อจำกัดในการเข้าถึงและใช้ข้อมูล ผู้ศึกษาจึงใช้ข้อมูลสำเร็จรูปของ German Credit ที่มาจาก UC Irvine Machine Learning Reposity โดยมีการใช้ R Programming เป็นเครื่องมือในการสร้างระเบียบวิธี ML เพื่อศึกษาแบบจำลองทางคณิตศาสตร์และเปรียบประสิทธิภาพความแม่นยำของการอนุมัติสินเชื่อทั้ง 3 แบบจำลอง และทำการศึกษาถึงโมเดลทางคณิตศาสตร์ที่อยู่ภายในระเบียบวิธี ML ของ แต่ละแบบจำลอง และ เปรียบเทียบผลลัพธ์ของตาราง Confusion Matrix ที่เป็นตัววัดการมีประสิทธิภาพของ ML ได้แก่ ค่าความถูกต้อง (Accuracy Score) ค่าความแม่นยำ (Precision Score) ค่าความถูกต้องของการทำนายว่าจะอนุมัติสินเชื่อเทียบกับจำนวนครั้งของเหตุการณ์ทั้งหมด (Recall) และค่าเฉลี่ยแบบฮาร์มอนิกระหว่าง Precision และ Recall (F1-Score) (Dr. N.D Lewis, 2017)

2. วัตถุประสงค์การวิจัย

1) เพื่อศึกษาประสิทธิภาพของแบบจำลองการพิจารณาการอนุมัติสินเชื่อจาก 3 แบบจำลอง ได้แก่ K-nearest neighbors algorithm Logistic Regression และ CART Model

2) เพื่อศึกษาแบบจำลองทางคณิตศาสตร์ที่อยู่เบื้องหลังระเบียบวิธี ML

3) เพื่อเปรียบเทียบผลลัพธ์ของ Confusion Matrix ของแต่ละแบบจำลอง

3. การดำเนินการวิจัย

1. เก็บรวบรวมข้อมูล ซึ่งในที่นี้ใช้ข้อมูลสำเร็จรูปของ German Credit จาก UC Irvine Machine Learning Reposity

2. ทำความสะอาดข้อมูล (Cleansing data) ให้ถูกต้องพร้อมใช้ Cleansing ข้อมูลให้ถูกต้องพร้อมใช้ โดยในชุดข้อมูลที่นำมาวิเคราะห์ อาจเกิดปัญหาโดยข้อมูลอาจจะไม่ครบหรือมีข้อมูลสูญหาย ซึ่งทำให้เกิดการสรุปหรือการวิเคราะห์ไม่ถูกต้อง หรือส่งผลให้เกิดค่า คาดเคลื่อน (Error) เมื่อนำข้อมูลไป Train เพื่อสร้าง model ในกระบวนการ Machine Learning

3. ทำ Data Dictionary

4. หา Feature ของแบบจำลอง โดยใช้เทคนิค low variance technique อธิบายข้อมูลเชิงสถิติของ feature หรือแต่ละตัวแปรที่ใช้ในแบบจำลองด้วยการทำ data visualizationหรือเรียกอีกอย่างว่า variable selection หรือ attribute selection เป็นการคัดเลือกหรือกรองเอาข้อมูลบางตัวแปร (บางคอลัมน์) ออกมาใช้ ซึ่งปัจจุบันมีหลายเทคนิคด้วยกัน วิธีที่นำมาใช้คือวิธี Low Varince Filter คือตัวแปรหรือคอลัมน์ใด มีข้อมูลที่ค่าความแปรปรวนต่ำที่สุด พิจารณาลบหรือไม่ใช้คอลัมน์นั้นไป ซึ่งข้อมูลที่มีสเกลขนาดใหญ่ มีค่าความแปรปรวนสูง ดังนั้นจึงต้องทำการ

normalize data เป็นการเปลี่ยนสเกลให้เป็น 0 – 1 เสียก่อน โดยการ normalization มีได้หลายวิธี Min – Max Normalization: ซึ่ง เป็นวิธีที่ง่ายที่สุด **ตาม**

สมการ 3.1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.1)$$

จากนั้นจึงทำการหาค่า ความแปรปรวน และเรียงลำดับจากค่ามากไปน้อย

จากขั้นตอนการ feature โดยใช้ Low Variance Filter ที่เลือกใช้คือ Telephone, Account Balance, Value Savings/Stocks , Instalment percent, Duration in Current address , No of dependents , Concurrent Credits , Most valuable available asset

Variable	Description
Creditability	1 = approve , 0 = not approved
Account Balance	1 = no running account , 2 = no balance or debit , 3 = 0-200 DM , 4 > 200 Dm or checking for at least 1 year
Value Savings/Stocks	1 = not available / no savings , 2 = < 100,- DM , 3 = 100,- <= ... < 500,- DM , 4 = 500,- <= ... < 1000,- DM, 5 = >= 1000,- DM
Instalment percent	1 = >= 35 , 2 = 25 <= ... < 35 , 3 = 20 <= ... < 25 , 4 = < 20
Duration in Current address	1 = < 1 year , 2 = 1 <= ... < 4 years , 3 = 4 <= ... < 7 years , 4 = >= 7 years
Most valuable available asset	1 = not available / no assets , 2 = Car / Other , 3 = Savings contract with a building society / Life insurance ,

	4 = Ownership of house or land
Concurrent Credits	1 = at other banks , 2 = at department store or mail order house , 3 = no further running credits
No of dependents	1 = 3 and more , 2 = 0 to 2
Telephone	1 = no , 2 = yes

5. สร้างการเรียนรู้ข้อมูล(Train Data) โดยใช้ สัดส่วน 70% ในการทำระเบียบวิธี ML แต่ละโมเดล

6. ทดสอบข้อมูล (Test Data) โดยใช้สัดส่วน 30% ในการทำระเบียบวิธี ML แต่ละโมเดลในที่นี้จะกำหนดค่าสุ่มเทียมเป็น 123

7 .เปรียบประสิทธิภาพความแม่นยำของการอนุมัติสินเชื่อทั้ง 3 แบบจำลองด้วยตาราง Confusion Matrix และทำการศึกษาถึงโมเดลทางคณิตศาสตร์ที่อยู่ภายในระเบียบวิธี ML แต่ละจำลองตาราง confusion matrix เป็นตารางเพื่อใช้ประเมินผลลัพธ์ของการทำงานของแบบจำลองโดยผลลัพธ์ของการทำนายว่าผลของการป้อน feature เข้าไปเพื่อหาการพิจารณาการอนุมัติสินเชื่อบุคคลจาก feature creditability ที่ label ไว้แล้ว ตามการเรียนรู้ของ ML แบบ Supervised Learning หรือการเรียนรู้แบบมีผู้สอน โดยมีค่าจาก Confusion Matrix ดังต่อไปนี้คือ Accuracy Score, Precision Score, Recall, F1-score

Accuracy Score คือสัดส่วนเปอร์เซ็นต์ความถูกต้อง คือ จำนวนที่ทำนายถูก/จำนวนทั้งหมด **ตามสมการ 3.2**

$$Accuracy = \frac{Correct\ Predictions}{All\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.2)$$

Precision Score คือความแม่นยำของผลทำนายจะสนใจผลทำนายหรือPredictionคำนวณเป็นค่าสัดส่วนกี่ % **ตามสมการ 3.3**

$$\textbf{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3.3)$$

Recall คือ การวัดค่าความแม่นยำอีกมิติหนึ่ง ที่สนใจผลลัพธ์กับที่เป็นของจริงด้วย เช่น มองว่าโมเดลนั้นทำนายถูกต้องกี่ % เมื่อเทียบกันกับของจริง **ตามสมการ 3.4**

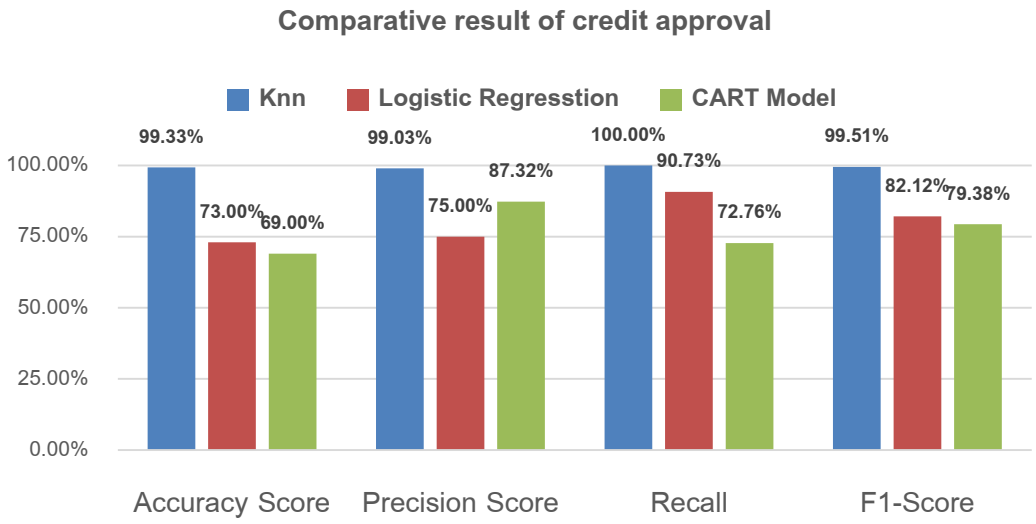
$$\textbf{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3.4)$$

F1-Score คือ ค่าที่แสดงประสิทธิภาพ โดยการนำค่า Precision และ Recall มาคำนวณหาค่าเฉลี่ย หรือเรียกว่า Harmonic Mean ซึ่งค่าสูงๆถือว่า Model มีประสิทธิภาพดี **ตามสมการ 3.5**

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (3.5)$$

4. ผลการวิจัย

จากการศึกษาโดยใช้ชุดข้อมูลสำเร็จรูป German Credit โดยมีการเก็บข้อมูลลูกค้าสินเชื่อกลุ่มหนึ่ง จากนั้นนำชุดข้อมูลที่ได้เช็คความถูกต้องของข้อมูลหาค่าผิดพลาด(Missing)ก่อนเข้าสู่การเลือก Feature จำเป็นต้องทำการปรับสเกลข้อมูลโดยการทำ Data Normalization ก่อนจากนั้นทำการเลือกใช้ Feature ด้วยวิธี Low Variance Filter จากขั้นตอนการ feature โดยใช้ Low Variance Filter Feature ที่เลือกใช้ได้แก่ Telephone, Account Balance, Value Savings/Stocks , Instalment percent, Duration in Current address , No of dependents , Concurrent Credits , Most valuable available asset และทำการแบ่งสัดส่วน Train 70% และ Test 30% โดยทำผ่าน ML แบบ Supervised Learning 3 model ได้แก่ KNN (k-Nearest Neighbors, Logistic Regression และ CART Model และวัดประสิทธิภาพของแบบจำลองโดยใช้ตาราง confusion matrix ที่ประกอบด้วย ค่าความถูกต้อง (Accuracy Score) ค่าความแม่นยำ (Precision Score) ค่าความถูกต้องของการทำนายว่าจะอนุมัติสินเชื่อเทียบกับจำนวนครั้งของเหตุการณ์ทั้งหมด (Recall) และค่าเฉลี่ยแบบฮาร์มอนิกระหว่าง Precision และ Recall (F1-Score) ได้ผลเปรียบเทียบในความแม่นยำของการพิจารณาอนุมัติสินเชื่อดังแสดงในรูป 1.22



รูปที่ 4.1 Comparative result of credit approval

ในรูป 4.1 แสดงให้เห็นว่า

ค่า Accuracy Score โดยวัดเป็นเปอร์เซ็น **KNN (k-Nearest Neighbors) 99.33% , Logistic Regression 73.00% และ CART Model 69.00 %** ตามลำดับ

ค่า Precision Score โดยวัดเป็นเปอร์เซ็นต์ **KNN (k-Nearest Neighbors) 99.03% , Logistic Regression 75.00% และ CART Model 87.32 %** ตามลำดับ

ค่า Recall โดยวัดเป็นเปอร์เซ็นต์ **KNN (k-Nearest Neighbors) 100.00% , Logistic Regression 90.73% และ CART Model 72.26%** ตามลำดับ

ค่า F1-Score โดยวัดเป็นเปอร์เซ็นต์ **KNN (k-Nearest Neighbors) 99.51% , Logistic Regression 82.12% และ CART Model 79.38%** ตามลำดับ

5. บทสรุปและข้อเสนอแนะ

ผลการศึกษาพบว่า แบบจำลอง **KNN (k-Nearest Neighbors)** มีประสิทธิภาพสูงที่สุด โดยมี Accuracy Score, Precision Score, Recall และ F1-Score คือ 99.33%, 99.03%, 100% และ 99.51% ตามลำดับ รองลงมาคือ แบบจำลอง **Logistic Regression** โดยมี Accuracy Score, Precision Score, Recall และ F1-Score คือ 73.00%, 75.00%, 90.73% และ 82.12% ตามลำดับ และ แบบจำลอง **CART Model** โดยมี Accuracy Score, Precision Score, Recall และ F1-Score คือ 69.00%, 87.32%, 72.76% และ 79.38% ตามลำดับ

จากบทความวิจัย“Credit Approval Analysis” โดย Abiola Smith, Brendan Maher และ Deepesh Khaneja พบว่า ML ที่มีความแม่นยำมากที่สุด 3 อันดับแรก คือ Classification and Regresstion Tree (CART) ,Logistic Regression และ K-Nearest Neighbors โดยวัดผลเพียงค่าเดียวคือ accuracy score แต่เนื่องจากข้อจำกัดในการเข้าถึงและใช้ข้อมูลในประเทศไทย การเลือกใช้ Data Set ที่นำมาใช้มีรายละเอียดแตกต่างกัน และการวัดผลจากทั้งสี่ค่าได้แก่ Accuracy Score, Precision Score, Recall และ F1-Score ผลการศึกษาจึงมีความแตกต่างกับจากบทความวิจัยดังกล่าว ผลของ KNN ที่ดีกว่าแบบจำลองอื่นคาดว่า KNN เป็นแบบจำลองที่ง่าย เหมาะกับข้อมูลแบบประเภท (nominal) เท่านั้นเช่น อนุมัติหรือไม่อนุมัติ หญิงหรือชาย เป็นต้น มีการพิจารณาจากข้อมูลเรียนรู้ที่อยู่ใกล้ที่สุด K ตัว ที่มักเป็นจำนวนคี่เป็นคำตอบ หรือ ให้ค่าน้ำหนักโดยการพิจารณาระยะห่างระหว่างข้อมูลที่สนใจกับข้อมูลที่อยู่ใกล้สุด K ตัวด้วย

ข้อเสนอแนะ ในการสร้างแบบจำลองต้องเลือกระเบียบวิธีหรือเทคนิคที่ใช้ให้เหมาะสมและเพื่อให้แบบจำลอง เกิดประสิทธิภาพที่ดีและนำไปใช้ได้จริง อาจต้องมีการใช้แบบจำลองอื่นที่มีมากกว่า 3 แบบจำลองดังกล่าว จากนั้นเลือกใช้ค่าที่ดีที่สุดของข้อมูล หรือการปรับจูนแบบจำลอง(Model Tuning) และเพื่อพัฒนาในการใช้งานได้จริงในอนาคต และในกรณีที่เกิดการสอนที่มากเกินไป (overfitting) อาจจะพิจารณาแบ่ง (split) เป็นหลายๆชุดเพื่อการประเมินผล

เอกสารอ้างอิง

กอบเกียรติ สระอุบล. (2563). *เรียนรู้ Data Science และ AI:Machine Learning ด้วย Python*. กรุงเทพฯ: หสก มีเดีย เน็ทเวิร์ค.

พูมใจ นาคสกุล. (2549). *สำรวจแบบจำลองความเสี่ยงด้านเครดิตในเชิงสัมพันธ์กับกรอบการดำรงเงินกองทุนของสถาบันการเงิน*. กรุงเทพฯ: สายนโยบายสถาบันการเงิน ธนาคารแห่งประเทศไทย

ศุภชัย สมพานิช. (2564). *เรียนรู้ Data Science ด้วย ภาษา R*. กรุงเทพฯ: บริษัท โปรวิชั่นจำกัด

Abiola Smith, Brendan Maher, and Deepesh Khaneja (2017). *Credit Approval using R.* Retrieved from https://www.researchgate.net/publication/321002603

Jayanthi, D. (2018). Credit Approval Data Analysis Using Classification And Regression Models. *IJAR September 2018, 5*(3), 162-169.

Lewis, N. D. (2017). *Machine Learning Made Easy With R***.** Retrieved from www.AusCov.com

Ryan Kuhn (2020). *Analysis of Credit Approval Data***.** Retrieved from/Rstudio-pubs-static.s3.amazonaws.com/73039_9949de135c0a49daa7a0a9eda4a67a72.html