การเปรียบเทียบตัวแบบการทำนายระดับและการเปลี่ยนระดับคุณค่าของลูกค้าในระยะยาว

A COMPARISON STUDY OF PREDICTIVE MODELS FOR CUSTOMER LIFETIME VALUE AND THE CHANGE OF ITS LEVEL

พรทิพย์ เดชพิชัย[1]

Pornthip Dechpichai

ลัทธพล โชครัตน์ประภา[2]

Latthapol Chokratprapa

สุพิชฌ์ ศรีแพทย์[3]

Supitch Sripath

ณรรฐคุณ วิรุฬห์ศรี [4*]

Nathakhun Wiroonsri[*]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้าง และเปรียบเทียบตัวแบบการทำนายระดับคุณค่าของลูกค้าในระยะยาว (CLV) และการเปลี่ยนระดับ CLV ในปีถัดไป โดยใช้ข้อมูลการซื้อขายผลิตภัณฑ์ประกันภัยของบริษัทกรณีศึกษาปีค.ศ. 2018-2020 ได้แก่ เพศ จำนวนปีที่เป็นลูกค้าต่อเนื่อง จำนวนกรมธรรม์ที่ยังมีผลคุ้มครอง เบี้ยประกันภัยรวม ทุนประกันรวม กำไรของกรมธรรม์รวม จำนวนครั้งที่เคลมรวม และระดับ CLV ตามเกณฑ์ของบริษัท (Standard, Plus, Extra และ Ultima) โดยใช้ ข้อมูลลูกค้าปีค.ศ. 2018 จำนวน 541,371 ราย สำหรับสร้างตัวแบบการถดถอยลอจิสติกเชิงอันดับ (OLR) และ Random Forest (RF) และใช้ข้อมูลลูกค้าปีค.ศ. 2019 เพื่อเปรียบเทียบประสิทธิภาพตัวแบบ และใช้ข้อมูลลูกค้าปีค.ศ. 2020 จำนวน 1,029,001 คน เพื่อทำนายระดับ CLV ในอนาคต

ผลการศึกษาพบว่า ตัวแบบ RF มีค่าความแม่นยำในการทำนายระดับ CLV โดยรวม คือ 75.01% ซึ่งมีประสิทธิภาพดีกว่าตัวแบบ OLR (ความแม่นยำ = 65.60%) อย่างไรก็ตาม ทั้งสองตัวแบบสามารถทำนายการเปลี่ยนระดับของลูกค้าได้แม่นยำเพียง 15.57% และ 25.74% ตามลำดับ ดังนั้นงานวิจัยนี้จึงเสนอการปรับค่าพารามิเตอร์ของตัวแบบ OLR คือ เกณฑ์ความน่าจะเป็น ซึ่งทำให้ตัวแบบมีความแม่นยำในการพยากรณ์ระดับ CLV โดยรวมเพิ่มสูงขึ้น (ความแม่นยำ = 74.60%) และได้ค่าความแม่นยำในการทำนายการเปลี่ยนระดับสูงถึง 52.16% ดังนั้นงานวิจัยนี้จึงใช้ตัวแบบ OLR ที่ปรับเกณฑ์ในการทำนายระดับ CLV ในอนาคต (ปีค.ศ. 2021) จำนวนทั้งสิ้น 1,029,001 ราย พบว่าจะมีลูกค้าอยู่ในระดับ Standard 40,786 ราย ระดับ Plus 809,951 ราย ระดับ Extra 168,389 ราย และ ระดับ Ultima 9,875 ราย

คำสำคัญ: คุณค่าของลูกค้าในระยะยาว การถดถอยลอจิสติกเชิงอันดับ Random Forest

[1,2] ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi

[3] บริษัท ธนชาตประกันภัย จำกัด (มหาชน)

Thanachart insurance public company limited

[4*]ผู้นิพนธ์ประสานงาน ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

อีเมล: nathakhun.wir@kmutt.ac.th

[*]Corresponding Author, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi,

E-mail: nathakhun.wir@kmutt.ac.th

วันที่รับบทความ (Received date): 26 January 2022 วันที่แก้ไขแล้วเสร็จ (Revised date): 26 August 2022

วันที่ตอบรับบทความ (Accepted date): 29 December 2022

Abstract

The main aims of this research are to construct and compare models to predict the change of Customer Lifetime Value or CLV level in two consecutive years. The data used in this work is collected by a case study insurance company from 2018 to 2020 containing the following features: gender, the number of consecutive years with the company, the number of active policies, the total insurance premium, the total coverage, the total policy profit and CLV level according to company criteria (Standard, Plus, Extra and Ultima). The data from 2018 and 2019 consisting of 541,371 customers are used to construct the ordinal logistic regression and Random Forest models and to check the model accuracies, respectively. Then the data from 2020 consisting of 1,029,001 customers is used for predicting future CLV.

The results show that, in term of predicting CLV, the Random Forest model (75.01%) has a higher accuracy than the ordinal logistic regression model (65.60%). However, both models cannot be used in detecting the change of CLV level providing the accuracies of 15.57% and 25.74%, respectively. After adjusting the predicting threshold criterion for the ordinal logistic regression model, the total accuracy and the change of CLV level accuracy are improved to 74.60% and 52.16%, respectively. Therefore, the ordinal logistic regression model with the new threshold criterion is selected as our final model. By using the final model, the 1,029,001 customers are classified as 40,786 Standard customers, 809,951 Plus customers, 168,389 Extra customers, and 9,875 Ultima customers.

Key words: Customer Lifetime Value, Ordinal Logistic Regression, Random Forest

บทนำ

ในปัจจุบันธุรกิจประกันวินาศภัยมีอัตราการเติบโตที่ดีมาอย่างต่อเนื่อง โดยข้อมูลจากสำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (คปภ.) พบว่าในปี 2019 มีอัตราการเติบโตของธุรกิจประกันวินาศภัยที่ลดลงจากปี 2018 ถึง 173.94% (Office of Insurance Commission, 2019) ในส่วนปริษัทกรณีศึกษา พบว่ามีอัตราการเติบโตอยู่ที่ ร้อยละ 28.08 และมีส่วนแบ่งตลาดประกันวินาศภัยเป็นอันดับ 10 โดยมีส่วนแบ่งอยู่ที่ 3.40% และในปี 2019 โดยมีสัดส่วนการรับประกันรถยนต์มากที่สุดอยู่ที่ 84.22% รองลงมาคือประกันเบ็ดเตล็ดอยู่ที่ 14.92% และ ประกันอัคคีภัยอยู่ที่ 0.86%

ในปัจจุบันลูกค้าส่วนใหญ่ได้รับการแนะนำผลิตภัณฑ์ของบริษัทจากช่องทางต่าง ๆ ของพันธมิตรของบริษัท มากกว่าจะสนใจทำประกันวินาศภัยที่เกิดจากความสมัครใจของตนเอง ซึ่งอาจจะเกิดจากเบี้ยประกันภัยหรือรูปแบบผลิตภัณฑ์ด้านประกันภัยของทางบริษัทยังไม่ตรงตามความต้องการหรือไม่ได้สะท้อนถึงพฤติกรรมการใช้ชีวิตของลูกค้าเท่าที่ควร และเนื่องจากในปัจจุบันนี้เป็นยุคที่ธุรกิจมีการแข่งขันกันสูง การใช้ข้อมูลขนาดใหญ่เข้ามาช่วยในการวิเคราะห์ทำนายและตัดสินใจในการประกอบธุรกิจจึงเป็นสิ่งที่มีความสำคัญมากและเป็นที่แพร่หลายมากขึ้นในภาคเอกชน

การแบ่งกลุ่มลูกค้า (Customer Segmentation) เป็นเทคนิคอย่างหนึ่งซึ่งทุกบริษัทนำเข้ามาใช้เพื่อช่วยสร้างกลยุทธ์ต่าง ๆ ในการดูแลลูกค้า ซึ่งในอดีตนั้นการจัดกลุ่มลูกค้าจะเป็นไปตามเกณฑ์ 1-2 มิติที่ดูได้ง่าย เช่น แบ่งจากประเภทประกันภัย แบ่งจากระดับประกันภัย แบ่งจากภูมิภาค เป็นต้น แต่ปัจจุบันเนื่องจากข้อมูลมหาศาลดังที่กล่าวมาแล้วนั้นการแบ่งกลุ่มลูกค้าจึงควรมาจากมิติของตัวแปรที่หลากหลายมากขึ้น ซึ่งบริษัทกรณีศึกษาได้ตั้งเกณฑ์การแบ่งระดับลูกค้าแบบใหม่ออกเป็น 4 ระดับ คือ Ultima, Extra, Plus และ Standard เพื่อวัดคุณค่าของลูกค้าในระยะยาว (Customer Lifetime Value: CLV) โดยใช้ตัวแปรการซื้อสินค้าแบบเชิงลึกมากขึ้น คือ ค่าเบี้ยประกันภัยรวม ยอดกำไรสุทธิ จำนวนกรมธรรม์ที่ยังดำเนินการอยู่ และระยะเวลาที่ทำประกันต่อเนื่อง

อย่างไรก็ตาม การที่ระดับของลูกค้าถูกกำหนดจากตัวแปรต่าง ๆ หลากหลายมิติ ประกอบกับจำนวนลูกค้าที่มาก ทำให้เป็นเรื่องยากที่จะทราบได้ว่าลูกค้าคนใดจะอยู่ระดับใดในปีถัดไป และยากยิ่งกว่าที่จะทราบถึงโอกาสในการเปลี่ยนระดับของลูกค้า ดังนั้นงานวิจัยนี้จึงสนใจศึกษาตัวแบบการเรียนรู้ของเครื่องโดยเปรียบเทียบ เทคนิคการวิเคราะห์การถดถอยลอจิสติกเชิงอันดับ (Ordinal Logistic Regression) และ Random Forest เพื่อทำนายระดับคุณค่าของลูกค้าในระยะยาว (CLV) และทำนายโอกาสในการเปลี่ยนระดับ CLV ในปีถัดไป ทั้งในกรณีที่เปลี่ยนระดับต่ำลง และเปลี่ยนระดับสูงขึ้น พร้อมทั้งศึกษาปัจจัยที่ส่งผลต่อการเปลี่ยนกลุ่มดังกล่าว ผลการศึกษาที่ได้น่าจะช่วยให้บริษัทสามารถวางแผนหรือออกแบบผลิตภัณฑ์ลูกค้าเฉพาะกลุ่ม (Niche market) ได้เหมาะสมยิ่งขึ้น เพื่อจะช่วยผลักดันลูกค้าให้มีการเปลี่ยนระดับไปยังระดับที่สูงขึ้น และป้องกันลูกค้าที่มีโอกาสในการเปลี่ยนระดับต่ำลงได้

ทบทวนวรรณกรรม

งานวิจัยนี้ได้ศึกษาเทคนิคการวิเคราะห์การถดถอยลอจิสติกเชิงอันดับ (Ordinal Logistic Regression) และ Random Forest เพื่อทำนายระดับคุณค่าลูกค้าระยะยาว และทำนายการเปลี่ยนระดับ CLV ของลูกค้า พร้อมทั้งศึกษาปัจจัยที่ส่งต่อการทำนายระดับ CLV ด้วย ดังมีรายละเอียดต่อไปนี้นี้

1. การถดถอยลอจิสติกเชิงอันดับ

การถดถอยลอจิสติกเชิงอันดับ (Ordinal Logistic Regression) เป็นตัวแบบที่ตัวแปรตาม (Y) มีการแจกแจงแบบมัลติโนเมียลที่เป็นอิสระกัน และลักษณะข้อมูลเป็นลำดับที่ต่อเนื่องกัน (โดยมีระดับการวัดแบบมาตรวัดเรียงลำดับ) โดยที่ระยะห่างของแต่ละลำดับที่ต่อเนื่องกันไม่จำเป็นต้องเท่ากัน เช่น ประเมินสุขภาพ (ไม่ดี/ปานกลาง/ดี) เกรดของลูกค้า (A/B/C) ระดับความเสี่ยง (น้อย/ปานกลาง/มาก) เป็นต้น

ให้ $1,2,\dots,J$ เป็นระดับของตัวแปรตาม Y เรียงจากน้อยไปมาก และ J คือจำนวนระดับของตัวแปร Y สมการลอจิสของการถดถอยลอจิสติกเชิงอันดับ มีดังนี้

$$\log\left(\frac{P(Y \le j)}{P(Y > j)}\right) = \log\left(\frac{\sum_{i=1}^{j} P(Y=i)}{\sum_{i=j+1}^{j} P(Y=i)}\right)$$

$$= \alpha_j - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_k X_k \quad (1)$$

ซึ่งจากการแก้สมการที่ (1) จะได้ว่า

$$P(Y \le j) = \frac{e^{\alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k}}{1 + e^{\alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k}} \quad (2)$$

และจะได้ว่า

$$P(Y = j) = P(Y \le j) - P(Y \le j-1) \quad (3)$$

สำหรับ $j = 1,2,\dots,J-1$ และ

$$P(Y = J) = 1 - P(Y \le J-1) \quad (4)$$

สำหรับ $j = J$

งานวิจัยนี้ใช้วิธี Maximum likelihood estimation (MLE) ในการประมาณค่าพารามิเตอร์ที่ไม่ทราบค่า (β_i) ตัวประมาณที่ได้จะมีคุณสมบัติของความคงเส้นคงวา (consistent) ความพอเพียง (sufficiency) และมีประสิทธิภาพ (efficient) (James, Witten, Tibshirani, 2013)

วิธีการถดถอยลอจิสติกไม่ใช่วิธีในการจำแนกกลุ่ม (Classification) แต่เพียงแค่หาความน่าจะเป็นดังสมการที่ (3-4) เท่านั้น ซึ่งจากความน่าจะเป็นดังกล่าว โดยทั่วไปจะทำนายไปยังกลุ่มที่ให้ความน่าจะเป็นสูงที่สุด อย่างไรก็ตามการปรับพารามิเตอร์ (Parameter Tuning) เช่นเกณฑ์ความน่าจะเป็น (Probability threshold) ในการทำนายสามารถปรับค่า

ได้) และทำการเปรียบเทียบเพื่อเลือกค่าที่เหมาะสมที่สุดได้ ซึ่งจะทำให้ประสิทธิภาพของตัวแบบเพิ่มขึ้น เช่น หากทำนายว่ายาชนิดใหม่จะปลอดภัยหรือไม่ เราจะสรุปว่ายาปลอดภัยก็ต่อเมื่อความน่าจะเป็นสูงมาก ๆ เป็นต้น นอกจากนี้ยังสามารถใช้แก้ปัญหาข้อมูลไม่สมดุล (unbalanced data) กล่าวคือหากการทำนายไปยังกลุ่มที่ความน่าจะเป็นสูงสุดทำให้ทำนายไปยังกลุ่มที่มีขนาดใหญ่มากจนเกินไป เราอาจปรับเกณฑ์ให้ทำนายไปยังกลุ่มใหญ่เมื่อความน่าจะเป็นสูงขึ้น เป็นต้น

2. **Random Forest**

แนวคิดของ Random Forest (RF) เกิดขึ้นโดย Ho ในปี 1995 (Ho,1995) และ ถูกนำไปต่อยอดเป็นอัลกอลิทึม และจดลิขสิทธิ์โดย Breiman และ Cutler ภายใต้ชื่อ Random Forest ในปี 2006 (Breiman, 2001) Random Forest เป็นวิธีการที่มีรากฐานจากต้นไม้ตัดสินใจ (Decision Tree) ที่ทำการแบ่งสเปซของตัวแปรอิสระออกเป็นหลาย ๆ ซับเซต (subset) และทำนายผลในแต่ละซับเซตไปยังค่า Y เดียวกันโดยท้ายสุดสามารถเขียนออกมาเป็นต้นไม้ได้ ซึ่งต้นไม้ตัดสินใจแบบปกตินั้นมีจุดอ่อนคือ ความแปรปรวนที่สูงมาก นั่นคือ เมื่อข้อมูลเปลี่ยนเพียงเล็กน้อยต้นไม้อาจต่างจากเดิมโดยสิ้นเชิง Random Forest แก้ปัญหานี้โดยการสร้างต้นไม้ตัดสินใจหลาย ๆ ต้น โดยในแต่ละต้นจะทำการสุ่มตัวแปรต้นมาเพียง m ตัว จากทั้งหมด k ตัว ซึ่งนิยมกำหนด $m \approx \sqrt{k}$ ซึ่งหลังจากได้ต้นไม้ตัดสินใจทั้งหมดแล้ว Random Forest จะเลือกกลุ่มที่ต้นไม้ส่วนใหญ่ทำนายได้

ในการสร้างต้นไม้ตัดสินใจนั้น ในแต่ละขั้นตอนจะทำการแบ่งเซตของตัวแปรต้น หรือ การแตกกิ่ง โดยทำให้ค่าดัชนีจีนี่ (Gini Index) มีค่าน้อยที่สุด โดยค่าดัชนีจีนี่ในเซตย่อย r (G_r) ในมีนิยามดังนี้

$$G_r = 1 - \sum_{j=1}^{J}[p_{rj}]^2 \qquad (5)$$

โดยที่ p_{rj} โดยที่คือสัดส่วนของจำนวนจุดที่มี Y อยู่ในกลุ่ม j ในเซตย่อย r

Predicted values	Actual Values			
	Ultima	Extra	Plus	Standard
Ultima	TP	FP	FP	FP
Extra	FN	TN	FN	FN
Plus	FN	FN	TN	TN
Standard	FN	FN	FN	TN

(1) Ultima

Predicted values	Actual Values			
	Ultima	Extra	Plus	Standard
Ultima	TN	FN	FN	FN
Extra	FP	TP	FP	FP
Plus	FN	FN	TN	FN
Standard	FN	FN	FN	TN

(2) Extra

Predicted values	Actual Values			
	Ultima	Extra	Plus	Standard
Ultima	TN	FN	FN	FN
Extra	FN	TN	FN	FN
Plus	FP	FP	TP	FP
Standard	FN	FN	FN	TN

(3) Plus

Predicted values	Actual Values			
	Ultima	Extra	Plus	Standard
Ultima	TN	FN	FN	FN
Extra	FN	TN	FN	FN
Plus	FN	FN	TN	FN
Standard	FP	FP	FP	TP

(4) Standard

ภาพที่ 1 การแสดง TP, FP, TN, และ FN สำหรับการจำแนกในกลุ่มที่สนใจโดย multi-class confusion matrix (Perea *et al*, 2013)

ทั้งนี้ค่าดัชนีจีนี่มีประโยชน์ที่สำคัญอีกอย่างคือ Random Forest สามารถตรวจสอบได้ว่าหากนำเอาตัวแปรต้นตัวใดตัวหนึ่งออก จะทำให้ Mean Gini ลดลงเท่าใด หรือที่เรียกว่า Mean Gini Decrease ซึ่งสามารถนำเอามาเรียงลำดับความสำคัญ (Importance) ของตัวแปรต้นได้ โดยหากค่า Mean Gini Decrease มาก ก็แปลว่าตัวแปรนั้นมีความสำคัญมาก

3. การวัดประสิทธิภาพของตัวแบบ

การวัดประสิทธิภาพของตัวแบบพยากรณ์เป็นการเปรียบเทียบผลของตัวแปรตามที่เกิดขึ้นจริงกับผลของตัวแปรตามที่พยากรณ์ขึ้นจากตัวแบบ โดยใช้ Confusion Matrix เพื่อสรุปจำนวนข้อมูลที่ตัวแบบจำแนกได้อย่างถูกต้องและไม่ถูกต้อง ในงานวิจัยได้แบ่งกลุ่มระดับคุณค่าของลูกค้าในระยะยาวออกเป็น 4 กลุ่มคือ Ultima, Extra, Plus และ Standard จึงสามารถสร้าง Confusion Matrix ดังแสดงในภาพที่ 1 โดยที่ True Positive (TP) คือ จำนวนข้อมูลที่จำแนกถูกในกลุ่มที่สนใจ False Positive (FP) คือ จำนวนข้อมูลที่จำแนกผิดในกลุ่มที่สนใจ True Negative (TN) คือ จำนวนข้อมูลที่จำแนกถูกในกลุ่มอื่นๆ และ False Negative (FN) คือ จำนวนข้อมูลที่จำแนกผิดในกลุ่มอื่น ๆ โดยจะนำค่าต่าง ๆ ดังกล่าวมาคำนวณเพื่อวัดประสิทธิภาพของตัวแบบด้วยค่าความแม่นยำรวม (Accuracy: Acc) ซึ่งเป็นการวัดประสิทธิภาพความถูกต้องของการพยากรณ์ด้วยตัวแบบโดยรวม ในรูปอัตราส่วน ดังสมการ (6)

$$Acc = \frac{TP(Ultima)+TP(Extra)+TP(Plus)+TP(Standard)}{n} \quad (6)$$

ในขณะที่ค่าความแม่นยำแต่ละระดับสามารถวัดได้ ดังนี้

$$Acc(j) = \frac{TP(j)}{n_j} \quad (7)$$

โดยที่ j เป็น Standard, Plus, Extra หรือ Ultima และ n_j เป็นจำนวนข้อมูลในระดับ j

นอกจากนี้ งานวิจัยนี้ได้พิจารณาประสิทธิภาพในการทำนายการเปลี่ยนระดับ โดยสรุปจำนวนข้อมูลที่ตัวแบบจำแนกการเปลี่ยนกลุ่มของลูกค้าได้อย่างถูกต้องและไม่ถูกต้องเป็นรายกลุ่มทั้งหมด 12 กลุ่ม นั่นคือ Standard->Plus, Standard->Extra, Standard->Ultima, Plus->Standard, Plus->Extra, Plus->Ultima, Extra->Standard, Extra->Plus, Extra->Ultima, Ultima->Standard, Ultima->Plus และ Ultima->Extra ซึ่งจะพิจารณาลูกค้าที่เปลี่ยนระดับเป็น Positive และลูกค้าที่ไม่เปลี่ยนระดับเป็น Negative โดยพิจารณา ลูกค้าที่มี CLV ปี 2020 ที่เปลี่ยนกลุ่มเป็นกลุ่มอื่นใน CLV ปี 2021 จริง และผลการทำนายมีกลุ่มตรงกับข้อมูล CLV ปี 2021 และนำมาสร้างตาราง Confusion Matrix ดังตารางที่ 1 และคำนวณเพื่อวัดประสิทธิภาพของตัวแบบด้วยค่าความแม่นยำซึ่งเป็นการวัดประสิทธิภาพความถูกต้องของการพยากรณ์ด้วยตัวแบบโดยรวม ในรูปอัตราส่วน ดังสมการที่ (8)

$$\frac{TP(A\rightarrow B)}{n(A\rightarrow B)} \quad (8)$$

โดยที่ A และ B เป็น Standard, Plus, Extra หรือ Ultima โดยที่ $A \neq B$ และ $n(A \rightarrow B)$ เป็นจำนวนของข้อมูลที่มีการเปลี่ยนระดับจาก A ไป B

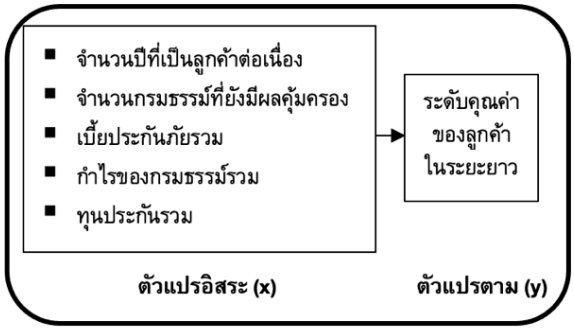
ตารางที่ 1 การแสดง TP, FP, TN, และ FN สำหรับการจำแนกการเปลี่ยนกลุ่มของลูกค้า

Actual \ Predict	ไม่เปลี่ยนกลุ่ม(A->B)	เปลี่ยนกลุ่ม (A->B)
ไม่เปลี่ยนกลุ่ม(A->B)	TN(A->B)	FN(A->B)
เปลี่ยนกลุ่ม (A->B)	FP(A->B)	TP(A->B)

กรอบแนวคิดการวิจัย

คุณค่าของลูกค้าในระยะยาว (CLV) คือมูลค่ารวมหลักหักค่าใช้จ่ายที่ลูกค้ารายหนึ่งสร้างให้กับบริษัทในระยะยาวนับตั้งแต่ก้าวเข้ามาเป็นลูกค้าของบริษัท โดยบริษัทอาจมีการตีมูลค่านั้น ๆ ออกเป็นระดับต่าง ๆ ซึ่งการศึกษา CLV มีความสำคัญอย่างมากกับธุรกิจเนื่องจากเป็นส่วนสำคัญในการบริหารจัดการการขายสินค้าและบริการให้กับลูกค้ารายสำคัญ (Jackson, 1989) รวมทั้งในวงการศึกษาวิจัย (Chang, Chang & Li, 2012)

บริษัทกรณีศึกษาได้ตั้งเกณฑ์การแบ่งระดับ CLV ของลูกค้าออกเป็น 4 ระดับ คือ Ultima, Extra, Plus และ Standard จากหลายปัจจัยที่ส่งผลต่อ CLV เช่น ค่าเบี้ยประกันภัยรวม ยอดกำไรสุทธิ จำนวนกรมธรรม์ที่ยังดำเนินการอยู่ และระยะเวลาที่ทำประกันต่อเนื่อง เป็นต้น ดังนั้นงานวิจัยจึงได้นำมาพัฒนากรอบแนวคิดการวิจัยดังภาพที่ 2



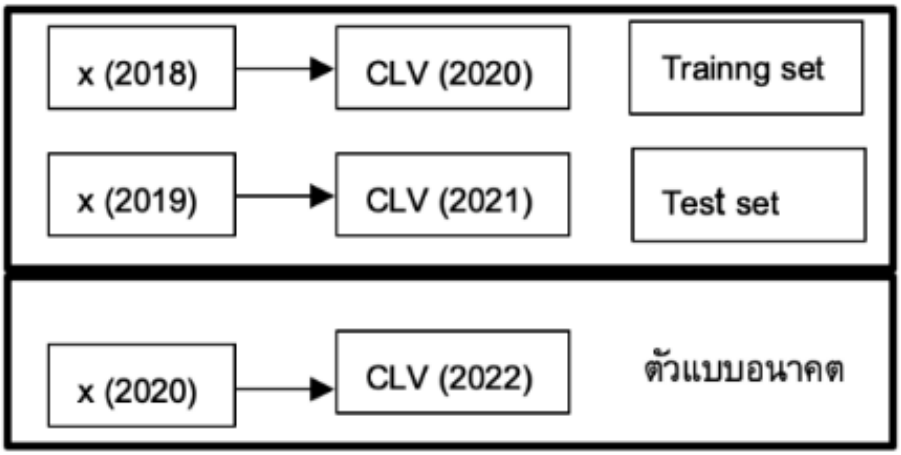
ภาพที่ 2 กรอบแนวคิดการวิจัย

วิธีดำเนินการวิจัย

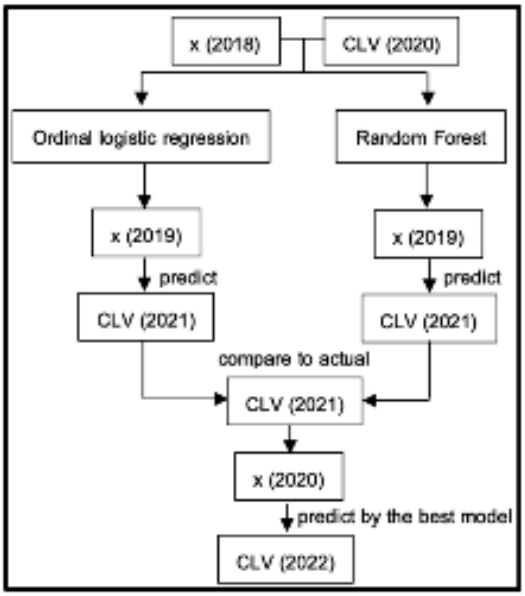
งานวิจัยนี้เป็นงานวิจัยเชิงปริมาณ โดยทำการศึกษาฐานข้อมูลการซื้อขายผลิตภัณฑ์ประกันของบริษัทกรณีศึกษาตั้งแต่ปี ค.ศ. 1997 ถึง ค.ศ. 2020 รวมทั้งสิ้น 23 ปี แล้วดำเนินการตรวจสอบข้อมูลที่สูญหาย (Missing Value) และค่าผิดปกติ (Outlier) และแก้ไขข้อมูลให้ถูกต้องและสมบูรณ์ ได้ข้อมูลจำนวนทั้งสิ้น 541,371 ราย โดยมีตัวแปรที่ศึกษาดังนี้ จำนวนปีที่เป็นลูกค้าต่อเนื่อง (YearContinue) จำนวนกรมธรรม์ที่ยังมีผลคุ้มครอง (ActivePol) เบี้ยประกันภัยรวม (Premium) กำไรของกรมธรรม์รวม (ProfitAmt) ทุนประกันรวม (SI) เพศ (Gender) และระดับคุณค่าของลูกค้าในระยะยาว (CLV) โดยแบ่งข้อมูลเป็น 2 ชุด คือ ชุดฝึกสอน (training set) ซึ่งจะเรียกว่าชุดข้อมูลปีค.ศ. 2018 ซึ่งใช้ข้อมูลปัจจัย (x) ปีค.ศ. 2018 และ CLV (y) ปีค.ศ. 2020 สำหรับสร้างตัวแบบการถดถอยลอจิสติกเชิงอันดับ และ ตัวแบบ Random Forest และชุดทดสอบ (test set) โดยชุดข้อมูลปีค.ศ. 2019 ซึ่งใช้ข้อมูลปัจจัย (x) ปีค.ศ. 2019 และ CLV (y) ปีค.ศ. 2021 เพื่อประเมินความเหมาะสมของตัวแบบอย่างที่อธิบายไว้ในการตรวจสอบความแม่นยำ ซึ่งสามารถสรุปได้ดังภาพที่ 3 และ 4

นอกจากนี้งานวิจัยนี้ยังใช้ตัวแบบข้างต้นในการวิเคราะห์ปัจจัยที่มีความสำคัญต่อระดับ CLV และการเปลี่ยนระดับ CLV โดยใช้การวิเคราะห์สัมประสิทธิ์ของวิธีการวิเคราะห์ถดถอยลอจิสติกเชิงอันดับ และ ค่า Mean Decrease Gini ของ ตัวแบบ Random Forest

หลังจากเปรียบเทียบประสิทธิภาพตัวแบบและได้ตัวแบบที่เหมาะสมแล้ว ผู้วิจัยจะทำการพยากรณ์ระดับ CLV ของปี 2022 โดยใช้ข้อมูลปัจจัย (x) ของปี 2020 ด้วยตัวแบบดังกล่าว ดังภาพที่ 4



ภาพที่ 3 การแบ่งชุดข้อมูล



ภาพที่ 4 ขั้นตอนในดำเนินการวิจัย

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบตัวแบบการถดถอยลอจิสติกเชิงอันดับและ Random Forest สำหรับการทำนายระดับคุณค่าของลูกค้าในระยะยาว และทำนายการเปลี่ยนระดับ CLV รวมทั้งศึกษาปัจจัยที่มีความสำคัญต่อระดับ CLV โดยใช้ข้อมูลลูกค้าจำนวน 541,371 ราย ของบริษัทกรณีศึกษา และโปรแกรม R-studio (RStudio Team, 2020) ผลการวิเคราะห์มีดังนี้

1. **สถิติพรรณนาของปัจจัยที่ศึกษา**

จากตารางที่ 2 ลูกค้าของบริษัทกรณีศึกษาในปีค.ศ. 2018 เป็นเพศชาย ร้อยละ 50.09 และเพศหญิง ร้อยละ49.91 และมีจำนวนกรมธรรม์ที่ยังมีผลคุ้มครองเฉลี่ย 1.30 กรมธรรม์ จำนวนปีที่เป็นลูกค้าต่อเนื่องเฉลี่ย 3.40 ปี กำไรของกรมธรรม์รวมเฉลี่ย 1,118 บาท เบี้ยประกันภัยรวมเฉลี่ย 31,925 บาท และทุนประกันรวมเฉลี่ย 1,749,132 บาท และเมื่อปีค.ศ. 2019 ลูกค้าชุดนี้ ยังคงมีจำนวนกรมธรรม์ที่ยังมีผลคุ้มครองเฉลี่ย 1.237 กรมธรรม์ จำนวนปีที่เป็นลูกค้าต่อเนื่องเฉลี่ย 4.39 ปี กำไรของกรมธรรม์รวมเฉลี่ย 13,931บาท เบี้ยประกันภัยรวมเฉลี่ย 37,061บาท และทุนประกันรวมเฉลี่ย 2,025,135 บาท

นอกจากนี้ลูกค้าของบริษัทกรณีศึกษา ในปี ค.ศ. 2018 ได้ถูกจัดระดับตามเกณฑ์คุณค่าของลูกค้าระยะยาว (CLV (2020)) พบว่าส่วนใหญ่จะอยู่ในระดับ Standard มากที่สุด ร้อยละ 42.88 รองลงมาคือ ระดับ Plus ร้อยละ 33.46, Extra ร้อยละ 20.00 และ Ultima ร้อยละ 3.66 ตามลำดับ และในปีค.ศ. 2019 (CLV (2021)) ลูกค้าส่วนใหญ่จะถูกจัดอยู่ในระดับ Plus มากที่สุด ร้อยละ 37.36 รองลงมาคือ ระดับ Standard ร้อยละ 36.00, Extra ร้อยละ 24.01 และ Ultima ร้อยละ 2.63 ตามลำดับ (ตารางที่ 3)

ส่วนลูกค้าของบริษัทกรณีศึกษา ในปีค.ศ. 2020 เป็นเพศชาย ร้อยละ 49.09 และเพศหญิง ร้อยละ 50.91 และมีจำนวนกรมธรรม์ที่ยังมีผลคุ้มครองเฉลี่ย 1.20 กรมธรรม์ จำนวนปีที่เป็นลูกค้าต่อเนื่องเฉลี่ย 3.516 ปี กำไรของกรมธรรม์รวมเฉลี่ย 9,722 บาท เบี้ยประกันภัยรวมเฉลี่ย 30,755 บาท และ ทุนประกันรวมเฉลี่ย 1,647,247บาท (ตารางที่ 2)

ตารางที่ 2 สถิติพรรณนาของปัจจัยจำแนกตามปี

ปัจจัย	2018		2019		2020	
	mean	sd	mean	sd	mean	sd
เพศ ชาย (n / %)	271,200	50.09	271,200	50.09	505,125	49.09
หญิง (n / %)	270,171	49.91	270,171	49.91	523,876	50.91
จำนวนกรมธรรม์ที่ยังมีผลคุ้มครอง	1.304	0.567	1.237	0.559	1.20	0.48
จำนวนปีที่เป็นลูกค้าต่อเนื่อง	3.397	2.679	4.389	2.674	3.516	2.82
กำไรของกรมธรรม์รวม	1,118	62,617.31	13,931	31,529.15	9,722	48,254.43
เบี้ยประกันภัยรวม	31,925	33,883.02	37,061	36,132.5	30,755	34,539.32
ทุนประกันรวม	1,749,132	3,142,649	2,025,135	3,129,640	1,647,247	3,099,994

ตารางที่ 3 จำนวนลูกค้าในแต่ละระดับกลุ่มตามเกณฑ์คุณค่าลูกค้าระยะยาว จำแนกตามปี

ระดับ CLV	2020		2021	
	n	%	n	%
Standard	232,138	42.88	194,893	36.00
Plus	181,129	33.46	202,241	37.36
Extra	108,234	20.00	130,003	24.01
Ultima	19,870	3.66	14,234	2.63

2. ผลการศึกษาปัจจัยที่ส่งผลต่อ ระดับคุณค่าของลูกค้าในระยะยาว

การศึกษาปัจจัยที่ส่งผลต่อระดับคุณค่าของลูกค้าในระยะยาวด้วยตัวแบบการถดถอยลอจิสติกเชิงอันดับ และ Random Forest พบว่าตัวแบบการถดถอยลอจิสติกเชิงอันดับ มีสัมประสิทธิ์ถดถอยดังตารางที่ 4 และ 6 และตัวแบบ Random Forest มีค่า Mean Decrease Gini ดังภาพที่ 5

จากตัวแบบถดถอยลอจิสติกเชิงอันดับ พบว่าทุกปัจจัยมีผลต่อการทำนาย CLV อย่างมีนัยสำคัญ และมีผลในเชิงบวกต่อระดับ CLV ยกเว้นปัจจัยทุนประกัน (SI) และปัจจัยที่สำคัญที่สุด คือ เปอร์เซ็นต์กำไรของกรมธรรม์รวมที่มีสัมประสิทธิ์ 7.8741 ส่วนปัจจัยที่ส่งผลน้อยที่สุด คือ เพศ (X_1) ซึ่งมีสัมประสิทธิ์ 0.0434 และสามารถเขียนเป็นสมการได้ดังสมการที่ (9), (10) และ (11) โดยให้ 1 แทนกลุ่ม Standard, 2 แทนกลุ่ม Plus, 3 แทนกลุ่ม Extra และ 4 แทนกลุ่ม Ultima และ ตัวแปร Gender (X_1) มีค่าเป็น 1 เมื่อเป็นเพศชาย และมีค่าเป็น 0 เมื่อเป็นเพศหญิง

สำหรับตัวแบบ Random Forest เมื่อพิจารณาความสำคัญของตัวแปรในการทำนายจากค่า Mean Decrease Gini แล้วพบว่าตัวแปรที่มีความสำคัญที่สุดคือ ทุนประกันรวม และ กำไรของกรมธรรม์รวม ในขณะที่ เพศ มีความสำคัญน้อยที่สุด ซึ่งสอดคล้องกับตัวแบบการถดถอยลอจิสติกเชิงอันดับ ดังภาพที่ 5

ตารางที่ 4 ผลการวิเคราะห์การถดถอยลอจิสติกเชิงอันดับ

Attribute	Value	Std. Error	t value
Gender:Male (X_1)	0.0434	0.005795	7.488*
ActivePol (X_2)	0.2035	0.003197	63.665*
YearContinue (X_3)	0.2017	0.004224	47.753*
SI (X_4)	-0.1622	0.004592	-35.316*
Premium (X_5)	0.6622	0.006646	99.645*
ProfitAmt (X_6)	7.8741	0.024651	319.387*

*ระดับนัยสำคัญ 0.05

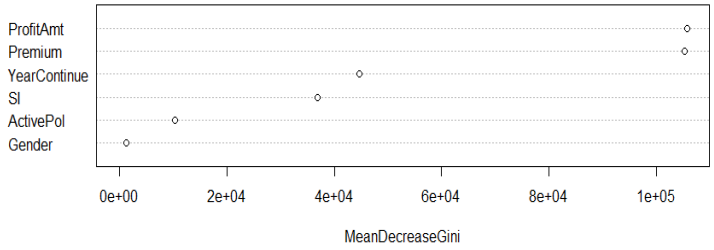
ตารางที่ 5 ผล Intercepts ของการวิเคราะห์การถดถอยลอจิสติกเชิงอันดับ

Case	Value	Std. Error	t value
Standard | Plus	-0.5566	0.0048	-114.9588
Plus | Extra	2.2724	0.0061	370.6365
Extra | Ultima	7.0610	0.0167	422.7790

$$\log\left(\frac{P(Y\leq 1)}{P(Y>1)}\right) = -0.56 - 0.04X_1 - 0.20X_2 - 0.20X_3 + 0.16X_4 - 0.66X_5 - 7.87X_6 \quad (9)$$

$$\log\left(\frac{P(Y\leq 2)}{P(Y>2)}\right) = 2.27 - 0.04X_1 - 0.20X_2 - 0.20X_3 + 0.16X_4 - 0.66X_5 - 7.87X_6 \quad (10)$$

$$\log\left(\frac{P(Y\leq 3)}{P(Y>3)}\right) = 7.06 - 0.04X_1 - 0.20X_2 - 0.20X_3 + 0.16X_4 - 0.66X_5 - 7.87X_6 \quad (11)$$



ภาพที่ 5 กราฟแสดงค่า Mean Decrease Gini จากตัวแบบ Random Forest

3. ผลการเปรียบเทียบตัวแบบการทำนายระดับคุณค่าของลูกค้าในระยะยาว

การเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการทำนายระดับคุณค่าของลูกค้าในระยะยาว ได้ใช้ข้อมูลลูกค้าจำนวน 541,371 ราย ปีค.ศ. 2018 เพื่อสร้างตัวแบบวิธีการถดถอยลอจิสติเชิงอันดับ (OLS) และ Random Forest (RF) และใช้ข้อมูลปีค.ศ. 2019 เพื่อเปรียบเทียบประสิทธิภาพตัวแบบ ได้ผลดังตารางที่ 6 และ ตารางที่ 7 พบว่า ค่าความแม่นยำในการทำนายระดับ CLV ในภาพรวมของตัวแบบ Random Forest เท่ากับ 75.01% ซึ่งมีประสิทธิภาพดีกว่าตัวแบบการถดถอยโลจิสติกส์เชิงอันดับ (ความแม่นยำ = 65.60%)

นอกจากนี้ตัวแบบ Random Forest มีค่าความแม่นยำที่สูงกว่าตัวแบบการถดถอยลอจิสติกเชิงอันดับในระดับ Plus และ Extra ส่วนตัวแบบการถดถอยลอจิสติกเชิงอันดับมีความแม่นยำมากกว่าในการทำนายระดับ Standard และ Ultima

ตารางที่ 7 ค่าความแม่นยำในการพยากรณ์ระดับคุณค่าของลูกค้าในระยะยาว

ระดับ CLV	OLR	RF
Standard	97.49%	95.31%
Plus	42.59%	59.63%
Extra	51.64%	68.79%
Ultima	83.46%	72.50%
รวม	**65.60%**	**75.01%**

ตารางที่ 6 Confusion Matrix สำหรับทำนายระดับคุณค่าของลูกค้าในระยะยาว 4 ระดับ

Actual CLV (2021)	Predicted CLV (2021) OLR				Predicted CLV (2021) RF			
	Ultima	Extra	Plus	Standard	Ultima	Extra	Plus	Standard
Ultima	11,880	2,263	70	21	10,319	2,444	1,456	15
Extra	43,845	67,134	15,649	3,375	14,468	89,425	24,761	1,349
Plus	2,885	21,896	86,144	91,316	921	19,613	120,598	61,109
Standard	640	1,805	2,453	189,995	378	1,592	7,168	185,755

หมายเหตุ: ตัวแบบการถดถอยลอจิสจิกเชิงอันดับจะถูกตั้งค่าให้ทำนายระดับที่มีความน่าจะเป็นสูงที่สุด

4. ผลการเปรียบเทียบตัวแบบทำนายการเปลี่ยนระดับคุณค่าของลูกค้าในระยะยาว

เมื่อพิจารณาความแม่นยำในการทำนายเฉพาะลูกค้าในปี 2020 ที่มีการเปลี่ยนระดับคุณค่าของลูกค้าในระยะยาวในปี 2020 จำนวน 103,178 คน พบว่า ตัวแบบการถดถอยลอจิสติกเชิงอันดับจะมีความแม่นยำเท่ากับ 25.74% ซึ่งสูงกว่าตัวแบบ Random Forest ที่มีความแม่นยำเพียง 15.57% อย่างไรก็ตาม

5. ผลการวิเคราะห์ตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์ เปรียบเทียบ

จากผลการเปรียบเทียบตัวแบบทำนายการเปลี่ยนระดับ CLV งานวิจัยนี้จึงพิจารณาการปรับค่าพารามิเตอร์ (Parameter tuning) ซึ่งมักจะนิยมในใช้การสร้างตัวแบบการเรียนรู้ด้วยเครื่อง (Machine learning) นั่นคือ จะมีการปรับค่าเกณฑ์ความน่าจะเป็นของตัวแบบการถดถอยลอจิสติกเชิงอันดับ โดยการพิจารณาผลความแม่นยำในการทำนายระดับ Plus และ Extra ของตัวแบบการถดถอยลอจิสติกเชิงอันดับซึ่งน้อยกว่าตัวแบบ Random Forest ดังตารางที่ 7 ซึ่งเนื่องมาจากตัวแบบลอจิสติกเชิงอันดับมักจะทำนายไปยังระดับ CLV ที่ความน่าจะเป็นสูงสุด จึงทำให้ตัวแบบพยากรณ์ไปยังระดับที่มีสัดส่วนสูงที่สุดซึ่งก็คือ Standard ค่อนข้างมาก ดังนั้นงานวิจัยนี้จึงปรับเกณฑ์ความน่าจะเป็นของตัวแบบการถดถอยลอจิสติกเชิงอันดับ ดังนี้

1. ตัวแบบจะทำนายระดับ Standard ก็ต่อเมื่อ ความน่าจะเป็นที่จะอยู่ในระดับนี้มากกว่า 0.8

2. หากความน่าจะเป็นระดับ Standard น้อยกว่า 0.8 จะทำการเปรียบเทียบความน่าจะเป็นระดับ Plus, Extra และ Ultima และทำนายไปยังระดับที่มีความน่าจะเป็นสูงที่สุด

หลักเกณฑ์นี้ถูกสร้างขึ้นเพื่อไม่ให้มีการทำนายไปยังระดับ Standard มากเกินไปเนื่องจากกลุ่มนี้มีจำนวนที่มากกว่ากลุ่มอื่นค่อนข้างมาก และ เพื่อให้ตัวแบบสามารถทำนายลูกค้าที่เปลี่ยนระดับ CLV ออกมาได้มาก

จากตารางที่ 8-9 แสดงผลความแม่นยำโดยรวมในการทำนายระดับ CLV (2021) ของตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์ความน่าจะเป็น ขึ้น พบว่า มีค่าความแม่นยำสูงขึ้น โดยเพิ่มขึ้นจาก 65.60% เป็น 74.60% ซึ่งเพิ่มขึ้น 9% จากตัวแบบเดิมรวมทั้งเพิ่มความแม่นยำในระดับ Plus จาก 42.59% เป็น 74.39%

ตารางที่ 10 แสดงผลการทำนายการเปลี่ยนระดับ CLV ซึ่งจะพิจารณาความแม่นยำในการทำนายการเปลี่ยนระดับแยกตามระดับของแต่ละกรณีได้ทั้งหมด 12 กรณี โดยมีจำนวนลูกค้าทั้งหมด 103,178 คนที่เปลี่ยนไปในแต่ละกลุ่ม เมื่อพิจารณาความแม่นยำของตัวแบบการเปลี่ยนกลุ่มจะพิจารณาเฉพาะลูกค้าที่ค่าจริงของ CLV (2021) มีการเปลี่ยนระดับไปจาก CLV (2020) และ คำนวณร้อยละที่การทำนายกลุ่ม CLV (2021) ตรงกับค่า CLV (2021) เฉพาะในกลุ่มนี้ ซึ่งได้ผลดังตารางที่ 10 พบว่าตัวแบบการถดถอยลอจิสติกเชิงอันดับ (ปรับเกณฑ์) สามารถทำนายลูกค้าที่เปลี่ยนระดับได้ถึง 52.61% ซึ่งพัฒนาขึ้นจากก่อนปรับเกณฑ์ที่ 25.74% รวมทั้งมีความแม่นยำกว่าตัวแบบ Random Forest ที่ทำนายการเปลี่ยนกลุ่มได้เพียง 15.57%

หากมองเจาะลึกลงไปจะพบว่าตัวแบบใหม่นี้สามารถทำนายลูกค้าที่เปลี่ยนระดับขึ้น 1 ระดับได้ในเกณฑ์ดีมาก นั่นคือ Standard -> Plus, Plus -> Extra และ Plus -> Ultima ได้ถึง 64.2%, 50.8% และ 61.9%

ตามลำดับ ส่วนการทำนายลูกค้าที่เปลี่ยนระดับลง 1 ระดับที่อยู่ในเกณฑ์ดี คือ Ultima->Extra, Extra->Plus และ Plus->Standard ได้ 42.2%, 34.5% และ 20.1% ตามลำดับ ในขณะที่ตัวแบบ Random Forest แทบจะไม่สามารถทำนายลูกค้าที่เปลี่ยนระดับได้เลย จะมีเพียงแค่ Ultima->Extra และ Extra->Ultima เท่านั้นที่มีให้ค่าความแม่นยำอยู่ในเกณฑ์ที่รับได้ คือ 21.5% และ 24.5% ตามลำดับ

ส่วนความแม่นยำในการทำนายลูกค้าที่เปลี่ยนระดับ 2 ระดับขึ้นไป ตัวแบบใหม่นี้สามารถทำนายได้เพียงการเปลี่ยนระดับ Plus->Ultima โดยทำได้ในระดับที่ดีคือ 60.0% ในขณะที่ตัวแบบ Random Forest แทบจะไม่สามารถทำนายลูกค้าที่เปลี่ยนระดับ 2 ระดับขึ้นไป ในเกณฑ์ที่ยอมรับได้เลย อย่างไรก็ตามการที่ไม่สามารถทำนายลูกค้าที่เปลี่ยน 2 ระดับได้นั้นอยู่ในความคาดหมายอยู่แล้ว เนื่องจากการเปลี่ยนแปลงอย่างก้าวกระโดดนี้ไม่น่าจะถูกทำนายได้ด้วยข้อมูลเชิงปริมาณของปีก่อนหน้าได้ และอาจจะเกิดจากปัจจัยภายนอกที่ไม่สามารถควบคุมได้

ตารางที่ 8 Confusion Matrix สำหรับทำนายระดับคุณค่าของลูกค้าในระยะยาว 4 ระดับโดยตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์ (OLR : ปรับเกณฑ์)

Actual CLV (2021)	**Predicted OLR (ปรับเกณฑ์) CLV (2021)**			
	Ultima	**Extra**	**Plus**	**Standard**
Ultima	11,879	2,264	79	12
Extra	43,845	67,134	18,125	899
Plus	2,885	21,896	150,439	27,021
Standard	640	1,805	17,864	174,584

ตารางที่ 9 ค่าความแม่นยำในการพยากรณ์ระดับคุณค่าของลูกค้าในระยะยาวสำหรับตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์ (OLR : ปรับเกณฑ์)

ระดับ CLV	**OLR (ปรับเกณฑ์)**
Standard	89.58%
Plus	74.39%
Extra	51.64%
Ultima	83.46%
รวม	**74.60%**

6. การทำนายระดับคุณค่าของลูกค้าในระยะยาวในอนาคต

จากผลการเปรียบเทียบตัวแบบที่เหมาะสมสำหรับการทำนายระดับ CLV พบว่า ตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์มีประสิทธิภาพโดยรวมเพิ่มขึ้นจนใกล้เคียงกับตัวแบบ Random Forest แต่ประสิทธิภาพในการทำนายการเปลี่ยนระดับดีกว่าตัวแบบ Random Forest มาก งานวิจัยนี้จึงใช้ตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์ในการทำนายระดับคุณค่าของลูกค้าในระยะยาวในอนาคต (CLV (2022)) ด้วยชุดข้อมูลปี 2020 จำนวนทั้งสิ้น 1,029,001 ราย พบว่าลูกค้าส่วนใหญ่ถูกทำนายว่าอยู่ในระดับ Plus มากสุด ร้อยละ 78.71 รองลงมาคือ ระดับ Extra ร้อยละ 16.36 Standard ร้อยละ 3.96 และ Ultima ร้อยละ 0.01 ตามลำดับ (ตารางที่ 11)

ตารางที่ 11 จำนวนลูกค้าในแต่ละระดับของตัวแบบในอนาคต

CLV (2022)	OLR (ปรับเกณฑ์)	ร้อยละ
Standard	40,786	3.96
Plus	809,951	78.71
Extra	168,389	16.36
Ultima	9,875	0.01

ตารางที่ 10 จำนวนสมาชิกที่มีการเปลี่ยนกลุ่มจากค่าจริง (CLV 2020) เป็น CLV20221 จำแนกตามตัวแบบ

Actual	Actual CLV 2021				Predicted OLR ปรับเกณฑ์ (CLV 2021)				Predicted RF (CLV 2021)			
CLV 2020	Ultima	Extra	Plus	Standard	Ultima	Extra	Plus	Standard	Ultima	Extra	Plus	Standard
Ultima	10,345	8,999	271	255	9,507 (91.9%)	3,797 (42.2%)	0 (0.0%)	0 (0.0%)	9,085 (87.8%)	1,935 (21.5%)	0 (0.0%)	0 (0.0%)
Extra	2,412	93,354	10,939	1,529	1,493 (61.9%)	48,916 (52.4%)	3,780 (34.5%)	0 (0.0%)	592 (24.5%)	85,637 (91.7%)	1,008 (9.2%)	0 (0.0%)
Plus	1,464	27,282	146,884	5,499	879 (60.0%)	14,421 (50.8%)	118,316 (80.5%)	1,106 (20.1%)	0 (0.0%)	2,316 (8.5%)	110,814 (75.4%)	1,632 (29.7%)
Standard	13	368	44,147	187,610	0 (0.0%)	0 (0.0%)	28,343 (64.2%)	173,478 (92.5%)	0 (0.0%)	0 (0.0%)	8,236 (18.7%)	184,178 (98.2%)

อภิปรายผล

งานวิจัยนี้ได้เปรียบเทียบตัวแบบการถดถอยลอจิสติกเชิงอันดับ และตัวแบบ Random Forest ในการพยากรณ์ระดับคุณค่าของลูกค้าในระยะยาว และการเปลี่ยนระดับ CLV ในอนาคต ผลการวิจัยพบว่า ตัวแบบ Random Forest มีความแม่นยำในการทำนายระดับ CLV ของลูกค้าโดยรวมมากกว่าตัวแบบการถดถอยลอจิสติกเชิงอันดับ โดยมีค่าความแม่นยำอยู่ที่ 75.01% ซึ่งอยู่ในเกณฑ์ที่ยอมรับได้ แต่ประสิทธิภาพการทำนายการเปลี่ยนระดับ CLV ต่ำกว่าตัวแบบการถดถอยลอจิสติกเชิงอันดับ และเนื่องจากตัวแบบการถดถอยลอจิสติกเชิงอันดับ โดยทั่วไปมักจะทำนายไปยังกลุ่มที่ให้ความน่าจะเป็นสูงที่สุด และเป็นตัวแบบสามารถปรับเกณฑ์ความน่าจะเป็นเพื่อเพิ่มประสิทธิภาพของตัวแบบได้ งานวิจัยนี้จึงเสนอการปรับเกณฑ์ความน่าจะเป็นของตัวแบบการถดถอยลอจิสติกเชิงอันดับเพื่อเพิ่มประสิทธิภาพในการทำนาย ซึ่งหลังจากที่ปรับเกณฑ์แล้ว พบว่าตัวแบบการถดถอยลอจิสติกเชิงอันดับที่ปรับเกณฑ์มีความแม่นยำในการพยากรณ์ระดับ CLV โดยรวมและการพยากรณ์การเปลี่ยนระดับ CLV เพิ่มสูงขึ้น ซึ่งสอดคล้องกับงานนอกจากนี้การปรับเกณฑ์ความน่าจะเป็นในการทำนายยังช่วยแก้ปัญหาข้อมูลไม่สมดุล (unbalanced data) โดยการปรับเกณฑ์ให้ทำนายไปยังกลุ่มใหญ่เมื่อความน่าจะเป็นสูงขึ้น เพื่อป้องกันการทำนายไปยังกลุ่มที่มีขนาดใหญ่มากจนเกินไป

ในการศึกษาครั้งต่อไป ควรจะมีการเก็บรวมรวบข้อมูลอื่น ๆ ที่น่าจะส่งผลต่อระดับกลุ่มคุณค่าของลูกค้าในระยะยาว ได้แก่ รายได้ สถานะภาพสมรส อาชีพ เป็นต้น เพื่อนำมาเพิ่มประสิทธิภาพของตัวแบบในการทำนายระดับกลุ่มลูกค้าและทำการศึกษาแนวโน้มการเปลี่ยนกลุ่มของลูกค้า

References

Breiman, L. (2001). *Random Forest s. Machine Learning*, 45, 5–32. doi: 10.1023/A:1010933 404324.

Chang, W., Chang, C., & Li, Q. (2012). Customer Lifetime Value: A Review. *Social Behavior and Personality: an international journal,* 40(7), 1057-1064. doi:10.2224/sbp.2012.40. 7.1057.

Ho, T.K. (1995). *Random Decision Forests*. Proceedings of the 3rd International Conference on Document Analysis and Recognition 1, 278-282. doi: 10.1109/ICDA R.1995.598994.

Jackson, D. (1989). Determining a customer's lifetime value. *Direct Marketing*, 51, 60-63.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Office of Insurance Commission. (2019). *Annual General Insurance Business Overview Report 2019.* Retrieved form https://www.oic.or. th/th/industry /statistic/data/39/2.

RStudio Team. (2020). *RStudio: Integrated Development for R, Boston: RStudio, PBC*. Retrieved form http://www.rstudio.com/.