การเปรียบเทียบตัวแบบการพยากรณ์สำหรับเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย

The Comparison of Forecasting Models for Total Premiums of Life Insurance Companies in Thailand

วิกานดา ผาพันธ์[1] และ วิราวรรณ พุทธมาตย์[2*]

[1]ภาควิชาสถิติประยุกต์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร 10800

[2]สาขาวิชาคณิตศาสตร์ คณะครุศาสตร์ มหาวิทยาลัยราชภัฏชัยภูมิ ชัยภูมิ 36000

Wikanda Phaphan[1] and Wirawan Puttamat[2*]

[1]Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800

[2]Department of Mathematics, Faculty of Education, Chaiyaphum Rajabhat University, Chaiyaphum, 36000

***Corresponding author:** pwirawan6251@gmail.com

Received: 7 June 2023/ Revised: 28 September 2023/ Accepted: 30 September 2023

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบตัวแบบการพยากรณ์ 4 ตัวแบบ คือ 1) การถดถอยต้นไม้ตัดสินใจ 2) การถดถอยแบบป่าสุ่ม 3) ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย และ 4) การถดถอยพหุนาม ในการศึกษาเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตรายเดือนในประเทศไทยตั้งแต่เดือนมกราคม พ.ศ. 2560 ถึง เดือนธันวาคม พ.ศ.2565 จำนวน 72 เดือน โดยเกณฑ์ที่ใช้ในการเปรียบเทียบ คือ เปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAPE) และ รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) จากการศึกษาพบว่า ตัวแบบการถดถอยต้นไม้ตัดสินใจ มีค่า RMSE เท่ากับ 1654.00 และ ค่า MAPE เท่ากับ 2.93% ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย มีค่า RMSE เท่ากับ 5560.59 และค่า MAPE เท่ากับ 9.03% ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย มีค่า RMSE เท่ากับ 6283.63 และ ค่า MAPE เท่ากับ 11.36% และ ตัวแบบการถดถอยต้นไม้ตัดสินใจ มีค่า RMSE เท่ากับ 6723.48 และ ค่า MAPE เท่ากับ 11.76% ซึ่งพบว่าตัวแบบการถดถอยต้นไม้ตัดสินใจมีประสิทธิภาพดีที่สุด แต่เนื่องจากตัวแบบการถดถอยต้นไม้ตัดสินใจ ไม่เหมาะสำหรับการพยากรณ์ในระยะยาว จึงเลือกใช้ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอยแทน เนื่องจากตัวแบบมีประสิทธิภาพรองลงมาและมีความเหมาะสมกับข้อมูลที่นำมาใช้งาน

คำสำคัญ: การพยากรณ์ การถดถอยต้นไม้ตัดสินใจ การถดถอยแบบป่าสุ่ม ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย การถดถอยพหุนาม

Abstract

The purpose of this research, in the study of total premiums of monthly life insurance companies in Thailand during a period of 72 months from January 2017 to December 2022, was to study and compare the forecasting performance of four forecasting models, 1) Decision Tree Regression model, 2) Random Forest Regression model, 3) Support Vector Regression (SVR) model and 4) Polynomial Regression model. The forecasting efficiency was compared in the four models using Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The results of the research showed that 1) the Decision Tree Regression model with 1654.00 of RMSE and 2.93% of MAPE, 2) SVR model with 5560.59 of RMSE and 9.03% of MAPE, 3) Polynomial Regression model with 6283.63 of RMSE and 11.36% of MAPE and 4) Random Forest Regression model with 6723.48 of RMSE and 11.76% of MAPE. The model with the suitable forecast performance is the Decision Tree Regression, but the Decision Tree Regression model is not suitable for long-term forecasting, so then the SVR model was chosen as a better fit for long-term forecasting.

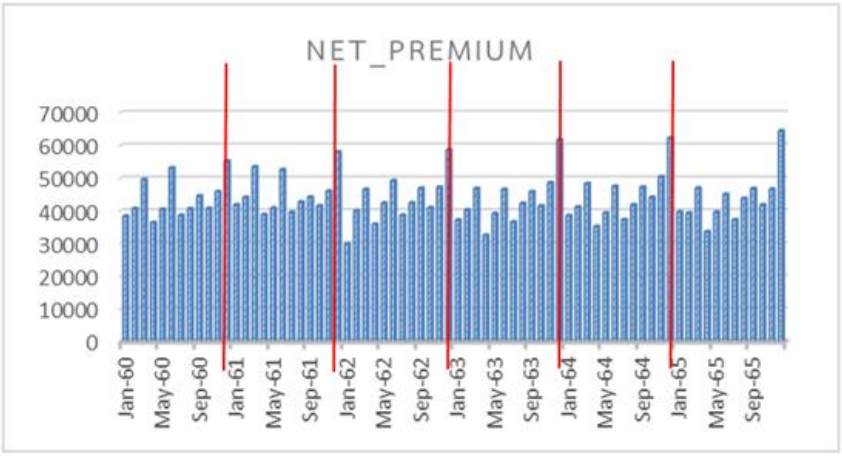
Keywords: Forecasting, Decision Tree Regression, Random Forest Regression, Support Vector Regression, Polynomial Regression model

บทนำ

ปัจจุบันอุตสาหกรรมประกันชีวิตเข้ามามีบทบาทในระบบเศรษฐกิจและสังคมในประเทศไทยอย่างมาก การทำประกันสุขภาพ ประกันชีวิตกับบริษัทประกันชีวิต จะช่วยบริหารความเสี่ยง บรรเทาความเดือดร้อน และคุ้มครองตั้งแต่การเจ็บป่วย การเกิดอุบัติเหตุ ไปจนถึงการเสียชีวิต ในปัจจุบันมีบริษัทประกันชีวิตในประเทศไทย 21 แห่ง โดยบริษัทที่มีสัดส่วนการตลาด (Market Share) มากที่สุดคือ บริษัท เอไอเอ (American International Assurance) 24.90% รองลงมาคือ บริษัท ไทยประกันชีวิต จำกัด (มหาชน) 14.41% และบริษัท เอฟดับบลิวดี ประกันชีวิต จำกัด (มหาชน) 13.65% ตามลำดับ [1]

เบี้ยประกันภัยรับรวม (Total Premium) คือจำนวนเบี้ยประกันภัยที่ผู้รับประกันภัยได้รับทั้งหมดก่อนหักเบี้ยประกันภัยต่อ โดยนำมาใช้คำนวณเปรียบเทียบกับความคุ้มครองที่ผู้เอาประกันภัยจะได้รับ จากภาพที่ 1 จะเห็นว่าเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตนั้นจะมียอดสูงสุดในช่วงเดือนธันวาคมของทุกปี และมียอดต่ำสุดในช่วงเดือนมกราคมของทุกปี นั่นคือข้อมูลมีฤดูกาลเข้ามาเกี่ยวข้อง ซึ่งส่งผลให้การประมาณการณ์ค่าเบี้ยประกันภัยรับรวมยากขึ้นด้วย เนื่องจากเบี้ยประกันภัยรับรวมถือเป็นรายได้ของบริษัท เพื่อให้บริษัทประกันชีวิตสามารถจัดสรรเบี้ยประกันภัยต่อและจำนวนเงินสำรองเพื่อการเสี่ยงภัยได้อย่างหมาะสม ดังนั้นบริษัทประกันชีวิตจึงจำเป็นที่จะต้องนำหลักการพยากรณ์ต่างๆ มาประยุกต์ใช้ในการพยากรณ์ทิศทางแนวโน้มการเปลี่ยนแปลงของเบี้ยประกันภัยรับรวม เพื่อนำมาใช้เป็นแนวทางในการวางแผนการดำเนินงานได้

มีนักวิจัยจำนวนมากให้ความสำคัญเกี่ยวกับการพยากรณ์โดยอาศัยวิธีการต่างๆ เช่น สร้างแบบจำลองการขายผลิตภัณฑ์และพยากรณ์ยอดขายประกันชีวิตโดยใช้เทคนิคการทำเหมืองข้อมูล [2] การพยากรณ์เบี้ยประกันภัยรับโดยตรงในกรมธรรม์หลัก (ประเภทสามัญ) ของบริษัทไทยประกันชีวิต [3] การพยากรณ์เบี้ยประกันชีวิตรายใหม่ประเภทสามัญด้วยวิธีปรับให้เรียบเอ็กซ์โปเนนเชียลแบบโฮลท์วินเทอร์วิธีบ็อกซ์-เจนกินส์ [4] การพยากรณ์เบี้ยประกันภัยรับโดยตรงโดยใช้ตัวแบบอนุกรมเวลา SARIMA และ SARIMAX [5] เป็นต้น โดยส่วนใหญ่เป็นวิธีการพยากรณ์แบบเชิงเส้นตรง (Linear approach) ซึ่งมีข้อเสียคือไม่สามารถอธิบายอนุกรมเวลาที่ซับซ้อนหรือมีความสัมพันธ์ไม่เป็นเส้นตรงได้ ทำให้ค่าพยากรณ์มีความคลาดเคลื่อนค่อนข้างมาก [6]



ภาพที่ 1 เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยรายเดือนตั้งแต่ พ.ศ.2560 ถึง พ.ศ.2565 [7]

จากเหตุผลดังกล่าวข้างต้น ผู้วิจัยจึงเลือกใช้ตัวแบบการพยากรณ์ 4 ตัวแบบ คือ 1) การถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression) 2) การถดถอยแบบป่าสุ่ม (Random Forest Regression) 3) ซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย (Support Vector Regression) และ 4) การถดถอยพหุนาม (Polynomial Regression) ในการพยากรณ์เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย เนื่องจากตัวแบบดังกล่าวมีประสิทธิภาพในการพยากรณ์อนุกรมเวลาที่ไม่เป็นเชิงเส้นตรงได้ดี โดยเปรียบเทียบความแม่นยำในการพยากรณ์ เพื่อหาตัวแบบที่เหมาะสมที่สุดด้วยเกณฑ์รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และค่าเฉลี่ยของค่าสัมบูรณ์เปอร์เซ็นต์ความคลาดเคลื่อน (MAPE)

วัตถุประสงค์ของการวิจัย

1) เพื่อศึกษาการนำตัวแบบการพยากรณ์ไปใช้กับปัญหาการประมาณการเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย
2) เพื่อเปรียบเทียบตัวแบบการพยากรณ์สำหรับเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย
3) เพื่อให้ได้ตัวแบบการพยากรณ์ที่เหมาะสมกับชุดข้อมูลมากที่สุด สำหรับนำไปพยากรณ์ล่วงหน้า

ทฤษฎีที่เกี่ยวข้อง

1) การถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression)

เป็นแบบจำลองทางคณิตศาสตร์เพื่อพยากรณ์ค่าต่อเนื่อง โดยการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ [8] โดยมีหลักการในการพยากรณ์ คือ 1) เลือกตัวแปรอิสระ X มา 1 ตัวแปร โดยเลือกจากตัวแปร X_j ที่ให้ค่า residual sum of squares (RSS) น้อยที่สุด มาทำการเรียงลำดับข้อมูล 2) หาจุดแบ่งข้อมูล (split point) ที่เป็นไปได้ทั้งหมด จากข้อมูล n ค่าสังเกต 3) สำหรับการแบ่งข้อมูลแต่ละแบบที่เป็นไปได้ ทำการคำนวณค่า residual sum of squares (RSS) ตามสูตรในสมการที่ (1) โดยกำหนดให้ค่าพยากรณ์ คือค่าเฉลี่ยของตัวแปรตาม Y ภายในโหนดของตัวเอง 4) เลือกจุดแบ่งข้อมูลที่ให้ค่า RSS น้อยที่สุด และสร้างเงื่อนไขจากจุดแบ่งข้อมูลในโครงสร้างต้นไม้ 5) หาจุดแบ่งข้อมูลของตัวแปรอิสระ X_i ที่เหลือจนครบ (ตามข้อ 1) – 4)) เมื่อสิ้นสุดการแบ่งข้อมูลแล้วจะพยากรณ์ค่าตัวแปรตาม Y จากค่าเฉลี่ยของตัวแปรตาม Y ภายใน

โหนดของตัวเอง สำหรับโหนดที่ยังสามารถแบ่งข้อมูลได้ต่อ จะแบ่งต่อไปจนได้เงื่อนไขที่กำหนด ตามสมการที่ (1)

$$RSS = \sum_{w=1}^{W} \sum_{i \in R_w}^{n} \left(Y_i - \hat{Y}_{R_w} \right)^2 \quad (1)$$

โดยที่ R_w คือ แต่ละกลุ่มของค่าสังเกตที่ถูกแบ่งออกมาเป็นทั้งหมด w กลุ่ม

Y_i คือ ตัวแปรตาม

$\hat{Y}_{R_w}$ คือ ค่าประมาณในแต่ละกลุ่ม คำนวณมาจากค่าเฉลี่ยของตัวแปรตามในกลุ่มนั้น ๆ

2) การถดถอยแบบป่าสุ่ม (Random Forest Regression)

เปนวิธีที่อิงจากแบบจำลองตนไมถดถอย (Regression tree model) ดังที่จะอธิบายตอไป แบบจำลองตนไมถดถอยเปนการประยุกตหลักการของแบบจำลองตนไมตัดสินใจ (Decision tree model) เพื่อใชในการพยากรณคาของตัวแปรที่พิจารณาโดยอาศัยวิธีการแบงกลุมของตัวแปรตน หลักการของแบบจำลองตนไม้ถดถอย สามารถอธิบายไดเปน 2 ขั้นตอนดังนี้ [9]

1. แบงปริภูมิของตัวแปรตน $X_1, X_2, \ldots, X_p$ ออกเปน j สวนที่ไมมีการซอนทับซึ่งกันและกัน ใหปริภูมิยอยนั้นเรียกวา $R_1, R_2, \ldots, R_j$ ซึ่งมีลักษณะเปน high-dimensional rectangles เพื่อใหไดปริภูมิยอย $R_1, R_2, \ldots, R_j$ ซึ่งใหคา RSS ที่นอยที่สุด กำหนดโดยสมการที่ (2)

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(Y_i - \hat{Y}_{R_j} \right)^2 \quad (2)$$

โดยที่ R_j คือ แต่ละกลุ่มของค่าสังเกตที่ถูกแบ่งออกมาเป็นทั้งหมด j กลุ่ม

Y_i คือ ตัวแปรตาม

$\hat{Y}_{R_j}$ คือ ค่าประมาณในแต่ละกลุ่ม คำนวณมาจากค่าเฉลี่ยของตัวแปรตามในกลุ่มนั้น ๆ

2. สำหรับทุกๆ ขอมูลของตัวแปรตนที่อยูใน R_j เราจะพยากรณคาของตัวแปรตามใหมีคาเทากับคาเฉลี่ยของคาตัวแปรตามในชุดขอมูลฝกทั้งหมด ซึ่งคาของตัวแปรตนตกอยูใน R_j

3) ซัพพอร์ตเวกเตอร์สำหรับการถดถอย (Support Vector Regression)

เป็นการดัดแปลงมาจากวิธีซัพพอร์ตเวกเตอร์แมชชีน โดยใช้สมการไฮเปอร์เพลน (hyperplane) เป็นสมการพยากรณ์ข้อมูลที่เป็นเลขจำนวนจริง หลักการคือจะลดความเสี่ยงเชิงโครงสร้างให้มีค่าต่ำที่สุด โดยใช้ฟังก์ชันในการปรับรูปแบบข้อมูล เรียกว่า ฟังก์ชันเคอร์เนล (kernel function) [10] โดยมีสมการถดถอยเป็นดังสมการที่ (3)

$$\hat{Y} = \langle w, x \rangle + b \quad (3)$$

โดยที่ w คือ เวกเตอร์น้ำหนักหรือความชันของเส้นถดถอย

b คือ ความคลาดเคลื่อนของเส้นถดถอย

x คือ เวกเตอร์ข้อมูลนำเข้า (Input data) แทนด้วยสัญลักษณ์ $(x_1, x_2, \ldots, x_n)^T$

ขนาดเท่ากับ n ; $x_i \in R^n; i = 1, 2, \ldots, n$

หลักการในการหาไฮเปอร์เพลนที่เหมาะสมที่สุด สำหรับกลุ่มของข้อมูล ในกรณีของซัพพอร์ตเวกเตอร์ จะหาตำแหน่งของข้อมูลที่เป็นซัพพอร์ตเวกเตอร์ ซึ่งเป็นข้อมูลในตำแหน่งที่อยู่ห่างจากไฮเปอร์เพลนมากที่สุด และตำแหน่งดังกล่าวจะอยู่บนเส้นแบ่งระยะขอบเขต (boundary line) ที่กำหนดขึ้น ในทางทฤษฎีจะพยายามให้ข้อมูลทั้งหมดอยู่ภายในเส้นแบ่งระยะขอบเขต และสร้างไฮเปอร์เพลนที่ใช้แทนกลุ่มข้อมูลขึ้น [11]

ฟังก์ชันเคอร์เนลที่นิยมใช้สำหรับซัพพอร์ตเวกเตอร์สำหรับการถดถอย มีดังนี้

- Radial basis function; $k(x,x_i)=\exp\left(-\gamma\|x-x_i\|^2\right),\gamma>0$
- Polynomial; $k(x,x_i)=\left(\left(x^T\cdot x_i\right)+\eta\right)^d$
- Sigmoidal; $k(x,x_i)=\tanh\left(\gamma\left(x^T\cdot x_i\right)+\eta\right),\gamma>0$
- Linear; $k(x,x_i)=x^T\cdot x_i$

ซึ่งในงานวิจัยนี้ใช้ฟังก์ชันเคอร์เนลที่นิยมใช้ในการสร้างตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย คือ radial basic function ดังนั้นสมการไฮเปอร์เพลนสามารถเขียนใหม่ในรูปแบบการถดถอยไม่เชิงเส้น โดยใช้ฟังก์ชันเคอร์เนลได้ดังสมการที่ (4)

$$\hat{Y}=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)k\left(x_i,x_j\right)+b \tag{4}$$

โดยที่ x_i คือ เวกเตอร์ข้อมูลนำเข้า (input data)

x_j คือ ซัพพอร์ตเวกเตอร์ (support vector)

α_i,α_i^* คือ ตัวคูณลากรานจ์ (lagrange multipliers) โดยกำหนดให้ค่าตัวคูณลากรานจ์มีค่ามากกว่าศูนย์ และ สามารถหาค่าโดยใช้ขั้นตอนวิธีการโปรแกรมกำลังสอง (quadratic programming)

4) การถดถอยพหุนาม (Polynomial Regression)

เป็นตัวแบบการถดถอยเชิงเส้นรูปแบบหนึ่งที่แสดงความสัมพันธ์ระหว่างตัวแปรที่มีความสัมพันธ์ไม่เป็นเส้นตรง (ความสัมพันธ์เป็นแบบเส้นโค้ง) แต่ใช้ตัวแปรกำหนดเพียง 1 ตัว ซึ่งขั้นตอนการวิเคราะห์จะมีความยากและซับซ้อนยิ่งขึ้น โดยรูปแบบความสัมพันธ์เขียนแสดงในรูปแบบสมการที่ (5)

$$y=\beta_0+\beta_1X+\beta_2X^2+\ldots+\beta_nX^n \tag{5}$$

การถดถอยพหุนามมีข้อตกลงเบื้องต้น คือ 1) ความคลาดเคลื่อนแต่ละค่าเป็นอิสระต่อกัน 2) ความคลาดเคลื่อนมีค่าเฉลี่ยเป็นศูนย์ 3) ความคลาดเคลื่อนมีความแปรปรวนคงที่ และ 4) ความคลาดเคลื่อนมีการแจกแจงแบบปกติ [12]

การถดถอยแบบพหุนามนี้สามารถพิจารณาได้ว่าเป็นรูปแบบพิเศษรูปแบบหนึ่งของการถดถอยเชิงเส้นพหุคูณ ดังนั้นเราจึงสามารถหาค่าพารามิเตอร์ที่เหมาะสมของตัวแบบได้โดยใช้วิธีกำลังสองน้อยสุด เช่นเดียวกับการหาค่าพารามิเตอร์ของตัวแบบการถดถอยเชิงเส้นพหุคูณ [13]

วิธีดำเนินการวิจัย

1) การเก็บรวบรวมข้อมูล

ข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลทุติยภูมิรายเดือน โดยรวบรวมข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย จากสมาคมประกันชีวิตไทย ตั้งแต่เดือนมกราคม พ.ศ.2560 ถึง เดือนธันวาคม พ.ศ.2565 จำนวน 72 เดือน [7] ด้วยการใช้โปรแกรมไมโครซอฟต์เอ็กเซล โดยกำหนดตัวแปรดังนี้

1) ตัวแปรตาม (Y) คือ เบี้ยประกันภัยรับรวม ประเภทสามัญ (Ordinary) ของบริษัทประกันชีวิตรายเดือน ในประเทศไทย ณ เดือนที่ t แทนด้วย Net_Premium มีหน่วยเป็นล้านบาท

2) ตัวแปรอิสระ (X) คือ ระยะเวลามีหน่วยเป็นเดือน ในที่นี้เก็บเป็นรายเดือน ตั้งแต่มกราคม 2560 – ธันวาคม 2565

2) การวิเคราะห์ข้อมูล

จากการเก็บรวบรวมข้อมูลจะแบ่งข้อมูลเป็น 2 ส่วน คือ ส่วนที่ 1 ข้อมูลที่ใช้สร้างตัวแบบจำนวน 57 รายการ (ชุดการเรียนรู้ 80%) และส่วนที่ 2 ข้อมูลที่ใช้ในการทดสอบจำนวน 15 รายการ (ชุดทดสอบ 20%) โดยใช้ฟังก์ชัน split ในโปรแกรม R จากนั้นทำการวิเคราะห์ข้อมูลด้วยตัวแบบการพยากรณ์ 4 ตัวแบบ คือ

ตัวแบบการถดถอยต้นไม้ตัดสินใจ

สร้างตัวแบบการถดถอยต้นไม้ตัดสินใจ ในโปรแกรม R โดยใช้คำสั่ง rpart( ) ในแพกเกจชื่อ rpart [14] กำหนดการแบ่งข้อมูลออกเป็นแบบไบนารี กำหนดให้ตัวแปรตามคือ เบี้ยประกันภัยรับรวม ประเภทสามัญ (Ordinary) ของบริษัทประกันชีวิตรายเดือน ในประเทศไทย ณ เดือนที่ t แทนด้วย Net_Premium มีหน่วยเป็น ล้านบาท และตัวแปรอิสระคือระยะเวลาในที่นี้เก็บเป็นรายเดือน ตั้งแต่มกราคม 2560 – ธันวาคม 2565

ตัวแบบการถดถอยแบบป่าสุ่ม

ในการสร้างตัวแบบการถดถอยแบบป่าสุ่ม ผู้วิจัยเลือกใช้แพคเกจของโปรแกรม R ในแพกเกจชื่อ randomForest [15] เพื่อให้ได้ตัวแบบที่มีประสิทธิภาพ ผู้วิจัยได้ทำการปรับแต่งไฮเปอร์พารามิเตอร์ในตัวแบบด้วยวิธีการค้นหาแบบกริด โดยกำหนดให้ช่วงของจำนวนต้นไม้ (ntrees) อยู่ระหว่าง 5 ถึง 500 ต้น โดยให้เพิ่มขึ้นทีละ 5 ต้น ช่วงของจำนวนตัวแปรที่สุ่มเลือกในแต่ละการแบ่ง (mtry) อยู่ระหว่าง 1 ถึง 12 ช่วงของจำนวนตัวอย่างขั้นต่ำในโหนดปลาย (nodesize) มีค่าอยู่ระหว่าง 3 ถึง 10 และช่วงของจำนวนสูงสุดของโหนดปลาย (maxnodes) มีค่าอยู่ระหว่าง 3 ถึง 10 จากการค้นหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม พบว่าไฮเปอร์พารามิเตอร์ที่ทำให้ค่า out-of-bag (OOB) RMSE ต่ำที่สุด คือ ntree = 500, mtry = 1, nodesize = 3, และ maxnodes = 4

ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย

หาตัวแบบพยากรณ์ซัพพอร์ตเวกเตอร์สำหรับการถดถอย ในโปรแกรม R โดยใช้คำสั่ง svm( ) ในแพกเกจชื่อ e1071 [16] โดยสร้างตัวแบบพยากรณ์จำนวน 6 ตัวแบบ เนื่องจากงานวิจัยนี้ทำการพยากรณ์ล่วงหน้า 6 เดือน ใช้ฟังก์ชัน tune () ค้นหาเคอร์เนลฟังก์ชันและกำหนดค่าพารามิเตอร์ที่เหมาะสม โดยฟังก์ชัน tune ( ) จะค้นหาค่าที่เหมาะสมแบบกริดเสิร์จ (Grid search) ตลอดช่วงของพารามิเตอร์ที่กำหนดไว้ งานวิจัยนี้กำหนดค่าไว้ที่ 0 -1 และให้เพิ่มค่าทีละ 0.001 และกำหนดค่า Cost ไว้ที่ 1-16 และประเมินความเหมาะสมด้วยการวัดค่า RMSE ผลลัพธ์การค้นหาเคอร์เนลฟังก์ชันที่เหมาะสม จากฟังก์ชัน tune ( ) พบว่า เคอร์เนลแบบเรเดียลเบซิส (Radial Basis Function) ให้ผลลัพธ์ดีที่สุดกับข้อมูลชุดนี้ ดังนั้นงานวิจัยนี้จึงใช้เคอร์เนลแบบเรเดียลเบซิสเป็นเคอร์เนลฟังก์ชัน

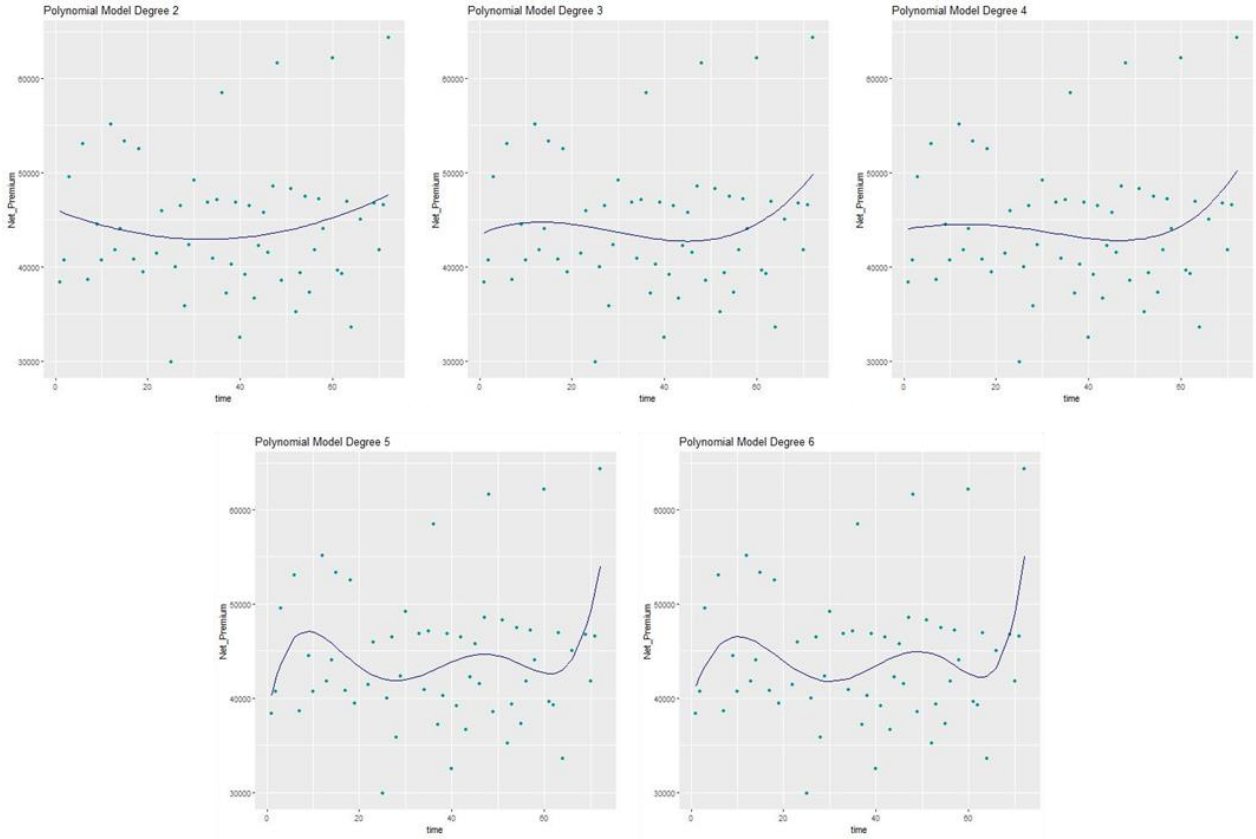
กำหนดตัวแปรตามคือค่า Net_Premium และตัวแปรที่เหลือคือตัวแปรอิสระ กำหนด type เป็น ‘eps-regression’ สำหรับตัวแปรตามที่เป็นจำนวนต่อเนื่อง และกำหนดค่าพารามิเตอร์เป็นค่า default

ตัวแบบการถดถอยพหุนาม

สำหรับการวิเคราะห์การถดถอยพหุนาม ผู้วิจัยได้ทำการตรวจสอบข้อตกลงเบื้องต้นของชุดข้อมูลที่ใช้สำหรับสร้างตัวแบบโดยตรวจสอบการแจกแจงของความคลาดเคลื่อน ตรวจสอบความเท่ากันของความแปรปรวนของความคลาดเคลื่อน และตรวจสอบความเป็นอิสระกันของความคลาดเคลื่อน ซึ่งมีการวิเคราะห์ดังนี้ ในการตรวจสอบการแจกแจงของความคลาดเคลื่อน ทำการทดสอบสมมติฐานด้วยสถิติทดสอบของ Kolmogorov-Smirnov (เนื่องจากจำนวนข้อมูลที่นำมาใช้ในการวิเคราะห์มากกว่า 50 ค่า) พบว่า ความคลาดเคลื่อนมีการแจกแจงแบบปกติที่ระดับนัยสำคัญ 0.05 โดยมีค่าสถิติ ทดสอบเท่ากับ 0.06 และ p-value เท่ากับ 0.20 และจากการทดสอบสมมติฐานด้วยสถิติทดสอบของ Durbin-Watson พบว่า ค่าสถิติทดสอบเท่ากับ 2.10 แสดงว่า

ความคลาดเคลื่อนเป็นอิสระกัน Breusch-Pagan พบว่า ค่าสถิติทดสอบเท่ากับ 1.17 และ p-value เท่ากับ 0.95 แสดงว่า ความแปรปรวนความคลาดเคลื่อนคงที่ ที่ระดับนัยสำคัญ 0.05 จากการตรวจสอบข้อตกลงเบื้องต้นของการวิเคราะห์การถดถอยพหุนาม พบว่า เป็นไปตามข้อตกลงเบื้องต้นทุกประการ

ในการสร้างตัวแบบการถดถอยพหุนามในโปรแกรม R ผู้วิจัยใช้คำสั่ง lm( ) ในแพกเกจชื่อ caTools [17] โดยกำหนดให้ตัวแปร X แทน ระยะเวลา ซึ่งในที่นี้เก็บเป็นรายเดือน ตั้งแต่มกราคม 2560 – ธันวาคม 2565 และทดสอบหากำลังที่เหมาะสมของตัวแบบว่าควรจะมีกำลังสูงสุดเป็นเท่าไร โดยทดสอบกำลังในระดับ 2, 3, 4, 5 และ 6 ผลการทดสอบเป็นดังภาพที่ 2



ภาพที่ 2 กราฟข้อมูลเบี้ยประกันภัยรับรวมจริง และเบี้ยประกันรับรวมที่ได้จากการพยากรณ์จากตัวแบบการถดถอยพหุนามที่มีกำลังเท่ากับ 2, 3, 4, 5 และ 6

พบว่าตัวแบบที่มีกำลังเท่ากับ 5 และ 6 มีลักษณะกราฟที่ข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยกระจายอยู่รอบ ๆ เส้นของตัวแบบพยากรณ์มากกว่าตัวแบบที่มีกำลังเท่ากับ 2, 3 และ 4 แม้ว่าตัวแบบยิ่งมีกำลังสูงจะทำให้ผลการพยากรณ์ชุดข้อมูลที่มีอยู่แม่นยำขึ้น แต่ตัวแบบที่ได้ออกมานั้นจะพยากรณ์ผลของข้อมูลชุดใหม่ได้ยากและซับซ้อนกว่าเดิม ซึ่งอาจจะทำให้ผลของการพยากรณ์ไม่ตรงกับข้อมูลความเป็นจริง ปัญหานี้เราเรียกว่า Model Overfitting ดังนั้นกำลังเหมาะสมสำหรับงานวิจัยนี้คือ กำลังเท่ากับ 5 จะได้สมการถดถอย ดังสมการที่ (6)

$$\hat{Y}_t = 0.12X + 0.09X^2 + 0.08X^3 + 0.08X^4 + 0.07X^5 \quad (6)$$

โดยที่ Y_t คือ ข้อมูล ณ เวลาที่ t

$\hat{Y}_t$ คือ ค่าพยากรณ์ของข้อมูล ณ เวลาที่ t

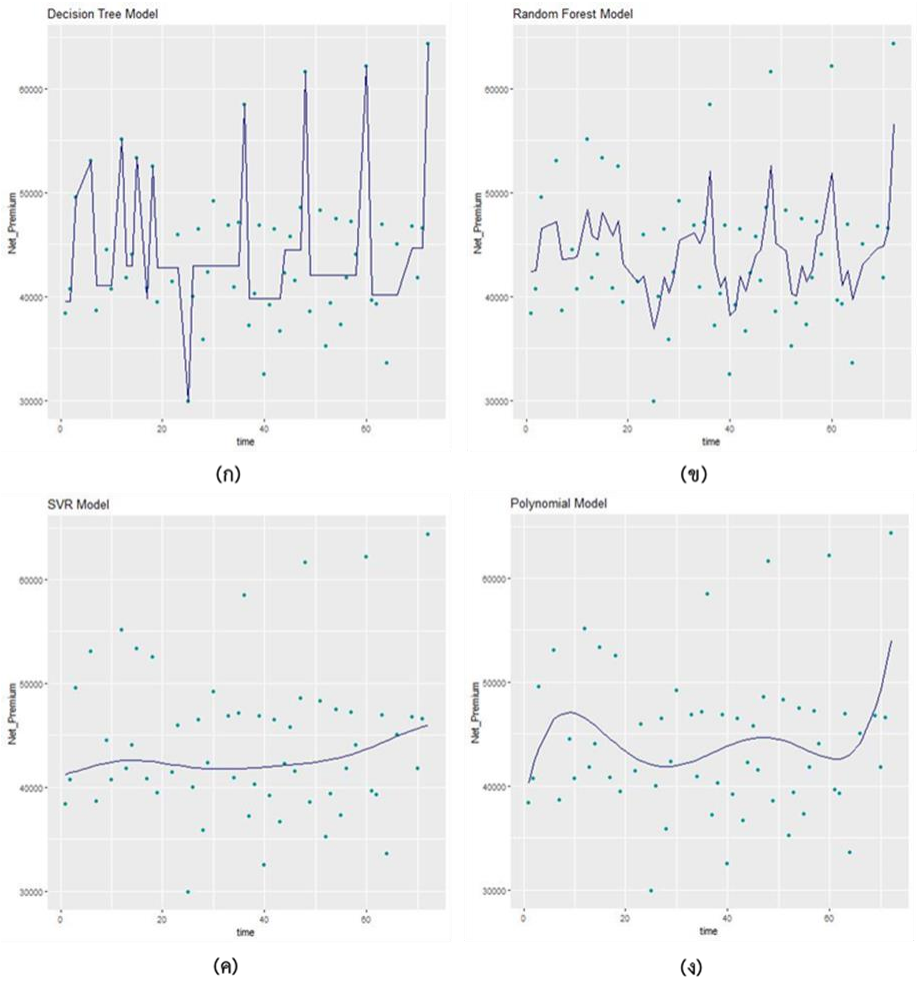
n คือ จำนวนข้อมูลตัวอย่าง

ผลการวิจัย

การศึกษาครั้งนี้ใช้เทคนิคการเรียนรู้ของเครื่อง 4 ตัวแบบ คือ 1) การถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression) 2) การถดถอยแบบป่าสุ่ม (Random Forest Regression) 3) ซัพพอร์ตเวกเตอร์สำหรับการถดถอย (Support Vector Regression) และ 4) การถดถอยพหุนาม (Polynomial Regression) ในการวิเคราะห์เพื่อหาวิธีพยากรณ์ที่เหมาะสมที่สุดสำหรับข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตรายเดือนในประเทศไทย โดยนำชุดข้อมูลส่วนที่ 1 มาสร้างตัวแบบการพยากรณ์ ผลการวิเคราะห์สรุปได้ดังนี้

การถดถอยต้นไม้ตัดสินใจ

จากผลการทดลอง ผู้วิจัยได้นำเสนอผลของเบี้ยประกันภัยรับรวมจริงในแต่ละเดือน รวม 72 เดือน และเบี้ยประกันรับรวมที่ได้จากการพยากรณ์จากตัวแบบการถดถอยต้นไม้ตัดสินใจ ตัวแบบการถดถอยแบบป่าสุ่ม ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย และตัวแบบการถดถอยพหุนาม ดังแสดงในภาพที่ 3



ภาพที่ 3 กราฟข้อมูลเบี้ยประกันภัยรับรวมจริง และเบี้ยประกันรับรวมที่ได้จากการพยากรณ์จาก (ก) ตัวแบบการถดถอยต้นไม้ตัดสินใจ (ข) ตัวแบบการถดถอยแบบป่าสุ่ม (ค) ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย และ (ง) ตัวแบบการถดถอยพหุนาม

จากภาพที่ 3 ลักษณะกราฟที่ได้จากการพยากรณ์จากตัวแบบทั้ง 4 ตัวแบบ เป็นดังนี้

ภาพที่ 3 (ก) ตัวแบบการถดถอยต้นไม้ตัดสินใจ ลักษณะกราฟมีการขึ้นลงและข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยส่วนใหญ่อยู่ตามเส้นของตัวแบบพยากรณ์ นั่นคือค่าพยากรณ์จากตัวแบบการถดถอยต้นไม้ตัดสินใจมีลักษณะเข้าใกล้กับค่าจริง ภาพที่ 3 (ข) ตัวแบบการถดถอยแบบป่าสุ่ม ข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยส่วนใหญ่อยู่นอกเส้นของตัวแบบพยากรณ์ นั่นคือค่าพยากรณ์จากตัวแบบการถดถอยแบบป่าสุ่ม มีลักษณะห่างจากค่าจริง ภาพที่ 3 (ค) ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย ข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยกระจายอยู่รอบ ๆ เส้นของตัวแบบพยากรณ์ นั่นคือค่าพยากรณ์จากตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย มีลักษณะห่างจากค่าจริง ภาพที่ 3 (ง) ตัวแบบการถดถอยพหุนาม ข้อมูลเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยกระจายอยู่รอบ ๆ เส้นของตัวแบบพยากรณ์ นั่นคือค่าพยากรณ์จากตัวแบบการถดถอยพหุนาม มีลักษณะห่างจากค่าจริง

สรุปได้ว่าค่าพยากรณ์จากตัวแบบการถดถอยต้นไม้ตัดสินใจ มีลักษณะเข้าใกล้กับค่าจริงมากที่สุด เมื่อเปรียบเทียบกับตัวแบบการถดถอยแบบป่าสุ่ม ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย และตัวแบบการถดถอยพหุนาม

ประเมินประสิทธิภาพของตัวแบบ

ผู้วิจัยประเมินประสิทธิภาพของตัวแบบโดยเปรียบเทียบ ค่า RMSE และ ค่า MAPE ของทั้ง 4 ตัวแบบ ได้ผลดังตารางที่ 1

ตารางที่ 1 ผลการประเมินประสิทธิภาพของตัวแบบการพยากรณ์

ตัวแบบ	ประสิทธิภาพของตัวแบบ	
	RMSE	MAPE (%)
การถดถอยต้นไม้ตัดสินใจ	1,654.00	2.93
การถดถอยแบบป่าสุ่ม	6,723.48	11.76
ซัพพอร์ตเวกเตอร์สำหรับการถดถอย	5,560.59	9.03
การถดถอยพหุนาม	6,283.63	11.36

จากตารางที่ 1 พบว่าตัวแบบการถดถอยต้นไม้ตัดสินใจ ที่ให้ความคลาดเคลื่อนต่ำสุด คือ ค่า RMSE = 1,654.00 และ ค่า MAPE = 2.93 รองลงมาคือตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย (RMSE = 5,560.59 , MAPE = 9.03) ตัวแบบการถดถอยพหุนาม (RMSE = 6,283.63 , MAPE = 11.36) และตัวแบบการถดถอยแบบป่าสุ่ม (RMSE = 6,723.48 , MAPE = 11.76) ตามลำดับ นั่นคือ ตัวแบบการถดถอยต้นไม้ตัดสินใจ มีประสิทธิภาพดีที่สุดเมื่อเทียบกับตัวแบบการถดถอยแบบป่าสุ่ม ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย และตัวแบบการถดถอยพหุนาม

อภิปรายและสรุปผลการวิจัย

จากการศึกษาพบว่า ตัวแบบการถดถอยต้นไม้ตัดสินใจมีประสิทธิภาพดีที่สุดเมื่อเทียบกับตัวแบบทั้ง 3 เนื่องจากให้ค่า RMSE และค่า MAPE ต่ำที่สุด และเหตุผลที่ทำให้ตัวแบบการถดถอยต้นไม้ตัดสินใจให้ค่า RMSE และค่า MAPE ต่ำที่สุด อาจเป็นเพราะการถดถอยต้นไม้ตัดสินใจจะเลือกตัวแปรที่นำมาใช้ในการสร้างตัวแบบทีละลำดับขั้นและถ้าได้ตัวแปรที่มีความสัมพันธ์กับคำตอบแล้ว ก็จะไม่ใช้ทุกตัวแปรในชุดการเรียนรู้ แต่เนื่องจากตัวแบบการถดถอยต้นไม้ตัดสินใจไม่เหมาะสำหรับการพยากรณ์ในระยะยาว เนื่องจากตัวแบบการถดถอยต้นไม้ตัดสินใจ ใช้ค่าเฉลี่ยในแต่ละกลุ่มที่แบ่งได้เป็นค่าพยากรณ์ และตัวแปรอิสระที่ใช้เป็นตัวเวลาซึ่งมีค่าเป็น 1, 2, 3, 4,... เมื่อทำการพยากรณ์ระยะยาวจะได้ค่าพยากรณ์เป็นค่าเดียวกันเสมอ แต่ตัวแบบการถดถอยต้นไม้

ตัดสินใจเหมาะสำหรับการทำนายในระยะสั้น 1 ถึง 2 ช่วงเวลา ดังนั้นสำหรับการประมาณการณ์เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทยในระยะยาว ผู้วิจัยจึงเลือกใช้ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอยแทน เนื่องจากตัวแบบมีประสิทธิภาพรองลงมา โดยให้ค่า RMSE เท่ากับ 5,560.59 และค่า MAPE เท่ากับ 9.03% ส่วนตัวแบบการถดถอยแบบป่าสุ่มไม่เหมาะกับการนำไปใช้ในการพยากรณ์เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิต เนื่องจากตัวแบบการถดถอยแบบป่าสุ่มให้ค่า RMSE และค่า MAPE สูงที่สุด และข้อมูลเบี้ยประกันภัยรับรวมที่ใช้ในงานวิจัยนี้มีค่าที่ค่อนข้างแตกต่างกัน จึงทำให้ตัวแบบการถดถอยแบบป่าสุ่มไม่เหมาะสมกับข้อมูลชุดนี้

จากนั้นผู้วิจัยจึงนำตัวแบบที่ได้ไปทำการพยากรณ์เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตล่วงหน้า 6 เดือน โดยแสดงดังตารางที่ 2

ตารางที่ 2 ค่าพยากรณ์ล่วงหน้า 6 เดือน

เดือน	ค่าพยากรณ์ (ล้านบาท)		ค่าจริง (ล้านบาท)
	ตัวแบบซัพพอร์ตเวกเตอร์สำหรับการถดถอย	ตัวแบบการถดถอยต้นไม้ตัดสินใจ	
มกราคม 2566	46,100.06	64,363.56	31,239.32
กุมภาพันธ์ 2566	46,200.81	64,363.56	31,174.93
มีนาคม 2566	46,278.93	64,363.56	37,782.62
เมษายน 2566	46,333.58	64,363.56	26,787.06
พฤษภาคม 2566	46,364.36	64,363.56	32,161.70
มิถุนายน 2566	46,371.31	64,363.56	36,639.68

ตัวแบบที่ใช้ในงานวิจัยนี้เป็นตัวแบบที่ง่ายใช้ตัวแปรเวลาแค่ตัวแปรเดียวในการพยากรณ์ ซึ่งมีความแม่นยำในระดับหนึ่ง สามารถนำผลที่ได้ไปใช้คาดคะเนสิ่งที่จะเกิดขึ้นในอนาคตได้โดยสังเขป

ข้อเสนอแนะ

ในการนำไปใช้ประโยชน์นักวิจัยหรือผู้ที่สนใจสามารถนำผลการวิจัยนี้ไปใช้ประโยชน์ในการวางแผนการวิจัยได้ สามารถนำตัวแบบที่มีประสิทธิภาพที่สุดไปประยุกต์ใช้ในการประมาณการณ์เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในอนาคตเพื่อใช้เป็นแนวทางในการวางแผนการดำเนินงานได้

สำหรับการทำวิจัยครั้งต่อไปควรเพิ่มตัวแปรอิสระอื่น ๆ เช่น จำนวนกรมธรรม์ มาใช้เป็นตัวแปรร่วมในการวิเคราะห์เพื่อให้ผลการวิเคราะห์ที่ได้มีประสิทธิภาพมากยิ่งขึ้น และนำข้อมูลมาประยุกต์ใช้กับการเรียนรู้เชิงลึก (Deep Learning) เพื่อให้ได้ตัวแบบการพยากรณ์ที่มีประสิทธิภาพมากขึ้น

เอกสารอ้างอิง

1. สมาคมประกันชีวิตไทย. รายงานธุรกิจประกันชีวิต ณ เดือนธันวาคม 2565. [อินเทอร์เน็ต]. 2565 [เข้าถึงเมื่อ 26 ก.พ. 2566]. เข้าถึงได้จาก: https://www.tlaa.org/page_statistics.php?cid=35
2. นิสานันท์ พลอาสา. การสร้างแบบจำลองการขายผลิตภัณฑ์และพยากรณ์ยอดขายประกันชีวิต โดยเทคนิคการทำเหมืองข้อมูล กรณีศึกษาบริษัทประกันชีวิตแห่งหนึ่ง. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาการบริหารเทคโนโลยี วิทยาลัยนวัตกรรม. มหาวิทยาลัยธรรมศาสตร์. กรุงเทพฯ; 2558.

3. ภิญญภา บุญเกษมธนกุล. การพยากรณ์เบี้ยประกันภัยรับโดยตรงในกรมธรรม์หลัก (ประเภทสามัญ) ของบริษัทไทยประกันชีวิต จำกัด (มหาชน). วิทยานิพนธ์ปริญญาวิทยาศาสตรบัณฑิต สาขาสถิติประยุกต์ คณะวิทยาศาสตร์. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. กรุงเทพฯ; 2559.
4. นฤดล พิทักษ์วิทยกูล. การเปรียบเทียบตัวแบบการพยากรณ์เบี้ยประกันชีวิตรายใหม่ ประเภทสามัญด้วยวิธีปรับให้เรียบเอ็กซ์โปเนนเชียลแบบโฮลท์– วินเทอร์วิธีบ็อกซ์-เจนกินส์และวิธีโครงข่ายประสาทเทียม. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติและการวิเคราะห์ธุรกิจ ภาควิชาสถิติ คณะวิทยาศาสตร์. สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. กรุงเทพฯ; 2563.
5. นิฉา แก้วหาวงษ์. ตัวแบบพยากรณ์เบี้ยประกันภัยรับโดยตรงของการประกันชีวิตแบบบำนาญ ในประเทศไทย ปี พ.ศ. 2564-2566. ใน: เอกสารประกอบการประชุมวิชาการระดับชาติ ครั้งที่ 17 และการประชุมวิชาการระดับนานาชาติ ครั้งที่ 7 วันที่ 27 ตุลาคม 2565. มหาวิทยาลัยศรีปทุม. กรุงเทพฯ; 2565. หน้า 2977-86.
6. ศุภามณ จันทร์สกุล. การพยากรณ์อนุกรมเวลาแบบเชิงเส้น แบบไม่ใช่เชิงเส้น และโมเดลผสมผสาน. วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเชีย ฉบับวิทยาศาสตร์และเทคโนโลยี 2558;9(2):50-63.
7. สมาคมประกันชีวิตไทย. รายงานธุรกิจประกันชีวิต ตั้งแต่เดือนมกราคม พ.ศ.2560 ถึง เดือนธันวาคม พ.ศ. 2565. [อินเทอร์เน็ต]. 2565 [เข้าถึงเมื่อ 26 ก.พ. 2566]. เข้าถึงได้จาก: https://www.tlaa.org/page_statistics.php?cid=35
8. อโณทัย ศิลเทพาเวทย์. แบบจำลองเพื่อพัฒนาคุณภาพของผลิตภัณฑ์เอชจีเอในโรงงานอุตสาหกรรมฮาร์ดดิสก์ ด้วยเทคนิคต้นไม้ตัดสินใจ. วิทยานิพนธ์ปริญญาวิทยา ศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์. จุฬาลงกรณ์มหาวิทยาลัย. กรุงเทพฯ; 2554.
9. ปริญญา สงวนสัตย์. Artificial Intelligence with Machine Learning, AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง. พิมพ์ครั้งที่ 1. นนทบุรี: ไอดีซี พรีเมียร์; 2562.
10. Vapnik, V. The Nature of Statistical Learning Theory. New York: Springer-Verlag New York, Inc; 1998.
11. Meyer, D. et al. Misc Functions of the Department of Statistics, Probability Theory Group [Internet]. 2015 [cited 2022 December 5]. Available from: http://cran.r-project.org/web/packages/e1071/index.html.
12. Aczel, A. Complete Business Statistics. 4th ed. University of California: Irwin; 1989.
13. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199-222.
14. Christophe D, Quentin G. An explicit split point procedure in model-based trees allowing for a quick fitting of GLM trees and GLM forests. Stat Comput 2021;32(1):1-27.
15. Gerard B. Analysis of a Random Forests Model. J Mach Learn Res 2012;13:1063-95.
16. Alexandros K, David M, Kurt H. Support Vector Machines in R. J Stat Softw 2006;15(9):1-28.
17. Louise S, Guillaume T, Shaun H, Olivier D, Alexander H, Abdou K, et al. Using R in hydrology: a review of recent developments and future directions. Hydrol Earth Syst Sci 2019;23(7):2939–63.