

การตรวจสอบความสมเหตุสมผลของตัวแบบต้นไม้ตัดสินใจ :
กรณีศึกษาปัจจัยที่มีผลต่อการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวัน
Validity Checks on Decision Trees Models :
A Case Study on Factors Affecting the Renewal of a Hospital Benefit Rider

กมล บุษบา* และกชกร ทุงสระโร

Kamon Budsaba* and Kotchakorn Thungsaro

สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

Department of Mathematics and Statistics, Faculty of Science and Technology,

Thammasat University, Klong-Neung, Klong-Luang, Phatumthani 12120

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบการตรวจสอบความสมเหตุสมผลของตัวแบบต้นไม้ตัดสินใจระหว่างวิธีการตรวจสอบความสมเหตุสมผลแบบไขว้และวิธีการตรวจสอบความสมเหตุสมผลโดยการแบ่งตัวอย่าง ใช้ข้อมูลสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันของผู้เอาประกันภัยของบริษัทประกันชีวิตแห่งหนึ่งจำนวน 25,249 สัญญา วิเคราะห์ข้อมูลด้วยโปรแกรม SPSS รุ่น 20 ตัวแปรตามคือ สถานะของสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวัน และตัวแปรทำนายคือ เพศ อายุ เงินผลประโยชน์ เบี้ยประกันภัยสำหรับสัญญาเพิ่มเติมรายปี ลักษณะการจ่ายเบี้ยประกันภัย ภูมิภาค สถานภาพสมรส ประเภทกรรมสิทธิ์ของสัญญาหลัก เบี้ยประกันภัยรายปีของสัญญาหลัก จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทน จำนวนเงินค่าสินไหมทดแทนรวม และระยะเวลาดำเนินการเฉลี่ยในการจ่ายค่าสินไหมทดแทน จากการทำซ้ำ 500 ครั้งพบว่าวิธีตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างด้วยวิธี QUEST 80:20 ให้ความแม่นยำในการทำนายการต่ออายุสัญญาสูงสุดเท่ากับ 62.374% และลักษณะการจ่ายเบี้ยประกันภัยมีผลต่อการต่ออายุสัญญามากที่สุด เมื่อใช้วิธี QUEST 80:20 โดยจำแนกตามลักษณะการจ่ายเบี้ยประกันภัยพบว่า สำหรับสัญญาที่จ่ายเบี้ยประกันภัยแบบรายปี ค่าความแม่นยำในการทำนายการต่ออายุสัญญาของสัญญาเท่ากับ 80.256% และจำนวนครั้งของการเคลมมีผลต่อการต่ออายุสัญญามากที่สุด

คำสำคัญ: การตรวจสอบความสมเหตุสมผลไขว้ การตรวจสอบความสมเหตุสมผลโดยแบ่งตัวอย่าง วิธีการลิดกิ่ง

* Corresponding author : kamon@mathstat.sci.tu.ac.th

Abstract

The objective of this paper is to compare validity checks on decision trees models between cross validation and split-sample validation on a case study of factors affecting the renewal of a hospital benefit rider of life insurance company in Thailand. 25,249 policies were analyzed by SPSS version 20. The dependent variable is a status of each hospital benefit rider which sex, age, face amount, annual life premium of rider, premium mode, domicile, marriage status, base policy type, annual life premium of base policy, total of claims, amount of claims and average duration of paid claims operation are independent variables. From 500 iterations, we found that the split-sample validation of QUEST 80:20 give 62.374% of the accuracy and the premium mode is the most influential factor to the renewal of a policy. QUEST 80:20 was used to predict the renewal of a policy classified by premium mode. The results found that the accuracy of a policy renewal prediction is 80.256% for annual life premium payment and total of claims is the most influential factor to the renewal of a policy.

Keywords: cross validation, split-sample validation, pruning

1. บทนำ

ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ที่นิยมมากที่สุดวิธีหนึ่งในการจำแนกประเภทข้อมูลแต่ละรายเป็นกลุ่มๆ หรือค่าทำนายของตัวแปรตามตัวหนึ่งโดยอาศัยค่าของตัวแปรทำนายหลาย ๆ ตัว การจำแนกประเภทข้อมูลอาจเป็น การแบ่งข้อมูลออกเป็นส่วน การแบ่งกลุ่มการพยากรณ์ การลดรูปข้อมูล การคัดกรองตัวแปร หรือการพิสูจน์ความสัมพันธ์

การตรวจสอบความสมเหตุสมผล (validation) ของต้นไม้ตัดสินใจเป็นวิธีการที่ใช้พิจารณาว่าโครงสร้างของต้นไม้ (tree structure) ที่ผู้วิจัยนำไปใช้โดยทั่วไปกับประชากรนั้นใช้ได้ดีเพียงใด การตรวจสอบความสมเหตุสมผลของต้นไม้ตัดสินใจในโปรแกรมสำเร็จรูป SPSS โดยทั่วไป ทำได้สองวิธี ได้แก่ (1) การตรวจสอบความสมเหตุสมผลไขว้ (cross validation) และ (2) การตรวจสอบความสมเหตุสมผลโดยแบ่งตัวอย่าง (split-sample validation)

การตรวจสอบความสมเหตุสมผลไขว้ทำโดยการแบ่งตัวอย่างออกเป็นตัวอย่างย่อยหลาย ๆ ชุด หรือเรียกว่าส่วนทบ (fold) จากนั้นทำการสร้างตัวแบบต้นไม้โดยไม่ใช้ข้อมูลจากตัวอย่างย่อยชุดหนึ่งเป็นลำดับไป ต้นไม้แรกสร้างขึ้นโดยอาศัยข้อมูลทุกตัวยกเว้นข้อมูลที่อยู่ในส่วนทบแรก ต้นไม้ต้นที่สองสร้างขึ้นโดยอาศัยข้อมูลทุกตัวยกเว้นข้อมูลที่อยู่ในส่วนทบที่สอง ดังนี้เรื่อย ๆ จากนั้นจึงทำการประมาณความเสี่ยงของการจำแนกผิดกลุ่ม (misclassification risk) ของต้นไม้แต่ละต้นโดยใช้ตัวแบบจากต้นไม้นั้นกับตัวอย่างย่อยที่ไม่ได้ใช้ในการสร้างต้นไม้นั้นๆ อนึ่งการตรวจสอบความสมเหตุสมผลไขว้ใช้กับวิธีการลิดกิ่ง (pruning) หรือวิธีการทำให้ต้นไม้

เจริญเติบโต (growing method) โดยวิธี CHAID (Chi-squared Automatic Interaction Detection) เท่านั้น (IBM, 2017)

การตรวจสอบความสมเหตุสมผลโดยแบ่งตัวอย่าง ทำโดยแบ่งข้อมูลเป็นสองชุด เรียกว่า ตัวอย่างฝึกฝน (training sample) และตัวอย่างทดสอบ (testing sample) จากนั้นทำการสร้างตัวแบบจากข้อมูลในชุดตัวอย่างฝึกฝน และทดสอบด้วยข้อมูลในชุดตัวอย่างทดสอบ (เอกสิทธิ์, 2558) การตรวจสอบความสมเหตุสมผลโดยแบ่งตัวอย่างใช้กับวิธีการลิดกิ่งหรือวิธีการทำให้ต้นไม้เจริญเติบโต โดยวิธี CHAID (Chi-squared Automatic Interaction Detection) CRT (Classification and Regression Tree) หรือ QUEST (Quick, Unbiased, Efficient Statistical Tree)

ในปัจจุบันการแข่งขันในตลาดที่สูงขึ้นส่งผลให้บริษัทประกันภัยจึงต้องมีการปรับตัว ปรับกลยุทธ์ในช่วงชิงพื้นที่ของตลาดมากขึ้น เพื่อการอยู่รอดในสภาวะการแข่งขันที่สูง สิ่งที่ได้อันดับแรกคือการรักษาลูกค้าโดยปกติเมื่อผู้เอาประกันภัยได้ทำสัญญากรมธรรม์ประกันชีวิตไม่ว่าจะเป็นแผนใดก็ตาม บริษัทจะสามารถเสนอสัญญาเพิ่มเติมแนบท้ายสัญญาประกันชีวิตหลักโดยสัญญาเพิ่มเติมเหล่านั้นจะเป็นสัญญาต่อปีผู้เอาประกันภัยจำเป็นต้องต่ออายุสัญญาเมื่อครบกำหนดเวลา และจะได้รับใบเตือนต่ออายุก่อนวันที่สัญญาจะหมดอายุ ดังนั้นหากบริษัททราบถึงปัจจัยที่ส่งผลให้ผู้เอาประกันภัยต่ออายุสัญญาถือเป็นสิ่งสำคัญที่สามารถทำให้บริษัทประกันชีวิตกำหนดกลยุทธ์ด้านการบริหารลูกค้าสัมพันธ์ บริการบริการลูกค้า เพื่อให้ตรงตามความต้องการของผู้เอาประกันภัยเพื่อที่ผู้เอาประกันภัยจะได้ต่ออายุสัญญากรมธรรม์กับบริษัทในอนาคต และทราบถึงพฤติกรรมความต้องการของผู้เอาประกันภัยที่แตกต่างกันในแต่ละกลุ่มได้อย่างมีประสิทธิภาพ

จากที่กล่าวมาข้างต้น ผู้วิจัยสนใจศึกษาเพื่อเปรียบเทียบความแม่นยำ (accuracy) ของวิธีการตรวจสอบความสมเหตุสมผลของต้นไม้ตัดสินใจด้วยวิธีการ CHAID, CRT และ QUEST โดยเฉพาะเมื่อนำมาประยุกต์ใช้กับข้อมูลจากบริษัทประกันชีวิตแห่งหนึ่งเพื่อศึกษาปัจจัยที่มีผลต่อการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวัน

2. วิธีดำเนินการวิจัย

ขอบเขตการวิจัย

1. ข้อมูลที่ใช้วิเคราะห์เป็นข้อมูลสถานะสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันของผู้เอาประกันภัยของบริษัทประกันชีวิตแห่งหนึ่งในประเทศไทยที่มีอายุตั้งแต่ 11 ปีขึ้นไป จำนวน 25,249 สัญญาและเป็นสัญญาที่เริ่มความคุ้มครองในปี พ.ศ. 2559 โดยไม่รวมกรมธรรม์ที่ทำสัญญาชดเชยค่ารักษาพยาบาลรายวันมากกว่า 1 สัญญา

2. ศึกษาวิธีการตรวจสอบความสมเหตุสมผลของตัวแบบต้นไม้ตัดสินใจด้วยโปรแกรม SPSS รุ่น 20 โดยเปรียบเทียบการตรวจสอบความสมเหตุสมผล 2 วิธี ได้แก่

(1) การตรวจสอบความสมเหตุสมผลไขว้ ใช้วิธีการลิดกิ่งวิธี CHAID และใช้ส่วนทศ (k) เท่ากับ 2 3

(2) การตรวจสอบความสมเหตุสมผลโดยแบ่งตัวอย่าง ใช้ใช้วิธีการลิตกิงวิธี CHAID CRT หรือ QUEST และใช้ตัวอย่างฝึกฝน ต่อ ตัวอย่างทดสอบ เป็น 50:50 60:40 70:30 และ 80:20

ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้ในการแบ่งกลุ่มปริภูมิตัวทำนาย (predictor space) ออกเป็นบริเวณหลายบริเวณเพื่อใช้ทำนายค่าของตัวแปรตามโดยใช้ค่าเฉลี่ย (กรณีที่ตัวแปรตามเป็นข้อมูลเชิงปริมาณ) หรือฐานนิยม (กรณีที่ตัวแปรตามเป็นข้อมูลเชิงคุณภาพ) ของค่าสังเกตฝึกฝนที่อยู่ในบริเวณนั้นเพื่อทำนายค่าตัวแปรตามในบริเวณเดียวกัน การแบ่งปริภูมิตัวทำนายสรุปในแผนภาพของต้นไม้ที่กลับหัว (James et al., 2017) สามารถใช้ได้กับปัญหาการถดถอยที่มีตัวแปรตามเป็นข้อมูลเชิงปริมาณซึ่งเรียกว่าต้นไม้การถดถอย (regression trees) และปัญหาการจำแนกที่มีตัวแปรตามเป็นข้อมูลจำแนกประเภท (categorical data) ซึ่งเรียกว่าต้นไม้การจำแนก (classification trees)

ต้นไม้ตัดสินใจง่ายในการอธิบายแก่บุคคลอื่นโดยเฉพาะถ้าเปรียบเทียบกับสมการถดถอยเนื่องจากเป็น การแสดงในรูปของกราฟ สามารถใช้กับตัวแปรทำนายเชิงคุณภาพได้ง่ายโดยไม่จำเป็นต้องสร้างตัวแปรหุ่น (dummy variable) อย่างไรก็ตามต้นไม้ตัดสินใจไม่มีความแกร่ง (robust) กล่าวคือถ้าข้อมูลมีการเปลี่ยนแปลงเล็กน้อยจะมีผลทำให้ค่าประมาณที่ได้ในขั้นสุดท้ายเปลี่ยนแปลงไปอย่างมาก

ขั้นตอนการสร้างต้นไม้ถดถอย

1. แบ่งปริภูมิตัวทำนายหรือชุดของค่าที่เป็นไปได้ของ $X_1, X_2, \dots, X_p$ เป็น j บริเวณที่ไม่ซ้อนทับกัน ได้แก่ $R_1, R_2, \dots, R_j$

2. สำหรับค่าสังเกตทุกค่าที่ตกอยู่ในบริเวณ R_j ให้ทำนายโดยใช้ค่าทำนายค่าเดียวกันซึ่งเป็นค่าเฉลี่ยของตัวแปรตามของค่าสังเกตฝึกฝนที่อยู่ใน R_j

การสร้าง $R_1, R_2, \dots, R_j$ หรือการแบ่งปริภูมิตัวทำนายเป็นกล่อง ๆ มีเป้าหมายคือการหากล่องที่ทำให้ผลบวกกำลังสองส่วนเหลือ (residual sum of squares: RSS) มีค่าต่ำสุด นิยามโดย

$$RSS = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

เมื่อ $\hat{y}_{R_j}$ คือค่าเฉลี่ยของตัวแปรตามสำหรับค่าสังเกตฝึกฝนที่อยู่ในบริเวณหรือกล่อง R_j

ในทางปฏิบัติวิธีนี้แบ่งที่เรียกว่า “การแบ่งทวิเวียนเกิด (recursive binary splitting)” โดยเริ่มจากด้านบนของต้นไม้ที่ค่าสังเกตทุกค่าอยู่ในบริเวณเดียวกัน จากนั้นทำการแบ่งปริภูมิตัวทำนายที่ละสองส่วน จุดที่แบ่งนี้เรียกว่า “โหนดภายใน (internal node)” ซึ่งทำให้ได้ “กิ่ง (branch)” ใหม่สองกิ่งด้านล่างของต้นไม้ โดยการแบ่งแต่ละครั้งเป็นการแบ่งนั้นดีที่สุดเฉพาะขั้นตอนนั้น ๆ ไม่ได้พิจารณาพร้อมกับขั้นตอนข้างหน้าขั้นตอนอื่น

การลิดกิ่งต้นไม้

กระบวนการสร้างต้นไม้ตัดสินใจที่กล่าวมาข้างต้นอาจใช้ในการทำนายข้อมูลชุดฝึกฝนได้ดี แต่มีแนวโน้มที่จะใช้ไม่ดีกับข้อมูลชุดทดสอบ เนื่องจากต้นไม้ที่ได้ซับซ้อนเกินไป ดังนั้นต้นไม้ที่เล็กกว่าหรือมีการแบ่งน้อยกว่าจะทำให้ความแปรปรวนมีค่าน้อยลงและง่ายในการตีความหมาย ดังนั้น วิธีการที่นิยมใช้คือ การสร้างต้นไม้ขนาดใหญ่มาต้นหนึ่ง จากนั้นทำการลิดกิ่งลงเพื่อให้ได้ต้นไม้ย่อย (subtree) ที่ดีที่สุดที่ให้อัตราความผิดพลาดต่ำสุดสำหรับข้อมูลชุดทดสอบ เมื่อกำหนดต้นไม้ย่อยต้นหนึ่ง สามารถประมาณค่าความผิดพลาดนั้นได้ โดยการทำการทดสอบความสมเหตุสมผลไขว้

ต้นไม้การจำแนก

หลักเกณฑ์ในการสร้างต้นไม้การจำแนกนั้นคล้ายกับต้นไม้ถดถอย เพียงแต่ใช้ฐานนิยามแทนค่าเฉลี่ยในการทำนายค่าตัวแปรตามที่เป็นข้อมูลเชิงคุณภาพ และใช้อัตราความผิดพลาดของการจำแนก (classification error rate: CER) แทน RSS ซึ่งเป็นสัดส่วนของ ค่าสังเกตชุดฝึกฝนในบริเวณนั้น ที่ไม่ใช่กลุ่มที่เกิดมากที่สุด หรือฐานนิยาม กล่าวคือ

$$CER = 1 - \max_k (\hat{p}_{mk})$$

โดยที่ $\hat{p}_{mk}$ คือสัดส่วนของค่าสังเกตชุดฝึกฝนในบริเวณที่ m ที่มาจากกลุ่มที่ k อย่างไรก็ตาม CER ไม่ไวเพียงพอในการสร้างต้นไม้ จึงนิยมใช้ตัววัดชนิดอื่นเช่น ดัชนีจินี (gini index: G) นิยามโดย

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

ซึ่งเป็นตัววัดความแปรปรวนรวมทั้ง K กลุ่ม ดัชนีนี้มีค่าน้อยถ้า $\hat{p}_{mk}$ ทุกค่าเข้าใกล้ 0 หรือ 1 ดังนั้นจึงใช้สำหรับวัดความบริสุทธิ์ของโหนด (node purity) ค่าที่น้อยทำให้เห็นว่าโหนดนั้น ๆ ประกอบด้วยค่าสังเกตที่มาจากกลุ่มเดียวกัน นอกจากใช้ดัชนีจินีแล้ว อาจใช้เครื่องวัดที่เรียกว่าเอนโทรปี (entropy: D) นิยามโดย

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

เนื่องจาก $0 \leq \hat{p}_{mk} \leq 1$ ทำให้ $-\hat{p}_{mk} \log(\hat{p}_{mk}) \geq 0$ ซึ่งถ้า $\hat{p}_{mk}$ ทุกค่าเข้าใกล้ 0 หรือ 1 เอนโทรปีจะมีค่าเข้าใกล้ศูนย์ หรือเอนโทรปีมีค่าน้อยถ้าโหนดที่ m มีความบริสุทธิ์

เนื่องจาก G และ D ให้ค่าค่อนข้างใกล้เคียงกันในการสร้างต้นไม้การจำแนกดังนั้นอาจใช้ตัววัดทั้งสองในการวัดคุณภาพของการแบ่งบริเวณ อย่างไรก็ตามในการลิดกิ่งต้นไม้สามารถใช้แนวทางใดก็ได้โดย CER มักนิยมใช้ถ้าจุดมุ่งหมายของการลิดกิ่งเพื่อต้องการดูความแม่นยำของการทำนายของต้นไม้ขั้นสุดท้ายที่ได้ลิดกิ่งแล้ว (James et al., 2017)

ขั้นตอนวิธี CRT

การทำให้ต้นไม้เจริญเติบโตโดยให้โหนดลูกที่เกิดใหม่มีความบริสุทธิ์มากที่สุดวิธีนี้ การแบ่งแต่ละครั้งขึ้นอยู่กับค่าของตัวแปรตัวทำนายเพียงตัวเดียว ถ้า X เป็นตัวแปรนามบัญญัติที่มี K กลุ่ม จะมีการแบ่งที่เป็นไปได้สำหรับตัวทำนายนี้ $2^{K-1} - 1$ วิธี ถ้า X เป็นตัวแปรจัดอันดับหรือต่อเนื่องที่มีค่าแตกต่างกัน K ค่า จะมีการแบ่งที่เป็นไปได้ทั้งหมด $K - 1$ วิธี ต้นไม้ต้นหนึ่งเริ่มจากโหนดที่เป็นรากโดยใช้ขั้นตอนต่อไปนี้เป็นแต่ละโหนดซ้ำ ๆ กันเพื่อสร้างต้นไม้

(1) ค้นหาตัวแปรทำนายที่แบ่งได้ดีที่สุด สำหรับตัวแปรทำนายนามบัญญัติให้พิจารณาเซตย่อยของกลุ่มที่เป็นไปได้ทั้งหมด ถ้า x เป็นสมาชิกของเซต A จะนำค่าสังเกตไปไว้ที่โหนดลูกด้านซ้าย แต่ถ้าหาก x ไม่เป็นสมาชิกของเซต A จะนำค่าสังเกตไปไว้ที่โหนดลูกด้านขวา สำหรับตัวแปรทำนายจัดอันดับหรือต่อเนื่องให้เรียงลำดับจากค่าน้อยสุดไปค่ามากที่สุด แล้วให้เริ่มใช้ค่าน้อยสุดเป็นจุดแบ่ง v ถ้า $x \leq v$ จะนำค่าสังเกตไปไว้ที่โหนดลูกด้านซ้าย แต่ถ้า $x > v$ จะนำค่าสังเกตไปไว้ที่โหนดลูกด้านขวา ทำในลักษณะเช่นเดียวกันนี้ โดยใช้ค่าน้อยสุดลำดับถัดไปเช่นนี้เรื่อย ๆ จุดแบ่งที่ดีที่สุดคือจุดที่ทำให้เกณฑ์การแบ่งมีค่าสูงสุดเมื่อโหนดนั้นถูกแบ่งด้วยค่านี

(2) ค้นหาการแบ่งที่ดีที่สุดของโหนด ในบรรดาตัวแบ่งที่ดีที่สุดที่พบในขั้นที่ (1) ให้เลือกการแบ่งที่ทำให้เกณฑ์การแบ่งมีค่ามากที่สุด

(3) แบ่งโหนดโดยใช้การแบ่งที่ดีที่สุดที่พบในขั้นที่ (2) ถ้ายังไม่ถึงเกณฑ์การหยุดแบ่ง (stopping rules)

ขั้นตอนวิธีการ CHAID

ขั้นตอนการรวม (merging step) สำหรับตัวแปรทำนาย X แต่ละตัว ให้รวมกลุ่มที่ไม่มีนัยสำคัญ กลุ่มสุดท้ายแต่ละกลุ่มของ X จะเป็นโหนดลูกของ X ถ้า X ถูกใช้ในการแบ่งโหนด ขั้นตอนการรวมจะมีการคำนวณค่า p-value ที่ปรับแล้วที่นำมาใช้ในขั้นตอนการแบ่ง

(1) ถ้า X มีเพียงกลุ่มเดียวให้หยุดและกำหนดให้ p-value ที่ปรับแล้วมีค่าเท่ากับ 1

(2) ถ้า X มี 2 กลุ่ม ให้ไปขั้นที่ (8)

(3) กรณีอื่น ๆ ให้หาคู่ที่ยอมรับได้ของกลุ่มของ X (คู่ที่ยอมรับได้ของกลุ่มสำหรับตัวแปรทำนายนามบัญญัติ คือกลุ่มสองกลุ่มใด ๆ สำหรับตัวแปรทำนายแบบจัดอันดับหรือต่อเนื่องได้แก่ กลุ่มที่อยู่ติดกัน 2 กลุ่ม) หรือคู่ที่คล้ายคลึงกันมากที่สุดซึ่งคือคู่ที่สถิติทดสอบให้ค่า p-value มากที่สุด เมื่อเทียบกับตัวแปรตาม Y

(4) สำหรับคู่ที่มี p-value มากที่สุด ให้ตรวจสอบว่า p-value มีค่ามากกว่าระดับนัยสำคัญที่ผู้กำหนด (α_{merge}) ถ้ามีค่ามากกว่าให้รวมคู่นี้เข้าด้วยกันทำให้ได้กลุ่มใหม่ของ X ถ้ามีค่าไม่มากกว่าให้ไปขั้นที่ (7)

(5) ถ้ากลุ่มที่จัดใหม่มีมากกว่าหรือเท่ากับ 3 กลุ่มย่อย ให้ทำการแบ่งเชิงทวิของกลุ่มนั้นซึ่งให้ค่า p -value น้อยที่สุด ให้ทำการแบ่งเชิงทวิแบบเดียวกันนี้ ถ้าค่า p -value ไม่มากกว่าระดับนัยสำคัญที่กำหนด ($\alpha_{split-merge}$)

(6) ไปขั้นที่ (2)

(7) ถ้ากลุ่มใดมีค่าสังเกตน้อยเกินไป (เมื่อเปรียบเทียบกับค่าที่ผู้ใช้ระบุ) ให้รวมกลุ่มกับกลุ่มอื่นที่คล้ายคลึงกันมากที่สุด วัดโดยใช้ค่ามากที่สุดของ p -value

(8) คำนวณค่า p -value ที่ปรับแล้วสำหรับกลุ่มที่รวมกัน โดยใช้การปรับแบบบอนแฟร์โรนี (Bonferroni)

ขั้นตอนการแบ่ง (splitting step) การแบ่งที่ดีที่สุดสำหรับตัวแปรทำนายแต่ละตัว ได้จากขั้นตอนการรวมที่กล่าวข้างต้น ขั้นตอนการแบ่งเป็นการเลือกตัวแปรทำนายที่จะใช้แบ่งโหนดแบบดีที่สุด การเลือกทำได้โดยการเปรียบเทียบค่า p -value ที่ปรับแล้วของตัวแปรทำนายแต่ละตัว ค่า p -value ที่ปรับแล้วคำนวณได้จากขั้นตอนการรวม

(1) เลือกตัวแปรทำนายที่มีค่า p -value ที่ปรับแล้วที่มีค่าน้อยที่สุด (มีนัยสำคัญมากที่สุด)

(2) ถ้าค่า p -value ที่ปรับแล้วนี้มีค่าน้อยกว่าหรือเท่ากับค่าระดับนัยสำคัญที่ผู้ใช้กำหนด (α_{split}) ให้แบ่งโหนดโดยใช้ตัวแปรทำนายตัวนี้ ถ้ามีค่ามากกว่าไม่ต้องทำการแบ่งและใช้โหนดนี้เป็นโหนดปลายสุด (terminal node)

ขั้นตอนวิธีการ QUEST

(1) สำหรับตัวแปรทำนายจำแนกประเภท ให้ทำการทดสอบสัดส่วนโดยใช้สถิติไคกำลังสองของเพียร์สัน (pearson's chi-square) และคำนวณค่า p -value ของสถิติไคกำลังสอง สำหรับตัวแปรทำนายต่อเนื่องให้ทำการวิเคราะห์ความแปรปรวนด้วยสถิติเอฟเพื่อใช้ทดสอบว่า แต่ละกลุ่มของตัวแปรตาม Y มีค่าเฉลี่ยของตัวแปร X เท่ากันหรือไม่ และคำนวณค่า p -value ของสถิติทดสอบเอฟ

(2) หาตัวแปรทำนายที่มีค่า p -value น้อยที่สุดกำหนดให้เป็น X^*

(3) ถ้าค่า p -value ที่น้อยที่สุดนี้มีค่าน้อยกว่า α/M เมื่อ α คือระดับนัยสำคัญที่ผู้ใช้กำหนดและ M คือ จำนวนตัวแปรทำนาย X^* จะใช้เป็นตัวแปรทำนายสำหรับโหนดนั้น แต่หากไม่เป็นเช่นนั้นไปขั้นที่ (4)

(4) สำหรับตัวแปรทำนายต่อเนื่องแต่ละตัวให้คำนวณสถิติทดสอบเอฟของเลวิน (Levene's F test) โดยใช้ค่าสัมบูรณ์ของความเบี่ยงเบนระหว่าง X กับค่าเฉลี่ยของกลุ่ม เพื่อทดสอบว่าความแปรปรวนของ X แตกต่างกันในแต่ละกลุ่มของ Y หรือไม่ และคำนวณค่า p -value ของการทดสอบ

(5) หาตัวแปรทำนายที่มีค่า p -value น้อยที่สุดและแทนด้วย X^{**}

(6) ถ้าค่า p -value ที่น้อยที่สุดนี้มีค่าน้อยกว่า $\alpha / (M + M1)$ เมื่อ $M1$ คือ จำนวนตัวแปรทำนายชนิดต่อเนื่องจะให้เลือก X^* ทำนายที่ใช้ในการแบ่งโหนดนั้น ถ้าไม่เป็นตามนี้ ไม่ต้องแบ่งโหนดนั้น (IBM, 2017)

การวัดประสิทธิภาพ

การวัดประสิทธิภาพของตัวแบบทำโดยคำนวณค่าความแม่นยำที่ได้จากเมทริกสับสน (confusion matrix) ซึ่งเป็นตารางสรุปจำนวนค่าสังเกตที่ตัวแบบจำแนกได้อย่างถูกต้องหรือไม่ถูกต้อง ดังนี้

		ค่าทำนาย	
		ต่อสัญญา	ไม่ต่อสัญญา
ค่าสังเกต	ต่อสัญญา	a	b
	ไม่ต่อสัญญา	c	d

ค่าความแม่นยำคำนวณโดย $\frac{a + d}{a + b + c + d}$

วิธีดำเนินการวิจัย

ใช้โปรแกรมสำเร็จรูป SPSS รุ่น 20 และใช้วิธีการลิดกึ่งต้นไม้แบบ CHAID, CRT และ QUEST โดยการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่าง โดยใช้ชุดตัวอย่างฝึกฝนต่อชุดตัวอย่างทดสอบคือ 50:50 60:40 70:30 และ 80:20 และวิธีการลิดกึ่งต้นไม้แบบ CHAID และการตรวจสอบความสมเหตุสมผลแบบไขว้โดยใช้ส่วนทศ (k) เท่ากับ 2 3 4 และ 5 เพื่อหาผลลัพธ์ที่มีความแม่นยำมากที่สุด โดยประมวลผลกับข้อมูลสถานะสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันของบริษัทประกันชีวิตแห่งหนึ่งในประเทศไทยที่มีอายุตั้งแต่ 11 ปีขึ้นไป จำนวน 25,249 สัญญา วิธีการวิเคราะห์ข้อมูลที่เป็นไปได้ทั้งหมด 16 วิธี ดังนี้

1. วิธีลิดกึ่งแบบ CHAID และแบ่งตัวอย่างเป็น 50:50
2. วิธีลิดกึ่งแบบ CHAID และแบ่งตัวอย่างเป็น 60:40
3. วิธีลิดกึ่งแบบ CHAID และแบ่งตัวอย่างเป็น 70:30
4. วิธีลิดกึ่งแบบ CHAID และแบ่งตัวอย่างเป็น 80:20
5. วิธีลิดกึ่งแบบ CRT และแบ่งตัวอย่างเป็น 50:50
6. วิธีลิดกึ่งแบบ CRT และแบ่งตัวอย่างเป็น 60:40
7. วิธีลิดกึ่งแบบ CRT และแบ่งตัวอย่างเป็น 70:30
8. วิธีลิดกึ่งแบบ CRT และแบ่งตัวอย่างเป็น 80:20

9. วิธีลิดกึ่งแบบ QUEST และแบ่งตัวอย่างเป็น 50:50
10. วิธีลิดกึ่งแบบ QUEST และแบ่งตัวอย่างเป็น 60:40
11. วิธีลิดกึ่งแบบ QUEST และแบ่งตัวอย่างเป็น 70:30
12. วิธีลิดกึ่งแบบ QUEST และแบ่งตัวอย่างเป็น 80:20
13. วิธีลิดกึ่งแบบ CHAID และใช้ส่วนทศ k เท่ากับ 2
14. วิธีลิดกึ่งแบบ CHAID และใช้ส่วนทศ k เท่ากับ 3
15. วิธีลิดกึ่งแบบ CHAID และใช้ส่วนทศ k เท่ากับ 4
16. วิธีลิดกึ่งแบบ CHAID และใช้ส่วนทศ k เท่ากับ 5

แต่ละวิธีทำซ้ำจำนวน 500 รอบ

ค่าที่เป็นไปได้ของตัวแปรทำนาย เป็นดังนี้

1. เพศ 0 (หญิง) 1 (ชาย)
2. อายุ 0 (0-23 ปี) 1 (24 ปีขึ้นไป)
3. เงินผลประโยชน์ (บาท)
4. เบี้ยประกันภัยสำหรับสัญญาเพิ่มเติมรายปี (บาท)
5. ลักษณะการจ่ายเบี้ยประกันภัย 1 (รายปี) 2 (รายครึ่งปี) 3 (รายไตรมาส) 4 (รายเดือน)
6. ภูมิลำเนา 0 (ต่างจังหวัด) 1 (กรุงเทพฯ และปริมณฑล)
7. สถานภาพสมรส 0 (โสด) 1 (แต่งงานและอื่น ๆ)
8. ประเภทกรรมสิทธิ์ของสัญญาหลัก 0 (ตลอดชีพ) 1 (อื่น ๆ)
9. เบี้ยประกันภัยรายปีของสัญญาหลัก (บาท)
10. จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทน 1 (ไม่มีการเรียกเคลม) 2 (เรียกเคลม 1 ครั้ง)
3 (เรียกเคลมมากกว่า 1 ครั้ง)
11. จำนวนเงินค่าสินไหมทดแทนรวม 0 (ไม่เคยได้รับเงินเคลม) 1 (เคยได้รับเงินเคลม)
12. ระยะเวลาดำเนินการเฉลี่ยในการจ่ายค่าสินไหมทดแทน (วัน)

จากนั้นจะเลือกวิธีการตรวจสอบความสมเหตุสมผลและวิธีการลิดกึ่งที่มีความแม่นยำมากที่สุดเพื่อหาความแม่นยำและหาปัจจัยที่มีผลต่อการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันจำแนกตามลักษณะการจ่ายเบี้ยประกันภัย

3. ผลการวิจัย

สถิติพรรณนาของตัวแปรทำนาย

ตารางที่ 1 สัดส่วนการต่ออายุสัญญาเพิ่มเติมชุดเซย์คำรักษาพยาบาลรายวันจำแนกตามตัวแปรทำนายที่เป็นลักษณะทั่วไป

ตัวแปรทำนาย		สถานะสัญญา (หน่วย:จำนวนสัญญา)		สัดส่วนการต่ออายุสัญญา
		ไม่ต่อสัญญา	ต่อสัญญา	
เพศ	หญิง	5,790 (39.55%)	8,849 (60.45%)	1.5283
	ชาย	4,282 (40.36%)	6,328 (59.64%)	1.4778
อายุ	0-23 ปี	1,828 (39.3%)	2,820 (60.7%)	1.5445
	24 ปีขึ้นไป	8,244 (40.0%)	12,357 (60.0%)	1.5000
สถานภาพสมรส	โสด	4,264 (37.94%)	6,975 (62.06%)	1.6358
	แต่งงานและอื่นๆ	5,675 (41.42%)	8,026 (58.58%)	1.4143
ภูมิลำเนา	กรุงเทพ-ปริมณฑล	1,596 (33.61%)	3,152 (66.39%)	1.9749
	อื่นๆ	8,476 (41.34%)	12,025 (58.66%)	1.4187

จากตารางที่ 1 จะเห็นว่าภูมิลำเนาน่าจะมีผลต่อการต่ออายุสัญญา โดยผู้เอาประกันอยู่ในกรุงเทพและปริมณฑลมีสัดส่วนการต่ออายุสัญญามากที่สุดคือ 1.975

ตารางที่ 2 สัดส่วนการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันจำแนกตามตัวแปร
ทำนายที่เป็นลักษณะเฉพาะทางประกันภัย

ปัจจัย		สถานะสัญญา (หน่วย:จำนวนสัญญา)		สัดส่วนการต่ออายุสัญญา
		ไม่ต่อสัญญา	ต่อสัญญา	
ลักษณะการจ่ายเบี้ยประกันภัย	รายปี	1,464 (21.07%)	5,485 (78.93%)	3.7466 (1)
	รายครึ่งปี	697 (37.09%)	1,182 (62.91%)	1.6958
	รายไตรมาส	1,437 (50.80%)	1,392 (49.20%)	0.9687
	รายเดือน	6,474 (47.63%)	7,118 (52.37%)	1.0995
ประเภทของสัญญาหลัก	ตลอดชีพ	7,480 (41.02%)	10,753 (58.98%)	1.4376
	อื่นๆ	2,592 (36.94%)	4,424 (63.06%)	1.7068
จำนวนครั้งของการเคลม	ไม่มีการเคลม	8,851 (42.70%)	11,877 (57.30%)	1.3419
	เคลม 1 ครั้ง	791 (27.24%)	2,113 (72.76%)	2.6713
	เคลมมากกว่า 1 ครั้ง	430 (26.59%)	1,187 (73.41%)	2.7605 (2)
จำนวนเงินเคลม	ไม่เคยได้รับเงินเคลม	8,851 (42.70%)	11,877 (57.30%)	1.3419
	เคยได้รับเงินเคลม	1,221 (27.01%)	3,300 (72.99%)	2.7027

จากตารางที่ 2 จะเห็นว่าลักษณะการจ่ายเบี้ยประกันภัยน่าจะมีผลต่อการต่ออายุสัญญาโดยการจ่ายเบี้ยประกันภัยแบบรายปีมีสัดส่วนการต่ออายุสัญญามากที่สุดคือ 3.747 รองลงมาคือจำนวนครั้งของการเคลม โดยผู้เอาประกันภัยที่มีจำนวนครั้งของการเคลมมากกว่า 1 ครั้งมีสัดส่วนการต่ออายุสัญญาคือ 2.761

ค่าความแม่นยำของต้นไม้การจำแนกในภาพรวม

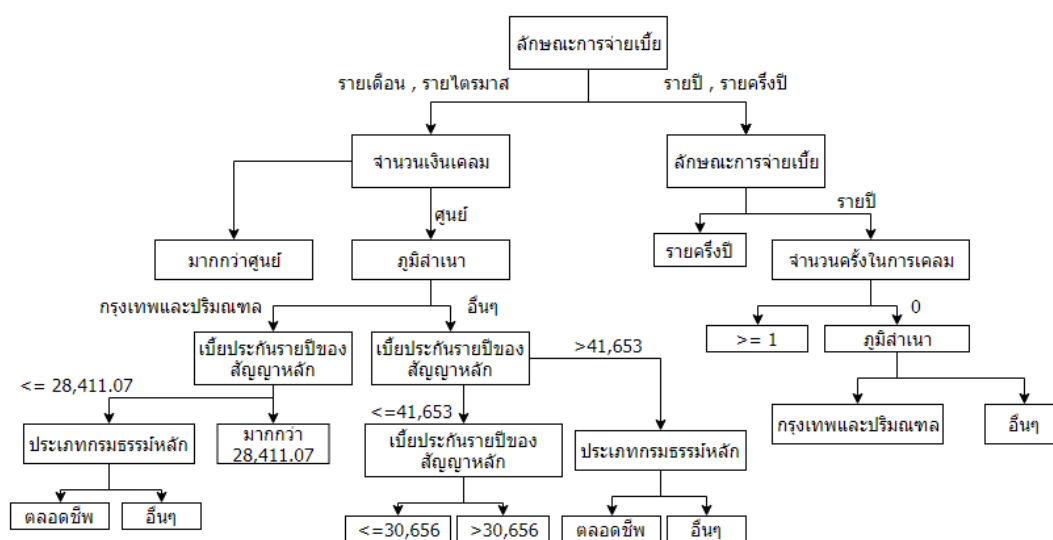
ค่าความแม่นยำเฉลี่ยของต้นไม้การจำแนกจากการใช้ตัวแปรทำนาย 12 ตัวและใช้วิธีการตรวจสอบความสมเหตุสมผลแบบไขว้ การตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างจำแนกตามวิธีการลิดกิ่งเมื่อทำซ้ำ 500 ครั้ง แสดงในตารางที่ 3

ตารางที่ 3 ค่าความแม่นยำเฉลี่ยของต้นไม้การจำแนกจากการทำซ้ำ 500 ครั้ง จำแนกตามวิธีการตรวจสอบความสมเหตุสมผลของตัวแบบและวิธีการลิดกิ่ง

			ค่าความแม่นยำ	
			ชุดตัวอย่างฝึกฝน	ชุดตัวอย่างทดสอบ
การตรวจสอบความสมเหตุสมผลแบบไขว้	CHAID	k=2	N/A	62.112%
		k=3	N/A	62.199%
		k=4	N/A	62.151%
		k=5	N/A	62.173%
การตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่าง	CHAID	50:50	62.380%	62.236% (3)
		60:40	62.592%	62.173%
		70:30	62.573%	62.174%
		80:20	62.523%	62.031%
	CRT	50:50	62.451%	62.167%
		60:40	62.492%	62.076%
		70:30	62.378%	62.237% (2)
		80:20	62.504%	61.806%

	QUEST	50:50	62.677%	62.033%
		60:40	62.496%	62.104%
		70:30	62.631%	62.117%
		80:20	62.654%	62.374% (1)

จากตารางที่ 3 จะเห็นว่าจากทำซ้ำ 500 ครั้ง ค่าความแม่นยำเฉลี่ยของตัวแบบต้นไม่การจำแนกโดยใช้การตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 80:20 และใช้การลิดกิ่งโดยวิธี QUEST ให้ค่าความแม่นยำเฉลี่ยสูงสุดเท่ากับ 62.374% รองลงมาคือวิธีการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 70:30 และใช้วิธีการลิดกิ่งโดยวิธี CRT ให้ค่าความแม่นยำเฉลี่ยเท่ากับ 62.237% ซึ่งไม่ต่างกันมากนักกับวิธีการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 50:50 และใช้วิธีการลิดกิ่งโดยวิธี CHAID ให้ค่าความแม่นยำเฉลี่ยเท่ากับ 62.236%



ภาพที่ 1 ต้นไม้การจำแนก (แสดงเฉพาะตัวแปรทำนาย) ที่ใช้วิธีการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 80:20 และใช้วิธีการลิดกิ่งโดยวิธี QUEST

ตัวอย่างการพิจารณาตัวแปรทำนายที่มีผลต่อการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวัน แสดงในภาพที่ 1 ซึ่งจะเห็นว่าลักษณะการจ่ายเบี้ยประกันภัยเป็นตัวแปรทำนายที่มีความสำคัญเป็นอันดับแรกในการทำนายการต่ออายุสัญญา สำหรับสัญญาที่จ่ายเบี้ยประกันภัยรายเดือนหรือรายไตรมาส ตัวแปรทำนายที่มีผลอันดับรองลงไปคือจำนวนเงินเคลม สำหรับสัญญาที่จ่ายเบี้ยประกันภัยรายปี ตัวแปรทำนายที่มีผลอันดับรองลงไปคือจำนวนครั้งในการเคลม เป็นต้น

ค่าความแม่นยำของต้นไม้การจำแนกตามลักษณะการจ่ายเบี้ยประกันภัย

ค่าความแม่นยำของต้นไม้การจำแนก เมื่อใช้วิธีการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 80:20 และใช้วิธีการลิดกิ่งโดยวิธี QUEST จำแนกตามลักษณะการจ่ายเบี้ยประกันภัย (รายปี รายครึ่งปี รายไตรมาส และรายเดือน) แสดงในตารางที่ 4

ตารางที่ 4 ค่าความแม่นยำของต้นไม้การจำแนกเมื่อใช้วิธีการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 80:20 และใช้วิธีการลิดกิ่งโดยวิธี QUEST จำแนกตามลักษณะการจ่ายเบี้ยประกันภัย

ปัจจัย		จำนวนสัญญา	ค่าความแม่นยำ
ลักษณะการจ่าย เบี้ยประกันภัย	รายปี	6,949 (27.52%)	80.256%
	รายครึ่งปี	1,879 (7.44%)	66.201%
	รายไตรมาส	2,829 (11.20%)	53.103%
	รายเดือน	13,592 (53.83%)	53.923%

จากตารางที่ 4 สัญญาที่จ่ายเบี้ยประกันภัยแบบรายปี ค่าความแม่นยำของต้นไม้การจำแนกในการทำนายการต่ออายุสัญญามีค่า 80.256% สัญญาที่จ่ายเบี้ยประกันภัยแบบรายครึ่งปีค่าความแม่นยำของต้นไม้การจำแนกในการทำนายการต่ออายุสัญญามีค่า 66.201% ส่วนสัญญาที่จ่ายเบี้ยประกันภัยแบบรายไตรมาสและรายเดือน ค่าความแม่นยำของต้นไม้การจำแนกในการทำนายการต่ออายุสัญญามีค่าประมาณ 53-54% และเมื่อพิจารณาจากแผนภาพต้นไม้พบว่าปัจจัยที่มีผลต่อการต่ออายุสัญญาของสัญญาที่จ่ายเบี้ยประกันภัยแบบรายปีได้แก่ จำนวนครั้งของการเคลม รองลงมาคือภูมิลำเนา

4. อภิปรายผลการวิจัยและข้อเสนอแนะ

จากทำซ้ำ 500 ครั้ง ค่าความแม่นยำของตัวแบบต้นไม้การจำแนกโดยใช้วิธีการตรวจสอบความสมเหตุสมผลแบบแบ่งตัวอย่างในสัดส่วน 80:20 และใช้วิธีการลิดกิ่งโดยวิธี QUEST ให้ค่าความแม่นยำสูงสุดเท่ากับ 62.374% เมื่อนำวิธีนี้มาหาค่าความแม่นยำและปัจจัยที่มีผลต่อการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันจำแนกตามลักษณะการจ่ายเบี้ยประกันภัย พบว่าสำหรับสัญญาที่จ่ายเบี้ยประกันภัยแบบรายปี ค่าความแม่นยำของต้นไม้การจำแนกในการทำนายการต่ออายุสัญญาเท่ากับ 80.26% สำหรับสัญญาที่จ่ายเบี้ยประกันภัยแบบรายครึ่งปี ค่าความแม่นยำของต้นไม้การจำแนกในการทำนายการต่ออายุสัญญาเท่ากับ 66.20% สอดคล้องกับการพิจารณาเบื้องต้นด้วยสถิติพรรณนาซึ่งพบว่าลักษณะการจ่ายเบี้ยประกันภัยน่าจะมีผลต่อการต่ออายุสัญญา โดยการจ่ายเบี้ยประกันภัยแบบรายปีมีสัดส่วนการต่ออายุสัญญามากที่สุดคือ 3.747 และการจ่ายเบี้ยประกันภัยแบบรายครึ่งปีมีสัดส่วนการต่ออายุสัญญาเท่ากับ 1.696 โดยปัจจัยที่มีผลต่อการต่ออายุสัญญาของสัญญาที่จ่ายเบี้ยประกันภัยแบบรายปีได้แก่ จำนวนครั้งของการเคลม รองลงมาคือภูมิภาค ซึ่งสอดคล้องกับสถิติพรรณนาที่พบว่าจำนวนครั้งของการเคลมน่าจะมีผลต่อการต่ออายุสัญญา รองจากปัจจัยลักษณะการจ่ายเบี้ยประกันภัยโดยผู้เอาประกันภัยที่มีจำนวนครั้งของการเคลมมากกว่า 1 ครั้งมีสัดส่วนการต่ออายุสัญญาเท่ากับ 2.761

สิ่งที่น่าสนใจศึกษาต่อไปในการวิจัยครั้งต่อไป ได้แก่ การศึกษาความแม่นยำของต้นไม้การจำแนกเมื่อพิจารณาจำแนกตามประเภทของกรมธรรม์หลัก ที่แบ่งเป็นแบบตลอดชีพและแบบอื่น ๆ รวมทั้งการจำแนกร่วมกันระหว่างประเภทของกรมธรรม์หลักและลักษณะการจ่ายเบี้ยประกัน ซึ่งเป็นประโยชน์ในการนำไปประยุกต์ใช้จริงตามชนิดของกรมธรรม์หลักและวิธีการจ่ายเบี้ยประกันที่บริษัทประกันภัยสามารถใช้ทำนายการต่ออายุสัญญาเพิ่มเติมชดเชยค่ารักษาพยาบาลรายวันได้ หรือนำไปประยุกต์กับผลิตภัณฑ์หรือกรมธรรม์ชนิดอื่น นอกจากนี้อาจใช้วิธีอื่น ๆ ในการทำนายการต่ออายุกรมธรรม์หรือการต่อสัญญา เช่น ตัวแบบป่าสุ่ม (random forest) หรือใช้วิธีบูสต์ทิงปรับตัว (adaptive boosting)

5. เอกสารอ้างอิง

- เอกสิทธิ์ พัทธวงศ์ศักดิ์. (2558). *การแบ่งข้อมูลเพื่อนำมาทดสอบประสิทธิภาพของโมเดล*.
<https://th.linkedin.com/pulse/การแบ่งข้อมูลเพื่อนำมาทดสอบประสิทธิภาพของโมเดล-eakasit-pacharawongsakda>
- IBM, (2017). *IBM SPSS Statistics Algorithms*. https://www.ibm.com/docs/en/SSLVMB_26.0.0/pdf/en/IBM_SPSS_Statistics_Algorithms.pdf
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. New York: Springer-Verlag New York.