การทำนายการเลือกประเภทประกันภัยของลูกค้า

Prediction of choosing types of insurance for the customers

ชนิตา หางแก้ว* จุฑาภรณ์ สินสมบูรณ์ทอง และ ธิดาพร ศุภภากร

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ กรุงเทพมหานคร 10900

Chanita Hangkaew*, Juthaphorn Sinsomboonthong and Thidaporn Supapakorn

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900

Received: 19 September 2020/ Revised: 20 December 2020/ Accepted: 25 December 2020

บทคัดย่อ

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาปัจจัยที่มีผลต่อการตัดสินใจการเลือกทำประเภทประกันภัยของลูกค้า และเปรียบเทียบประสิทธิภาพเทคนิคการวิเคราะห์จำแนกกลุ่ม 4 วิธี คือ วิธีการวิเคราะห์การถดถอยลอจิสติกพหุ (multinomial logistic regression) วิธีการวิเคราะห์จำแนกกลุ่ม (discriminant analysis) วิธีต้นไม้ตัดสินใจ (decision tree) และวิธีโครงข่ายประสาทเทียม (artificial neural network) โดยพิจารณาจากตัวแปรอิสระ 6 ตัว ได้แก่ เพศ อายุ รายได้ต่อเดือน วุฒิการศึกษา สถานภาพครอบครัว และอาชีพ ผลการวิจัยสรุปได้ว่า จากการวิเคราะห์การถดถอยลอจิสติกพหุ ปัจจัยที่ส่งผลต่อการเลือกทำประเภทประกันภัยของลูกค้ามี 5 ปัจจัย คือ อายุ รายได้ต่อเดือน วุฒิการศึกษา สถานภาพครอบครัว และอาชีพ และจากการวิเคราะห์จำแนกกลุ่มพบว่า ปัจจัยที่ส่งผลต่อการตัดสินใจการเลือกทำประเภทประกันของลูกค้ามี 3 ปัจจัย คือ เพศ อายุ และสถานภาพครอบครัว นอกจากนี้ในการเปรียบเทียบประสิทธิภาพเทคนิคการวิเคราะห์จำแนกกลุ่มทั้ง 4 วิธี พบว่า วิธีโครงข่ายประสาทเทียมเป็นวิธีที่มีประสิทธิภาพในการจำแนกกลุ่มมากที่สุด โดยให้ค่าความแม่นร้อยละ 85.75 รองลงมาคือ วิธีต้นไม้ตัดสินใจ วิธีการวิเคราะห์การถดถอยลอจิสติกพหุ และวิธีการวิเคราะห์จำแนกกลุ่ม โดยให้ค่าความแม่นร้อยละ 76.50, 70.00 และ 61.80 ตามลำดับ

คำสำคัญ: การวิเคราะห์การถดถอยลอจิสติกพหุ การวิเคราะห์จำแนกกลุ่ม ต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม ประเภทของการประกันภัย

Corresponding author: chanita.ha@ku.th

Abstract

The objectives of this research are to study the factors that have an influence on the decision making for choosing types of insurance of customers and to compare the efficiency of four classification methods which are multinomial logistic regression, discriminant analysis, decision tree and artificial neural network. By considering of six independent variables, namely gender, age, monthly income, education, marital and occupation. This study is found that multinomial logistic regression analysis indicates that five factors affecting the decision making are age, monthly income, education, marital, and occupation. In addition, the study of discriminant analysis indicates that three factors; gender, age and marital are significantly affecting the decision making. For the efficiency comparison of four methods, it's shown that neural network is the most efficiency method which have the accuracy is 85.57% and followed by decision tree, multinomial logistic regression analysis, and discriminant analysis which have the accuracy are 76.50%, 70.00% and 61.80%, respectively.

Keywords: Multinomial logistic regression, Discriminant analysis, Decision tree, Artificial neural network, Insurance types

บทนำ

ธุรกิจประกันภัยจัดเป็นหนึ่งอุตสาหกรรมสำคัญของโลก ด้วยเหตุผลที่ช่วยเพิ่มความมั่นคงให้กับผู้ซื้อประกันภัย ช่วยแบ่งเบาภาระ และเพิ่มความมั่นใจในการใช้ชีวิตในโลกปัจจุบันที่ทุกอย่างล้วนไม่แน่นอน ซึ่งอาจจะเป็นภัยอันตรายที่เกิดจากการกระทำของตนเอง และภัยอันตรายที่เกิดขึ้นเองตามธรรมชาติ เช่น น้ำท่วม เป็นต้น ดังนั้นการประกันภัยจึงเป็นการบริหารความเสี่ยงภัยวิธีหนึ่ง ซึ่งจะโอนความเสี่ยงภัยของผู้เอาประกันภัยไปสู่บริษัทประกันภัย และเมื่อเกิดความเสียหายขึ้น บริษัทประกันภัยจะชดใช้ค่าสินไหมทดแทนตามที่ได้รับความคุ้มครองในกรมธรรม์ประกันภัยให้แก่ผู้เอาประกันภัย โดยที่ผู้เอาประกันภัยจะต้องเสียเบี้ยประกันภัยให้แก่บริษัทประกันภัยตามที่ได้ตกลงกันไว้ ซึ่งรูปแบบการประกันภัยมาตรฐาน แบ่งออกเป็นสองสายหลัก คือ การประกันชีวิตและการประกันวินาศภัย และในงานวิจัยนี้ ผู้วิจัยต้องการมุ่งเน้นไปที่การศึกษารูปแบบการประกันของการประกันชีวิตประเภทต่าง ๆ ซึ่งรวมไปถึงประกันสุขภาพด้วย

การประกันชีวิตมีหลายแบบ ซึ่งแต่ละแบบจะมีลักษณะการคุ้มครองและผลประโยชน์ต่างกันออกไป ซึ่งมีพื้นฐานอยู่ด้วยกัน 4 แบบ คือ

1) แบบตลอดชีพ เป็นการประกันชีวิตที่ให้ความคุ้มครองตลอดชีพ โดยบริษัทประกันภัยจะจ่ายเงินให้ตามจำนวนที่ระบุไว้ให้แก่ผู้รับประโยชน์ เมื่อผู้เอาประกันเสียชีวิต โดยไม่คำนึงว่าจะเสียชีวิตลงเมื่อใด

2) แบบสะสมทรัพย์ เป็นการประกันแบบคุ้มครองและออมทรัพย์ โดยที่บริษัทประกันภัยจะจ่ายเงินให้ผู้รับประโยชน์ หากผู้เอาประกันเสียชีวิตภายในระยะเวลาที่ได้กำหนดไว้ในสัญญา การทำประกันประเภทนี้จะกำหนดระยะเวลาในการคุ้มครองไว้แน่นอน เช่น 10, 15 และ 20 ปี เป็นต้น

3) แบบชั่วระยะเวลา เป็นการประกันแบบที่บริษัทประกันภัยจะจ่ายเงินให้กับผู้รับประโยชน์เมื่อผู้เอาประกันเสียชีวิตภายในระยะเวลาที่กำหนด เป็นประกันที่ให้คุ้มครองการเสี่ยงภัยอันเกิดจากการเสียชีวิตเพียงอย่างเดียว ไม่มีการสะสมทรัพย์ใด ๆ เบี้ยประกันภัยจึงต่ำกว่าแบบอื่น ดังนั้นเมื่อครบกำหนดสัญญาแล้วจะไม่มีมูลค่าใด ๆ คืนให้แก่ผู้เอาประกัน

4) แบบได้เงินประจำ เป็นการประกันที่ผู้เอาประกันต้องการรายได้เมื่อยามชรา โดยบริษัทประกันภัยจะจ่ายเงินเป็นรายงวดให้แก่ผู้เอาประกันเป็นประจำนับตั้งแต่ครบกำหนดสัญญา เป็นประกันที่เน้นการออมทรัพย์ไว้ใช้จ่ายยามเกษียณมากกว่าคุ้มครองชีวิต

ทั้งนี้ประกันอีกประเภทหนึ่งที่ได้รับความสนใจมากเช่นกัน คือ การประกันสุขภาพ ซึ่งจะแบ่งประเภทไปตามระดับความคุ้มครอง 5 ระดับ ดังนี้ ผู้ป่วยใน (inpatient care) ผู้ป่วยนอก (outpatient care) โรคร้ายแรง อุบัติเหตุ และชดเชยรายได้ ซึ่งบริษัทประกันภัยที่ผู้วิจัยได้นำข้อมูลมาศึกษานั้น มีผลิตภัณฑ์หลายแบบเพื่อให้ผู้ทำประกันได้ตัดสินใจเลือกทำเพื่อให้ตรงตามความต้องการที่สุด คือ ประกันชีวิตแบบตลอดชีพ ประกันชีวิตแบบสะสมทรัพย์ และประกันสุขภาพ เนื่องจากความหลากหลายของประเภทประกัน ผู้วิจัยจึงต้องการใช้เทคนิคการวิเคราะห์ทางสถิติต่าง ๆ มาศึกษาปัจจัยที่ส่งผลต่อการเลือกทำผลิตภัณฑ์ประกันภัยแต่ละประเภท รวมทั้งสร้างตัวแบบที่เหมาะสมในการพยากรณ์การเลือกทำประเภทประกันภัยของลูกค้า จากการศึกษางานวิจัยที่เกี่ยวข้องกับปัจจัยที่ส่งผลต่อการเลือกทำประเภทประกันภัยนั้น พบว่างานวิจัยของ นพินดา [1] ซึ่งได้ทำการศึกษาปัจจัยที่มีผลต่อการตัดสินใจซื้อประกันภัยของผู้ที่อยู่ในวัยทำงานในเขตกรุงเทพฯ พบว่ารายได้ต่อเดือนและอาชีพมีผลต่อการตัดสินใจซื้อประกันภัย แต่เพศ อายุ และวุฒิการศึกษานั้นไม่มีผลต่อการตัดสินใจซื้อประกันภัย ต่อมา อุสมาน และคณะ [2] ได้ทำการศึกษาปัจจัยที่ผลต่อการเลือกทำประกันชีวิตกับบริษัทกรุงไทย แอกซ่า ในเขตเทศบาลนครหาดใหญ่พบว่า วุฒิการศึกษาและอาชีพมีผลต่อการเลือกทำประกัน และพัสวี [3] ได้ศึกษาปัจจัยที่มีอิทธิพลต่อการเลือกซื้อประกันชีวิตของข้าราชการครูพบว่าอายุและวุฒิการศึกษาที่แตกต่างกันนั้นมีผลต่อการเลือกซื้อประกัน และยังได้เสนอให้ศึกษาปัจจัยสถานภาพครอบครัวเพิ่มเติม ผู้วิจัยจึงเลือกปัจจัย เพศ อายุ รายได้ต่อเดือน วุฒิการศึกษา สถานภาพครอบครัว และอาชีพ มาทำการศึกษาและสร้างตัวแบบที่เหมาะสมในการพยากรณ์การตัดสินใจเลือกทำประกันภัยประเภทต่าง ๆ ของลูกค้า โดยวัตถุประสงค์ในการวิจัย คือ ศึกษาปัจจัยที่มีผลต่อการตัดสินใจเลือกทำประกันภัยในผลิตภัณฑ์แต่ละประเภท และเปรียบเทียบประสิทธิภาพเทคนิคการวิเคราะห์การจำแนกกลุ่ม 4 วิธี คือ เทคนิคการวิเคราะห์การถดถอยลอจิสติกพหุ (multinomial logistic regression) เทคนิคการวิเคราะห์จำแนกกลุ่ม (discriminant analysis) ต้นไม้ตัดสินใจ (decision tree) และโครงข่ายประสาทเทียม (artificial neural network)

วิธีดำเนินการวิจัย

ข้อมูลและตัวแปรที่ใช้ในการวิจัย

งานวิจัยนี้ใช้ข้อมูลทุติยภูมิซึ่งเก็บรวบรวมโดยบริษัทประกันภัยแห่งหนึ่งในช่วงปี พ.ศ. 2557-2562 และผู้วิจัยได้สุ่มเลือกหน่วยตัวอย่าง จำนวน 400 ราย โดยสุ่มเลือกหน่วยตัวอย่างที่ซื้อประกันภัยเพียงประเภทเดียวเท่านั้น โดยตัวแปรที่อิสระใช้ในการศึกษามีจำนวนทั้งสิ้น 6 ตัวแปร ได้แก่ เพศ อายุ รายได้ต่อเดือน วุฒิการศึกษา สถานภาพครอบครัว และอาชีพ ซึ่งตัวแปรอิสระเพศ วุฒิการศึกษา สถานภาพครอบครัว และอาชีพเป็นข้อมูลเชิงคุณภาพ โดยตัวแปรเพศจะแบ่งออกเป็น 2 ประเภท คือ เพศชายและหญิง ตัวแปรวุฒิการศึกษาแบ่งออกเป็น 8 ประเภท คือ ประถมศึกษา มัธยมศึกษาตอนต้น มัธยมศึกษาตอนปลาย ปวช. ปวส. ปวท. อนุปริญญา ปริญญาตรี และปริญญาโท หรือสูงกว่า ตัวแปรสถาพภาพครอบครัว แบ่งออกเป็น 3 ประเภท คือ สมรส โสด และหย่าร้าง ส่วนตัวแปรอิสระอายุและรายได้ต่อเดือนนั้นเป็นข้อมูลเชิงปริมาณ ตัวแปรตาม คือ ประเภทของประกันภัย ซึ่งแบ่งออกเป็น 3 กลุ่ม คือ ประกันชีวิตแบบสะสมทรัพย์ ประกันชีวิตแบบตลอดชีวิต และประกันสุขภาพ

การวิเคราะห์ข้อมูล

ในการวิเคราะห์ข้อมูลจะแบ่งเป็น 2 แบบ คือ การใช้ตัวแบบทางสถิติ ได้แก่ วิธีการวิเคราะห์การถดถอยลอจิสติกพหุและการวิเคราะห์จำแนกกลุ่ม ซึ่งจะทำการวิเคราะห์ด้วยโปรแกรม SPSS และการใช้ตัวแบบที่ไม่ได้อยู่บนข้อสมมติเชิงสถิติ ได้แก่ ต้นไม้ตัดสินใจและโครงข่ายประสาทเทียม ซึ่งจะทำการวิเคราะห์ด้วยโปรแกรม Weka

1. การวิเคราะห์ด้วยตัวแบบทางสถิติ

1.1 การวิเคราะห์การถดถอยลอจิสติกพหุ

วัตถุประสงค์ของการวิเคราะห์การถดถอยลอจิสติก คือ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ รวมถึงระดับความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม และเพื่อพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยใช้ตัวแบบที่ได้สร้างขึ้นมาจากตัวแปรอิสระที่ส่งผลกับตัวแปรตามที่ได้ศึกษาไว้ก่อนหน้า [4]

1.1.1 ข้อสมมติเบื้องต้น มีดังนี้

1) ตัวแปรอิสระ (x) เป็นตัวแปรที่อยู่ในระดับมาตราช่วง (interval scale) เป็นอย่างต่ำ หากเป็นตัวแปรที่เป็นข้อมูลเชิงกลุ่มให้แปลงเป็นตัวแปรหุ่น (dichotomous variable) ที่มีค่าเป็น 0 กับ 1 ส่วนตัวแปรตาม (y) ในกรณีการวิเคราะห์ พหุ จะกำหนดค่าตามจำนวนกลุ่มของตัวแปรตาม

2) ความคลาดเคลื่อนสุ่ม (ε_{ij}) เป็นอิสระกัน โดย E (ε_{ij}) =0

3) ตัวแปรอิสระเป็นอิสระกัน

4) ตัวแปรอิสระต้องไม่มีความสัมพันธ์กันหรือไม่มีปัญหาความสัมพันธ์เชิงเส้นพหุ ซึ่งจะพิจารณาจากค่าความคลาดเคลื่อนยินยอม (tolerance) และค่าองค์ประกอบการขยายความแปรปรวน (Variance Inflation Factor : VIF) โดยค่าความคลาดเคลื่อนยินยอมต้องมีค่าเข้าใกล้ 1 และค่าองค์ประกอบการขยายความแปรปรวนต้องมีค่าไม่เกิน 10

5) การวิเคราะห์ถดถอยลอจิสติกต้องใช้ขนาดตัวอย่าง (n) มากกว่าการวิเคราะห์การถดถอยแบบปกติ โดยจะใช้ขนาดตัวอย่างไม่น้อยกว่า 30p ซึ่ง p คือ จำนวนตัวแปรอิสระ

1.1.2 การเลือกตัวแปรอิสระเข้าสมการถดถอยลอจิสติก โดยในงานวิจัยนี้เลือกใช้วิธี Enter

1.1.3 ตัวแบบการวิเคราะห์การถดถอยลอจิสติกพหุ [5, 6] ในกรณีตัวแปรตามมีค่ามากกว่า 2 ค่า เช่น มี k ค่าโดย k > 2 จะได้สมการลอจิต จำนวนเท่ากับ k-1 สมการ โดยที่แต่ละสมการจะเปรียบเทียบกับลอจิตของกลุ่มที่เป็นฐาน (baseline category) เช่น ถ้าให้กลุ่มที่เป็นฐาน คือ k จะได้สมการลอจิตของกลุ่มที่ i ดังนี้

$$\log\left(\frac{P(\text{กลุ่ม i})}{P(\text{กลุ่ม k})}\right)=b_{i0}+b_{i1}x_1+...+b_{ip}x_p \qquad (1)$$

เมื่อ P(กลุ่ม i) คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (y=i)

P(กลุ่ม k) คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ (y=k หรือกลุ่มฐาน)

x_j คือ ตัวแปรอิสระ ตัวที่ j โดยที่" j = 1, 2,..., p"

b_{ij} คือ ตัวประมาณของค่าสัมประสิทธิ์ของตัวแปรอิสระที่ j ในกลุ่มที่ i โดยที่ j = 1, 2,..., p และ i = 1, 2,..., k

สำหรับค่าของตัวประมาณของค่าสัมประสิทธิ์ของตัวแปรอิสระแต่ละตัวในกลุ่มฐานหรือกลุ่มที่ k นั้น จะเท่ากับ 0 เพื่อเป็นฐานในการเปรียบเทียบให้กับลอจิตของกลุ่มอื่น ตัวอย่างเช่น ถ้าตัวแปรตามมี 3 ค่า หรือเมื่อ k=3 กลุ่มที่เป็นฐานคือ กลุ่มที่ 3 (y=3) โดยชุดที่ 1 แสดงค่าประมาณของค่าสัมประสิทธิ์ของ y=1 เปรียบเทียบกับ y=3 และชุดที่ 2 จะแสดงค่าประมาณของค่าสัมประสิทธิ์ของ y=2 เปรียบเทียบกับ y=3

1.2 การวิเคราะห์จำแนกกลุ่ม

เป็นเทคนิคทางสถิติที่ใช้ในการจำแนกกลุ่มตั้งแต่ 2 ประเภท ขึ้นไป โดยใช้ตัวแปรอิสระ (x) ตั้งแต่ 1 ตัวขึ้นไปในการพยากรณ์ค่าของตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพหรือเชิงกลุ่ม (y) จำนวน 1 ตัว ซึ่งอาจเรียกได้ว่าเป็นตัวแปรที่แสดงกลุ่ม ซึ่งจะมีได้หลายค่า แต่ละค่าจะแสดงกลุ่มที่สังกัด และตัวแปรอิสระหรือตัวแปรที่ทำให้กลุ่มต่างกันจะเรียกว่าตัวแปรจำแนกกลุ่ม (discriminant variable) ซึ่งควรเป็น

ตัวแปรเชิงปริมาณ ดังนั้นหากเกิดกรณีที่ตัวแปรอิสระเป็นตัวแปรเชิงกลุ่มหรือเชิงคุณภาพจะต้องเปลี่ยนให้อยู่ในรูปตัวแปรหุ่น (dummy variable) ก่อน ซึ่งในกรณีที่มีตัวแปรตามมากกว่า 2 กลุ่ม หรือแบ่งเป็น k กลุ่มนั้น จะสามารถสร้างสมการจำแนกกลุ่มได้เท่ากับ min{p, k-1} ซึ่งสมการจำแนกกลุ่มกรณีตัวแปรตามมี k กลุ่ม จะได้ดังนี้

$$\hat{D}_i = b_{i1}x_1 + b_{i2}x_2 + \ldots + b_{ip}x_p \qquad (2)$$

เมื่อ i = 1, 2,..., k-1

โดย $\hat{D}_i$ แทน discriminant score กลุ่มที่ i โดยที่ i = 1, 2,..., k-1

x_j แทน ตัวแปรอิสระ ตัวที่ j โดยที่ j = 1, 2,..., p

b_{ij} แทน ตัวประมาณสัมประสิทธิ์ของสมการจำแนกกลุ่ม โดยที่ i = 1, 2,..., k-1 และ j = 1, 2,..., p

k แทน จำนวนกลุ่มของตัวแปรตาม

การประมาณค่าสัมประสิทธิ์ของสมการจำแนกกลุ่มมีเป้าหมายเพื่อทำให้ความแตกต่างระหว่างกลุ่มมีค่ามากที่สุด นั่นคือ ต้องทำให้ค่าไอเก้นมีค่าสูงสุด หรือทำให้ร้อยละของการจัดกลุ่มผิดมีค่าน้อยที่สุด ซึ่งเรียกค่าไอเก้นว่า เกณฑ์การจำแนกกลุ่ม (discriminant criterion) หรือรากลักษณะเฉพาะ (characteristic roots) หรือรากแฝง (latent roots) เขียนเป็นสัญลักษณ์ได้ด้วย $\boldsymbol{\lambda}$ ซึ่งค่า $\boldsymbol{\lambda}$ มีได้หลายค่า โดยจำนวนของค่า $\boldsymbol{\lambda}$ จะเท่ากับจำนวนกลุ่ม (k) ลบด้วย 1 หรือเท่ากับจำนวนตัวแปร (p) แล้วแต่จำนวนใดน้อยกว่ากัน หรือ min{k-1, p}

วัตถุประสงค์ของการวิเคราะห์จำแนกกลุ่มเพื่อศึกษาความแตกต่างของตัวแปรอิสระระหว่างกลุ่ม และตรวจสอบว่ากลุ่มที่ได้แบ่งไว้นั้น มีความแตกต่างของตัวแปรอิสระอย่างมีนัยสำคัญหรือไม่ รวมถึงยังใช้ในการพยากรณ์หน่วยใหม่ว่าควรอยู่กลุ่มใด

1.2.1 ข้อสมมติเบื้องต้นของการวิเคราะห์จำแนกกลุ่ม [7] มีดังนี้

1) ตัวแปรอิสระต้องมีการแจกแจงปกติหลายตัวแปร

2) เมทริกซ์ค่าแปรปรวนร่วมของตัวแปรอิสระทั้ง p ตัว ของทุกกลุ่มต้องมีค่าเท่ากัน

3) ตัวแปรอิสระและตัวแปรตาม (D) ต้องมีความสัมพันธ์เชิงเส้นกัน

4) ตัวแปรอิสระต้องไม่มีความสัมพันธ์เชิงเส้นแบบพหุ (multicollinearity)

1.2.2 การคัดเลือกตัวแปรอิสระเข้าสมการจำแนกกลุ่ม โดยในงานวิจัยนี้เลือกใช้วิธีแบบขั้นตอน (stepwise method)

1.2.3 สถิติสำคัญของการวิเคราะห์จำแนกกลุ่ม ได้แก่

1) ค่าไอเก้น (Eigen value) เป็นค่าที่แสดงอัตราส่วนการผันแปรระหว่างกลุ่มต่อการผันแปรภายในกลุ่ม ใช้วัดความสำคัญเชิงเปรียบเทียบของสมการว่า สมการที่ได้มีอำนาจการแบ่งแยกกลุ่มได้ดีแค่ไหน โดยถ้าค่าไอเก้นสูงแสดงว่าสมการจำแนกกลุ่มดีหรือมีค่าจำแนกสูง

2) ร้อยละสัมพัทธ์ (relative pecentage) หรือร้อยละสะสม (cumulative percentage) เป็นค่าที่ใช้แสดงความสัมพันธ์ระหว่างค่าไอเก้นกับสมการที่ได้ว่า มีอำนาจในแบ่งแยกคิดเป็นร้อยละเท่าใดของอำนาจในการจำแนกรวม

3) ค่าวิลกส์แลมบ์ดา (Wilks' Lambda) คือ ค่าสถิติที่ใช้ทดสอบความเท่ากันของค่าเฉลี่ยของตัวแปรอิสระแต่ละกลุ่มและยังเป็นมาตราวัดอำนาจในการจำแนกกลุ่มของตัวแปรด้วย ซึ่งถ้าค่าวิลกส์แลมบ์ดามีค่ามาก ตัวแปรจะอธิบายการเป็นสมาชิกของกลุ่มได้น้อย แต่ถ้าหากค่าวิลกส์แลมบ์ดามีค่าน้อย ตัวแปรอิสระจะอธิบายการเป็นสมาชิกของกลุ่มได้มาก

2. การวิเคราะห์ด้วยตัวแบบที่ได้อยู่บนข้อสมมติเชิงสถิติ

2.1 การวิเคราะห์ต้นไม้ตัดสินใจ

การจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ [8] เป็นกระบวนการสร้างต้นไม้เพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลแนบอยู่ด้วย ซึ่งเรียกการจำแนกข้อมูลด้วยข้อมูลลักษณะนี้ว่า การเรียนรู้แบบมีผู้สอน (supervised learning) ซึ่งจะสามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จากกลุ่มตัวอย่างของข้อมูลที่ได้กำหนดให้เป็นข้อมูลฝึกฝน (training data) ได้โดยอัตโนมัติแล้วยังสามารถนำมาพยากรณ์กลุ่มของรายการใหม่ที่ยังไม่เคยนำมาใช้ในการจัดหมวดหมู่ได้อีกด้วย

2.1.1 โครงสร้างของต้นไม้ตัดสินใจ

1) โหนดราก (root node) คือ โหนดแรกของต้นไม้ตัดสินใจ

2) โหนดภายใน (internal node) คือ คุณลักษณะต่าง ๆ ของข้อมูล ซึ่งเมื่อข้อมูลมาถึงโหนดจะใช้คุณลักษณะเหล่านี้ ในการตัดสินใจว่าข้อมูลจะไปในทิศทางใดต่อไป เรียกได้อีกอย่างว่าโหนดลูก ซึ่งโหนดภายในจะมีจุดเริ่มต้นคือโหนดราก

3) กิ่ง (branch) คือ ค่าคุณลักษณะของแต่ละโหนดที่มีการแตกกิ่งออกไป โดยจะแตกจำนวนกิ่งไปเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายในนั้น

4) โหนดใบ (leaf node) คือ กลุ่มของผลลัพธ์ในการจำแนกประเภทข้อมูล

2.1.2 ขั้นตอนวิธีการสร้างต้นไม้ตัดสินใจ เริ่มต้นการสร้างด้วยการสร้างจากโหนดราก ใช้โหนดรากเป็นโหนดปัจจุบันที่มีข้อมูลเรียนรู้ทั้งหมด หลังจากนั้นทำตามขั้นตอนดังต่อไปนี้

1) ตรวจสอบเงื่อนไขการหยุดการสร้างโหนดลูก โดยถ้าเป็นจริงจะหยุดการสร้างโหนดลูกที่โหนดปัจจุบัน และจะกำหนดให้ที่โหนดปัจจุบันเป็นโหนดใบที่มีผลของการจำแนกประเภทของข้อมูล ซึ่งในโหนดนั้นจะเป็นประเภทของข้อมูลที่มีสัดส่วนมากที่สุดในข้อมูลชุดนั้น แต่ถ้าเงื่อนไขการหยุดการสร้างโหนดลูกไม่เป็นจริง ให้ทำขั้นตอนต่อไป

2) สร้างเงื่อนไขการตัดสินใจที่เป็นไปได้ทั้งหมดของโหนดปัจจุบัน

3) ในแต่ละเงื่อนไขตัดสินใจ คำนวณค่ามาตราวัดที่ไว้ใช้สำหรับเลือกเงื่อนไข โดยเลือกเงื่อนไขที่ให้ค่ามาตราวัดที่ดีที่สุด

4) แยกชุดข้อมูลเรียนรู้ของโหนดปัจจุบันออกเป็นส่วน ๆ ตามเงื่อนไขที่ได้เลือกไว้ในขั้นที่ 3 แล้วทำการสร้างโหนดลูกสำหรับชุดข้อมูลแต่ละส่วน หลังจากนั้นกลับไปทำขั้นตอน 1 โดยจะกำหนดให้แต่ละโหนดลูกเป็นโหนดปัจจุบัน

2.1.3 เงื่อนไขการหยุดการสร้างโหนดลูก คือ จะหยุดการสร้างโหนดเมื่อเกิดกรณีดังนี้

1) เมื่อโหนดปัจจุบันประกอบไปด้วยข้อมูลที่มีประเภทเดียวกันเป็นส่วนใหญ่

2) เมื่อจำนวนข้อมูลในโหนดปัจจุบันน้อยกว่าค่าที่กำหนด การที่มีข้อมูลน้อยเกินไปที่โหนดปัจจุบันจะทำให้การตัดสินประเภทของข้อมูลที่โหนดใบที่จะสร้างขึ้นต่อจากนี้ขึ้นอยู่กับข้อมูลเพียงไม่กี่จำนวน ซึ่งอาจกล่าวได้ว่า เป็นข้อมูลรบกวน (noise data)

3) เมื่อข้อมูลทั้งหมดในโหนดปัจจุบันมีค่าตัวแปรที่เหลือสำหรับการตัดสินใจเหมือนกันทั้งหมด ซึ่งสุดท้ายแล้วข้อมูลเหล่านี้ไม่มีทางที่จะสามารถแบ่งไปยังคนละโหนดได้อีก ดังนั้นจึงไม่จำเป็นต้องสร้างโหนดลูกอีกแล้ว

4) อัลกอริทึมที่ใช้ในการสร้างตัวแบบต้นไม้ตัดสินใจนั้นจะใช้อัลกอริทึม C4.5(J48) ซึ่งเป็นอัลกอริทึมในกลุ่มของการจำแนกที่พัฒนามาจาก ID3 ที่เป็นอัลกอริทึมพื้นฐานที่ใช้ในการสร้างต้นไม้ตัดสินใจ โดยใช้หลักการของทฤษฎีสารสนเทศ (information theory) โดยจะเลือกเงื่อนไขในการตัดสินใจที่ทำให้ค่าเกรนเอนโทรปี (Entropy Gain) มีค่ามากที่สุดแต่ใน C4.5 นั้น จะใช้ค่า อัตราส่วนเกรน (Gain Ratio) มาใช้เปรียบเทียบแทน [9] ดังนี้

$$\text{Gain Ratio} = \text{Entropy gain}/\text{SpiltInfo} \quad (3)$$

โดยที่

$$\text{Entropy Gain} = \text{Entropy(โหนดแม่หรือโหนดปัจจุบัน)} - \sum_{i=1}^{m} \frac{n_i}{N} \text{Entropy}(N_i)$$

$$\text{SplitInfo} = -\sum_{j}^{m} p(j)\log(p(j))$$

ซึ่ง P(j) คือ สัดส่วนของข้อมูลประเภท j

N_i คือ โหนดลูก i ของโหนดปัจจุบัน

m คือ จำนวนโหนดลูก

n_i คือ ขนาดของชุดข้อมูลที่แบ่งไปยังโหนดลูกที่ i

N คือ ขนาดของชุดข้อมูลของโหนดปัจจุบัน ซึ่งเท่ากับ $\sum_{i=1}^{m} n_i$

ค่าสารสนเทศการแบ่งกลุ่ม (SplitInfo) จะทำหน้าที่เป็นปัจจัยในการทำให้เป็นปรกติ (normalization factor) ซึ่งจะมีค่ามากขึ้นตามจำนวนโหนดลูก ดังนั้นค่าอัตราส่วนจีนี (Gini Ratio) ของเงื่อนไขที่มีจำนวนโหนดลูกมาก จะถูกปรับให้มีค่าลดลงด้วยค่าสารสนเทศการแบ่งกลุ่ม [9] และในอัลกอริทึม C4.5 ยังสามารถหลีกเลี่ยงการสร้างโครงสร้างต้นไม้ที่ใหญ่เกินไปเนื่องจากมีข้อมูลจำนวนมาก มีการตัดทอนความผิดพลาดออกไปจึงทำให้ความผิดพลาดลดลง และยังสามารถใช้กับข้อมูลที่มีความต่อเนื่อง (continuous attributes) ได้อีกด้วย

2.2 วิธีโครงข่ายประสาทเทียม

โครงข่ายที่ใช้ในการจำแนกประเภทของข้อมูลจะเป็นแบบโครงข่ายป้อนไปหน้า (feed forward network) [9] ที่มีกระบวนการฝึกฝนแบบมีผู้สอน (supervise) ซึ่งประกอบด้วยชั้นของเซลล์ประสาท (neurons) หลายชั้น เซลล์ประสาทชั้นหนึ่งจะรับค่าของข้อมูลจากเซลล์ประสาทที่อยู่ในชั้นก่อนหน้า แล้วทำการคำนวณด้วยฟังก์ชันแบบไม่เชิงเส้น และให้ผลลัพธ์ที่จะส่งออกไปยังเซลล์ประสาทถัดไป โครงข่ายใยประสาทจะต้องมีอย่างน้อย 2 ชั้น คือชั้นรับข้อมูล ซึ่งประกอบด้วยเซลล์ประสาท ทำหน้าที่รับข้อมูลและส่งต่อไปยังเซลล์ประสาทชั้นถัดไป และชั้นส่งผลลัพธ์ ซึ่งจะประกอบด้วยเซลล์ประสาทที่รับผลลัพธ์จากเซลล์ประสาทชั้นก่อนหน้า แล้วทำการคำนวณแล้วส่งผลลัพธ์ออกไปเพื่อเป็นผลลัพธ์ (output) สุดท้ายของการคำนวณทั้งหมดของเครือข่าย สำหรับชั้นของเครือข่ายที่อยู่ตรงกลางของทั้งสองชั้นนั้นจะเรียกว่า ชั้นปกปิด (hidden layer) ซึ่งอาจมีหรือไม่มีก็ได้และจะมีกี่ชั้นก็ได้ จะเป็นชั้นที่ใช้เพิ่มประสิทธิภาพในการจำแนกประเภทข้อมูล การเชื่อมโยงเซลล์ประสาทระหว่างชั้นนั้น จะมีลักษณะที่เป็นการเชื่อมโยงแบบสมบูรณ์ (fully connected) นั่นคือ เซลล์ประสาทแต่ละตัวจะส่งผลลัพธ์จากการคำนวณของตัวเองไปยังเซลล์ประสาททุกตัวที่อยู่ในชั้นถัดไป ซึ่งเส้นเชื่อมโยงต่าง ๆ ระหว่างเซลล์ประสาทนั้นจะประกอบด้วยค่าน้ำหนักคูณกับค่าของผลลัพธ์ที่ถูกส่งผ่านมาจากเส้นเชื่อมโยงของมัน แล้วค่าผลคูณนี้จะถูกส่งเข้าสู่เซลล์ประสาทในชั้นถัดไป และโครงข่ายประสาทเทียมแบบป้อนไปหน้าจะมีการใช้อัลกอริทึมแบบแพร่กลับ (back-propagation) เพื่อใช้ในการปรับปรุงน้ำหนักคะแนนของเครือข่าย (network weight) โดยหลังจากใส่รูปแบบข้อมูลสำหรับฝึกให้แก่โครงข่ายในแต่ละครั้งแล้ว ค่าที่ได้รับ (output) จากโครงข่ายจะถูกนำไปเปรียบเทียบกับผลที่คาดหวังแล้วคำนวณหาความผิดพลาดแล้วค่าความผิดพลาดนี้จะถูกส่งกลับเข้าสู่โครงข่ายเพื่อใช้แก้ไขค่าน้ำหนักคะแนนของเครือข่าย

การประเมินตัวแบบการจำแนกประเภทข้อมูล

ใช้ค่าความแม่น (accuracy) เป็นเครื่องมือหรือตัววัดในการประเมินประสิทธิภาพของแต่ละตัวแบบซึ่งคำนวณได้จากเมทริกซ์ความสับสน (confusion matrix) ที่เป็นตารางที่แสดงจำนวนของข้อมูลในแต่ละประเภทที่ถูกจำแนกด้วยตัวแบบ สำหรับการจำแนกข้อมูล 3 ประเภท แสดงดังตารางที่ 1 ดังนี้

ตารางที่ 1 เมทริกซ์ความสับสน (confusion matrix)

		Actual Values		
		A	B	C
Predicted Values	A	TA	FBA	FCA
	B	FAB	TB	FCB
	C	FAC	FBC	TC

TA เป็นจำนวนข้อมูลที่เป็นประเภท A และถูกจำแนกประเภทว่าเป็น A
TB เป็นจำนวนข้อมูลที่เป็นประเภท B และถูกจำแนกประเภทว่าเป็น B
TC เป็นจำนวนข้อมูลที่เป็นประเภท C และถูกจำแนกประเภทว่าเป็น C
FBA เป็นจำนวนข้อมูลที่เป็นประเภท B แต่ถูกจำแนกว่าเป็น A
FCA เป็นจำนวนข้อมูลที่เป็นประเภท C แต่ถูกจำแนกว่าเป็น A
FAB เป็นจำนวนข้อมูลที่เป็นประเภท A แต่ถูกจำแนกว่าเป็น B
FCB เป็นจำนวนข้อมูลที่เป็นประเภท C แต่ถูกจำแนกว่าเป็น B
FAC เป็นจำนวนข้อมูลที่เป็นประเภท A แต่ถูกจำแนกว่าเป็น C
FBC เป็นจำนวนข้อมูลที่เป็นประเภท B แต่ถูกจำแนกว่าเป็น C

และค่าความแม่นจะพิจารณาจากสัดส่วนระหว่างจำนวนข้อมูลทั้งหมดที่จำแนกประเภทถูกต้องกับจำนวนข้อมูลทั้งหมดที่ถูกจำแนกประเภท ดังนี้

$$\text{Accuracy} = \frac{T}{T + FBA + FCA + FAB + FCB + FAC + FBC} \times 100\% \quad (4)$$

โดย T=TA+TB+TC

ผลการวิจัย

1. การวิเคราะห์การถดถอยลอจิสติกพหุ

1.1 ตรวจสอบข้อสมมติเบื้องต้นของการวิเคราะห์การถดถอยลอจิสติกพหุ

1) ค่าเฉลี่ยของความคลาดเคลื่อนสุ่มเป็นศูนย์ โดยจากสถิติทดสอบ Z พบว่า ค่าพี (*p*-value) มีค่ามากกว่า 0.05 จึงยอมรับสมมติฐานหลัก นั่นคือ ค่าเฉลี่ยของความคลาดเคลื่อนสุ่มมีค่าเท่ากับศูนย์ที่ระดับนัยสำคัญ 0.05

2) ตัวแปรอิสระต้องไม่มีความสัมพันธ์กัน ซึ่งจะพิจารณาจากค่าความคลาดเคลื่อนยินยอมและค่าองค์ประกอบการขยายความแปรปรวนซึ่งผลการวิเคราะห์แสดงดังตารางที่ 2

ตารางที่ 2 ค่าความคลาดเคลื่อนยินยอมและค่าองค์ประกอบการขยายความแปรปรวนของตัวแปรอิสระ

ตัวแปรอิสระ	ค่าความคลาดเคลื่อนยินยอม	ค่าองค์ประกอบการขยายความแปรปรวน
x_1 (gender)	0.970	1.030
x_2 (age)	0.840	1.191
x_3 (monthly income)	0.863	1.159
x_4 (education)	0.821	1.218
x_5 (marital)	0.887	1.127
x_6 (occupation)	0.914	1.094

พบว่าค่าความคลาดเคลื่อนยินยอมของตัวแปรเพศ (x_1), อายุ (x_2), รายได้ต่อเดือน (x_3), วุฒิการศึกษา (x_4), สถานภาพครอบครัว (x_5) และอาชีพ (x_6) มีค่าเข้าใกล้ 1 นอกจากนี้ค่า VIF ของตัวแปรเพศ (x_1), อายุ (x_2), รายได้ต่อเดือน (x_3), วุฒิการศึกษา (x_4), สถานภาพครอบครัว (x_5) และอาชีพ (x_6) ไม่ได้มีค่าเข้าใกล้ 10 จึงสรุปได้ว่า ตัวแปรอิสระทั้ง 6 ตัว ไม่มีความสัมพันธ์กัน

3) การวิเคราะห์การถดถอยลอจิสติกจะต้องใช้ขนาดตัวอย่าง (n) มากกว่าการวิเคราะห์การถดถอยแบบปกติ โดยจะใช้ขนาดตัวอย่าง (n) ไม่น้อยกว่า 30p โดยที่ p คือ จำนวนตัวแปรอิสระ ในงานวิจัยนี้มีจำนวนตัวแปรอิสระ 6 ตัว และ n เท่ากับ 400 ดังนั้น 400 ไม่น้อยกว่า 30(6) จึงเป็นไปตามข้อตกลงเบื้องต้น

1.2 การเลือกตัวแปรอิสระ

1) การเลือกตัวแปรอิสระเข้าสู่สมการถดถอยลอจิสติกพหุจะใช้วิธีทางตรง (enter) และทำการทดสอบความสัมพันธ์ระหว่างชุดตัวแปรอิสระกับตัวแปรตาม ผลการวิเคราะห์พบว่า ถ้าสมการถดถอยลอจิสติกประกอบด้วยค่าคงที่เพียงอย่างเดียวนั้นจะให้ค่า -2LL เท่ากับ 820.356 และถ้าสมการถดถอยลอจิสติกประกอบด้วยค่าคงที่และชุดตัวแปรอิสระจะให้ค่า -2LL เท่ากับ 567.558 ดังนั้นสมการถดถอยลอจิสติกที่มีชุดตัวแปรอิสระรวมอยู่ด้วยจะมีความเหมาะสมมากกว่า นั่นคือ การมีตัวแปรอิสระอย่างน้อย 1 ตัว มีความสัมพันธ์กับตัวแปรตาม อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05 (p-value = 0.000 <0.05)

2) การทดสอบความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามของสมการถดถอยลอจิสติกพหุ พบว่า ตัวแปรอายุ (x_2), รายได้ต่อเดือน (x_3), วุฒิการศึกษา (x4), สถานภาพครอบครัว (x_5) และอาชีพ (x_6) มีค่าพี (p-value) น้อยกว่า 0.05 จึงปฏิเสธสมมติฐานหลัก นั่นคือ ตัวแปรอิสระ 5 ตัว นี้มีความสัมพันธ์กับการทำนายการตัดสินใจในการเลือกทำประเภทประกันของลูกค้าที่ระดับนัยสำคัญ 0.05

ตารางที่ 3 ค่าประมาณสัมประสิทธิ์การถดถอยของสมการลอจิสติกพหุ

Variable	β	Wald	*p*-value
whole life insurance (y_1)			
x_2 (age)	0.090	20.782	0.000*
endowment insurance (y_2)			
constant	-18.587	135.644	0.000*
x_2 (age)	0.089	12.606	0.000*
x_3 (monthly income)	0.000	6.628	0.010*
x_4 (education) (reference: master degree or higher)			
elementary school	-3.760	10.753	0.001*
junior high school	-3.398	15.395	0.000*
senior high school	-1.669	5.954	0.015*
vocational school	-2.282	5.353	0.021*
bachelor degree	-1.021	0.000	0.999
x_5 (marital) (reference: divorced)			
married	14.618	1175.894	0.000*

reference group: health insurace

p-value < 0.05

และจากผลการวิเคราะห์ในตารางที่ 3 จะได้ว่าสมการถดถอยลอจิสติกพหุของการทำนายการตัดสินใจในการเลือกทำประเภทประกันของลูกค้าโดยมีประเภทประกันสุขภาพเป็นกลุ่มอ้างอิงของตัวแปรตาม คือ

$$\log\left(\frac{P(y=1)}{P(y=3)}\right) = 0.090x_2 \quad (5)$$

$$\log\left(\frac{P(y=2)}{P(y=3)}\right) = -18.587+0.089x_2-3.760(x_4=\text{ประถมศึกษา})-3.398(x_4=\text{มัธยมศึกษาตอนต้น})-1.669(x_4=\text{มัธยมศึกษาตอนปลาย}) -2.282(x_4=\text{ปวช.})+14.618(x_5=\text{สมรส}) \quad (6)$$

โดยที่ y=1 คือ กลุ่มของประกันชีวิตแบบตลอดชีพ y=2 คือ กลุ่มของประกันชีวิตแบบสะสมทรัพย์และ y=3 คือ กลุ่มของประกันสุขภาพ

1.3 จากการทดสอบประสิทธิภาพความถูกต้องในการจำแนกประเภทข้อมูลของสมการลอจิสติกพหุพบว่าตัวแบบให้ค่าความแม่น ร้อยละ 70.00

2. การวิเคราะห์จำแนกกลุ่ม

2.1 ตรวจสอบข้อสมมติเบื้องต้นของการวิเคราะห์จำแนกกลุ่ม

1) ตัวแปรอิสระต้องมีการแจกแจงปกติหลายตัวแปร พิจารณาจากข้อมูลสุดโต่งแบบหลายตัวแปร (multivariate outliers) โดยการวิเคราะห์จากค่าระยะทางมาฮาลาโนบิส (Mahalanobis distances) พบว่า ค่าระยะทางมาฮาลาโนบิสที่มากที่สุดเท่ากับ 12.228 มีค่าน้อยกว่าค่าสถิติทดสอบไคว์แควร์ที่องศาเสรี เท่ากับ 6 คือ 12.592 แสดงว่า มีความน่าจะเป็นสูงที่ตัวแปรอิสระจะมีการแจกแจงปรกติหลายตัวแปรที่ระดับนัยสำคัญ เท่ากับ 0.05

2) เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระทุกกลุ่มต้องเท่ากัน ตรวจสอบจากค่า Box's M พบว่า ค่าพี (*p*-value) เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่า 0.05 ดังนั้นความแปรปรวนร่วมของประชากรแต่ละกลุ่มไม่เท่ากันที่ระดับนัยสำคัญ 0.05 จึงไม่เป็นไปตามข้อสมมติเบื้องต้น ดังนั้นในโปรแกรม SPSS จะใช้คำสั่ง separated groups แทนเพื่อให้มีการแยกวิเคราะห์ความแปรปรวนร่วมของแต่ละกลุ่ม [5]

3) ตัวแปรอิสระและตัวแปรตามต้องมีความสัมพันธ์กันแบบเชิงเส้น โดยจะพิจารณาจากสัมประสิทธิ์สหสัมพันธ์พอยท์ไบซีเรียล (Point Biserial Correlation Coefficient) ซึ่งจากค่าพี (*p*-value) พบว่า ตัวแปรวุฒิการศึกษา (x_4) ไม่มีความสัมพันธ์เชิงเส้นกับตัวแปรประเภทประกัน ที่ระดับนัยสำคัญ 0.05 ดังนั้นในการวิเคราะห์จึงตัดตัวแปรวุฒิการศึกษา (x_4) ออกจากการวิเคราะห์

4) ตัวแปรอิสระต้องไม่มีความสัมพันธ์เชิงเส้นแบบพหุ โดยพิจารณาจากค่าความคลาดเคลื่อนยินยอมและค่าองค์ประกอบการขยายความแปรปรวนซึ่งจากผลการวิเคราะห์ในตารางที่ 4 พบว่า จากค่าความคลาดเคลื่อนยินยอมของตัวแปรอิสระทั้งหมดพบว่า มีค่าเข้าใกล้ 1 นอกจากนี้ค่าองค์ประกอบการขยายความแปรปรวนของตัวแปรไม่ได้มีค่าเข้าใกล้ 10 จึง สรุปได้ว่า ตัวแปรอิสระทั้ง 5 ตัว ไม่มีความสัมพันธ์เชิงเส้นแบบพหุ

2.2 การสร้างสมการจำแนกกลุ่ม

1) นำข้อมูลทั้งหมดมาใช้สร้างสมการจำแนกกลุ่มและทำการทดสอบความเท่ากันของค่าเฉลี่ยของตัวแปรอิสระแต่ละกลุ่ม คือ กลุ่มประกันชีวิตแบบตลอดชีพ กลุ่มประกันชีวิตแบบสะสมทรัพย์ และกลุ่มประกันสุขภาพ โดยพิจารณาจากสถิติทดสอบวิลก์สแลมบ์ดา (Wilks' Lambda) พบว่าสถิติการทดสอบค่าเฉลี่ยของตัวแปรอิสระต่าง ๆ ที่แบ่งกลุ่มตัวแปรตาม โดยใช้สถิติทดสอบวิลกส์แลมบ์ดาพบว่า มีค่าเฉลี่ยของตัวแปรเพศ อายุ อาชีพ สถานภาพครอบครัว รายได้ต่อเดือน ใน 3 กลุ่มแตกต่างกันที่ระดับนัยสำคัญ 0.05 (*p*-value < 0.05 แสดงว่า มีนัยสำคัญทางสถิติที่ระดับ 0.05)

2) การนำเสนอค่าไอเก้นและความสามารถในการอธิบายความผันแปรของข้อมูล

ตารางที่ 4 ค่าไอเก้นและความสามารถในการอธิบายความผันแปรของข้อมูลของการตรวจสอบสมการจำแนกกลุ่ม

Function	Eigenvalue	% of Variance	Canonical Correlation
1	0.388	88.9	0.529
2	0.048	11.1	0.215

จากผลการวิเคราะห์ในตารางที่ 4 พบว่า เนื่องจากมี 3 กลุ่ม จึงมีฟังก์ชันแบ่งกลุ่ม 2 ฟังก์ชัน คือ Function 1 และ Function 2 ซึ่งมีค่าไอเก้นเท่ากับ 0.388 และ 0.048 ตามลำดับ และมีค่าร้อยละของความแปรปรวน (% of Variance) เท่ากับ 88.9 และ 11.1 ตามลำดับ จึงควรใช้สมการที่ 1 (Function 1) ในการทำนายกลุ่มให้กับหน่วยต่าง ๆ เพราะสมการที่ 1 สามารถอธิบายความผันแปรของข้อมูลได้ร้อยละ 88.9 ในขณะที่สมการที่ 2 (Function 2) สามารถอธิบายได้เพียง ร้อยละ 11.1 เท่านั้น

3) การทดสอบความมีนัยสำคัญของสมการจำแนกกลุ่ม จากผลการวิเคราะห์ด้วยสถิติทดสอบวิลกส์แลมบ์ดาจะใช้ในการทดสอบสมมติฐานเกี่ยวกับค่ากลางทั้ง 3 กลุ่ม ว่าเท่ากันหรือไม่ โดยจากค่าวิลกส์แลมบ์ดาของ 1 through 2 ซึ่งหมายถึงค่าวิลกส์แลมบ์ดาของสมการ 1 และ 2 ซึ่งใช้ทดสอบสมมติฐานหลักที่ว่าค่ากลาง (centroid) ของทั้ง 2 สมการ มีค่าเท่ากันทั้ง 3 กลุ่ม พบว่าค่าพี (*p*-value) = 0.000 ซึ่งมีค่าน้อยกว่าระดับนัยสำคัญ 0.05 ดังนั้นจึงปฏิเสธสมมติฐานที่ว่าค่ากลางของทั้ง 2 สมการ เท่ากันทั้ง 3 กลุ่ม และจากในค่าวิลกส์แลมบ์ดาของสมการที่ 2 ที่มีค่าพี (*p*-value) = 0.000 จึงปฏิเสธสมมติฐานที่ว่าค่ากลางของสมการที่ 2 เท่ากันทั้ง 3 กลุ่ม แสดงว่าสมการจำแนกกลุ่มที่ได้สามารถจำแนกออกเป็น 3 กลุ่ม ได้อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05

4) ค่าประมาณสัมประสิทธิ์ของสมการจำแนกกลุ่มโดยแยกเป็นกลุ่มตามวิธีของ Fisher's

ตารางที่ 5 ค่าสัมประสิทธิ์ของสมการจำแนกกลุ่ม

	Insurance Types		
	Whole Life Insurance	Endowment Insurance	Health Insurance
constant	-26.302	-27.890	-23.912
x_1 (gender)	6.067	5.192	5.490
x_2 (age)	0.700	0.668	0.614
x_5 (marital)	6.862	9.671	8.351

จากผลการวิเคราะห์ในตารางที่ 5 จะได้จำนวนสมการจำแนกกลุ่มเท่ากับจำนวนกลุ่ม นั่นคือ 3 สมการ และสามารถเขียนสมการจำแนกกลุ่มได้ดังนี้

สมการจำแนกกลุ่มของประกันชีวิตแบบตลอดชีพ คือ

$$\hat{D}_1 = -26.302+6.067x_1+0.700x_2+6.862x_5 \quad (7)$$

สมการจำแนกกลุ่มของประกันชีวิตแบบสะสมทรัพย์ คือ

$$\hat{D}_2 = -27.890+5.192x_1+0.668x_2+9.671x_5 \quad (8)$$

สมการจำแนกกลุ่มของประกันสุขภาพ คือ

$$\hat{D}_3 = -23.912+5.490x_1+0.614x_2+8.351x_5 \quad (9)$$

2.3 จากการทดสอบประสิทธิภาพความถูกต้องในการจำแนกประเภทข้อมูลของสมการจำแนกกลุ่มพบว่าตัวแบบให้ค่าความแม่น ร้อยละ 61.80

3. วิธีการวิเคราะห์ต้นไม้ตัดสินใจ

นำข้อมูลทั้งหมดมาใช้สร้างตัวแบบต้นไม้ตัดสินใจด้วยอัลกอริทึม C4.5 (J48) ในโปรแกรม Weka ผลจากการ

ทดสอบประสิทธิภาพความถูกต้องในการจำแนกประเภทข้อมูลของตัวแบบพบว่าตัวแบบต้นไม้ตัดสินใจให้ค่าความแม่นร้อยละ 76.50

4. วิธีโครงข่ายประสาทเทียม

จากการนำข้อมูลทั้งหมดที่มาสร้างตัวแบบในโปรแกรม Weka โดยใช้อัลกอริทึมแบบเพอร์เซปตรอนหลายชั้น (multilayer perceptron) ซึ่งจากการทำการทดลองในหลาย ๆ ครั้ง พบว่าเมื่อค่าอัตราการเรียนรู้ (learning rate) มีค่าเป็น 0.1 และกำหนดให้ค่าโมเมนตัม (momentum) มีค่าเป็น 0.15 โดยมีจำนวนรอบการสอน (training time) เท่ากับ 50,000 รอบ โดยที่โครงข่ายประสาทเทียมจะมีชั้นปกปิด (hidden layer) 1 ชั้น และมีจำนวนเซลล์ประสาท (neurons) เท่ากับ 12 จะให้ตัวแบบโครงข่ายประสาทเทียมที่ดีที่สุด นั่นคือ ให้ค่าความแม่น ร้อยละ 85.75

5. การเปรียบเทียบประสิทธิภาพการทำนาย

จากเทคนิคการวิเคราะห์จำแนกกลุ่มทั้ง 4 วิธี ในตารางที่ 6 พบว่า วิธีโครงข่ายประสาทเทียมเป็นวิธีที่มีประสิทธิภาพมากที่สุดในการจำแนกกลุ่ม โดยให้ค่าความแม่น ร้อยละ 85.75 รองลงมาคือ วิธีการวิเคราะห์ต้นไม้ตัดสินใจ วิธีการวิเคราะห์การถดถอยลอจิสติกพหุ และการวิเคราะห์จำแนกกลุ่ม ตามลำดับ ซึ่งผลการวิเคราะห์แสดงดังตารางที่ 6

ตารางที่ 6 การเปรียบเทียบค่าความแม่น (accuracy) ของการทำนายสำหรับการวิเคราะห์จำแนกกลุ่ม 4 วิธี

เทคนิคการวิเคราะห์จำแนกกลุ่ม	ค่าความแม่น (accuracy) (ร้อยละ)
การวิเคราะห์การถดถอยลอจิสติกพหุ	70.00
การวิเคราะห์จำแนกกลุ่ม	61.80
ต้นไม้ตัดสินใจ	76.50
โครงข่ายประสาทเทียม	85.75

อภิปรายและสรุปผลการวิจัย

จากการเปรียบเทียบประสิทธิภาพในการทำนายการเลือกทำประเภทประกันภัยของลูกค้าของทั้ง 4 วิธี พบว่าโครงข่ายประสาทเทียมให้ค่าความแม่นในการทำนายการเลือกทำประเภทประกันภัยของลูกค้าได้ถูกต้อง ร้อยละ 85.75 รองลงมาคือ วิธีต้นไม้ตัดสินใจที่ให้ค่าความแม่นในการทำนายถูกต้อง ร้อยละ 76.50 และวิธีการวิเคราะห์การถดถอยลอจิสติกพหุ การวิเคราะห์จำแนกกลุ่มให้ค่าความแม่น ร้อยละ 70.00 และ 61.80 ตามลำดับ ดังนั้นวิธีโครงข่ายประสาทเทียมให้ประสิทธิภาพในทำนายการเลือกทำประเภทประกันภัยของลูกค้าได้มีประสิทธิภาพมากที่สุด สอดคล้องกับงานของ Ansari และ Riasi [10] ที่ได้สร้างตัวแบบทำนายความจงรักภักดีของลูกค้า (customer loyalty) ที่มีต่อบริษัทประกัน ซึ่งการทำนายให้ประสิทธิภาพดีที่สุดเมื่อใช้วิธีโครงข่ายประสาทเทียม หรืองานวิจัยของ Weerasinghe และ Wijegunasekara [11] ที่ทำการเปรียบเทียบวิธีการวิเคราะห์การถดถอยลอจิสติกพหุ วิธีโครงข่ายประสาทเทียม และวิธีต้นไม้สินใจในการสร้างตัวแบบทำนายค่าสินไหมทดแทนของประกันรถยนต์ ซึ่งพบว่า โครงข่ายประสาทเทียมให้ประสิทธิภาพในการทำนายดีที่สุด รองลงมาคือ วิธีต้นไม้ตัดสินใจและวิธีการวิเคราะห์การถดถอยลอจิสติกพหุ ตามลำดับ และจากงานวิจัยของ Huang และ Wang [12] ที่ได้สร้างตัวแบบทำนายการล้มละลายของบริษัทประกันในประเทศจีนด้วยวิธีโครงข่ายประสาทเทียม วิธีการถดถอยลอจิสติก และวิธีการวิเคราะห์จำแนกกลุ่ม พบว่าวิธีโครงข่ายประสาทเทียมให้ประสิทธิภาพดีที่สุด รวมถึงสอดคล้องกับงานวิจัยของ Baione และ Angelis [13] ที่ได้สร้างตัวแบบทำนายความมั่นคงของบริษัทประกันแล้วพบว่า วิธีต้นไม้ตัดสินใจให้ประสิทธิภาพในการทำนายดีกว่าวิธีการวิเคราะห์จำแนกกลุ่ม นอกจากนี้ยังสอดคล้องกับงานวิจัยของ Gulsun

และ Umit [14] ที่พบว่าวิธีการถดถอยลอจิสติกให้ประสิทธิภาพในการทำนายความเสี่ยงของการเกิดสภาวะล้มเหลวทางการเงินของบริษัทประกันสูงกว่าวิธีการวิเคราะห์จำแนกกลุ่ม

จากการวิเคราะห์การถดถอยลอจิสติกพหุ พบว่าปัจจัยที่ส่งผลต่อการตัดสินใจการเลือกทำประกันในผลิตภัณฑ์แต่ละประเภทของลูกค้าคือ อายุ รายได้ต่อเดือน วุฒิการศึกษา สถานภาพครอบครัว และอาชีพ และจากการวิเคราะห์จำแนกกลุ่มพบว่า ปัจจัยที่ส่งผลต่อการตัดสินใจ คือ เพศ อายุ และสถานภาพครอบครัว เมื่อจากการวิเคราะห์การถดถอยลอจิสติกพหุ ปัจจัยที่ส่งผลต่อการเลือกทำประเภทประกันภัยของลูกค้าคือ อายุ รายได้ต่อเดือน วุฒิการศึกษา สถานภาพครอบครัว และอาชีพ ซึ่งสอดคล้องกับงานวิจัยของนพินดา [1] ที่พบว่าอาชีพและรายได้ต่อเดือนมีผลต่อการเลือกทำประเภทประกันภัยของลูกค้า และสอดคล้องกับงานวิจัยของ พัสวี [3] ที่ว่า อายุและวุฒิการศึกษานั้นส่งผลต่อการเลือกทำประเภทประกันภัยของลูกค้าเช่นกัน และในงานวิจัยนี้ยังได้ศึกษาปัจจัยสถานภาพครอบครัวเพิ่มเติม ซึ่งพบว่าจากทั้ง 2 วิธี ได้แก่ วิธีการวิเคราะห์การถดถอยลอจิสติกพหุและวิธีการวิเคราะห์จำแนกกลุ่มนั้น สถานภาพทางครอบครัวส่งผลต่อการเลือกทำประเภทประกันภัยของลูกค้า

ข้อเสนอแนะ

จากการวิจัยครั้งนี้พบว่า ในการวิเคราะห์ตัวแปรหลายตัวแปร (multivariate analysis) นั้น วิธีที่ควรนำไปใช้ในการจำแนกกลุ่มให้ได้ประสิทธิภาพดีที่สุดคือ วิธีโครงข่ายประสาทเทียม แต่ข้อเสียของวิธีนี้คือ ไม่มีตัวแบบทางสถิติให้นำไปใช้ ดังนั้นผู้วิจัยจึงเสนอให้นำวิธีวิเคราะห์การถดถอยลอจิสติกพหุที่ให้ตัวแบบทางสถิติไปใช้ในการจำแนกกลุ่มแทน เพราะถึงจะให้ค่าความแม่นต่ำกว่าวิธีโครงข่ายประสาทเทียม แต่เมื่อเทียบกับวิธีการวิเคราะห์จำแนกกลุ่มที่ให้ตัวแบบทางสถิติเช่นเดียวกันนั้น พบว่ามีประสิทธิภาพดีกว่าวิธีการวิเคราะห์จำแนกกลุ่ม ซึ่งจากงานวิจัยบริษัทประกันต่าง ๆ สามารถนำผลการวิจัยนี้ไปใช้ในการเพิ่มยอดขายให้แก่บริษัท โดยใช้เป็นแนวทางว่าจะเลือกเสนอขายผลิตภัณฑ์ประกันประเภทไหนให้แก่ลูกค้าแต่ละคนที่มีลักษณะต่างกันออกไป หรือนำไปออกแบบผลิตภัณฑ์ประกันให้ครอบคลุมกับความต้องการของลูกค้ามากขึ้น เช่น จากงานวิจัยพบว่า เมื่อใช้การเลือกทำประกันสุขภาพเป็นกลุ่มอ้างอิง ลูกค้าที่มีสถานภาพสมรสแล้วนั้นจะเลือกซื้อประกันออมทรัพย์มากกว่าประกันสุขภาพ และลูกค้าที่มีวุฒิการศึกษาระดับมัธยมศึกษาตอนปลายจะเลือกซื้อประกันสุขภาพมากกว่าประกันออมทรัพย์ เป็นต้น ทั้งนี้ในการวิจัยครั้งต่อไปอาจทดลองนำวิธีการค้นหาเพื่อนบ้านใกล้สุด k อันดับ (K-Nearest Neighbors) มาใช้ในการจำแนกกลุ่มได้เช่นกัน

เอกสารอ้างอิง

1. นพินดา หาญจริง. ปัจจัยที่มีผลต่อการตัดสินใจซื้อกรมธรรม์ประกันชีวิตของผู้ที่อยู่ในวัยทำงานในเขตกรุงเทพมหานคร. วิทยานิพนธ์ปริญญาเศรษฐศาสตรมหาบัณฑิต สาขาเศรษฐศาสตร์, บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์. กรุงเทพฯ; 2549.
2. อุสมาน ฮะบีบุรราห์มาน, นิวัตน์ สวัสดิ์แก้ว, พรชัย ลิขิตธรรมโรจน์, ศรัญลักษณ์ เทพวารินทร์, จิตกรี บุญโชติ. ปัจจัยที่มีผลต่อการตัดสินใจในการเลือกทำประกันชีวิตกับบริษัทกรุงไทย แอกซา จำกัด (มหาชน) ในเขตเทศบาลนครหาดใหญ่ จังหวัดสงขลา. ใน: เอกสารประกอบการประชุมหาดใหญ่วิชาการ ครั้งที่ 4 วันที่ 10 พฤษภาคม 2556. มหาวิทยาลัยหาดใหญ่. สงขลา; 2556. หน้า 137-47.
3. พัสวี ไข่มุกข์. ปัจจัยที่มีอิทธิพลต่อการเลือกซื้อประกันชีวิตของข้าราชการครูในสังกัดสำนักงานเขตยานนาวา กรุงเทพมหานคร. วารสารมนุษยศาสตร์และสังคมศาสตร์ มรพ. 2553;2(1):135-47.
4. กัลยา วานิชย์บัญชา. การวิเคราะห์ข้อมูลหลายตัวแปร. พิมพ์ครั้งที่ 4. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย; 2554.

5. กัลยา วานิชย์บัญชา. การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. พิมพ์ครั้งที่ 7. กรุงเทพฯ: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย; 2552.
6. ยุทธ ไกยวรรณ์. หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย. วารสารวิจัย มทร. ศรีวิชัย 2555:4(1):1-12.
7. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 7th ed. New Jersey: Pearson Prentice Hall; 2010.
8. เอกรัฐ บุญเชียง. การแบ่งกลุ่มและการแบ่งประเภทข้อมูล. เชียงใหม่: สยามพิมพ์นานา; 2561.
9. สุรพงค์ เอื้อวัฒนามงคล. การทำเหมืองข้อมูล. กรุงเทพฯ: บางกอกบล็อก; 2557.
10. Ansari A, Riasi A. Modeling and evaluating customer loyalty using neural networks: evidence from startup insurance companies. Future Bus J 2016;2(1):15-30.
11. Weerasinghe KPMLP, Wijegunasekara MC. A comparative study of data mining algorithms in the prediction of auto insurance claims. Eur Int J Sci Technol 2016;5(1):47-54.
12. Huang Y, Wang H. A comparative analysis of solvency prediction models based on China`s property insurance market. In: proceedings of 2012 Annual Conference of Asia-Pacific Risk and Insurance Association, August 6-8, 2012; Seoul, Korea; 2012. p. 1-23.
13. Baione F, Angelis PD. A review on statistical and probabilistic models for the control of insurance companies. Invest Manag Financial Innov 2006;3(4):65-78.
14. Gulsun S, Umit G. Early warning model with statistical analysis procedures in Turkish insurance companies. Afr J Bus Manage 2010;3(12):1-10.