รายงานวิจัยฉบับสมบูรณ์

การเปรียบเทียบตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตในประเทศไทย

Model Comparison for Forecasting the Number of Life Insurance Polices in Thailand

โดย

จันทร์จิรา พิลาแดง

สาขาวิชาสถิติ ภาควิชาคณิตศาสตร์

โครงการวิจัยประเภทงบประมาณเงินรายได้ คณะวิทยาศาสตร์

ประจำปีงบประมาณ พ.ศ. ๒๕๖๗

มหาวิทยาลัยบูรพา

สัญญาเลขที่ SC–NO1/2567

รายงานวิจัยฉบับสมบูรณ์

การเปรียบเทียบตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตในประเทศไทย

Model Comparison for Forecasting the Number of Life Insurance Polices in Thailand

โดย

จันทร์จิรา พิลาแดง

สาขาวิชาสถิติ ภาควิชาคณิตศาสตร์

กิตติกรรมประกาศ

โครงการวิจัยนี้ ได้รับการสนับสนุนทุนอุดหนุนการวิจัยจากคณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา สำหรับนักวิจัยรุ่นใหม่คณะวิทยาศาสตร์ ประจำปีงบประมาณ พ.ศ. 2567 ผู้วิจัยขอขอบพระคุณคณะวิทยาศาสตร์ที่เล็งเห็นความสำคัญของการวิจัยทางด้านวิทยาศาสตร์ และให้ทุนการวิจัยจนดำเนินการสำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณสำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (คปภ.) ธนาคารแห่งประเทศไทย (ธปท.) และสำนักงานนโยบายและยุทธศาสตร์การค้า (สนค.) เป็นอย่างยิ่ง ที่ให้การสนับสนุนและเอื้อเฟื้อข้อมูลที่มีคุณค่าอันเป็นประโยชน์ต่อการดำเนินงานวิจัย จนสำเร็จลุล่วงตามวัตถุประสงค์ที่ตั้งไว้

สุดท้ายนี้ ขอขอบคุณทุกท่านที่มีส่วนเกี่ยวข้องและให้การสนับสนุนทั้งทางตรงและทางอ้อม ซึ่งมีส่วนสำคัญยิ่งต่อความสำเร็จของโครงการวิจัยนี้

จันทร์จิรา พิลาแดง

บทคัดย่อ

ชื่อโครงการ การเปรียบเทียบตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตในประเทศไทย

ชื่อผู้วิจัย จันทร์จิรา พิลาแดง

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของตัวแบบทางสถิติและตัวแบบการเรียนรู้ของเครื่องสำหรับการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี พร้อมทั้งศึกษาตัวแปรทำนายที่มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี โดยใช้ข้อมูลรายเดือนตั้งแต่เดือนมกราคม พ.ศ. 2555 ถึงตุลาคม พ.ศ. 2567 รวมระยะเวลา 154 เดือน ตัวแปรที่นำมาวิเคราะห์ประกอบด้วยตัวแปรตาม นั่นคือ จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี และตัวแปรทำนาย 5 ตัวแปร ได้แก่ จำนวนเงินเอาประกันชีวิตเบี้ยประกันภัยรับโดยตรง อัตราดอกเบี้ยนโยบาย สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP และดัชนีราคาผู้บริโภคทั่วไป ตัวแบบที่นำมาใช้ในการศึกษาและเปรียบเทียบมีทั้งหมด 5 ตัวแบบ ได้แก่ ตัวแบบการถดถอยเชิงเส้นพหุคูณ ตัวแบบการถดถอยเกาส์เซียนผกผัน ตัวแบบการถดถอยต้นไม้ตัดสินใจ ตัวแบบการถดถอยป่าสุ่ม และตัวแบบการถดถอยแบบเอ็กซ์จีบูสต์ ทั้งในรูปแบบ Tree Booster และ Linear Booster โดยใช้ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAPE) เป็นเกณฑ์ในการประเมินประสิทธิภาพและความแม่นยำของตัวแบบ

ผลการวิเคราะห์ชี้ให้เห็นว่า ตัวแบบการถดถอยเอ็กซ์จีบูสต์ ในรูปแบบ Linear Booster ให้ผลการพยากรณ์ที่ดีที่สุด ทั้งในด้านความแม่นยำและความสอดคล้องกับลักษณะข้อมูล ในขณะที่ตัวแบบการถดถอยเชิงเส้นพหุคูณให้ผลใกล้เคียงกัน แต่มีข้อจำกัดด้านทฤษฎี นั่นคือ ปัญหาสหสัมพันธ์ในตัวของส่วนเหลือ ซึ่งอาจกระทบต่อความน่าเชื่อถือของตัวแบบ ส่วนตัวแบบการถดถอยเกาส์เซียนผกผันแม้จะสามารถพยากรณ์ข้อมูลชุดใหม่ได้ในระดับหนึ่ง แต่ข้อมูลจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีไม่ได้มีลักษณะสอดคล้องกับการแจกแจงแบบเกาส์เซียนผกผันอย่างแท้จริง ในทางกลับกัน ตัวแบบการถดถอยต้นไม้ตัดสินใจและตัวแบบการถดถอยป่าสุ่มมีประสิทธิภาพต่ำ โดยมีค่า RMSE และ MAPE สูงมากในข้อมูลชุดทดสอบ ซึ่งแสดงถึงปัญหา Overfitting ที่ตัวแบบสามารถทำนายข้อมูลชุดฝึกสอนได้แม่นยำ แต่ทำนายได้ไม่แม่นยำเมื่อเจอกับข้อมูลที่ไม่เคยเห็นมาก่อน ก่อให้เกิดการพยากรณ์ที่ผิดพลาด

นอกจากนี้ การศึกษาความสำคัญของตัวแปรทำนายจากตัวแบบที่เหมาะสม พบว่า สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP มีความสำคัญมากที่สุดและมีอิทธิพลในทิศทางตรงกันข้ามกับจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ซึ่งหมายความว่า ยิ่งสัดส่วนเงินให้กู้ยืมของภาคครัวเรือนต่อ GDP สูงขึ้น จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีก็จะลดลง รองลงมาคือ เบี้ยประกันภัยรับโดยตรงและอัตราดอกเบี้ยนโยบาย ซึ่งมีอิทธิพลในทิศทางเดียวกันกับจำนวนกรมธรรม์ ในขณะที่ดัชนีราคาผู้บริโภคและจำนวนเงินเอาประกันมีอิทธิพลหรือมีความสำคัญน้อยกว่า จากผลการศึกษานี้ สามารถสรุปได้ว่า ปัจจัยทางเศรษฐกิจโดยเฉพาะภาระหนี้ครัวเรือน มีบทบาทสำคัญต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ซึ่งสะท้อนถึงข้อจำกัดด้านการใช้จ่ายและการตัดสินใจซื้อประกันชีวิตของครัวเรือน อันสามารถใช้เป็นแนวทางในการวางนโยบายและออกแบบผลิตภัณฑ์ประกันชีวิตให้สอดคล้องกับสภาพเศรษฐกิจและตอบสนองความต้องการของผู้บริโภคในปัจจุบัน

คำสำคัญ: กรมธรรม์ประกันชีวิต ตัวแบบทางสถิติ ตัวแบบการเรียนรู้ของเครื่อง ปัจจัยทางเศรษฐกิจ เอ็กซ์จีบูสต์

Abstract

Project Title Model Comparison for Forecasting the Number of Life Insurance Polices in Thailand

Investigators Janjira Piladaeng

This study aims to develop and compare the performance of statistical and machine learning models for forecasting the number of new life insurance policies issued during the year, as well as to examine the predictors influencing this number. The analysis uses monthly data from January 2012 to October 2024, covering 154 months. The dependent variable is the number of new life insurance policies issued during the year, while the five predictors are the sum assured, direct premium written, policy interest rate, household debt-to-GDP ratio, and the general consumer price index. The study evaluates five forecasting models, namely multiple linear regression, inverse Gaussian regression, decision tree regression, random forest regression, and extreme gradient boosting (XGBoost) in both tree booster and linear booster forms. The model performance and accuracy are assessed using root mean square error (RMSE) and mean absolute percentage error (MAPE).

The analysis results indicate that the XGBoost regression model with a linear booster provides the best forecasting performance in terms of accuracy and consistency with the data characteristics, whereas the multiple linear regression model shows similar performance. However, the multiple linear regression model has a theoretical limitation, specifically the autocorrelation in its residuals, which may affect its reliability. Although the inverse Gaussian regression model performs reasonably well on new data, the number of new life insurance policies issued during the year does not fully conform to the inverse Gaussian distribution. In contrast, the decision tree regression and random forest regression models perform poorly, with very high RMSE and MAPE values on the test set. This suggests overfitting, where the models can accurately predict the training data but produce inaccurate predictions on new data, resulting in unreliable forecasts.

Furthermore, the analysis of predictor importance from the optimal model reveals that the household debt-to-GDP ratio is the most influential factor, negatively associated with the number of new life insurance policies issued during the year. In other words, a higher household debt-to-GDP ratio corresponds to a lower number of new life insurance policies. The next most influential factors are direct premium written and the policy interest rate, both positively associated with the number of policies. Meanwhile, the consumer price index and sum assured have relatively smaller effects. From this study, it can be concluded that economic factors, particularly household debt, play a significant role in the number of new life insurance policies issued during the year, reflecting household spending constraints and decisions regarding life insurance purchases. These findings can serve as a guideline for policy formulation and product design, helping align offerings with current economic conditions and consumer needs.

Keywords: Life Insurance Policy, Statistical Models, Machine Learning Models, Economic Factor, XGBoost

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

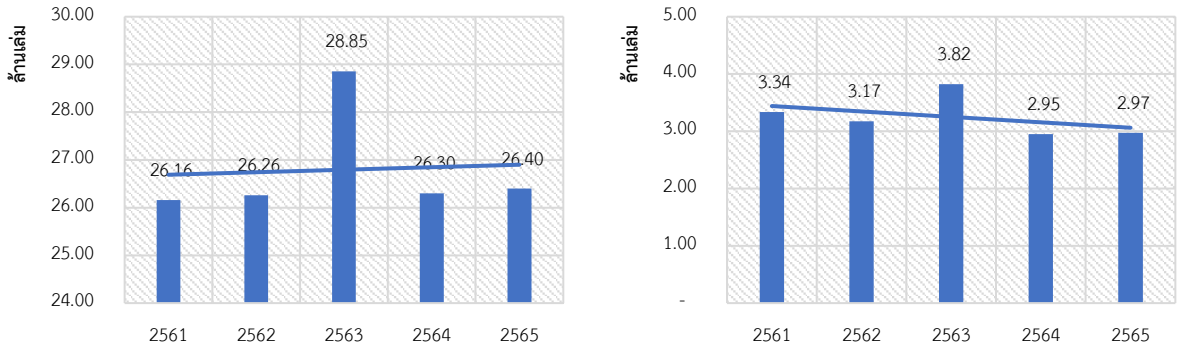
การประกันชีวิต คือ การทำประกันที่ผู้เอาประกันภัยจ่ายเงินให้กับบริษัทประกันชีวิต ตามการทรงชีพหรือมรณกรรมของตนเอง ซึ่งเป็นการชดเชยรายได้อันเนื่องมาจากความเสี่ยงที่เกี่ยวข้องจากการเสียชีวิต สูญเสียอวัยวะ ทุพพลภาพ และสูญเสียรายได้หลังวัยเกษียณอายุ ซึ่งมีลักษณะความเสี่ยงภัยที่คล้ายคลึงกัน เพื่อช่วยลดความเดือดร้อนในกรณีที่เกิดความเสียหายและทำให้ผู้เอาประกันภัยมีสภาพการเงินกลับคืนหรือใกล้เคียงสภาพเดิม โดยบริษัทประกันชีวิตจะจ่ายเงินตามจำนวนที่ระบุไว้ให้แก่ผู้เอาประกัน หรือผู้รับประโยชน์ ตามที่กำหนดไว้ในกรมธรรม์ประกันชีวิต (สำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย, ม.ป.ป.; สมาคมประกันชีวิตไทย, 2550) นอกจากนี้ การประกันชีวิตยังมีส่วนที่เป็นการออมทรัพย์ ผู้เอาประกันภัยสามารถซื้อความคุ้มครองในรูปแบบสะสมทรัพย์ ซึ่งเป็นส่วนหนึ่งของการออมทรัพย์ และหากมีชีวิตจนครบกำหนดสัญญา จะได้รับผลตอบแทน ซึ่งเป็นส่วนของการคุ้มครองของการประกันชีวิตแบบสะสมทรัพย์แท้จริง (pure endowment) (ฐิติวดี ชัยวัฒน์, 2560)

ในปัจจุบัน การประกันชีวิตในประเทศไทยสามารถแบ่งได้เป็น 3 ประเภทหลัก ๆ ด้วยกัน คือ การประกันชีวิตประเภทสามัญ (Ordinary Life Insurance) การประกันชีวิตประเภทอุตสาหกรรม (Industrial Life Insurance) และการประกันชีวิตประเภทกลุ่ม (Group Life Insurance) (สมาคมประกันชีวิตไทย, 2550) โดยที่กรมธรรม์ประกันชีวิตนั้นมีการเปลี่ยนแปลงตลอดเวลาเพื่อให้ทันยุคสมัย และตอบสนองความต้องการของลูกค้า ในปัจจุบัน การพยากรณ์จำนวนกรมธรรม์ประกันชีวิตจึงเป็นปัจจัยสำคัญที่ช่วยในการวางแผนธุรกิจ โดยเฉพาะอย่างยิ่งในการจัดทำแผนการตลาด การเลือกกลยุทธ์ทางธุรกิจ และการปรับตัวตามแนวโน้มของตลาด ทำให้มีความสามารถในการแข่งขันได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

เพื่อศึกษาถึงความสัมพันธ์ระหว่างจำนวนกรมธรรม์ประกันชีวิตหรือตัวแปรตาม (Dependent Variable; *Y*) และตัวแปรทำนาย (Predicted Variable; *X*) ต่าง ๆ ที่เกี่ยวข้อง หนึ่งในเทคนิคทางสถิติที่มีความเหมาะสมสำหรับการทำนายและอธิบายความสัมพันธ์ดังกล่าว คือ การวิเคราะห์การถดถอย (Regression Analysis) โดยตัวแบบการถดถอย (Regression Models) ที่มีความสำคัญและเกี่ยวข้องกับธุรกิจประกันภัย (สำนักงานอัตราเบี้ยประกันวินาศภัย, 2559) ประกอบไปด้วยตัวแบบที่อยู่ในรูปของตัวแบบเชิงเส้น (Linear Model) รวมไปถึงตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Models) ที่มีการขยายจากตัวแบบเชิงเส้นไปสู่ตัวแบบที่การแจกแจงของตัวแปรตามที่ไม่จำกัดอยู่ที่การแจกแจงปรกติ เช่น การแจกแจงทวินาม การแจกแจงปัวซง การแจกแจงทวินามลบ หรือการแจกแจงอื่น ๆ ที่อยู่ในวงศ์เลขชี้กำลัง (Exponential Family) (สำนักงานราชบัณฑิตยสภา, 2558) นอกจากการวิเคราะห์การถดถอยที่กล่าวมา ในปัจจุบันยังมีการใช้เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) เพื่อช่วยในการศึกษาปัจจัยและสร้างตัวแบบการทำนาย เช่น การถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression) การถดถอยแบบป่าสุ่ม (Random Forest Regression) การถดถอย XGBoost (eXtreme Gradient Boosting Regression) เป็นต้น

จากข้อมูลสถิติธุรกิจประกันชีวิต ในช่วงเดือนมกราคม – มิถุนายน พ.ศ. 2566 มีจำนวนกรมธรรม์ที่มีผลบังคับใช้ จำนวน 26,464,783 ฉบับ และมีการสมัครทำประกันใหม่ระหว่างปี จำนวน 1,471,352 ฉบับ ซึ่งลดลงจากปี 2561 จำนวน 1,864,692 ฉบับ คิดเป็นร้อยละ 55.9 (สำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย, 2566) นอกจากนี้ หากพิจารณาการเติบโตของธุรกิจประกันชีวิตย้อนหลัง 5 ปี ภาพที่ 1(ก.) แสดงให้เห็นว่าจำนวนกรมธรรม์ที่มีผลบังคับใช้นั้นมีแนวโน้มที่จะเพิ่มสูงขึ้น ในขณะที่ในภาพที่ 1(ข.) จำนวนกรมธรรม์ที่ทำใหม่ระหว่างปี มีแนวโน้มที่จะลดลง ซึ่งสะท้อนให้เห็นถึงความสำคัญในการทำประกันชีวิตที่ลดลง

ด้วยเหตุนี้ ผู้วิจัยจึงสนใจที่จะศึกษา เพื่อศึกษาตัวแปรที่มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี และหาตัวแบบที่เหมาะสมสำหรับพยากรณ์จำนวนกรมธรรม์ที่ทำใหม่ระหว่างปี โดยการเปรียบเทียบตัวแบบทางสถิติ ที่ตัวแปรตามมีการแจกแจงแบบต่าง ๆ รวมถึงตัวแบบการเรียนรู้ของเครื่อง เพื่อเป็นแนวทางในการวางแผน และปรับปรุงการดำเนินงานของธุรกิจประกันชีวิตในอนาคต หลังจากสถานการณ์การแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (COVID-19)



(ก.) (ข.)

ภาพที่ 1-1 (ก.) จำนวนกรมธรรม์ที่มีผลบังคับใช้ และ (ข.) จำนวนกรมธรรม์ที่ทำใหม่ระหว่างปี

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อเปรียบเทียบและหาตัวแบบที่เหมาะสมสำหรับการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี

1.2.2 เพื่อศึกษาตัวแปรทำนายที่มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีจากตัวแบบที่เหมาะสม

1.2.3 เพื่อนำตัวแบบที่เหมาะสมไปประยุกต์ใช้ในการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

ตัวแบบการพยากรณ์ที่ได้จะเป็นเครื่องมือที่เหมาะสมในการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตในประเทศไทย ซึ่งจะช่วยให้ทราบถึงปัจจัยที่มีอิทธิพลต่อจำนวนกรมธรรม์ดังกล่าว ทำให้หน่วยงานหรือบุคคลที่เกี่ยวข้องสามารถวางแผนการตลาดที่เป็นไปตามทิศทางที่เป็นประโยชน์สำหรับธุรกิจประกันชีวิตได้

1.4 ขอบเขตของการวิจัย

1.4.1 ศึกษาเกี่ยวกับตัวแบบทางสถิติและเปรียบเทียบตัวแบบดังกล่าว เพื่อหาตัวแบบที่เหมาะสมที่สุดสำหรับการทำนายจำนวนกรมธรรม์ประกันชีวิตในประเทศไทย

1.4.2 ใช้ข้อมูลเกี่ยวกับธุรกิจประกันภัยจากสำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (คปภ.) และข้อมูลอื่น ๆ ที่เกี่ยวข้อง จากธนาคารแห่งประเทศไทย และสำนักงานนโยบายและยุทธศาสตร์การค้า

1.4.3 ใช้โปรแกรม R สำหรับปฏิบัติการ Windows ในการประมวลผล

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

ในส่วนของทฤษฎีที่เกี่ยวข้องจะกล่าวถึงตัวแบบเชิงเส้น ตัวแบบเชิงเส้นนัยทั่วไป ตัวแบบการถดถอยต้นไม้ตัดสินใจ ตัวแบบการถดถอยป่าสุ่ม ตัวแบบการถดถอย XGBoost และเกณฑ์ที่ใช้ในการประเมินประสิทธิภาพของตัวแบบ

2.1.1 ตัวแบบเชิงเส้น

วัตถุประสงค์ของตัวแบบเชิงเส้น (Linear Models; LM) และตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Models; GLM) คือ การศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Y) และตัวแปรทำนาย (X) สำหรับตัวแบบเชิงเส้นนั้น มีแนวคิดที่ว่าตัวแปรตามสามารถเขียนให้อยู่ในรูปของผลรวมของค่าเฉลี่ย (μ) กับความคลาดเคลื่อน (ε) นั่นคือ (สำนักงานอัตราเบี้ยประกันวินาศภัย, 2559)

$$Y = \mu + \varepsilon \tag{2-1}$$

โดยที่ พจน์ความคลาดเคลื่อน หรือ ε มีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนเท่ากับ σ^2 และค่าคาดหมายของ Y หรือ μ สามารถเขียนให้อยู่ในรูปของการรวมเชิงเส้นของตัวแปรทำนายได้ โดยตัวแบบในสมการที่ (2-1) อาจเขียนให้อยู่ในรูปแบบดังต่อไปนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{2-2}$$

โดยที่ X_i คือ ตัวแปรทำนายที่ i และ β_i คือ ค่าพารามิเตอร์ที่ไม่ทราบค่า เมื่อ i = 0, 1, 2, ... , p และ p คือ จำนวนตัวแปรทำนายในตัวแบบ

เพื่อให้เขียนได้ง่ายและดูเป็นระเบียบ สามารถเขียนตัวแบบเชิงเส้นในสมการที่ (2-2) ให้อยู่ในรูปของเมทริกซ์ได้ ดังนี้

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2-3}$$

โดยที่ $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ ทั้งนี้ ตัวแบบเชิงเส้นประกอบไปด้วยส่วนประกอบ 3 ส่วน ได้แก่ ส่วนประกอบเชิงสุ่ม ส่วนประกอบเชิงระบบ และฟังก์ชันเชื่อมโยง ซึ่งมีรายละเอียดดังนี้

2.1.1.1 ส่วนประกอบเชิงสุ่ม (Random Component) จะแสดงการแจกแจงของตัวแปรตาม โดยในที่นี้ค่าของตัวแปรตาม Y จะต้องเป็นอิสระต่อกัน และมีการแจกแจกปรกติ โดยที่ค่าเฉลี่ย μ_i ไม่จำเป็นต้องมีค่าเท่ากัน แต่จะต้องมีความแปรปรวนร่วมเดียวกัน (Common Variance) นั่นคือเท่ากับ σ^2

2.1.1.2 ส่วนประกอบเชิงระบบ (Systematic Component) จะแสดงตัวแปรทำนายในรูปแบบของผลรวมเชิงเส้นของพารามิเตอร์ (η) นั่นคือ $\eta = X\beta$ โดยที่ X คือ เมทริกซ์ของตัวแปรทำนาย และ β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอย ที่ประมาณด้วยวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares; OLS)

2.1.1.3 ฟังก์ชันเชื่อมโยง (Link Function) เป็นฟังก์ชันของค่าเฉลี่ยของตัวแปรตาม ซึ่งเป็นฟังก์ชันที่ใช้เชื่อมโยงระหว่างส่วนประกอบเชิงสุ่มและส่วนประกอบเชิงระบบเข้าด้วยกัน นั่นคือ $\eta = g(\mu)$ โดยที่ $g(\mu)$ คือ ฟังก์ชันเชื่อมโยง ซึ่งตัวแบบเชิงเส้นจะมีฟังก์ชันเชื่อมโยงเอกลักษณ์ (Identity) และมีค่าคาดหมายของตัวแปรตาม คือ $E(Y) \equiv \mu = g^{-1}(\eta) = \eta$

2.1.2 ตัวแบบเชิงเส้นนัยทั่วไป

เมื่อกล่าวถึงข้อมูลที่เกี่ยวกับธุรกิจประกันภัย สมมุติฐานของตัวแบบที่ข้อมูลต้องมีการแจกแจงปรกติอาจไม่สามารถใช้ได้เสมอไป ยกตัวอย่างเช่น ขนาดของการเรียกร้อง (Claim Sizes) ความถี่ของการเรียกร้อง (Claim Frequencies) และการเกิดเหตุการณ์การเรียกร้องบนกรมธรรม์เดียวกัน ซึ่งลักษณะของข้อมูลมักจะไม่มีการแจกแจงปรกติ (สำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย, 2566) ตัวแบบเชิงเส้นนัยทั่วไปจึงเป็นการศึกษาตัวแปรตามที่มีความสัมพันธ์กับตัวแปรทำนาย ที่ขยายจากตัวแบบเชิงเส้นไปสู่การแจกแจงของตัวแปรตามที่ไม่จำกัดอยู่ที่การแจกแจงปรกติ เช่น การแจกแจงทวินาม การแจกแจงปัวซง หรือการแจกแจงอื่น ๆ ที่อยู่ในวงศ์เลขชี้กำลัง โดยตัวแบบจะประกอบไปด้วย 3 ส่วน คือ (สำนักงานอัตราเบี้ยประกันวินาศภัย, 2559; สำนักงานราชบัณฑิตยสภา, 2558)

2.1.2.1 ส่วนประกอบเชิงสุ่ม (Random Component) จะแสดงการแจกแจงของตัวแปรตาม โดยฟังก์ชันความน่าจะเป็นที่อยู่ในวงศ์เลขชี้กำลังสามารถเขียนอยู่ในรูป

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{2-4}$$

เมื่อ $a(\cdot)$, $b(\cdot)$, และ $c(\cdot)$ แทน ฟังก์ชันต่าง ๆ โดยที่ θ คือ พารามิเตอร์บ่งตำแหน่ง (Location Parameter), ϕ คือ พารามิเตอร์การกระจาย (Dispersion Parameter) หรือพารามิเตอร์บ่งขนาด (Scale Parameter), และฟังก์ชัน $a(\phi)$ มีรูปแบบทั่วไปเป็น $a(\phi) = \phi\omega$ ซึ่ง ω เป็นค่าคงที่ที่ทราบค่า ทั้งนี้ สามารถกำหนดรูปแบบต่าง ๆ ของการแจกแจงในวงศ์เลขชี้กำลังโดยการระบุฟังก์ชัน $a(\cdot)$, $b(\cdot)$, และ $c(\cdot)$ ที่แตกต่างกันออกไป (Myers et.al., 2012; Faraway, 2016) โดยมีการแจกแจงในวงศ์เลขชี้กำลังที่ใช้ในศึกษา ได้แก่

2.1.2.1.1 การแจกแจงปรกติ (Normal Distribution) หรือการแจกแจงเกาส์เซียน (Gaussian Distribution)

กำหนดให้ y เป็นตัวแปรสุ่มที่มีการแจกแจงปรกติ ที่มีพารามิเตอร์ μ และ σ^2 สามารถเขียนฟังก์ชันความหนาแน่นความน่าจะเป็น (Probability Density Function; p.d.f.) ของตัวแปรสุ่ม y ได้ ดังนี้

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{y\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right] \\ &= \exp\left[\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right)\right] \end{aligned} \tag{2-5}$$

เมื่อ $-\infty<y<\infty$, $-\infty<\mu<\infty$, และ $\sigma^2>0$ ทั้งนี้ การแจกแจงปรกติสามารถเขียนให้อยู่ในวงศ์เลขชี้กำลังได้ โดยกำหนดให้ $\theta=\mu$ และ $\phi=\sigma^2$ จะได้ $a(\phi)=\phi$, $b(\theta)=\frac{\theta^2}{2}$, และ $c(y,\phi)=-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$ ในกรณีนี้ อนุพันธ์อันดับหนึ่งและอนุพันธ์อันดับสองของ $b(\theta)$ คือ $b'(\theta)=\theta$ และ $b''(\theta)=1$ ตามลำดับ ซึ่งนำไปสู่ผลลัพธ์ที่ว่า ค่าเฉลี่ยของการแจกแจง คือ $\mathrm{E}(Y)=b'(\theta)=\theta=\mu$ และความแปรปรวนของการแจกแจง คือ $\mathrm{Var}(Y)=b''(\theta)a(\phi)=(1)(\sigma^2)=\sigma^2$

2.1.2.1.2 การแจกแจงเกาส์เซียนผกผัน (Inverse Gaussian Distribution)

กำหนดให้ y เป็นตัวแปรสุ่มที่มีการแจกแจงเกาส์เซียนผกผัน ที่มีพารามิเตอร์ μ และ λ สามารถเขียนฟังก์ชันความหนาแน่นความน่าจะเป็นของตัวแปรสุ่ม y ได้ ดังนี้

$$\begin{aligned} f(y;\mu,\lambda) &= \sqrt{\frac{\lambda}{2\pi y^3}}\exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right] \\ &= \exp\left[-\frac{\lambda y}{2\mu^2}+\frac{\lambda}{\mu}-\frac{\lambda}{2y}+\frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)\right] \\ &= \exp\left[\frac{y\left(\frac{1}{2\mu^2}\right)-\sqrt{\frac{1}{\mu^2}}}{-\frac{1}{\lambda}}-\frac{\lambda}{2y}+\frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)\right] \end{aligned} \tag{2-6}$$

เมื่อ $y>0$, $\mu>0$, และ $\lambda>0$ ทั้งนี้ การแจกแจงเกาส์เซียนผกผันสามารถเขียนให้อยู่ในวงศ์เลขชี้กำลังได้ โดยกำหนดให้ $\theta=\frac{1}{2\mu^2}$ และ $\phi=\frac{1}{\lambda}$ จะได้ $a(\phi)=-\phi$, $b(\theta)=\sqrt{2\theta}$, และ $c(y,\phi)=-\frac{\lambda}{2y}+\frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)$ โดยที่ในกรณีนี้ อนุพันธ์อันดับหนึ่งและอนุพันธ์อันดับสองของ $b(\theta)$ คือ $b'(\theta)=(2\theta)^{1/2}=\frac{1}{\sqrt{2\theta}}=\mu$ และ

$b''(\theta) = -(2\theta)^{-3/2} = -\dfrac{1}{\left(\sqrt{2\theta}\right)^3} = -\mu^3$ ตามลำดับ ซึ่งนำไปสู่ ค่าเฉลี่ยของการแจกแจง คือ $E(Y) = \mu$ และความแปรปรวนของการแจกแจง คือ $Var(Y) = b''(\theta)a(\phi) = (-\mu^3)\left(-\dfrac{1}{\lambda}\right) = \dfrac{\mu^3}{\lambda}$

2.1.2.2 ส่วนประกอบเชิงระบบ (Systematic Component) จะแสดงตัวแปรทำนายในรูปแบบของผลรวมเชิงเส้นของพารามิเตอร์ (η) นั่นคือ $\eta = X\beta$ โดยที่ X คือ เมทริกซ์ของตัวแปรทำนาย และ β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอย ที่ประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Method) โดยการหาอนุพันธ์ของฟังก์ชันล็อกภาวะน่าจะเป็น (Log – Likelihood Function) เทียบกับ β แล้วกำหนดให้เท่ากับศูนย์ โดยผลลัพธ์ที่ได้จะอยู่ในรูปแบบไม่เชิงเส้น ซึ่งไม่มีผลเฉลยในรูปแบบปิดสำหรับประมาณค่าพารามิเตอร์ในตัวแบบเชิงเส้นนัยทั่วไป จึงได้มีการนำวิธีการทางตัวเลข (Numerical Method) มาช่วยในการประมาณค่าพารามิเตอร์ ยกตัวอย่างเช่น วิธี Iteratively Reweighted Least Squares วิธี Newton-Raphson หรือวิธี EM (Expectation Maximization) Algorithm เป็นต้น ซึ่งมีอยู่ในโปรแกรมพื้นฐานทั่วไป

2.1.2.3 ฟังก์ชันเชื่อมโยง (Link Function) เป็นฟังก์ชันของค่าเฉลี่ยของตัวแปรตาม โดยที่ $E(Y) \equiv \mu = g^{-1}(\eta)$ ในการใช้ตัวแบบเชิงเส้นนัยทั่วไป จำเป็นต้องทำการแปลงค่าตัวแปรตามให้สอดคล้องกับสมมุติฐาน โดยใช้ฟังก์ชันเชื่อมโยง เพื่อทำให้ตัวแปรตามสามารถแสดงอยู่ในรูปของฟังก์ชันผลรวมของตัวแปรทำนายได้ ซึ่งฟังก์ชันเชื่อมโยงที่ใช้ต้องเป็นฟังก์ชันทางเดียว (Monotonic) และสามารถหาอนุพันธ์ได้ (Differentiable) สำหรับฟังก์ชันเชื่อมโยงที่ได้จากการพิจารณาพารามิเตอร์ธรรมชาติ หรือกำหนดให้ $\eta = \theta$ จะเรียกฟังก์ชันเชื่อมโยงนี้ว่า ฟังก์ชันเชื่อมโยงคานอนิคัล (Canonical Link Function) โดยรูปแบบฟังก์ชันเชื่อมโยงคานอนิคัลสำหรับตัวแบบเชิงเส้นนัยทั่วไปสำหรับการแจกแจงปรกติหรือการแจกแจงเกาส์เซียน และการแจกแจงเกาซ์เซียนผกผัน สามารถสรุปรายละเอียดได้ดังตารางที่ 2-1 (Myers et.al., 2012; Faraway, 2016)

ตารางที่ 2-1 รูปแบบของฟังก์ชันเชื่อมโยงคานอนิคัลสำหรับตัวแบบเชิงเส้นนัยทั่วไป

การแจกแจง	ฟังก์ชันเชื่อมโยง	$\eta = X\beta = g(\mu)$	$\mu = g^{-1}(\eta)$
ปรกติ หรือเกาส์เซียน	เอกลักษณ์ (Identity)	$X\beta = \mu$	$\mu = X\beta$
เกาส์เซียนผกผัน	กำลังสองผกผัน (Inverse squared)	$X\beta = \dfrac{1}{\mu^2}$	$\mu = \dfrac{1}{\sqrt{X\beta}}$

อย่างไรก็ตาม สำหรับการแจกแจงเกาซ์เซียนผกผัน เพื่อให้มั่นใจว่าค่าที่ประมาณได้ของตัวแปรตอบสนองมีค่าเป็นบวกตามสมบัติของการแจกแจง จึงนิยมใช้ฟังก์ชันเชื่อมโยงในรูปของลอการิทึม นั่นคือ $X\beta = \ln(\mu)$ ซึ่งสามารถเขียนได้ว่า $\mu = \exp(X\beta)$ การเลือกใช้ฟังก์ชันเชื่อมโยงในรูปนี้ นอกจากจะรับประกันค่าบวกของ μ แล้ว แต่ยังช่วยให้การตีความผลลัพธ์ของตัวแบบทำได้ง่ายขึ้น

2.1.3 ตัวแบบการถดถอยต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นเครื่องมือที่ใช้ในการประมวลผลข้อมูลและการทำนายด้วยการสร้างโครงสร้างต้นไม้จากข้อมูลที่มีอยู่ เทคนิคนี้ใช้การเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยมีสองประเภทหลัก คือ ต้นไม้ตัดสินใจสำหรับการจำแนก (Classification Trees) และต้นไม้ตัดสินใจสำหรับการถดถอย (Regression Trees) หรือที่เรียกว่า ตัวแบบต้นไม้ตัดสินใจสำหรับการจำแนกและการถดถอย (CART – Classification and Regression Trees) โดยต้นไม้ตัดสินใจสำหรับการจำแนกใช้ในการทำนายผลลัพธ์ที่อยู่ในรูปแบบกลุ่มหรือประเภทต่าง ๆ ในขณะที่ต้นไม้ตัดสินใจสำหรับการถดถอยใช้ในการทำนายผลลัพธ์ที่เป็นค่าต่อเนื่อง

การใช้งานต้นไม้ตัดสินใจเริ่มต้นที่โหนดราก (Root Node) ซึ่งเป็นโหนดแรกที่ตรวจสอบคุณลักษณะ (Attribute: A) ของข้อมูล แล้วแบ่งข้อมูลออกเป็นกลุ่มย่อยโดยใช้เงื่อนไขที่เกี่ยวข้องกับคุณลักษณะนั้น ๆ กลุ่มย่อยจะแบ่งออกเป็นโหนดกิ่ง (Branch Node) ซึ่งจะเป็นโหนดถัดไปในต้นไม้ กระบวนการนี้จะทำซ้ำไปเรื่อย ๆ จนกว่าข้อมูลในกลุ่มย่อยไม่สามารถแบ่งได้แล้ว และโหนดสุดท้ายจะเรียกว่า โหนดใบ (Leaf Node) ซึ่งในโหนดนี้จะเป็นผลลัพธ์ที่เกี่ยวข้องกับปัญหาของต้นไม้ตัดสินใจนั้น ๆ ในแต่ละขั้นตอนของการสร้างโหนด จะใช้เกณฑ์การคำนวณค่าเกนสารสนเทศ (Information Gain) เพื่อเลือกคุณลักษณะหรือตัวแปรอิสระที่สามารถจัดหมวดหมู่ของข้อมูลได้ดีที่สุด โดยค่าเกนสารสนเทศสามารถคำนวณได้จากสมการที่ (2-7) (Dang & Nguyen, 2022)

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{a} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \qquad (2\text{-}7)$$

โดยที่ $S = \{S_1, S_2, \ldots, S_a\}$ คือ การแบ่งข้อมูล S ออกเป็นกลุ่มย่อยตามค่าของคุณลักษณะ a คือ จำนวนคุณลักษณะ $|S_i|$ คือ จำนวนข้อมูลในกลุ่มย่อย S_i และ $|S|$ คือ จำนวนข้อมูลทั้งหมดใน S นอกจากนี้ ยังสามารถคำนวณหาเอนโทรปี (Entropy) ของข้อมูลใน S และกลุ่มย่อย S_i ซึ่งเป็นการวัดความไม่บริสุทธิ์ (Impurity) ของข้อมูล ได้จากสมการที่ (2-8)

$$\text{Entropy}(S) = \sum_{j=1}^{g} -\pi_j \log_2(\pi_j) \quad \text{และ} \quad \text{Entropy}(S_i) = \sum_{j=1}^{g} -\pi_{ij} \log_2(\pi_{ij}) \qquad (2\text{-}8)$$

เมื่อ π_j คือ สัดส่วนของข้อมูล S ที่ผลลัพธ์มีค่า j และ π_{ij} คือ สัดส่วนของข้อมูลในกลุ่มย่อย S_i ที่ผลลัพธ์มีค่า j โดยที่เอนโทรปีจะมีค่าอยู่ระหว่าง 0 ถึง 1 ทั้งนี้ เอนโทรปีจะมีค่าเป็นศูนย์ เมื่อข้อมูลใน S หรือ S_i มีผลลัพธ์แบบเดียวกัน และจะมีค่ามากขึ้น เมื่อมีผลลัพธ์ที่แตกต่างกันมากขึ้น (หรือมีค่ามากขึ้น หากข้อมูลไม่บริสุทธิ์)

2.1.4 ตัวแบบการถดถอยป่าสุ่ม

การถดถอยป่าสุ่ม (Random Forest Regression) เป็นเทคนิคการถดถอยที่อยู่ในกลุ่มของการเรียนรู้ของเครื่อง (Machine Learning) ที่ใช้วิธีการแบบไม่อิงพารามิเตอร์ (Non-parametric) โดยใช้แนวคิด

จากเทคนิค Bagging (หรือมีชื่อเต็มว่า Bootstrap Aggregating) และ Random Subspace เพื่อเพิ่มความหลากหลายของตัวแบบ จะมีการสร้างต้นไม้การถดถอย (Regression Tree) จำนวนมากด้วยการสุ่มตัวอย่างบูตสแทร็ป (Bootstrap Sampling) นั่นคือ แต่ละชุดข้อมูลย่อยที่ได้จะถูกสร้างขึ้นโดยการสุ่มตัวอย่างจำนวน n ตัวอย่างจากข้อมูลชุดฝึกสอน (Training Data) ซึ่งมีทั้งหมด N ตัวอย่าง โดยทั่วไปชุดข้อมูลที่สุ่มมานี้จะมีประมาณ 2 ใน 3 ของข้อมูลชุดฝึกสอน และไม่ซ้ำกัน

จากชุดข้อมูลย่อยที่ถูกสร้างขึ้นจำนวน K ชุดข้อมูล จะถูกนำมาสร้างต้นไม้การถดถอยที่เป็นอิสระกันจำนวน K ต้น ซึ่งแต่ละต้นจะรับข้อมูลเข้า ($\boldsymbol{X}$) และให้ค่าทำนาย ($\hat{Y}_k$) ที่แตกต่างกัน จากนั้นนำค่าทำนายของต้นไม้ทั้งหมดทั้ง K ต้น มาหาค่าเฉลี่ย เพื่อใช้เป็นค่าทำนายสุดท้ายของตัวแบบ ซึ่งสามารถเขียนได้ดังสมการที่ (2-9)

$$\hat{\boldsymbol{Y}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{Y}}_k \tag{2-9}$$

โดยที่ $\hat{\boldsymbol{Y}}' = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)'$ การใช้ค่าเฉลี่ยนี้จะช่วยลดความแปรปรวนของตัวแบบ และลดความเสี่ยงจากปัญหา Overfitting นั่นคือ ตัวแบบสามารถทำนายข้อมูลชุดฝึกสอนได้แม่นยำ แต่ตัวแบบทำงานได้แย่หรือทำนายได้ไม่แม่นยำเมื่อเจอกับข้อมูลใหม่หรือข้อมูลที่ไม่เคยเห็นมาก่อน

สำหรับตัวอย่างที่ไม่ถูกเลือกในการนำไปสร้างต้นไม้การถดถอยต้นที่ k เมื่อ k = 1, 2, ... , K จะถูกเก็บไว้ในชุดข้อมูลอีกชุดหนึ่งที่เรียกว่า ชุดข้อมูล Out-of-Bag (OOB) โดยทั่วไป ชุดข้อมูล OOB จะมีประมาณ 1 ใน 3 ของข้อมูลชุดฝึกสอนทั้งหมด ถึงแม้ว่าข้อมูลชุดนี้จะไม่ได้ถูกนำมาใช้ในการสร้างตัวแบบ แต่สามารถนำมาใช้ในการประเมินประสิทธิภาพของต้นไม้ต้นนั้นได้ว่าสามารถทำนายได้แม่นยำแค่ไหน โดยการคำนวณค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE) จากชุดข้อมูล OOB ดังนี้

$$\mathrm{MSE}_{\mathrm{OOB}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}_{i,\mathrm{OOB}}\right)^2 \tag{2-10}$$

เมื่อ $\hat{Y}_i$ คือ ค่าพยากรณ์ของค่าสังเกตที่ i เมื่อ i = 1, 2, ... , n และ $\overline{\hat{Y}}_{i,\mathrm{OOB}}$ คือ ค่าเฉลี่ยของค่าพยากรณ์ของค่าสังเกตที่ i จากต้นไม้ทุกต้น (Jain et al., 2021; Seo et al., 2017)

2.1.5 ตัวแบบการถดถอยเอ็กซ์จีบูสต์

ในโลกของการเรียนรู้ของเครื่อง ตัวแบบดั้งเดิมอย่างตัวแบบต้นไม้ตัดสินใจ หรือตัวแบบป่าสุ่มเป็นตัวแบบที่เข้าใจง่ายและตีความได้ไม่ยาก แต่ในหลายกรณีโดยเฉพาะเมื่อทำงานกับชุดข้อมูลที่มีความซับซ้อนสูง ตัวแบบเหล่านี้มักไม่สามารถให้ความแม่นยำที่เพียงพอได้ เพื่อแก้ปัญหาดังกล่าว จึงได้มีการพัฒนาขั้นตอนวิธี (Algorithm) ที่เรียกว่า เอ็กซ์จีบูสต์ (XGBoost) ที่ย่อมาจาก eXtreme Gradient Boosting ซึ่งเป็นขั้นตอนวิธีที่สามารถประยุกต์ใช้ได้ทั้งในปัญหาเกี่ยวกับการจำแนก (Classification) และการถดถอย (Regression) โดยที่เอ็กซ์จีบูสต์เป็นการนำแนวคิดของ Gradient Boosting มาพัฒนาให้มีประสิทธิภาพสูงขึ้นภายใต้แนวคิดของการ

เรียนรู้แบบผสมผสาน (Ensemble Learning) ซึ่งหมายถึงการรวมตัวแบบย่อยหลาย ๆ ตัวแบบเข้าด้วยกัน โดยตัวแบบย่อยที่ใช้ คือ ตัวแบบต้นไม้ตัดสินใจที่ถูกสร้างขึ้นอย่างต่อเนื่องเป็นลำดับ โดยต้นไม้แต่ละต้นจะพยายามแก้ไขข้อผิดพลาดของต้นก่อนหน้า ซึ่งกระบวนการนี้เรียกว่า Boosting ซึ่งเป็นหัวใจสำคัญในการปรับปรุงประสิทธิภาพของตัวแบบอย่างต่อเนื่องที่ทำให้เอ็กซ์จีบูสต์มีความรวดเร็ว มีประสิทธิภาพสูง และสามารถทำงานกับข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ

ในกระบวนการของเอ็กซ์จีบูสต์จะนำตัวแบบต้นไม้หลายต้นมารวมกันโดยใช้วิธีการบวกแบบสะสม (Additive Method) โดยสมมติว่าใช้ต้นไม้ทั้งหมดจำนวน K ต้น และให้ F แทน ตัวแบบต้นไม้พื้นฐาน (Basic Tree Model) ดังนั้น ตัวแบบที่ได้จากการประมาณค่าสามารถเขียนในรูปของสมการทั่วไปได้ดังสมการที่ (2-11)

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{2-11}$$

เมื่อ $f_k \in F$ ทั้งนี้ เพื่อป้องกันไม่ให้ตัวแบบการถดถอยเอ็กซ์จีบูสต์ (XGBoost Regression) เกิดปัญหา Overfitting จึงมีการกำหนดฟังก์ชันจุดประสงค์ (Objective Function) ดังแสดงในสมการที่ (2-12)

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2-12}$$

เมื่อ l คือ ฟังก์ชันการสูญเสีย (Loss Function) ซึ่งใช้วัดความคลาดเคลื่อนระหว่างค่าจริงและค่าพยากรณ์ และ Ω คือ ฟังก์ชันที่ช่วยควบคุมไม่ให้ตัวแบบมีความซับซ้อนมากเกินไป (Regularization Function) เพื่อป้องกันไม่ให้เกิดปัญหา Overfitting โดยที่

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{2-13}$$

เมื่อ γ คือ ค่าสูญเสียต่ำสุด ซึ่งใช้เป็นเกณฑ์สำหรับการตัดสินใจว่าจะสร้างกิ่งใหม่หรือไม่ T คือ จำนวนใบของต้นไม้แต่ละต้น λ คือ พารามิเตอร์ปรับโทษ (Penalty Parameter) หรือพารามิเตอร์ที่ใช้ลดความซับซ้อนของตัวแบบ (Regularization Parameter) λT คือ ค่าที่แสดงการควบคุมการตัดแต่งต้นไม้ (Tree Pruning) และ w คือ น้ำหนักของใบของต้นไม้แต่ละต้น (Chen & Guestrin, 2016; Lai et al., 2021; Li et al., 2019; Wu et al., 2022)

2.1.6 เกณฑ์ที่ใช้ในการประเมินประสิทธิภาพของตัวแบบ

ในการศึกษาครั้งนี้ จะใช้รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAPE) ในการเปรียบเทียบตัวแบบต่าง ๆ ที่แตกต่างกัน และตรวจสอบประสิทธิภาพของการทำนายของตัวแบบ ดังมีรายละเอียดต่อไปนี้

2.1.6.1 รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย

รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error; RMSE) เป็นค่าที่ใช้วัดความแตกต่างระหว่างค่าจริงกับค่าพยากรณ์ เพื่อตรวจสอบประสิทธิภาพของการทำนาย หาก RMSE มีค่าน้อย แสดงว่าค่าพยากรณ์สามารถประมาณค่าได้ใกล้เคียงกับค่าจริง โดยที่คำนวณหาค่า RMSE ได้จาก

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{2-9}$$

เมื่อกำหนดให้ Y_i และ $\hat{Y}_i$ คือ ค่าที่เกิดขึ้นจริงและค่าพยากรณ์ของค่าสังเกตที่ i ในขณะที่ n คือ จำนวนข้อมูลที่ใช้ในการวิเคราะห์

2.1.6.2 ร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย

ร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error; MAPE) จะใช้ในการวัดความถูกต้องของการพยากรณ์พร้อมทั้งเปรียบเทียบตัวแบบ โดยตัวแบบที่เหมาะสมที่สุดจะให้ MAPE ต่ำที่สุดซึ่งสามารถคำนวณได้จากสมการต่อไปนี้

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100 \tag{2-10}$$

เมื่อกำหนดให้ Y_i และ $\hat{Y}_i$ คือ ค่าที่เกิดขึ้นจริงและค่าพยากรณ์ของค่าสังเกตที่ i ในขณะที่ n คือ จำนวนข้อมูลที่ใช้ในการวิเคราะห์ นอกจากนี้ ยังสามารถแบ่งระดับความแม่นยำในการพยากรณ์จากเกณฑ์ MAPE (Lewis, 1982) ได้ดังตารางที่ 2-2

ตารางที่ 2-2 เกณฑ์ความแม่นยำสำหรับการพยากรณ์

เกณฑ์	ความแม่นยำ
MAPE < 10%	การพยากรณ์ที่ความแม่นยำสูงมาก
10% ≤ MAPE < 20%	การพยากรณ์ที่ใช้ได้ดี
20% ≤ MAPE < 50%	การพยากรณ์ที่พอยอมรับได้
≥ 50%	การพยากรณ์ที่ไม่แม่นยำ

2.2 งานวิจัยที่เกี่ยวข้อง

ในปัจจุบัน อัตราผู้ถือกรมธรรม์ประกันชีวิตในประเทศไทยเมื่อเทียบกับจำนวนประชากรทั้งหมด มีเพียงร้อยละ 39.74 (สำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย, 2566) โดยในการศึกษาของ

ทิพวัลย์ อิ่มรส (2562) ได้ศึกษาถึงปัจจัยที่ส่งผลต่อการตัดสินใจซื้อประกันชีวิตของประชากร ในเขตกรุงเทพมหานคร พบว่า เพศ จำนวนบุตร รายได้เฉลี่ยต่อเดือน รวมไปถึงปัจจัยทางด้านผลิตภัณฑ์ การส่งเสริมการตลาด และราคา มีผลต่อการตัดสินใจซื้อกรมธรรม์ประกันชีวิต ซึ่งสอดคล้องกับการศึกษาของ ไตรรัตน์ ศานนิกรภาพ และณัฐพล พันธุ์ภักดี (2565) ที่ได้ศึกษาเกี่ยวกับปัจจัยที่มีผลต่อความตั้งใจซื้อประกันชีวิตของผู้บริโภคในเขตกรุงเทพมหานครและปริมณฑล โดยพบว่าผู้บริโภคให้ความสำคัญต่อความตั้งใจซื้อประกันชีวิตโดยภาพรวมอยู่ในระดับมาก อีกทั้งการตั้งใจซื้อประกันชีวิตยังได้รับผลกระทบจากปัจจัยต่าง ๆ อีกด้วย ซึ่งรวมถึงเพศ รายได้ต่อเดือน อาชีพ และปัจจัยทางด้านสภาพแวดล้อม ได้แก่ ด้านพฤติกรรม ด้านบริการ และด้านสภาพแวดล้อมภายนอก ยังเป็นตัวทำนายที่มีอิทธิพลต่อพฤติกรรมความตั้งใจซื้อประกันชีวิตของผู้บริโภคอีก

ในทำนองเดียวกัน การศึกษาของ สุธรรม เลิศพงษ์ประเสริฐ และวิธาวีรัฐ ประทุมสวัสดิ์ (2566) ได้ศึกษาถึงปัจจัยที่มีอิทธิพลต่อพฤติกรรมการซื้อประกันชีวิตของผู้บริโภคในจังหวัดนนทบุรี พบว่า พฤติกรรมการซื้อประกันชีวิตส่วนใหญ่ เป็นการซื้อประกันชีวิตแบบสะสมทรัพย์ ที่ตนเองเป็นผู้ตัดสินใจซื้อ โดยมีระยะเวลาคุ้มครอง 16 – 20 ปี และมีการชำระเบี้ยประกันปีละ 25,001 – 50,000 บาท โดยส่วนใหญ่ซื้อประกันชีวิตกับบริษัท สยามซัมซุง ประกันชีวิต จำกัด ผ่านตัวแทนบริษัท และช่วงอายุที่ซื้อประกันชีวิตอยู่ระหว่าง 31 – 40 ปี โดยปัจจัยส่วนประสมทางการตลาดในการซื้อประกันชีวิตมีความสำคัญอยู่ในระดับมากที่สุด อีกทั้งปัจจัยส่วนบุคคลและปัจจัยส่วนประสมทางการตลาดมีอิทธิพลต่อพฤติกรรมการซื้อประกันชีวิตของผู้บริโภคในเขตจังหวัดนนทบุรี นอกจากนี้ การศึกษาของ สรฤทธิ์ ทรัพย์สมบัติ และอารีย์ นัยพินิจ (2562) ได้วิเคราะห์แนวทางการตลาดที่มีผลต่อการตัดสินใจซื้อกรมธรรม์ประกันชีวิตของผู้บริโภคในเขตจังหวัดขอนแก่น พบว่ามีปัจจัยทางการตลาดที่มีผลกระทบต่อการตัดสินใจซื้อประกันชีวิต เช่น มีแบบการประกันชีวิตให้เลือกหลายรูปแบบ (ปัจจัยด้านผลิตภัณฑ์) อัตราค่าเบี้ยประกันชีวิตมีหลากหลายราคาตามช่วงอายุ (ปัจจัยด้านราคา) มีตัวแทนและสาขากระจายอยู่ทั่วประเทศ (ปัจจัยด้านช่องทางการจัดจำหน่าย) ความน่าเชื่อถือของตัวแทน (ปัจจัยด้านบุคคล) ผู้บริหารเป็นที่รู้จักทางสังคมและบริษัทประกันมีความมั่นคง (ปัจจัยด้านลักษณะทางกายภาพ) และมีระบบบริการการอนุมัติกรมธรรม์ที่รวดเร็ว และถูกต้องสมบูรณ์ (ปัจจัยด้านกระบวนการ)

อย่างไรก็ตาม การทำประกันชีวิตในประเทศไทยมีสัดส่วนการเพิ่มขึ้นที่ไม่มากเมื่อเทียบกับประเทศเพื่อนบ้าน และการศึกษาของ สุนันทา นิยมศิลป์ และสัญญา สัญญาวิวัฒน์ (2555) ได้พบว่า ปัจจัยที่มีผลต่อการเพิ่มการทำประกันชีวิตในประเทศไทย ได้แก่ 1.) การมีสัมพันธภาพอันดีกับลูกค้า ซึ่งจะทำให้ผู้บริโภคมีความเชื่อมั่นและสบายใจในการทำประกันชีวิต 2.) การยอมรับบริษัทประกันชีวิต ซึ่งทำให้เกิดความไว้วางใจในความเสี่ยงและความน่าเชื่อถือของบริษัท 3.) ความเต็มใจในรูปแบบประกันชีวิตของผู้เอาประกัน 4.) การยอมรับตัวแทนประกันชีวิต ซึ่งเป็นการเชื่อในตัวแทนที่นำเสนอและให้คำแนะนำกับผู้เอาประกันชีวิต และ 5.) การใช้หลักธรรมภิบาล นอกจากนี้ การทำประกันชีวิตในปัจจุบัน ยังมีบทบาทที่สำคัญต่อการเปลี่ยนแปลงทางสังคมที่กำลังเกิดขึ้นอย่างรวดเร็ว โดยเฉพาะในยุคที่เข้าสู่สังคมผู้สูงอายุ และแนวโน้มของสังคมผู้สูงอายุในปัจจุบันและอนาคตที่ส่งผลต่อธุรกิจประกันชีวิต สามารถแบ่งได้ 3 กลุ่ม คือ กลุ่มที่ 1 กลุ่มที่มีความพร้อมในฐานะทางการเงินที่สามารถทำประกันชีวิตได้ กลุ่มที่ 2 กลุ่มที่สนใจการออมเงินและการลงทุนผ่านประกันชีวิต และกลุ่มที่ 3 กลุ่มที่ยังต้องพึ่งพิงลูกหลานไม่มีรายได้ อยู่ได้ด้วยเงินสวัสดิการของภาครัฐและลูกหลาน ในการปรับตัวของธุรกิจประกันชีวิตมุ่งเน้นในด้านการสื่อสารและการจัดจำหน่ายที่หลากหลาย ด้านผลิตภัณฑ์ใหม่ที่คุ้มครองการประกันสุขภาพและรองรับการเกษียณ และด้านกลยุทธ์ในการสร้างการรับรู้และสื่อสารเกี่ยวกับความสำคัญและความจำเป็นของการทำประกันชีวิต (สุกัญญา ชำนาญ, 2564)

นอกจากนี้ ในการศึกษาปัจจัยที่มีผลต่อการทำประกันชีวิตแบบบำนาญของคนทำงานในเขตพื้นที่กรุงเทพมหานคร จากงานวิจัยของ พงศกร สัตยพานิช (2562) พบว่า คนที่ทำงานในเขตกรุงเทพมหานครจะมีพฤติกรรมในการออมเงินด้วยวิธีนำไปฝากกับธนาคารเป็นส่วนใหญ่ ทั้งนี้ปัจจัยส่วนบุคคลที่มีผลต่อการตัดสินใจทำประกันชีวิตแบบบำนาญ ได้แก่ เพศ อายุ สถานภาพ ระดับการศึกษา จำนวนเงินออมต่อเดือน และอัตราการเสียภาษีที่แตกต่างกัน ซึ่งส่งผลให้มีแนวโน้มการทำประกันชีวิตแบบบำนาญที่แตกต่างกัน และเมื่อวิเคราะห์ความสัมพันธ์ระหว่างปัจจัยจูงใจกับการตัดสินใจทำประกันชีวิตแบบบำนาญ พบว่า การทำประกันชีวิตแบบบำนาญมีความสัมพันธ์กับความมั่นคงหลังเกษียณอายุและผลิตภัณฑ์ การลดหย่อนภาษี และการตัดสินใจออม

ในการศึกษาที่ผ่านมา ได้มีการนำตัวแบบทางสถิติมาใช้ในการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับธุรกิจประกันในรูปแบบต่าง ๆ อย่างแพร่หลาย เช่น ตัวแบบอนุกรมเวลา (Time Series Models) ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Models) และตัวแบบการเรียนรู้ของเครื่อง (Machine Learning Models) เป็นต้น จากการศึกษาของเรณู นิยมเดชา (2554) ได้นำตัวแบบเชิงเส้นนัยทั่วไปมาใช้ในการวิเคราะห์ต้นทุนความเสียหายของการประกันอัคคีภัยที่อยู่อาศัย โดยแบ่งการวิเคราะห์เป็น 2 ตัวแบบ คือ ตัวแบบความถี่และตัวแบบความรุนแรง โดยการใช้ตัวแบบเชิงเส้นนัยทั่วไปนี้ช่วยในการประมาณค่าต้นทุนความเสียหายที่เกิดขึ้น ซึ่งสามารถคำนวณได้จากผลคูณของความถี่และความรุนแรง และทั้ง 2 ตัวแบบ นี้มีการใช้ตัวแปรทำนายชุดเดียวกัน ได้แก่ ปีที่เริ่มรับประกันภัย เขตพื้นที่/ที่ตั้ง ลักษณะภัย และชั้นของสิ่งปลูกสร้าง โดยตัวแบบความถี่นั้นจะมีตัวแปรตอบสนองเป็นจำนวนครั้งการเกิดความเสียหาย (Claims) ที่มีการแจกแจงปัวซง ในขณะที่ตัวแบบความรุนแรงจะมีค่าสินไหมทดแทนเฉลี่ยต่อครั้ง (Average Loss) เป็นตัวแปรตอบสนอง ที่มีการแจกแจงแกมมา ในการคำนวณต้นทุนความเสียหายของการประกันอัคคีภัยที่อยู่อาศัย การนำตัวแบบความถี่และตัวแบบความรุนแรงมารวมกันบนสมมุติฐานว่าปัจจัยเสี่ยงแต่ละปัจจัยจากทั้งสองตัวแบบเป็นอิสระต่อกัน มีไว้ใช้ในการกำหนดอัตราเบี้ยประกันภัย และในการวิเคราะห์ความแตกต่างของปัจจัยเสี่ยง จากการศึกษา พบว่า ปัจจัยเสี่ยงที่ใช้ในการกำหนดอัตราเบี้ยประกันภัย ในพิกัดอัตราเบี้ยประกันอัคคีภัยไม่เป็นไปตามโอกาสในการเกิดความเสียหายหรือเสี่ยงภัยที่แท้จริง เนื่องจากข้อมูลที่นำมาใช้มีการรับประกันภัยบ้ายที่อยู่อาศัยที่ตั้งอยู่ในบริเวณพื้นที่ที่มีโอกาสในการเกิดภัยสูง และไม่มีการกระจายความเสี่ยงภัยที่เหมาะสม

Alshamsi (2014) ได้ศึกษาการทำนายกรมธรรม์ประกันรถยนต์ด้วยวิธีป่าสุ่ม (Random Forest) ซึ่งเป็นเทคนิคการเหมืองข้อมูล (Data Mining) ที่ถูกนำมาใช้ในการทำนายกรมธรรม์ประกันรถยนต์ที่ลูกค้าน่าจะเลือก โดยที่ประสิทธิภาพวิธีการแบบป่าสุ่ม จะถูกนำมาเปรียบเทียบกับเทคนิคการทำเหมืองข้อมูลวิธีอื่น ๆ เช่น ZeroR classifier, Simple Logistics Function, Decision Tree, และ Naïve Bayes โดยในการศึกษานี้ จะใช้ชุดข้อมูลที่ประกอบไปด้วย 7 กรมธรรม์ที่แตกต่างกัน ซึ่งผลลัพธ์พบว่า วิธีการแบบป่าสุ่มเป็นเทคนิคที่มีความแม่นยำโดยรวมสูงที่สุด

Kafková & Křivánková (2014) ได้ศึกษาการประมาณความถี่ของการเรียกร้องค่าสินไหมทดแทนต่อปี (Annual Frequency of Claims; AFC) ซึ่งเป็นที่มาของเบี้ยประกันภัยรถยนต์ ที่คาดว่ามีความสัมพันธ์กับปัจจัยเสี่ยงหลายประการ ได้แก่ ประเภทตัวถัง อายุรถ พื้นที่ที่อยู่อาศัย เพศของผู้ถือกรมธรรม์ และช่วงอายุของผู้ถือกรมธรรม์ ผู้วิจัยได้ทำการเปรียบเทียบตัวแบบการถดถอยเชิงเส้น (Linear Regression) แบบทั่วไป กับตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model; GLM) เพื่อหาตัวแบบ GLM ที่เหมาะสมที่สุด โดยการวิเคราะห์ข้อมูลกรมธรรม์ประกันภัยรถยนต์หนึ่งปี จำนวน 57,410 ฉบับ และใช้การวิเคราะห์ความเบี่ยงเบน (Analysis of Deviance) และเกณฑ์สารสนเทศของอะกะอิเกะ (Akaike’s Information Criterion; AIC) ในการเปรียบเทียบตัวแบบ

Noor & Isa (2014) ได้ใช้วิธีการโซ่มาร์คอฟ (Markov Chain) ซึ่งเป็นตัวแบบอนุกรมเวลา มาใช้ในการทำนายจำนวนการซื้อประกันชีวิต จากข้อมูลการซื้อประกันชีวิตของบริษัท General Assurance Berhad ในช่วงปี 2003 ถึง 2006 ที่ช่วยอธิบายถึงแนวโน้มและสถานการณ์ของการซื้อประกันชีวิต โดยการศึกษานี้ได้แนะนำว่าควรพิจารณาถึงจำนวนลูกค้าที่ซื้อกรมธรรม์ในแต่ละเดือน ซึ่งจะทำให้ตัวแบบทำนายได้แม่นยำมากขึ้น เนื่องจากสามารถมองเห็นการเปลี่ยนแปลงที่เกิดขึ้นในแต่ละเดือนได้

ชนกนาต สุขประยูร และคณะ (2560) ได้ศึกษาการใช้ตัวแบบเชิงเส้นวางนัยทั่วไปในการประมาณค่าเบี้ยประกันภัยรถยนต์ โดยใช้ข้อมูลกรมธรรม์รถยนต์ประเภทชั้น 1 ของบริษัทประกันแห่งหนึ่งในประเทศไทย ด้วยวิธีการวิเคราะห์ตัวแบบการถดถอยทวินามลบ (Negative Binomial Regression Model) และตัวแบบการถดถอยแกมมา (Gamma Regression Model) ซึ่งพบว่ามีค่าส่วน ลด-เพิ่ม จากประวัติการเรียกร้องสินไหมทดแทน และลักษณะการใช้งานรถยนต์ มีผลกระทบต่อการประมาณเบี้ยประกันภัยแท้จริง

Boodhun & Jayabalan (2018) ได้ศึกษาการทำนายความเสี่ยงในอุตสาหกรรมประกันชีวิตด้วยอัลกอริทึมการเรียนรู้แบบมีผู้สอน จากข้อมูลคำขอเอาประกันจำนวน 59,381 รายการ 128 ตัวแปร ซึ่งอธิบายเกี่ยวกับคุณลักษณะของผู้ขอเอาประกันชีวิต โดยผู้วิจัยได้ทำการลดมิติของข้อมูลเพื่อเลือกคุณลักษณะที่สำคัญ ที่สามารถปรับปรุงความสามารถในการทำนายของตัวแบบ โดยใช้เทคนิค Correlation-Based Feature Selection (CFS) และ Principal Components Analysis (PCA) และใช้ตัวแบบการเรียนรู้ของเครื่อง (Machine Learning) ได้แก่ การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression), Artificial Neural Network (ANN), Reduced Error Pruning Tree (REPTree), และ Random Tree Classifiers ในการทำนายระดับความเสี่ยงของผู้ขอเอาประกันชีวิต โดยพบว่า ในการลดมิติของข้อมูลด้วยวิธี CFS วิธีการ Reduced Error Pruning Tree จะมีประสิทธิภาพในการทำนายดีที่สุด และเมื่อใช้ PCA ในการลดมิติของข้อมูล ตัวแบบการถดถอยเชิงเส้นพหุคูณ จะมีประสิทธิภาพมากที่สุด

อริญชย์ และคณะ (2562) ได้มีการศึกษาการวิเคราะห์การขาดอายุของกรมธรรม์ประกันชีวิต โดยใช้ตัวแบบถดถอยลอจิสติก (Logistic Regression) ที่มีตัวแปรตาม คือ การขาดอายุกรมธรรม์ และมีตัวแปรทำนายที่เป็นตัวแปรเชิงคุณภาพ และตัวแปรเชิงปริมาณ รวมทั้งหมด 18 ตัวแปร ในการวิเคราะห์ข้อมูล ผู้วิจัยได้แบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ชุดข้อมูลฝึกหัดสำหรับสร้างตัวแบบทำนาย และชุดข้อมูลทดสอบ โดยผลการวิจัยพบว่า มีตัวแปรที่มีผลต่อการขาดอายุของกรมธรรม์จำนวน 6 ตัวแปร ได้แก่ อายุ จำนวนเงินเอาประกันชีวิตระหว่าง 50,001 – 100,000 บาท ระยะเวลาการชำระเบี้ยประกันภัยมากกว่า 3 ปี รากที่สองของระยะเวลาความคุ้มครอง ระดับชั้นอาชีพแบบอาชีพชั้น 3 และระดับชั้นอาชีพแบบอาชีพชั้น 4

Das et al. (2022) ได้ทำการศึกษาการทำนายค่าเบี้ยประกันชีวิตระหว่างระหว่างก่อนและหลังโควิด-19 ด้วยวิธีการ Higher-Order Neural Network โดยใช้ข้อมูลจากบริษัทประกันชีวิต 24 แห่ง ในอินเดีย ตั้งแต่เดือนมกราคม 2558 ถึงเดือนธันวาคม 2563 พร้อมทั้งได้มีการคาดการณ์เบี้ยประกันภัยในอนาคต โดยที่ผู้วิจัยได้นำเสนอ คือ ตัวแบบ Functional Link Neural Network (FLANN) ด้วยการใช้วิธีการเชิงพันธุกรรม (Genetic Algorithm; GA) ในการหาโครงสร้าง FLANN ที่เหมาะสมที่สุด หรือเรียกว่าตัวแบบ FLANN-GA และถูกนำไปเปรียบเทียบกับตัวแบบอื่น ๆ ได้แก่ FLANN, Multi-Layer Perceptron (MLP), Autoregressive Integrated Moving Average (ARIMA), และตัวแบบการถดถอยเชิงเส้นพหุคุณ (Multiple Linear Regression; MLR)

วิกานดา และวิราวรรณ (2566) ได้ศึกษาเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตแบบรายเดือนในประเทศไทย โดยใช้ข้อมูลตั้งแต่เดือนมกราคม พ.ศ. 2560 ถึงเดือนธันวาคม พ.ศ. 2565 จำนวน 72 เดือน และเปรียบเทียบตัวแบบการพยากรณ์ 4 ตัวแบบ ได้แก่ ตัวแบบการถดถอยต้นไม้ตัดสินใจ ตัวแบบการถดถอยแบบ

ป่าสุ่ม ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย และตัวแบบการถดถอยพหุนาม ผลการศึกษาพบว่า ตัวแบบการถดถอยต้นไม้ตัดสินใจมีประสิทธิภาพดีที่สุดในการพยากรณ์เบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย แต่เนื่องจากตัวแบบการถดถอยต้นไม้ตัดสินใจไม่เหมาะสำหรับการพยากรณ์ในระยะยาว จึงเลือกใช้ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอยแทน เนื่องจากตัวแบบมีประสิทธิภาพรองลงมา และมีความเหมาะสมกับข้อมูลที่นำมาใช้งาน

นอกจากนี้ จากการศึกษาตัวแบบที่เหมาะสมที่สุดสำหรับการพยากรณ์การเติบโตของอุตสาหกรรมประกันชีวิตในประเทศไทย โดยใช้ข้อมูลตั้งแต่เดือนมกราคม พ.ศ. 2546 ถึงเดือนธันวาคม พ.ศ. 2565 รวมทั้งสิ้น 240 เดือน พบว่ามีการเปรียบเทียบประสิทธิภาพของตัวแบบการพยากรณ์ที่แตกต่างกัน โดย Huadsri et al. (2024) ได้ศึกษาตัวแบบ SARIMAX ตัวแบบ RNN (Recurrent Neural Network) และตัวแบบผสม RNN-SARIMAX ซึ่งเป็นการรวมจุดเด่นของทั้งสองวิธีเข้าด้วยกัน ผลการศึกษาพบว่า ตัวแบบ SARIMAX มีความแม่นยำและมีประสิทธิภาพในการพยากรณ์ดีกว่า ในขณะที่งานของ Huadsri & Phaphan (2024) ที่ศึกษาประสิทธิภาพของตัวแบบ SARIMAX ตัวแบบ SVR (Support Vector Regression) และตัวแบบผสม SVR-SARIMAX ผลการศึกษาพบว่า ตัวแบบ SARIMAX ยังคงให้ผลแม่นยำสูงสุด เนื่องจากข้อมูลมีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระและตัวแปรตาม อย่างไรก็ตาม ตัวแบบ SVR-SARIMAX ก็แสดงให้เห็นถึงความแม่นยำของการพยากรณ์ที่ดีกว่าเมื่อเทียบกับตัวแบบไม่เชิงเส้น (ตัวแบบ SVR) และต่อมา Phaphan et al. (2024) ได้ศึกษาและเปรียบเทียบตัวแบบ SARIMAX ตัวแบบโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) และตัวแบบผสมที่รวมทั้ง SARIMAX และ MLP เข้าด้วยกัน (ตัวแบบ MLP-SARIMAX) ผลการศึกษาพบว่า ตัวแบบ MLP ให้ผลการพยากรณ์ที่แม่นยำที่สุด

ในการทบทวนงานวิจัยเกี่ยวกับการตัดสินใจทำประกันชีวิตข้างต้น พบว่า มีปัจจัยที่มีอิทธิพลต่อพฤติกรรมการซื้อประกันชีวิตของประชากรในพื้นที่ต่าง ๆ ของประเทศหลายปัจจัย ซึ่งรวมถึง เพศ จำนวนบุตร รายได้ อาชีพ และปัจจัยทางด้านสภาพแวดล้อมมีผลต่อการตัดสินใจทำประกันชีวิต โดยการปรับตัวของธุรกิจประกันชีวิตควรมุ่งเน้นในด้านการสื่อสารและการจัดจำหน่ายที่หลากหลาย ผลิตภัณฑ์ใหม่ที่คุ้มครองการประกันสุขภาพและรองรับการเกษียณ และกลยุทธ์ในการสร้างการรับรู้ถึงความสำคัญของการทำประกันชีวิต อย่างไรก็ตาม การทำประกันชีวิตในประเทศไทยยังต้องพัฒนาต่อไปเพื่อทำให้ประชากรเข้าใจและรับรู้ความสำคัญของการมีประกันชีวิตในชีวิตประจำวัน การปรับตัวต่อการเปลี่ยนแปลงทางสังคม โดยเฉพาะในสังคมผู้สูงอายุเป็นสำคัญ

ในการศึกษาตัวแบบที่เกี่ยวข้องกับธุรกิจประกันภัยข้างต้น พบว่า มีการนำตัวแบบทางสถิติและการเรียนรู้ของเครื่องมาใช้มากมายในอุตสาหกรรมนี้ ด้วยความสามารถในการวิเคราะห์ข้อมูลและความแม่นยำในการทำนาย ตัวแบบเหล่านี้จึงเป็นเครื่องมือสำคัญที่ช่วยธุรกิจประกันภัยในการตัดสินใจและวางกลยุทธ์ทางธุรกิจ ด้วยเหตุนี้ ผู้วิจัยจึงสนใจที่จะทำการเปรียบเทียบตัวแบบทางสถิติ เพื่อหาตัวแบบที่เหมาะสมและมีประสิทธิภาพในการทำนายจำนวนกรมธรรม์ประกันชีวิตในประเทศไทย เพื่อช่วยลดความเสี่ยงและเพิ่มความแม่นยำในการตัดสินใจการดำเนินงานและการพัฒนาต่อไปของธุรกิจประกันชีวิต เพื่อให้ธุรกิจประกันชีวิตทำงานได้ต่อเนื่องในสภาวะทางเศรษฐกิจและสังคมในปัจจุบันที่มีเปลี่ยนแปลงอย่างรวดเร็ว

บทที่ 3
วิธีการดำเนินการวิจัย

3.1 การเก็บรวบรวมข้อมูล

ข้อมูลที่นำมาใช้ในการศึกษาครั้งนี้ เป็นข้อมูลทุติยภูมิแบบรายเดือน ตั้งแต่เดือนมกราคม พ.ศ. 2555 ถึงเดือนตุลาคม พ.ศ. 2567 รวมทั้งสิ้น 154 เดือน ที่ได้รับมาจากหลายแหล่ง ได้แก่ คณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย (คปภ.) ธนาคารแห่งประเทศไทย (ธปท.) และสำนักงานนโยบายและยุทธศาสตร์การค้า (สนค.) ทั้งนี้ ข้อมูลประกอบไปด้วยตัวแปรตาม และตัวแปรทำนาย ดังตารางที่ 3-1 โดยข้อมูลที่นำมาใช้ในการวิเคราะห์จะแบ่งออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 ข้อมูลที่ใช้ในการสร้างตัวแบบ หรือข้อมูลชุดฝึกสอน (Training Data) โดยใช้ข้อมูลตั้งแต่เดือนมกราคม พ.ศ. 2555 ถึงเดือนธันวาคม พ.ศ. 2566 รวมทั้งสิ้น 144 เดือน คิดเป็น 93.51% ของข้อมูลทั้งหมด

ส่วนที่ 2 ข้อมูลที่ใช้ในการตรวจสอบความแม่นยำของตัวแบบ หรือข้อมูลชุดทดสอบ (Testing Data) โดยใช้ข้อมูลตั้งแต่เดือนมกราคม พ.ศ. 2567 ถึงเดือนตุลาคม พ.ศ. 2567 รวมทั้งสิ้น 10 เดือน คิดเป็น 6.49% ของข้อมูลทั้งหมด

ตารางที่ 3-1 ตัวแปรตามและตัวแปรทำนายที่เกี่ยวข้อง

ตัวแปร	คำอธิบายตัวแปร	แหล่งข้อมูล
Y	จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (ฉบับ)	คณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย
X_1	จำนวนเงินเอาประกันชีวิต (บาท)	
X_2	เบี้ยประกันภัยรับโดยตรง (บาท)	
X_3	อัตราดอกเบี้ยนโยบาย	ธนาคารแห่งประเทศไทย
X_4	สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP	
X_5	ดัชนีราคาผู้บริโภคทั่วไป (ปีฐาน พ.ศ. 2566)	สำนักงานนโยบายและยุทธศาสตร์การค้า

3.2 การจัดการข้อมูล

ข้อมูลตัวแปรตามและตัวแปรทำนายที่ใช้ในการศึกษา ได้มาจากหน่วยงานที่เกี่ยวข้องหลายแห่ง ซึ่งมีรูปแบบและโครงสร้างข้อมูลแตกต่างกัน เพื่อให้สามารถนำข้อมูลไปวิเคราะห์ได้อย่างมีประสิทธิภาพ จึงต้องดำเนินการจัดการข้อมูลให้เหมาะสม ดังนี้

3.2.1 รวมข้อมูลให้เป็นแฟ้มข้อมูลเดียวกัน

ในการศึกษาครั้งนี้ ข้อมูลที่ใช้ประกอบการวิเคราะห์ถูกจัดเก็บจากหลายหน่วยงาน ได้แก่ คณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย ธนาคารแห่งประเทศไทย และสำนักงานนโยบายและยุทธศาสตร์การค้า เพื่อให้สามารถนำข้อมูลดังกล่าวมาใช้ในการวิเคราะห์ร่วมกันอย่างมีประสิทธิภาพ จึงได้

ดำเนินการตรวจสอบความสอดคล้องของรูปแบบข้อมูลจากแต่ละแหล่ง และทำการรวมข้อมูลทั้งหมดตามช่วงเวลาที่ข้อมูลจากทุกแหล่งมีร่วมกัน ผลลัพธ์ที่ได้คือแฟ้มข้อมูลชุดเดียวที่ประกอบด้วยข้อมูลจากทุกแหล่งในแต่ละเดือนอย่างครบถ้วน

3.2.2 แปลงข้อมูลให้เป็นค่ามาตรฐาน

เนื่องจากตัวแปรทำนายในชุดข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลเชิงปริมาณ ที่มีช่วงของข้อมูลและหน่วยวัดที่แตกต่างกัน ซึ่งอาจส่งผลต่อประสิทธิภาพของตัวแบบเชิงสถิติและตัวแบบการเรียนรู้ของเครื่อง โดยเฉพาะเมื่อตัวแปรมีหน่วยหรือช่วงของข้อมูลที่แตกต่างกันอย่างมาก จึงจำเป็นต้องทำการแปลงข้อมูลหรือการทำให้เป็นมาตรฐาน (Standardization) เพื่อให้ข้อมูลอยู่ในรูปแบบที่สามารถเปรียบเทียบกันได้และลดอิทธิพลของความแตกต่างในหน่วยวัด

ในการศึกษานี้ จะทำการแปลงค่าเฉพาะตัวแปรทำนายให้เป็นค่ามาตรฐาน โดยใช้เทคนิคการแปลงให้อยู่ในรูปคะแนน *z* (*z* score) ที่แต่ละตัวแปรทำนายมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 ซึ่งคิดจากสูตรในสมการที่ (3-1)

$$X_i' = z = \frac{X_i - \overline{X}}{S} \tag{3-1}$$

โดยที่ X_i' หรือ z คือ ค่าของตัวแปรที่แปลงข้อมูลให้เป็นค่ามาตรฐาน ในขณะที่ X_i, $\overline{X}$, และ S คือ ค่าจริงของข้อมูล ค่าเฉลี่ยของตัวแปร และค่าเบี่ยงเบนมาตรฐานของตัวแปร ตามลำดับ

3.3 การวิเคราะห์ข้อมูล

ในการศึกษาครั้งนี้ได้ดำเนินการวิเคราะห์ข้อมูลด้วยโปรแกรม R ซึ่งเป็นเครื่องมือที่มีความสามารถสูงในการจัดการข้อมูล รวมไปถึงสร้างตัวแบบทางสถิติและตัวแบบการเรียนรู้ของเครื่อง โดยการวิเคราะห์จะครอบคลุมทั้งการสร้างตัวแบบการพยากรณ์ การพยากรณ์ และการเปรียบเทียบประสิทธิภาพของตัวแบบ

3.3.1 การสร้างตัวแบบการพยากรณ์

ในการศึกษานี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีที่หลากหลาย โดยได้เลือกตัวแบบการพยากรณ์ 5 ตัวแบบ ที่ครอบคลุมทั้งตัวแบบเชิงสถิติและตัวแบบการเรียนรู้ของเครื่อง ได้แก่ ตัวแบบการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Model; MLR) ตัวแบบการถดถอยเกาส์เซียนผกผัน (Inverse Gaussian Regression Model; IGR) ตัวแบบการถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression Model, DTR) ตัวแบบการถดถอยป่าสุ่ม (Random Forest Regression Model; RFR) และตัวแบบการถดถอยเอ็กซ์จีบูสต์ (XGBoost Regression Model; XGBR)

ก่อนดำเนินการสร้างตัวแบบ ข้อมูลตัวแปรทำนายทั้งหมดได้ถูกแปลงให้อยู่ในรูปค่ามาตรฐาน เพื่อให้ค่าของแต่ละตัวแปรอยู่ในช่วงเดียวกัน จากนั้นดำเนินการสร้างตัวแบบด้วยโปรแกรม R โดยใช้ฟังก์ชันที่เหมาะสมกับแต่ละตัวแบบดังตารางที่ 3-2

ตารางที่ 3-2 ฟังก์ชันและแพ็คเกจสำหรับการวิเคราะห์ข้อมูลแต่ละวิธี

วิธีการวิเคราะห์	ฟังก์ชัน	แพ็คเกจ
1. การถดถอยเชิงส้นพหุคูณ	lm()	stats
2. การถดถอยเกาส์เซียนผกผัน	glm2()	glm2
3. การถดถอยต้นไม้ตัดสินใจ	rpart()	rpart
4. การถดถอยป่าสุ่ม	randomForest()	randomForest
5. การถดถอยเอ็กซ์จีบูสต์	xgboost()	xgboost

ในการวิเคราะห์จะใช้ชุดข้อมูลส่วนที่ 1 ตั้งแต่เดือนมกราคม พ.ศ. 2555 ถึงเดือนธันวาคม พ.ศ. 2566 จำนวน 144 เดือน ในการสร้างตัวแบบพยากรณ์ สำหรับตัวแบบการเรียนรู้ของเครื่องจะมีการปรับแต่งพารามิเตอร์ตามความเหมาะสมของแต่ละตัวแบบด้วยการค้นหาแบบกริด (Grid Search)

3.3.2 การพยากรณ์ข้อมูล

หลังจากที่ได้ดำเนินการสร้างตัวแบบการพยากรณ์จากแต่ละวิธีเรียบร้อยแล้ว ขั้นตอนถัดไปคือการนำตัวแบบเหล่านั้นไปใช้ในการพยากรณ์ค่าของตัวแปรตาม โดยดำเนินการพยากรณ์ข้อมูล 2 ส่วน คือ ข้อมูลที่ใช้ในการสร้างตัวแบบ และข้อมูลที่ใช้ในการตรวจสอบความแม่นยำของตัวแบบ โดยการพยากรณ์ข้อมูลชุดฝึกสอนมีวัตถุประสงค์เพื่อประเมินความสามารถของตัวแบบในการเรียนรู้รูปแบบของข้อมูล ในขณะที่การพยากรณ์ด้วยข้อมูลชุดทดสอบมีวัตถุประสงค์เพื่อตรวจสอบประสิทธิภาพของตัวแบบในการทำนายค่าข้อมูลใหม่ที่แบบจำลองไม่เคยเห็นมาก่อน

3.3.3 การเปรียบเทียบประสิทธิภาพของตัวแบบ

เพื่อระบุว่าตัวแบบใดมีความเหมาะสมสำหรับการนำไปพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ในประเทสไทย การศึกษานี้จะดำเนินการเปรียบเทียบประสิทธิภาพของตัวแบบการพยากรณ์ทั้ง 5 ตัวแบบ โดยใช้เกณฑ์การประเมินประสิทธิภาพของตัวแบบ คือ รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAPE) ซึ่งทั้งสองเกณฑ์ดังกล่าวสามารถสะท้อนถึงความคลาดเคลื่อนระหว่างค่าจริงกับค่าที่พยากรณ์ได้

ทั้งนี้ ค่าที่ได้จากการพยากรณ์ข้อมูลชุดฝึกสอนและข้อมูลชุดทดสอบจะถูกนำมาคำนวณหาค่า RMSE และ MAPE สำหรับแต่ละตัวแบบ เพื่อเปรียบเทียบประสิทธิภาพโดยรวมและพิจารณาว่าตัวแบบใดให้ผลการพยากรณ์ที่แม่นยำและเหมาะสมที่สุดกับข้อมูลในบริบทของการวิจัยนี้

3.4 การสรุปผลการวิเคราะห์ข้อมูล

ในขั้นตอนการสรุปผลการวิเคราะห์ข้อมูลนี้ จะเป็นการนำเสนอผลการดำเนินงานจากกระบวนการวิเคราะห์ โดยแบ่งเนื้อหาออกเป็น 3 ส่วนหลัก ได้แก่ (1) ผลการวิจัย ซึ่งนำเสนอผลการวิเคราะห์สถิติเชิงพรรณนา ผลการแปลงข้อมูลให้เป็นค่ามาตรฐานของตัวแปรทำนาย และผลการวิเคราะห์จากวิธีการสร้างตัวแบบพยากรณ์ทั้ง 5 วิธี (2) สรุปและอภิปรายผลการวิจัย โดยแสดงผลการประเมินตัวแบบ รวมถึงวิเคราะห์และเปรียบเทียบประสิทธิภาพของตัวแบบ และ (3) ข้อเสนอแนะที่มุ่งเน้นแนวทางการนำผลการวิจัยไปประยุกต์ใช้และพัฒนาต่อยอดในเชิงวิชาการและการใช้งานจริงในอนาคต

บทที่ 4
ผลการวิจัย

4.1 ผลการวิจัย

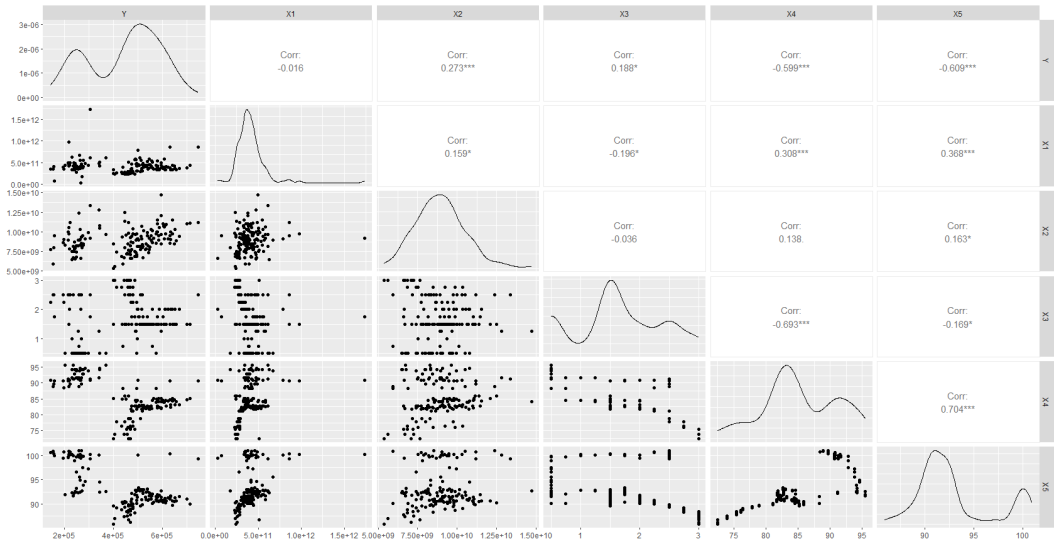
ผลการศึกษาวิจัยสำหรับการเปรียบเทียบตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตในประเทศไทยนั้น จะนำเสนอผลการวิจัยออกเป็นผลการวิเคราะห์สถิติเชิงพรรณนาของตัวแปรที่ศึกษา ผลการแปลงข้อมูลให้เป็นค่ามาตรฐานของตัวแปรทำนาย และผลการสร้างตัวแบบการพยากรณ์ของตัวแบบต่าง ๆ ดังนี้

4.1.1 ผลการวิเคราะห์สถิติเชิงพรรณนาของตัวแปรที่ศึกษา

ในการดำเนินการสร้างตัวแบบการพยากรณ์สำหรับการศึกษานี้ ได้เริ่มต้นด้วยการวิเคราะห์สถิติเชิงพรรณนา (Descriptive Statistics) ของตัวแปรที่ใช้ในการศึกษา ซึ่งมีวัตถุประสงค์เพื่อทำความเข้าใจลักษณะเบื้องต้นของข้อมูลก่อนนำไปใช้ในการพัฒนาตัวแบบ โดยผลการวิเคราะห์ประกอบด้วยการนำเสนอค่าสถิติพื้นฐาน ได้แก่ ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน ค่าต่ำสุด ค่าสูงสุด ความเบ้ และความโด่ง ของตัวแปรตามและตัวแปรทำนาย นอกจากนี้ ยังมีการนำเสนอผลการวิเคราะห์ในรูปแบบของแผนภาพความหนาแน่น (Density Plot) ของแต่ละตัวแปร เพื่อแสดงการกระจายของข้อมูลแต่ละตัวแปร และช่วยให้มองเห็นลักษณะของข้อมูลได้อย่างชัดเจน รวมไปถึงแผนภาพการกระจาย (Scatter Plot) และค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient) เพื่อแสดงความสัมพันธ์ระหว่างตัวแปรแต่ละคู่ ผลการวิเคราะห์สถิติเบื้องต้นนี้ครอบคลุมข้อมูลจำนวน 154 เดือน โดยสรุปผลลัพธ์ของค่าทางสถิติดังกล่าวได้แสดงไว้ในตารางที่ 4-1 และภาพรวมของลักษณะข้อมูลในแต่ละตัวแปรได้แสดงไว้ในภาพที่ 4-1.

ตารางที่ 4-1 สรุปค่าสถิติเชิงพรรณนาของตัวแปรที่ศึกษา

ตัวแปร	ค่าเฉลี่ย ($\overline{X}$)	ค่าเบี่ยงเบนมาตรฐาน (S.D.)	ค่าต่ำสุด (Minimum)	ค่าสูงสุด (Maximum)
Y	443,980.30	148,406.70	143,908.00	747,211.00
X_1	407,938,894,937.00	169,186,150,901.00	25,305,868,036.00	1,739,773,000,000.00
X_2	8,936,520,930.00	1,565,710,580.00	5,387,983,328.00	14,720,662,000.00
X_3	1.66	0.74	0.50	3.00
X_4	85.36	5.72	72.50	95.50
X_5	92.95	3.93	85.90	100.98



ภาพที่ 4-1 แผนภาพความหนาแน่น แผนภาพการกระจาย และค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของตัวแปรที่ศึกษา

จากผลการวิเคราะห์สถิติเชิงพรรณนาของตัวแปรที่ใช้ในการศึกษาจากตารางที่ 4-1 และแผนภาพความหนาแน่นจากภาพที่ 4-1 สามารถสรุปลักษณะของตัวแปรแต่ละตัวที่ศึกษาได้ ดังนี้

ตัวแปรตาม (*Y*) หรือจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีในประเทศไทย มีค่าเฉลี่ยอยู่ที่ 443,980 ฉบับ และค่าเบี่ยงเบนมาตรฐานเท่ากับ 148,407 ฉบับ โดยมีจำนวนกรมธรรม์ต่ำสุดและสูงสุดเท่ากับ 143,908 และ 747,211 ฉบับ ตามลำดับ นอกจากนี้ จะเห็นได้ว่าตัวแปรตาม *Y* มีลักษณะการแจกแจงเป็นแบบสองยอด (Bimodal Distribution) แสดงว่าข้อมูลมีการกระจุกตัวอยู่ในสองช่วงนี้ ซึ่งอาจสะท้อนถึงเหตุการณ์สำคัญที่ส่งผลต่อพฤติกรรมของผู้ทำประกันชีวิต

สำหรับตัวแปรทำนายตัวแรก คือ จำนวนเงินเอาประกันชีวิต (X_1) ซึ่งมีค่าเฉลี่ยอยู่ที่ประมาณ 407,938 ล้านบาท และค่าเบี่ยงเบนมาตรฐานประมาณ 169,186 ล้านบาท โดยมีค่าต่ำสุดประมาณ 25,306 ล้านบาท และค่าสูงสุดมากถึง 1,739,773 ล้านบาท สะท้อนถึงความหลากหลายของมูลค่าประกันชีวิตในแต่ละช่วงเวลา อีกทั้งข้อมูลแสดงให้เห็นถึงการแจกแจงที่เบ้ขวา (Right-Skewed) อย่างชัดเจน ซึ่งบ่งชี้ว่ามีเพียงบางช่วงเวลาที่มีจำนวนเงินเอาประกันชีวิตในมูลค่าสูงเป็นพิเศษ

ในส่วนของตัวแปรเบี้ยประกันภัยรับโดยตรง (X_2) เป็นเบี้ยประกันภัยที่บริษัทประกันภัยได้รับก่อนหักเบี้ยประกันภัยต่อ (เป็นเบี้ยประกันภัยหลังหักภาษีอากร) ซึ่งพบว่ามีค่าเฉลี่ยอยู่ที่ประมาณ 8,937 ล้านบาท และค่าเบี่ยงเบนมาตรฐานประมาณ 1,566 ล้านบาท โดยมีค่าต่ำสุดและค่าสูงสุดประมาณ 5,388 และ 14,721 ล้านบาท ตามลำดับ นอกจากนี้ จะเห็นได้ว่าการแจกแจงข้อมูลมีลักษณะเบ้ทางขวาเพียงเล็กน้อย อย่างไรก็ตาม โดยรวมแล้วข้อมูลมีการกระจายตัวที่ค่อนข้างสมมาตร

สำหรับตัวแปรอัตราดอกเบี้ยนโยบาย (X_3) ซึ่งเป็นตัวแปรด้านนโยบายการเงินของประเทศ มีค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานเท่ากับ 1.66 และ 0.74 ตามลำดับ โดยมีอัตราดอกเบี้ยต่ำสุดและสูงสุด คือ 0.50 และ 3.00 ตามลำดับ จากการพิจารณากราฟความหนาแน่น พบว่าข้อมูลมีการกระจุกตัวในระดับอัตราดอกเบี้ยต่ำ และมีลักษณะเบ้ขวาเล็กน้อย ซึ่งสอดคล้องกับสภาวะอัตราดอกเบี้ยต่ำที่เกิดขึ้นในช่วงหลายปีที่ผ่านมา

ในกรณีของตัวแปรสัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) ที่สะท้อนถึงระดับหนี้สินของภาคครัวเรือนเมื่อเทียบกับขนาดของเศรษฐกิจโดยรวม ซึ่งพบว่ามีค่าเฉลี่ยสัดส่วนเงินให้กู้ยืมอยู่ที่ 85.36 และค่าเบี่ยงเบนมาตรฐาน 5.72 โดยมีสัดส่วนเงินให้กู้ยืมต่ำสุดและสูงสุดอยู่ที่ 72.50 และ 95.50 ตามลำดับ อีกทั้งข้อมูลค่อนข้างมีการกระจายที่สมมาตร ซึ่งสะท้อนว่าสัดส่วนหนี้ครัวเรือนต่อ GDP ในช่วงเวลาศึกษาไม่มีแนวโน้มของการเปลี่ยนแปลงอย่างเฉียบพลัน

สุดท้ายคือตัวแปรดัชนีราคาผู้บริโภคทั่วไป (X_5) ซึ่งเป็นตัวชี้วัดระดับราคาสินค้าและบริการในแต่ละช่วงเวลา ที่แสดงให้เห็นถึงภาวะเงินเฟ้อที่เกิดขึ้นในระบบเศรษฐกิจ จากผลการวิเคราะห์พบว่ามีค่าเฉลี่ยอยู่ที่ 92.95 ค่าเบี่ยงเบนมาตรฐาน 3.93 โดยมีดัชนีราคาต่ำสุดเท่ากับ 85.90 และดัชนีราคาสูงสุดเท่ากับ 100.98 ซึ่งค่อนข้างใกล้เคียงกับปีฐาน พ.ศ. 2566 แสดงว่าหมายความว่าราคาสินค้าและบริการเพิ่มขึ้นไม่มากนักเมื่อเทียบกับปีฐาน ทั้งนี้ ยังเห็นได้ว่าลักษณะการแจกแจงของข้อมูลค่อนข้างเบ้ขวา หมายความว่าเมื่อเทียบกับปีฐาน ดัชนีราคาผู้บริโภคส่วนใหญ่ต่ำกว่าปีฐาน ซึ่งบ่งชี้ว่าระดับราคาสินค้าและบริการในช่วงก่อนถึงปีฐานอยู่ในระดับที่ต่ำกว่าเกณฑ์อ้างอิงปัจจุบัน

นอกจากนี้ เมื่อพิจารณาแผนภาพการกระจายและค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรตามและตัวแปรทำนายทั้ง 5 ตัวแปร ดังภาพที่ 4-1 พบว่า จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) มีความสัมพันธ์เชิงเส้นตรงกับเบี้ยประกันภัยรับโดยตรง (X_2) และอัตราดอกเบี้ยนโยบาย (X_3) ในระดับน้อยและไปในทิศทางเดียวกัน ซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันเท่ากับ 0.273 และ 0.188 ที่ระดับนัยสำคัญ 0.001 และ 0.05 ตามลำดับ ในขณะที่จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) มีความสัมพันธ์เชิงเส้นตรงกับสัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) และดัชนีราคาผู้บริโภคทั่วไป (X_5) ในทิศทางตรงกันข้าม ที่ระดับนัยสำคัญ 0.001 มีค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันเท่ากับ −0.599 และ −0.609 แสดงว่ามีตัวแปรตามและตัวแปรทำนายทั้งสองมีความสัมพันธ์กันในระดับปานกลางและมาก ตามลำดับ ในขณะที่มีเพียงจำนวนเงินเอาประกันชีวิต (X_1) เท่านั้น ที่ไม่มีความสัมพันธ์กับจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) ที่ระดับนัยสำคัญ 0.05

4.1.2 ผลการแปลงข้อมูลให้เป็นค่ามาตรฐานของตัวแปรทำนาย

ก่อนทำการแปลงข้อมูลให้อยู่ในรูปแบบค่ามาตรฐาน หรือให้อยู่ในรูปของคะแนน z จำเป็นต้องพิจารณาลักษณะของข้อมูลเบื้องต้นผ่านค่าสถิติเชิงพรรณนา โดยตัวแปรทำนายที่ใช้ในการศึกษา ได้แก่ จำนวนเงินเอาประกันชีวิต (X_1) เบี้ยประกันภัยรับโดยตรง (X_2) อัตราดอกเบี้ยนโยบาย (X_3) สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) และดัชนีราคาผู้บริโภคทั่วไป (X_5)

จากข้อมูลของตัวแปรทำนายที่เกี่ยวข้องในตารางที่ 3-1 จะเห็นได้ว่าตัวแปรทำนายทั้ง 5 ตัวแปรมีหน่วยที่แตกต่างกัน โดยที่จำนวนเงินเอาประกันชีวิต (X_1) และเบี้ยประกันภัยรับโดยตรง (X_2) จะมีหน่วยเหมือนกัน คือ บาท ในขณะที่อีก 3 ตัวแปร คือ อัตราดอกเบี้ยนโยบาย (X_3) สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) และดัชนีราคาผู้บริโภคทั่วไป (X_5) จะไม่มีหน่วย นอกจากนี้ ผลการวิเคราะห์ในหัวข้อ 4.1.1 ยังแสดงให้เห็นว่า ตัวแปรในชุดข้อมูลที่ใช้ในการศึกษานี้มีช่วงของค่าที่แตกต่างกันอย่างมาก โดยจะเห็นได้ว่าตัวแปรอย่างจำนวนเงินเอาประกันชีวิต (X_1) และเบี้ยประกันภัยรับโดยตรง (X_2) มีค่าเฉลี่ยอยู่ในระดับพันล้านและ

แสนล้านบาท และมีค่าเบี่ยงเบนมาตรฐานที่สูงมาก ในขณะที่ตัวแปรอื่น ๆ อย่างอัตราดอกเบี้ยนโยบาย (X_3) สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) และดัชนีราคาผู้บริโภคทั่วไป (X_5) มีค่าเฉลี่ยอยู่ในช่วงหลักหน่วยหรือหลักสิบเท่านั้น ซึ่งความแตกต่างในขนาดของตัวแปรเหล่านี้อาจส่งผลต่อการวิเคราะห์ได้

การแปลงข้อมูลตัวแปรทำนายให้เป็นคะแนน z ช่วยให้ข้อมูลของทุกตัวแปรมีค่าเฉลี่ยเท่ากับ 0 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 1 ซึ่งทำให้ข้อมูลของตัวแปรทั้งหมดอยู่ในมาตรฐานเดียวกัน และสามารถนำข้อมูลไปวิเคราะห์ร่วมกันได้โดยไม่มีผลกระทบจากหน่วยของแต่ละตัวแปร ซึ่งเป็นประโยชน์อย่างยิ่งในกรณีที่ตัวแปรมีช่วงของค่าที่แตกต่างกันมาก โดยผลการแปลงค่าของตัวแปรทำนาย 12 แถวแรก และสถิติเชิงพรรณนาของตัวแปรที่แปลงเป็นค่ามาตรฐานแล้วแสดงดังตารางที่ 4-2

ตารางที่ 4-2 ผลการแปลงค่าของตัวแปรทำนาย 12 แถวแรก และค่าสถิติเชิงพรรณนาของตัวแปรทำนายที่แปลงเป็นค่ามาตรฐาน

แถวที่	ปี	เดือน	X_1	X_2	X_3	X_4	X_5
1	2555	มกราคม	−0.9759	−2.2664	1.8140	−2.2494	−1.7944
2	2555	กุมภาพันธ์	−0.9016	−1.1735	1.8140	−2.2494	−1.7156
3	2555	มีนาคม	0.6173	−0.8203	1.8140	−2.2494	−1.5858
4	2555	เมษายน	−1.0480	−2.1260	1.8140	−2.0045	−1.4916
5	2555	พฤษภาคม	−0.9169	−1.4332	1.8140	−2.0045	−1.4051
6	2555	มิถุนายน	−0.8464	−1.0680	1.8140	−2.0045	−1.3669
7	2555	กรกฎาคม	−0.7177	−1.4786	1.8140	−1.7421	−1.2931
8	2555	สิงหาคม	−1.0225	−1.3305	1.8140	−1.7421	−1.2015
9	2555	กันยายน	−1.0402	−1.1124	1.8140	−1.7421	−1.1302
10	2555	ตุลาคม	−0.9492	−0.9053	1.4766	−1.6196	−1.0997
11	2555	พฤศจิกายน	−0.7447	−0.2901	1.4766	−1.6196	−1.1786
12	2555	ธันวาคม	−1.0696	0.5861	1.4766	−1.6196	−1.0920
ค่าเฉลี่ย ($\overline{X}$)			0	0	0	0	0
ค่าเบี่ยงเบนมาตรฐาน (S.D.)			1	1	1	1	1
ค่าต่ำสุด (Minimum)			−2.26	−2.27	−1.56	−2.25	−1.79
ค่าสูงสุด (Maximum)			7.87	3.69	1.81	1.77	2.04

ในขั้นตอนถัดไปของการศึกษา จะเป็นการนำข้อมูลที่ผ่านการแปลงแล้วในส่วนของตัวแปรทำนายมาใช้ในการวิเคราะห์เชิงปริมาณ นั่นคือ สร้างตัวแบบทางสถิติและตัวแบบการเรียนรู้ของเครื่อง เพื่อศึกษาตัวแปรที่มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) รวมถึงนำไปประยุกต์ใช้ในการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีต่อไป

4.1.3 ผลการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

สำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ได้ดำเนินการตรวจสอบข้อตกลงเบื้องต้นของข้อมูลที่ใช้ในการสร้างตัวแบบ หรือข้อมูลชุดฝึกสอน โดยการตรวจสอบความเป็นอิสระต่อกันของตัวแปรทำนาย การตรวจสอบการแจกแจงของส่วนเหลือ (Residual) การตรวจสอบความเท่ากันของความแปรปรวนของส่วนเหลือ การตรวจสอบความเป็นอิสระต่อกันของส่วนเหลือ และการตรวจสอบค่าผิดปกติ

เพื่อตรวจสอบความเป็นอิสระต่อกันของตัวแปรทำนาย ในการวิเคราะห์นี้ใช้ค่า Variance Inflation Factor (VIF) ในการตรวจสอบ โดยหากค่า VIF เกิน 10 แสดงว่าตัวแปรทำนายมีความสัมพันธ์กัน จากผลลัพธ์ในตารางที่ 4-3 พบว่า ค่า VIF ของตัวแปรทำนายทั้งหมดน้อยกว่า 10 แสดงว่า ไม่มีความสัมพันธ์กันระหว่างตัวแปรทำนายแต่ละตัว ในขณะที่ ภาพที่ 4-2(ก.) แสดงให้เห็นว่า ส่วนเหลือกระจายรอบ ๆ ศูนย์ในภาพรวม แต่มีลักษณะการกระจายที่ไม่สม่ำเสมอ และมีความกระจุกตัวในบางช่วง จากการทดสอบสมมุติฐานด้วยสถิติทดสอบของ Breusch-Pagan ในตารางที่ 4-4 พบว่า ความแปรปรวนของส่วนเหลือไม่เท่ากัน ที่ระดับนัยสำคัญ 0.05 โดยมีค่าสถิติทดสอบเท่ากับ 14.334 นอกจากนี้ จากการตรวจสอบค่ามาตรฐานของส่วนเหลือในภาพที่ 4-2(ก.) พบว่า มีค่าไม่เกินค่า |3| ซึ่งแสดงว่า ข้อมูลไม่มีค่าผิดปกติ (ฉัตรศิริ ปิยะพิมลสิทธิ์, 2547)

จากภาพที่ 4-2(ข.) แสดงให้เห็นว่าค่ามาตรฐานของส่วนมีการกระจายบ่ายเบนไปจากแนวเส้นอ้างอิงเพียงเล็กน้อย และผลการทดสอบสมมุติฐานด้วยสถิติทดสอบของ Anderson-Darling ในตารางที่ 4-4 (เนื่องจากมีจำนวนข้อมูลมากกว่า 50) พบว่า ส่วนเหลือมีการแจกแจงแบบปรกติ ที่ระดับนัยสำคัญ 0.05 โดยมีค่าสถิติทดสอบเท่ากับ 0.675 นอกจากนี้ ในการตรวจสอบความเป็นอิสระต่อกันของส่วนเหลือด้วยการทดสอบสมมุติฐาน พบว่า ส่วนเหลือมีปัญหาเรื่องสหสัมพันธ์ในตัว (Autocorrelation) หรือส่วนเหลือไม่เป็นอิสระต่อกันที่ระดับนัยสำคัญ 0.001 โดยมีค่าสถิติทดสอบของ Durbin-Watson เท่ากับ 0.8148

จากการตรวจสอบข้อสมมุติเบื้องต้นของการวิเคราะห์การถดถอยพหุคูณ พบว่า ไม่เป็นไปตามข้อสมมุติเบื้องต้นในเรื่องของความเท่ากันของความแปรปรวนและความเป็นอิสระต่อกันของส่วนเหลือ ซึ่งอาจส่งผลกระทบต่อความถูกต้องและความน่าเชื่อถือของผลการประมาณค่าสัมประสิทธิ์ถดถอยได้

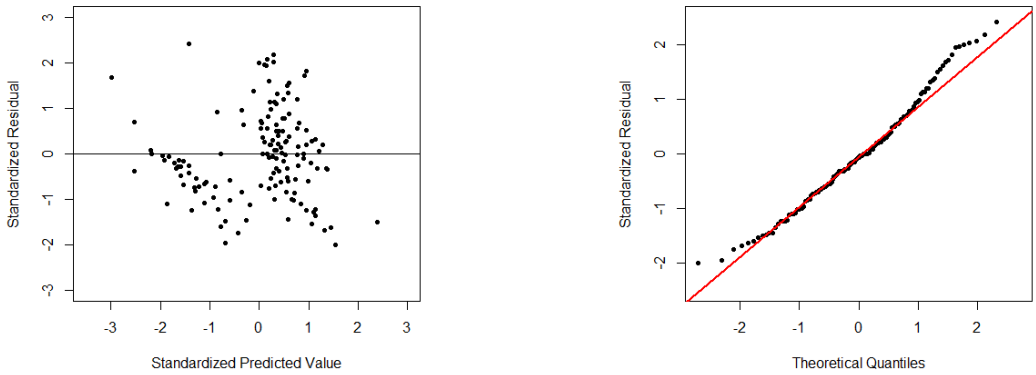
ตารางที่ 4-3 ค่า VIF ของตัวแปรทำนายแต่ละตัว

ตัวแปรทำนาย				
X_1	X_2	X_3	X_4	X_5
1.2220	1.0298	3.6617	6.3756	2.8896

ตารางที่ 4-4 ตรวจสอบการแจกแจงปกติ ความเท่ากันของความแปรปรวน และความเป็นอิสระกันของส่วนเหลือ

การทดสอบสมมุติฐาน	สถิติทดสอบ	ค่าพี
1. ตรวจสอบการแจกแจงปกติ	Anderson-Darling = 0.675	0.076
2. ตรวจสอบความเท่ากันของความแปรปรวน	Breusch-Pagan = 14.334	0.015 *
3. ตรวจสอบความเป็นอิสระกัน (Lag = 1)	Durbin-Watson = 0.815	< 0.001 ***

* $p < 0.05$, *** $p < 0.001$



(ก.) (ข.)

ภาพที่ 4-2 (ก.) แผนภาพการกระจายระหว่างค่ามาตรฐานของส่วนเหลือและค่ามาตรฐานของค่าทำนาย และ (ข.) แผนภาพ Normal Quantile-Quantile (Normal Q-Q Plot)

ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยการนำเข้าตัวแปรทำนายทั้งหมด (วิธี Enter) เพื่อประเมินอิทธิพลของตัวแปรทำนายแต่ละตัวที่มีต่อตัวแปรตามหรือจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) ผลลัพธ์ที่ได้จากการวิเคราะห์จะสรุปรวมไว้ในตารางที่ 4-5

ตารางที่ 4-5 ผลการประมาณค่าประมาณพารามิเตอร์ ความคลาดเคลื่อนมาตรฐาน และการทดสอบสมมุติฐานตัวแปรทำนายแต่ละตัว จากการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

ตัวแปรทำนาย	ค่าประมาณพารามิเตอร์	ค่ามาตรฐานของค่าประมาณพารามิเตอร์	ความคลาดเคลื่อนมาตรฐาน	ค่าสถิติทดสอบ t	ค่าพี
ค่าคงที่	443,908	–	7,260	61.147	< 0.001 ***
X_1	19,402	0.141	7,812	2.483	0.014 *
X_2	56,789	0.403	7,340	7.737	< 0.001 ***
X_3	−54,041	−0.385	13,768	−3.925	< 0.001 ***
X_4	−118,395	−0.865	17,730	−6.678	< 0.001 ***
X_5	−31,861	−0.206	13,508	−2.359	0.020 *
R^2 = 0.6369, Adjusted R^2 = 0.6238, $F_{(5,138)}$ = 48.41***					
AIC = 3685.725, Log Likelihood = −1835.862					

* $p < 0.05$, *** $p < 0.001$

ผลลัพธ์ในตารางที่ 4-5 พบว่า ตัวแบบการถดถอยเชิงเส้นพหุคูณมีค่าสัมประสิทธิ์แสดงการตัดสินใจที่ปรับค่าแล้ว (Adjusted R^2) เท่ากับ 0.6238 หมายความว่า ตัวแปรทำนายทั้งหมดในตัวแบบนี้สามารถอธิบายความผันแปรทั้งหมดของจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) ได้ร้อยละ 62.38

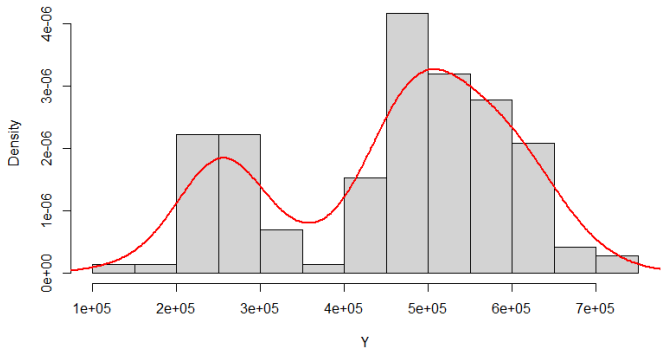
ทั้งนี้ จากการนำตัวแปรทำนายทุกตัวเข้ามาในตัวแบบ พบว่า ตัวแปรทำนายทั้ง 5 ตัวแปร มีความสัมพันธ์กับจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) อย่างมีนัยสำคัญทางสถิติ และสามารถเขียนสมการถดถอยได้ ดังนี้

$$Y = 443{,}908 + 19{,}402X_1 + 56{,}789X_2 - 54{,}041X_3 - 118{,}395X_4 - 31{,}861X_5 \qquad (4\text{-}1)$$

นอกจากนี้ เมื่อเรียงลำดับความสำคัญของตัวแปรทำนายที่มีอิทธิพลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) จากมากไปน้อย โดยการพิจารณาค่ามาตรฐาน (Standardized) ของค่าประมาณพารามิเตอร์ พบว่า สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) มีอิทธิพลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีมากที่สุด เนื่องจากค่าสัมบูรณ์ของค่ามาตรฐานของสัมประสิทธิ์การถดถอยเท่ากับ 0.865 รองลงมา คือ เบี้ยประกันภัยรับโดยตรง (X_2) อัตราดอกเบี้ยนโยบาย (X_3) ดัชนีราคาผู้บริโภคทั่วไป (X_5) และจำนวนเงินเอาประกันชีวิต (X_1) ตามลำดับ

4.1.4 ผลการวิเคราะห์การถดถอยเกาส์เซียนผกผัน

เพื่อประเมินความเหมาะสมของการใช้ตัวแบบการถดถอยเกาส์เซียนผกผัน ก่อนที่จะดำเนินการวิเคราะห์ข้อมูลด้วยการวิเคราะห์การถดถอยเกาส์เซียนผกผัน จึงเริ่มจากการตรวจสอบลักษณะการแจกแจงของข้อมูลของตัวแปรตาม นั่นคือ จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) เพื่อพิจารณาว่าข้อมูลดังกล่าวมีลักษณะสอดคล้องกับการแจกแจงเกาส์เซียนผกผัน (Inverse Gaussian; IG) หรือไม่ จากการพิจารณาเบื้องต้นด้วยกราฟฮิสโทแกรม (Histogram) ร่วมกับเส้นความหนาแน่น (Density Line) ของข้อมูลชุดฝึกสอน ดังภาพที่ 4-3 แสดงให้เห็นว่าจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) มีลักษณะการแจกแจงเป็นแบบสองยอด โดยมีช่วงค่าที่กระจุกตัวอยู่ราว 250,000 และ 500,000 ฉบับ อย่างไรก็ตาม เมื่อพิจารณารูปแบบของการกระจายโดยรวมของข้อมูลยังคงใกล้เคียงสมมาตร ถึงแม้จะมีลักษณะเป็นสองยอดอยู่บ้าง แต่อาจมีแนวโน้มเบ้ซ้ายเล็กน้อย หมายความว่า ข้อมูลดังกล่าวยังไม่สอดคล้องกับแจกแจงเกาส์เซียนผกผันที่มักมีลักษณะเบ้ขวา



ภาพที่ 4-3 ฮิสโทแกรมและเส้นความหนาแน่นของจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (ข้อมูลชุดฝึกสอน)

นอกจากนี้ สำหรับการตรวจสอบลักษณะการแจกแจงของข้อมูลด้วยการทดสอบสมมุติฐานทางสถิติ ได้มีการกำหนดสมมุติฐานไว้ (Fathurahman, 2022) ดังนี้

$H_0 : Y_1 = Y_2 = \cdots = Y_n \sim \text{IG}(\mu, \lambda)$
(Y_i มีการแจกแจงเกาส์เซียนผกผัน)
$H_1 : Y_i$ อย่างน้อยหนึ่งค่าไม่เป็นไปตามการแจกแจง สำหรับ $i = 1, 2, \ldots, n$

ในการดำเนินการทดสอบสมมุติฐานจะใช้ฟังก์ชัน ig_test() ในแพ็คเกจ goft ซึ่งประกอบด้วย 2 วิธีหลัก คือ วิธีการแปลงข้อมูล และวิธีการเปรียบเทียบอัตราส่วนความแปรปรวน ทั้งนี้ ได้ทำการทดสอบสมมุติฐาน 3 วิธี ได้แก่ การแปลงข้อมูลให้อยู่ในรูปแบบใกล้เคียงการแจกแจงปรกติ การแปลงข้อมูลให้อยู่ในรูปแบบการแจกแจงแกมมา ที่มีพารามิเตอร์บ่งรูปร่างเท่ากับ 0.5 และการเปรียบเทียบอัตราส่วนความแปรปรวน โดยมีผลการทดสอบสมมุติฐานดังตารางที่ 4-6

ตารางที่ 4-6 ผลการตรวจสอบลักษณะการแจกแจงเกาส์เซียนผกผันของจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y)

วิธี		สถิติทดสอบ	ค่าพี
แปลงข้อมูล	ใกล้เคียงการแจกแจงปรกติ	Shapiro-Wilk = 0.990	0.394
	การแจกแจงแกมมา	Anderson-Darling = 1.867	0.021 *
เปรียบเทียบอัตราส่วนความแปรปรวน		T_1 = −4.241	< 0.001 ***

* $p < 0.05$, *** $p < 0.001$

จากตารางที่ 4-6 พบว่า การแปลงข้อมูลให้อยู่ในรูปแบบใกล้เคียงการแจกแจงปกติ แล้วทดสอบด้วยสถิติทดสอบ Shapiro-Wilk ได้ค่าสถิติทดสอบเท่ากับ 0.990 และค่าพีเท่ากับ 0.3941 ซึ่งมากกว่าระดับนัยสำคัญ 0.05 ดังนั้น จึงไม่สามารถปฏิเสธสมมุติฐานที่ว่า Y_i มีการแจกแจงเกาส์เซียนผกผันได้ ในขณะที่การแปลงข้อมูลให้อยู่ในรูปของการแจกแจงแบบแกมมา แล้วใช้สถิติทดสอบ Anderson-Darling ในการตรวจสอบ ซึ่งให้ค่าสถิติทดสอบเท่ากับ 1.8674 และมีค่าพีเท่ากับ 0.021 ซึ่งมากกว่าระดับนัยสำคัญ 0.05 ทำให้ปฏิเสธสมมุติฐานที่ว่า Y_i มีการแจกแจงเกาส์เซียนผกผัน ในขณะเดียวกัน ผลจากการทดสอบด้วยวิธีการเปรียบเทียบอัตราส่วนความแปรปรวน ให้ค่าสถิติทดสอบเท่ากับ −4.2412 และค่าพีน้อยกว่า 0.001 จึงสรุปได้ว่าข้อมูลดังกล่าวไม่เป็นไปตามการแจกแจงเกาส์เซียนผกผัน

ถึงแม้ว่าผลจากการทดสอบสมมุติฐานจากการแปลงข้อมูลให้อยู่ในรูปแบบใกล้เคียงการแจกแจงปกติ จะไม่สามารถปฏิเสธสมมุติฐานหลักที่ว่าจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) มีการแจกแจงเกาส์เซียนผกผันได้ แต่จากฮิสโทแกรมและเส้นความหนาแน่น รวมถึงผลการทดสอบสมมุติฐานจากการแปลงข้อมูลให้อยู่ในรูปของการแจกแจงแบบแกมมาและการเปรียบเทียบอัตราส่วนความแปรปรวนกลับให้ผลในทางตรงกันข้าม กล่าวคือ ปฏิเสธสมมุติฐานหลักที่ว่าจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) มีการ

แจกแจงเกาส์เซียนผกผัน ดังนั้น จึงควรพิจารณาอย่างรอบคอบในการเลือกใช้ตัวแบบการถดถอยเกาส์เซียนผกผัน และในการศึกษานี้จึงพิจารณาเปรียบเทียบตัวแบบการถดถอยเกาส์เซียนผกผันกับตัวแบบทางเลือกอื่น ๆ ได้แก่ ตัวแบบการถดถอยพหุคูณ ตัวแบบการถดถอยต้นไม้ตัดสินใจ ตัวแบบการถดถอยป่าสุ่ม และตัวแบบการถดถอยเอ็กซ์จีบูสต์ เพื่อหาตัวแบบที่เหมาะสมสำหรับการพยากรณ์และศึกษาตัวแปรทำนายที่มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีต่อไป

ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยการวิเคราะห์การถดถอยเกาส์เซียนผกผัน โดยใช้ฟังก์ชันเชื่อมโยงแบบลอการิทึม และนำเข้าตัวแปรทำนายทั้งหมด (วิธี Enter) เพื่อประเมินอิทธิพลของตัวแปรทำนายแต่ละตัวที่มีต่อตัวแปรตามหรือจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) ผลลัพธ์ที่ได้จากการวิเคราะห์จะสรุปรวมไว้ในตารางที่ 4-7

ตารางที่ 4-7 ผลการประมาณค่าประมาณพารามิเตอร์ ความคลาดเคลื่อนมาตรฐาน และการทดสอบสมมุติฐานตัวแปรทำนายแต่ละตัว จากการวิเคราะห์การถดถอยเกาส์เซียนผกผัน

ตัวแปรทำนาย	ค่าประมาณพารามิเตอร์ ($\hat{\beta}_j$)	$\exp(\hat{\beta}_j)$	ค่ามาตรฐานของค่าประมาณพารามิเตอร์	ความคลาดเคลื่อนมาตรฐาน	ค่าสถิติทดสอบ t	ค่าพี
ค่าคงที่	12.960	426992.200	–	0.019	700.542	< 0.001 ***
X_1	0.067	1.069	0.00000046	0.018	3.618	< 0.001 ***
X_2	0.132	1.141	0.00000083	0.019	6.863	< 0.001 ***
X_3	−0.117	0.890	−0.00000079	0.032	−3.637	< 0.001 ***
X_4	−0.331	0.718	−0.00000228	0.041	−8.027	< 0.001 ***
X_5	−0.082	0.921	−0.00000054	0.030	−2.761	0.007 **
AIC = 3706.139, Log Likelihood = −1846.069						

* $p < 0.05$, *** $p < 0.001$

จากผลการวิเคราะห์ข้อมูลด้วยตัวแบบการถดถอยเกาส์เซียนผกผันในตารางที่ 4-7 พบว่า ตัวแปรทำนายทั้ง 5 ตัวแปรที่นำเข้ามาในตัวแบบการถดถอยเกาส์เซียนผกผัน มีความสัมพันธ์กับจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) อย่างมีนัยสำคัญทางสถิติ และสามารถเขียนให้อยู่ในรูปสมการถดถอยได้ ดังนี้

$$\ln(\hat{\mu}) = 12.960 + 0.067X_1 + 0.132X_2 - 0.117X_3 - 0.331X_4 - 0.082X_5 \quad (4\text{-}2)$$

หรือ

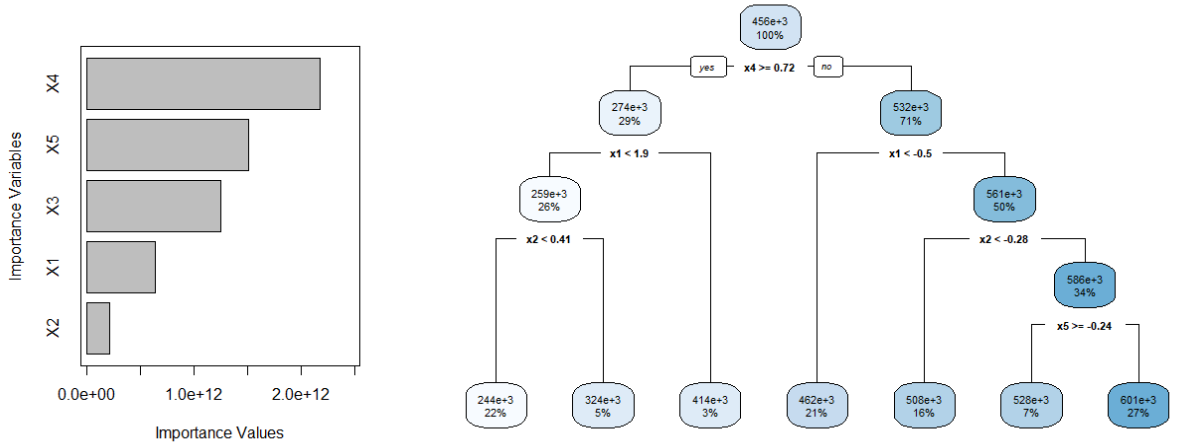
$$\hat{\mu} = \exp(12.960 + 0.067X_1 + 0.132X_2 - 0.117X_3 - 0.331X_4 - 0.082X_5) \quad (4\text{-}3)$$

นอกจากนี้ เมื่อเรียงลำดับความสำคัญของตัวแปรทำนายที่มีอิทธิพลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) จากมากไปน้อย โดยการพิจารณาค่ามาตรฐาน (Standardized) ของค่าประมาณพารามิเตอร์ พบว่า สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) มีอิทธิพลต่อจำนวนกรมธรรม์

ประกันชีวิตที่ทำใหม่ระหว่างปีมากที่สุด เนื่องจากค่าสัมบูรณ์ของค่ามาตรฐานของสัมประสิทธิ์การถดถอยเท่ากับ 0.041 รองลงมา คือ อัตราดอกเบี้ยนโยบาย (X_3) ดัชนีราคาผู้บริโภคทั่วไป (X_5) เบี้ยประกันภัยรับโดยตรง (X_2) และจำนวนเงินเอาประกันชีวิต (X_1) ตามลำดับ

4.1.5 ผลการวิเคราะห์การถดถอยต้นไม้ตัดสินใจ

ในการวิเคราะห์ข้อมูลด้วยต้นไม้ตัดสินใจ เพื่อให้ได้แบบตัวแบบที่มีประสิทธิภาพดี จึงได้ทำการปรับแต่งไฮเปอร์พารามิเตอร์ (Hyperparameter Tuning) ในแบบตัวแบบ ด้วยวิธีการค้นหาแบบกริด (Grid Search) โดยกำหนดให้ช่วงของจำนวนข้อมูลต่ำสุดที่จำเป็นต้องมีก่อนที่จะทำการสร้างโหนด (minsplit) อยู่ระหว่าง 2 ถึง 20 และช่วงของจำนวนโหนดภายในสูงสุดระหว่างโหนดรากและโหนดปลาย (maxdepth) อยู่ระหว่าง 2 ถึง 15 จากการค้นหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม พบว่า ไฮเปอร์พารามิเตอร์ที่ทำให้ค่าความคลาดเคลื่อนต่ำที่สุด คือ minsplit = 10 และ maxdepth = 10 ซึ่งทำให้ได้ผลลัพธ์ดังแสดงในรูปภาพที่ 4-4



(ก.) (ข.)

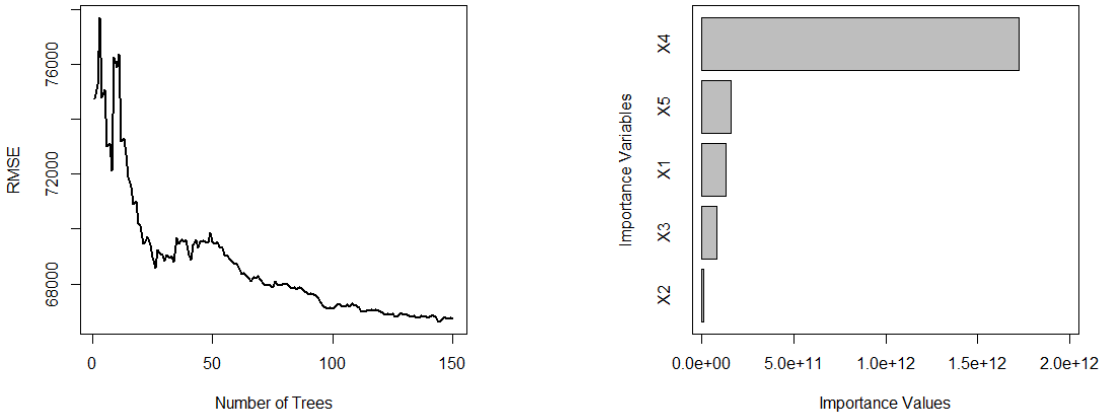
ภาพที่ 4-4 (ก.) แผนภูมิแท่งแสดงค่าความสำคัญของตัวแปรทำนายที่ได้จากการวิเคราะห์การถดถอยต้นไม้ตัดสินใจ และ (ข.) แผนภาพต้นไม้ตัดสินใจ

จากภาพที่ 4-4(ก.) พบว่า ตัวแปรทำนายทั้ง 5 ตัวแปร มีความสำคัญกับตัวแบบการถดถอยต้นไม้ตัดสินใจ โดยตัวแปรที่มีความสำคัญต่อการพยากรณ์มากที่สุด คือ สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) รองลงมา คือ ดัชนีราคาผู้บริโภคทั่วไป (X_5) อัตราดอกเบี้ยนโยบาย (X_3) จำนวนเงินเอาประกันชีวิต (X_1) และเบี้ยประกันภัยรับโดยตรง (X_2) ตามลำดับ ทั้งนี้ จากตัวแปรที่มีความสำคัญทั้ง 5 ตัวแปร จะมีตัวแปรทำนายที่ถูกนำมาใช้ในการสร้างต้นไม้ตัดสินใจเพียง 4 ตัวแปร ซึ่งประกอบไปด้วย จำนวนเงินเอาประกันชีวิต (X_1) เบี้ยประกันภัยรับโดยตรง (X_2) สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) และดัชนีราคาผู้บริโภคทั่วไป (X_5) ดังแสดงในภาพที่ 4-4(ข.) โดยที่สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) เป็นตัวแปรแรกที่ใช้แบ่งข้อมูลในโหนดรากของต้นไม้ ซึ่งแสดงถึงอิทธิพลสูงสุดในการจำแนกกลุ่มย่อยของข้อมูล และโครงสร้าง

ของต้นไม้แสดงให้เห็นการแบ่งกลุ่มข้อมูลตามเงื่อนไขของตัวแปรต่าง ๆ ซึ่งส่งผลให้ได้กลุ่มย่อยที่มีค่าพยากรณ์ของจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) แตกต่างกัน

4.1.6 ผลการวิเคราะห์การถดถอยป่าสุ่ม

เพื่อให้ได้ตัวแบบที่มีประสิทธิภาพดี ในการวิเคราะห์การถดถอยป่าสุ่มจึงได้ทำการปรับแต่งไฮเปอร์พารามิเตอร์ในตัวแบบด้วยวิธีการค้นหาแบบกริดเช่นเดียวกันกับการวิเคราะห์การถดถอยต้นไม้ตัดสินใจ โดยกำหนดให้ช่วงของจำนวนต้นไม้ (ntree) อยู่ระหว่าง 5 ถึง 500 ต้น โดยให้เพิ่มขึ้นทีละ 5 ต้น ช่วงของจำนวนตัวแปรที่สุ่มเลือกในแต่ละการแบ่ง (mtry) อยู่ระหว่าง 1 ถึง 5 ในขณะที่ช่วงของจำนวนตัวอย่างขั้นต่ำในโหนดปลาย (nodesize) และช่วงของจำนวนสูงสุดของโหนดปลาย (maxnodes) มีค่าอยู่ระหว่าง 3 ถึง 10 จากการค้นหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม พบว่า ไฮเปอร์พารามิเตอร์ที่ทำให้ค่า Out-of-Bag (OOB) RMSE ต่ำที่สุด คือ ntree = 150, mtry = 4, nodesize = 10, และ maxnodes = 3 ซึ่งทำให้ได้ผลลัพธ์ดังแสดงในภาพที่ 4-5



(ก.) (ข.)

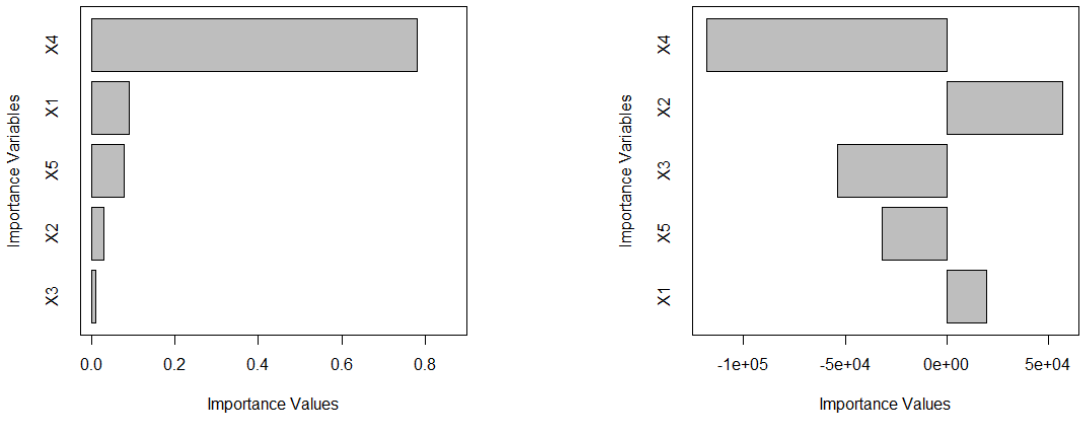
ภาพที่ 4-5 (ก.) แผนภาพแสดงระหว่างค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองและจำนวนต้นไม้ และ (ข.) แผนภูมิแท่งแสดงค่าความสำคัญของตัวแปรทำนายที่ได้จากการวิเคราะห์การถดถอยป่าสุ่ม

จากความสัมพันธ์ระหว่างจำนวนต้นไม้กับค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ดังภาพที่ 4-5(ก.) แสดงให้เห็นว่า ค่า RMSE มีแนวโน้มลดลงเมื่อมีจำนวนต้นไม้มากขึ้น โดยจำนวนต้นไม้ที่ให้ค่า RMSE ต่ำที่สุด คือ จำนวน 144 ต้น ซึ่งมีค่าเท่ากับ 66,632.63 ในขณะที่ผลจากการวิเคราะห์ด้วยต้นไม้จำนวน 150 ต้น พบว่ามีค่า RMSE เท่ากับ 66,714.97 และตัวแปรทำนายทั้ง 5 ตัวแปรต่างก็มีความสำคัญต่อตัวแบบการถดถอยป่าสุ่ม โดยตัวแปรที่มีความสำคัญมากที่สุด คือ สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) ดังแสดงในภาพที่ 4-5(ข.) รองลงมา คือ ดัชนีราคาผู้บริโภคทั่วไป (X_5) จำนวนเงินเอาประกันชีวิต (X_1) อัตราดอกเบี้ยนโยบาย (X_3) และเบี้ยประกันภัยรับโดยตรง (X_2) ตามลำดับ

4.1.7 ผลการวิเคราะห์การถดถอยเอ็กซ์จีบูสต์

ในทำนองเดียวกันกับการวิเคราะห์การถดถอยต้นไม้ตัดสินใจ และการวิเคราะห์การถดถอยป่าสุ่ม เพื่อให้ได้ตัวแบบที่มีประสิทธิภาพดี ในการวิเคราะห์การถดถอยเอ็กซ์จีบูสต์จึงได้ทำการปรับแต่งไฮเปอร์พารามิเตอร์สำหรับตัวแบบโดยใช้ Tree Booster และ Linear Booster ด้วยวิธีการค้นหาแบบกริด โดยพารามิเตอร์สำหรับ Tree Booster มีการกำหนดช่วงของความลึกสูงสุดของต้นไม้ (max_depth) อยู่ระหว่าง 1 ถึง 10 ในขณะที่ช่วงของสัดส่วนข้อมูลตัวอย่างที่ใช้ในการฝึกสอนแต่ละรอบ (subsample) ช่วงของสัดส่วนของตัวแปรทำนาย (คอลัมน์) ที่ใช้ในการสร้างต้นไม้แต่ละต้น (colsample_bytree) และช่วงของอัตราการเรียนรู้ (eta หรือ learning_rate) อยู่ระหว่าง 0.1 ถึง 0.9 โดยให้เพิ่มขึ้นทีละ 0.1 ทั้งนี้ จากการค้นหาไฮเปอร์พารามิเตอร์ที่เหมาะสม พบว่า ไฮเปอร์พารามิเตอร์สำหรับ Tree Booster ที่ให้ค่า RMSE ต่ำที่สุด คือ max_depth = 10, subsample = 0.9, colsample_bytree = 0.7, และ eta = 0.5 ซึ่งทำให้ได้ผลลัพธ์ดังแสดงในภาพที่ 4-6(ก.)

ในขณะที่พารามิเตอร์สำหรับ Linear Booster มีการกำหนดช่วงของ L_1 Regularization (alpha) และ L_2 Regularization (lambda) อยู่ระหว่าง 0 ถึง 10 ในขณะที่ช่วงของจำนวนการวนซ้ำสูงสุดของการบูสต์ (nrounds) อยู่ระหว่าง 50 ถึง 150 โดยให้เพิ่มขึ้นทีละ 50 และช่วงของอัตราการเรียนรู้ (eta) อยู่ระหว่าง 0.1 ถึง 0.9 โดยให้เพิ่มขึ้นทีละ 0.1 ทั้งนี้ จากการค้นหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสม พบว่าไฮเปอร์พารามิเตอร์สำหรับ Linear Booster ที่ให้ค่า RMSE ต่ำที่สุด คือ L_1 Regularization = 9, L_2 Regularization = 0, nrounds = 150, และ eta = 0.7 ซึ่งทำให้ได้ผลลัพธ์ดังแสดงในภาพที่ 4-6(ข.)



(ก.) (ข.)

ภาพที่ 4-6 (ก.) แผนภูมิแท่งแสดงค่าความสำคัญของตัวแปรทำนายที่ได้จากการวิเคราะห์การถดถอยเอ็กซ์จีบูสต์ในรูปแบบ Tree Booster และ (ข.) ในรูปแบบ Linear Booster

จากแผนภูมิแท่งแสดงค่าความสำคัญของตัวแปรทำนายที่ได้จากการวิเคราะห์ด้วยตัวแบบการถดถอยเอ็กซ์จีบูสต์ ทั้งในรูปแบบ Tree Booster และ Linear Booster ในภาพที่ 4-6 พบว่าในทั้งสองกรณีสัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) เป็นตัวแปรที่มีความสำคัญสูงที่สุดในการพยากรณ์จำนวน

กรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี อย่างไรก็ตาม ลักษณะของความสำคัญมีความแตกต่างกันระหว่างสองตัวแบบ กล่าวคือ ในรูปแบบ Tree Booster ค่าความสำคัญของสัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) มีค่าสูงมากอย่างโดดเด่นเหนือกว่าตัวแปรอื่น ซึ่งบ่งชี้ว่ามีบทบาทสำคัญอย่างยิ่งต่อกระบวนการตัดสินใจของตัวแบบ ในขณะที่ตัวแปรอื่น ๆ อย่าง จำนวนเงินเอาประกันชีวิต (X_1) ดัชนีราคาผู้บริโภคทั่วไป (X_5) เบี้ยประกันภัยรับโดยตรง (X_2) และอัตราดอกเบี้ยนโยบาย (X_3) มีบทบาทน้อยมากหรือแทบไม่มีผลเลย ในขณะที่รูปแบบ Linear Booster พบว่า สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) ยังคงเป็นตัวแปรสำคัญสูงสุดเช่นกัน แต่มีค่าความสำคัญในทางลบ ซึ่งสะท้อนว่ามีผลกระทบในทิศทางตรงข้ามกับจำนวนกรมธรรม์ โดยที่เบี้ยประกันภัยรับโดยตรง (X_2) อัตราดอกเบี้ยนโยบาย (X_3) และดัชนีราคาผู้บริโภคทั่วไป (X_5) มีความสำคัญรองลงมา และจำนวนเงินเอาประกันชีวิต (X_1) มีความสำคัญต่ำสุด ซึ่งสอดคล้องกับผลลัพธ์ที่ได้จากตัวแบบการถดถอยเชิงเส้นพหุคูณ

4.2 สรุปและอภิปรายผลการวิจัย

ในการศึกษาครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของตัวแบบทางสถิติและการเรียนรู้ของเครื่องสำหรับใช้ในการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี โดยได้พิจารณาตัวแบบทั้งหมด 5 ตัวแบบ ได้แก่ ตัวแบบการถดถอยเชิงเส้นพหุคูณ ตัวแบบการถดถอยเกาส์เซียนผกผัน ตัวแบบการถดถอยต้นไม้ตัดสินใจ ตัวแบบการถดถอยป่าสุ่ม และตัวแบบการถดถอยเอ็กซ์จีบูสต์ ทั้งในรูปแบบ Tree Booster และ Linear Booster ทั้งนี้ นอกจากการประเมินประสิทธิภาพของตัวแบบดังกล่าวแล้ว ยังได้ดำเนินการวิเคราะห์เพื่อศึกษาตัวแปรทำนายที่มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่จากตัวแบบที่เหมาะสมอีกด้วย

ข้อมูลที่นำมาใช้ในการศึกษาครั้งนี้เป็นข้อมูลทุติยภูมิแบบรายเดือน ตั้งแต่เดือนมกราคม พ.ศ. 2555 ถึงเดือนตุลาคม พ.ศ. 2567 รวมระยะเวลา 154 เดือน โดยรวบรวมมาจากหลายหน่วยงาน ได้แก่ คณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย ธนาคารแห่งประเทศไทย และสำนักงานนโยบายและยุทธศาสตร์การค้า โดยตัวแปรที่ใช้ในการวิเคราะห์ประกอบด้วยตัวแปรตาม คือ จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี (Y) และตัวแปรทำนายจำนวน 5 ตัว ได้แก่ จำนวนเงินเอาประกันชีวิต (X_1) เบี้ยประกันภัยรับโดยตรง (X_2) อัตราดอกเบี้ยนโยบาย (X_3) สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) และดัชนีราคาผู้บริโภคทั่วไป (X_5) ทั้งนี้ได้แบ่งข้อมูลที่ใช้ในการวิเคราะห์ออกเป็น 2 ส่วน คือ ข้อมูลชุดฝึกสอน (ร้อยละ 93.51) และข้อมูลชุดทดสอบ (ร้อยละ 6.49) เพื่อใช้ในการสร้างและประเมินความแม่นยำของตัวแบบ

อย่างไรก็ตาม ตัวแปรทำนายที่นำมาใช้ในการศึกษานี้มีช่วงของค่าที่แตกต่างกันอย่างมาก ซึ่งอาจส่งผลกระทบต่อกระบวนการเรียนรู้ของบางตัวแบบ เพื่อแก้ไขปัญหาดังกล่าว งานวิจัยนี้จึงได้ดำเนินการแปลงค่าของตัวแปรทำนายโดยการทำให้เป็นมาตรฐาน (Standardization) ด้วยการแปลงให้อยู่ในรูปของคะแนน z ซึ่งเป็นการแปลงข้อมูลให้แต่ละตัวแปรมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 เพื่อให้ข้อมูลอยู่ในรูปแบบที่สามารถเปรียบเทียบกันได้และลดอิทธิพลของความแตกต่างในหน่วยวัด

ภายหลังจากการเตรียมข้อมูลและการแปลงตัวแปรทำนายให้อยู่ในรูปของคะแนน z แล้ว จึงดำเนินการสร้างตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีโดยใช้ข้อมูลชุดฝึกสอน จากนั้นนำตัว

แบบที่ได้ไปพยากรณ์ข้อมูลชุดทดสอบ ซึ่งเป็นข้อมูลในช่วงเดือนมกราคม พ.ศ. 2567 ถึงเดือนตุลาคม พ.ศ. 2567 พร้อมทั้งประเมินประสิทธิภาพและความแม่นยำของแต่ละตัวแบบโดยใช้ตัวชี้วัด คือ รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และร้อยละของค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAPE) ทั้งในชุดฝึกสอนและชุดทดสอบ ดังแสดงในตารางที่ 4-8

ตารางที่ 4-8 ผลการประเมินประสิทธิภาพและความแม่นยำของตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี

ตัวแบบ		ข้อมูลชุดฝึกสอน		ข้อมูลชุดทดสอบ	
		RMSE	MAPE	RMSE	MAPE
1. การถดถอยเชิงเส้นพหุคูณ		83291.7900	15.604800000	139572.10	30.08
2. การถดถอยเกาส์เซียนผกผัน		94863.1900	15.844370000	140248.10	32.14
3. การถดถอยต้นไม้ตัดสินใจ		45475.3700	7.812286000	233388.20	87.12
4. การถดถอยแบบป่าสุ่ม		60554.9500	10.758210000	260041.60	105.47
5. การถดถอยเอ็กซ์จีบูสต์	Tree Booster	0.0270	0.000001861	235715.70	98.32
	Linear Booster	83291.7900	15.607070000	139573.60	30.09

จากผลการประเมินประสิทธิภาพและความแม่นยำของตัวแบบการพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ในตารางที่ 4-8 พบว่า ตัวแบบการถดถอยเชิงเส้นพหุคูณและตัวแบบการถดถอยเอ็กซ์จีบูสต์ในรูปแบบ Linear Booster เป็นตัวแบบที่มีประสิทธิภาพและความแม่นยำสูง ซึ่งทั้งสองตัวแบบนี้มีค่า RMSE และ MAPE ในข้อมูลชุดทดสอบต่ำที่สุดเมื่อเทียบกับตัวแบบอื่น ๆ โดยตัวแบบการถดถอยเอ็กซ์จีบูสต์ในรูปแบบ Linear Booster ให้ค่า RMSE เท่ากับ 139,573.60 ฉบับ และ MAPE เท่ากับ 30.09% ส่วนตัวแบบการถดถอยเชิงเส้นพหุคูณให้ค่า RMSE เท่ากับ 139,572.10 ฉบับ และ MAPE เท่ากับ 30.08% ซึ่งถือว่าการพยากรณ์มีความแม่นยำอยู่ในระดับที่พอยอมรับได้ ทั้งนี้ เมื่อพิจารณาจากข้อมูลชุดฝึกสอนพบว่าทั้งสองตัวแบบมีค่า RMSE เท่ากัน นั่นคือ 83,291.79 ฉบับ แต่มีค่า MAPE ต่างกันเพียงเล็กน้อย นอกจากนี้ ทั้งสองตัวแบบยังมีค่า MAPE จากข้อมูลชุดทดสอบแตกต่างจากข้อมูลชุดฝึกสอนประมาณ 2 เท่า เมื่อเทียบกับตัวแบบอื่น ๆ ซึ่งสะท้อนถึงการเรียนรู้จากข้อมูลได้ดีกว่า

อย่างไรก็ตาม ถึงแม้ว่าตัวแบบการถดถอยเชิงเส้นพหุคูณจะมีประสิทธิภาพในการพยากรณ์ แต่จากการตรวจสอบข้อสมมุติเบื้องต้นทางสถิติพบว่า พบว่า ส่วนเหลือของตัวแบบมีปัญหาเรื่องสหสัมพันธ์ในตัว (Autocorrelation) ซึ่งถือว่าเป็นการละเมิดข้อสมมุติเบื้องต้นของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยเฉพาะอย่างยิ่งในข้อมูลอนุกรมเวลาที่มีลักษณะต่อเนื่องรายเดือนเช่นในงานวิจัยนี้ ซึ่งเป็นข้อจำกัดสำคัญที่อาจส่งผลต่อความน่าเชื่อถือของค่าพารามิเตอร์และความสามารถในการนำตัวแบบไปใช้งานจริง ดังนั้น ถึงแม้ว่าตัวแบบจะมีประสิทธิภาพเชิงตัวเลขดี แต่อาจไม่เหมาะสมในแง่ของความถูกต้องทางทฤษฎี

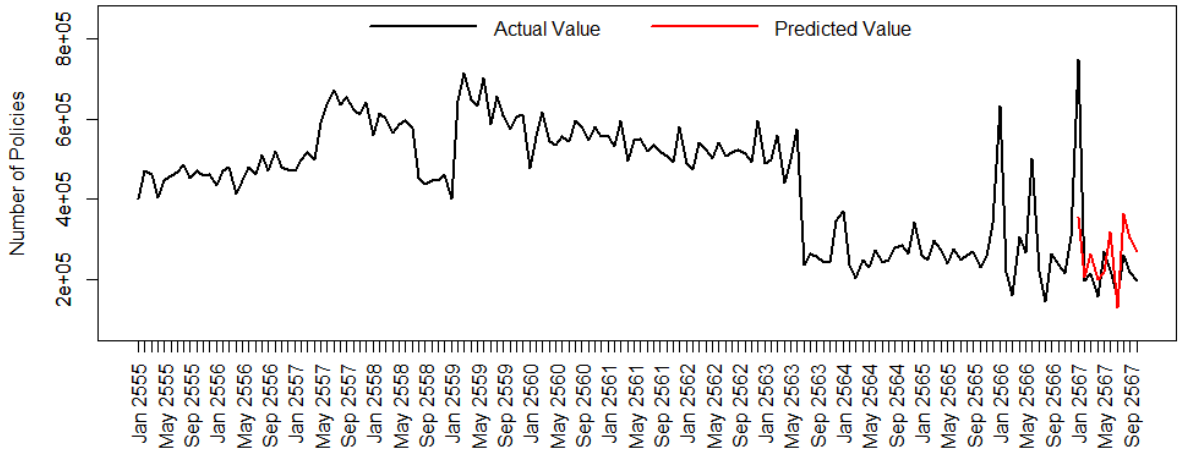
สำหรับตัวแบบการถดถอยเกาส์เซียนผกผัน ถึงแม้จะสามารถพยากรณ์ข้อมูลชุดใหม่ได้ในระดับหนึ่ง โดยมีค่า MAPE อยู่ที่ 32.14% ซึ่งไม่แตกต่างจากตัวแบบการถดถอยเชิงเส้นพหุคูณมากนัก แต่เนื่องจากข้อมูล

จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ไม่ได้มีลักษณะสอดคล้องกับการแจกแจงแบบเกาส์เซียนผกผันอย่างแท้จริง จึงทำให้ตัวแบบนี้ยังไม่เหมาะสำหรับการนำมาใช้ในการพยากรณ์ข้อมูล

ในส่วนของตัวแบบการถดถอยต้นไม้ตัดสินใจและตัวแบบการถดถอยป่าสุ่ม พบว่า ทั้งสองตัวแบบมีค่า RMSE และ MAPE ในข้อมูลชุดทดสอบสูงมาก โดยเฉพาะการถดถอยแบบป่าสุ่มที่มีค่า MAPE สูงถึง 105.47% ซึ่งสูงเกินกว่าระดับที่ยอมรับได้ และแสดงถึงความผิดพลาดในการพยากรณ์อย่างรุนแรง ซึ่งสะท้อนถึงปัญหา Overfitting ที่ตัวแบบเรียนรู้ข้อมูลชุดฝึกสอนมากเกินไป จึงทำให้การพยากรณ์ข้อมูลที่ไม่เคยเห็นมาก่อนได้ไม่แม่นยำ

ในขณะที่ตัวแบบการถดถอยเอ็กซ์จีบูสต์ ซึ่งเป็นตัวแบบที่พัฒนาต่อยอดจากแนวคิด Gradient Boosting ได้แสดงให้เห็นถึงความแตกต่างของผลลัพธ์ระหว่างการใช้ Tree Booster และ Linear Booster โดยที่ Tree Booster ให้ค่า MAPE สูงมาก (98.32%) และ RMSE สูงถึง 235,715.70 ฉบับ ซึ่งบ่งชี้ถึงความไม่เหมาะสมอย่างชัดเจนในการนำไปใช้พยากรณ์ข้อมูลจริง ในขณะที่ Linear Booster ซึ่งมีโครงสร้างเชิงเส้นคล้ายกับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ กลับให้ผลการพยากรณ์ที่ดีกว่าอย่างชัดเจน ซึ่งสอดคล้องกับลักษณะของข้อมูลที่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรทำนายและจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี

เมื่อพิจารณาโดยรวม ทั้งในด้านของประสิทธิภาพในเชิงตัวเลข ความแม่นยำในการพยากรณ์ และความสอดคล้องกับลักษณะของข้อมูลแล้ว ตัวแบบการถดถอยเอ็กซ์จีบูสต์ในรูปแบบ Linear Booster จึงเป็นตัวแบบที่เหมาะสมที่สุดในการนำไปใช้พยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี เนื่องจากสามารถให้ผลการพยากรณ์ที่แม่นยำ และไม่มีข้อจำกัดทางทฤษฎีเหมือนกับการถดถอยเชิงเส้นพหุคูณ จึงสามารถนำไปใช้งานในเชิงปฏิบัติได้อย่างมีประสิทธิภาพและเชื่อถือได้ โดยผลการเปรียบเทียบระหว่างค่าจริง (Actual Value) และค่าพยากรณ์ (Predicted Value) ข้อมูลชุดทดสอบ แสดงดังภาพที่ 4-7 และตารางที่ 4-9 พบว่า ค่าพยากรณ์มีแนวโน้มเคลื่อนไหวสอดคล้องกับค่าจริง แต่มีความแตกต่างจากค่าจริงในบางช่วงเวลา โดยเฉพาะในเดือนมกราคม และสิงหาคม พ.ศ. 2567



ภาพที่ 4-7 ค่าจริงและค่าพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ตั้งแต่เดือนมกราคม พ.ศ. 2567 ถึงเดือนตุลาคม พ.ศ. 2567

ตารางที่ 4-9 ค่าจริงและค่าพยากรณ์จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ตั้งแต่เดือนมกราคม พ.ศ. 2567 ถึงเดือนตุลาคม พ.ศ. 2567

ปี พ.ศ.	เดือน	ค่าจริง	ค่าพยากรณ์
2567	มกราคม	747,211	354,041
	กุมภาพันธ์	196,006	206,232
	มีนาคม	216,274	264,411
	เมษายน	156,812	199,277
	พฤษภาคม	270,736	218,843
	มิถุนายน	224,248	319,780
	กรกฎาคม	153,318	128,897
	สิงหาคม	260,232	365,238
	กันยายน	217,497	303,667
	ตุลาคม	198,449	270,096

นอกจากนี้ จากตัวแบบที่เหมาะสม คือ ตัวแบบการถดถอยเอ็กซ์จีบูสต์ในรูปแบบ Linear Booster พบว่า ตัวแปรทำนายทั้ง 5 ตัวแปร มีผลต่อจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี จากค่าความสำคัญของตัวแปรทำนายแสดงให้เห็นว่า สัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP (X_4) เป็นตัวแปรที่มีอิทธิพลมากที่สุดต่อการพยากรณ์ และมีความสัมพันธ์ในทิศทางตรงกันข้ามกับจำนวนกรมธรรม์ กล่าวคือ เมื่อสัดส่วนเงินให้กู้ยืมแก่ภาคครัวเรือนต่อ GDP เพิ่มขึ้น มีแนวโน้มที่จำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปีจะลดลง ซึ่งสะท้อนให้เห็นว่าภาระหนี้สินของภาคครัวเรือนที่เพิ่มขึ้นอาจเป็นอุปสรรคต่อการตัดสินใจซื้อกรมธรรม์ประกันชีวิต เนื่องจากครัวเรือนมีข้อจำกัดด้านกระแสเงินสดและความสามารถในการใช้จ่าย

ในขณะที่เบี้ยประกันภัยรับโดยตรง (X_2) แสดงค่าความสำคัญรองลงมาและมีความสัมพันธ์ในเชิงบวก ซึ่งสามารถอธิบายได้ว่าการเพิ่มขึ้นของเบี้ยประกันภัยรับโดยตรงจากผู้เอาประกันภัย สัมพันธ์กับการเติบโตของจำนวนกรมธรรม์ประกันชีวิตที่ทำใหม่ระหว่างปี ในส่วนของอัตราดอกเบี้ยนโยบาย (X_3) มีอิทธิพลในระดับปานกลางและแสดงความสัมพันธ์เชิงบวกกับจำนวนกรมธรรม์ประกันชีวิต อาจเป็นเพราะการปรับระดับดอกเบี้ยนโยบายส่งผลต่อความคาดหวังและพฤติกรรมการวางแผนทางการเงินระยะยาวของประชาชน ซึ่งรวมถึง การเลือกซื้อผลิตภัณฑ์ประกันชีวิต

นอกจากนี้ ดัชนีราคาผู้บริโภคทั่วไป (X_5) มีค่าความสำคัญในระดับต่ำ ซึ่งแสดงให้เห็นว่าภาวะเงินเฟ้อหรือระดับราคาสินค้าในภาพรวมอาจไม่ได้ส่งผลโดยตรงต่อการตัดสินใจซื้อกรมธรรม์ประกันชีวิตอย่างชัดเจนและสุดท้าย จำนวนเงินเอาประกันชีวิต (X_1) เป็นตัวแปรทำนายที่มีอิทธิพลน้อยที่สุดในตัวแบบ อาจเนื่องจากความหลากหลายของพฤติกรรมผู้บริโภคที่เลือกวงเงินเอาประกันไม่เท่ากัน ทำให้ตัวแปรนี้ไม่สามารถอธิบายความเปลี่ยนแปลงของจำนวนกรมธรรม์ได้อย่างเด่นชัด

จากผลการวิเคราะห์สรุปได้ว่า ผลการศึกษาที่ได้ช่วยเน้นย้ำถึงความสำคัญของปัจจัยทางเศรษฐกิจ โดยเฉพาะภาระหนี้ของครัวเรือน ที่มีอิทธิพลต่อการตัดสินใจด้านการทำประกันชีวิตอย่างชัดเจน ซึ่งสามารถสนับสนุนการวางนโยบายและการออกแบบผลิตภัณฑ์ประกันชีวิตเพื่อให้ตอบโจทย์กับภาวะเศรษฐกิจได้อย่างมีประสิทธิภาพ

4.3 ข้อเสนอแนะ

4.3.1 เพื่อให้การศึกษามีความครอบคลุมและสามารถพัฒนาต่อยอดได้อย่างเหมาะสม ควรพิจารณาเพิ่มเติมตัวแปรทำนายอื่น ๆ ที่มีความเกี่ยวข้องกับการตัดสินใจซื้อกรมธรรม์ประกันชีวิต ซึ่งอาจช่วยเพิ่มความแม่นยำของตัวแบบได้มากยิ่งขึ้น

4.3.2 งานวิจัยในครั้งนี้มุ่งเน้นที่ตัวแบบการถดถอยที่มีตัวแปรทำนายเป็นหลัก การนำตัวแบบอนุกรมเวลา เช่น ARIMAX SARIMAX หรือ Prophet มาทดลองเปรียบเทียบ อาจช่วยให้ได้มุมมองที่หลากหลายยิ่งขึ้นในการพยากรณ์ข้อมูลที่มีโครงสร้างเชิงเวลา

บรรณานุกรม

ฉัตรศิริ ปิยะพิมลสิทธิ์. (2547). ค่าผิดปกติ. *วารสาร มฉก.วิชาการ*, *8*(15), 106-110.

ชนกนาต สุขประยูร, ตวงพร พัฒนบัณฑิต, ภัทรพงศ์ สกุลวโรภาส และอุไรวรรณ เจริญกีรติกุล. (2560). การใช้ตัวแบบเชิงเส้นวางนัยทั่วไปในการประมาณค่าเบี้ยประกันภัยรถยนต์. *วารสารคณิตศาสตร์*, *62*(692), 31-38.

ฐิติวดี ชัยวัฒน์. (2560). *การบริหารความเสี่ยงภัย และการประกันภัยในศตวรรษที่ 21*. กรุงเทพฯ : โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย

ทิพวัลย์ อิ่มรส. (2562). *ปัจจัยที่มีผลต่อกระบวนการตัดสินใจซื้อประกันชีวิตของประชากร ในเขตกรุงเทพมหานคร*. สาขาการตลาด คณะบริหารธุรกิจ มหาวิทยาลัยรามคำแหง.

ธนาคารแห่งประเทศไทย. (2568). *เงินให้กู้ยืมแก่ภาคครัวเรือน*. https://www.bot.or.th/th/statistics/monetary-statistic.html

ธนาคารแห่งประเทศไทย. (2568). *อัตราดอกเบี้ยนโยบาย*. https://www.bot.or.th/th/our-roles/monetary-policy/mpc-publication/policy-interest-rate.html

เรณู นิยมเดชา. (2554). *การคำนวณอัตราเบี้ยประกันอัคคีภัยที่อยู่อาศัยด้วยตัวแบบเชิงเส้นนัยทั่วไป*. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์และและเทคโนโลยีสารสนเทศ คณะสถิติประยุกต์ สถาบัณพัฒนบริหารศาสตร์

ไตรรัตน์ ศานนิกรภาพ และณัฐพล พันธุ์ภักดี. (2565). ปัจจัยที่มีผลต่อความตั้งใจซื้อประกันชีวิตของผู้บริโภค ในเขตกรุงเทพมหานครและปริมณฑล. *วารสารมหาจุฬานาครทรรศน์*, *9*(11), 18-29.

พงศกร สัตยพานิช. (2562). *ปัจจัยที่มีความสัมพันธ์ต่อการทำประกันชีวิตแบบบำนาญ*. สารนิพนธ์รัฐประศาสนศาสตรมหาบัณฑิต สาขาวิชารัฐประศาสนศาสตร์ คณะรัฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

วิกานดา ผาพันธ์ และวิราวรรณ พุทธมาตย์. (2566). การเปรียบเทียบตัวแบบการพยากรณ์สำหรับเบี้ยประกันภัยรับรวมของบริษัทประกันชีวิตในประเทศไทย. *วารสารวิทยาศาสตร์และเทคโนโลยีหัวเฉียวเฉลิมพระเกียรติ*, *9*(2), 64-74.

สมาคมประกันชีวิตไทย. (2550). *รอบรู้เรื่องประกันชีวิต*. https://lifeif.or.th/wp-content/uploads/2018/05/รอบรู้เรื่องประกันชีวิต.pdf.

สำนักงานนโยบายและยุทธศาสตร์การค้า. (2568). *ดัชนีราคาผู้บริโภคชุดทั่วไป และอัตราการเปลี่ยนแปลง : ประเทศไทย*. https://index.tpso.go.th/cpi

สรฤทธิ์ ทรัพย์สมบัติ, & อารีย์ นัยพินิจ. (2562). การกำหนดแนวทางการตลาดเพื่อการตัดสินใจเลือกซื้อกรมธรรม์ประกันชีวิตของผู้บริโภคในเขตจังหวัดขอนแก่น. *วารสารวิชาการและวิจัยมหาวิทยาลัยภาคตะวันออกเฉียงเหนือ*, *9*(3), 65-79.

สุกัญญา ชำนาญ. (2564). *การปรับตัวของธุรกิจประกันชีวิตในยุคสังคมผู้สูงอายุ*. สารนิพนธ์ปริญญารัฐประศาสนศาสตรมหาบัณฑิต, สาขาวิชารัฐประศาสนศาสตร์, ภาควิชารัฐประศาสนศาสตร์, คณะรัฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

สุธรรม เลิศพงษ์ประเสริฐ และวิธาวีรัฐ ประทุมสวัสดิ์ (2566). ปัจจัยที่มีอิทธิพลต่อพฤติกรรมการซื้อประกันชีวิตของผู้บริโภคในเขตจังหวัดนนทบุรี. *วารสารการบริหารและสังคมศาสตรปริทรรศน, 6*(2), 119-130.

สุนันทา นิยมศิลป์ และสัญญา สัญญาวิวัฒน์. (2555). การเพิ่มการทำประกันชีวิตในประเทศไทย. วารสารศรีนครินทรวิโรฒวิจัยและพัฒนา (สาขามนุษยศาสตร์และสังคมศาสตร์), *4*(8), 177-189.

สำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย. (ม.ป.ป.). *คู่มือปฏิบัติงานสำหรับตัวแทนประกันชีวิต*. https://www.tlaa.org/download/file/คู่มือหลักสูตรขอรับฯ.pdf

สำนักงานคณะกรรมการกำกับและส่งเสริมการประกอบธุรกิจประกันภัย. (2566). *สถิติธุรกิจประกันชีวิต*. https://www2.oic.or.th/th/industry/statistic/data/31/2.

สำนักงานราชบัณฑิตยสภา. (2558). *พจนานุกรมศัพท์สถิติศาสตร์ ฉบับราชบัณฑิตสถาน*. กรุงเทพฯ : สำนักงานราชบัณฑิตยสภา.

สำนักงานอัตราเบี้ยประกันวินาศภัย. (2559). *คู่มือสำหรับผู้ปฏิบัติงานด้านคณิตศาสตร์ประกันภัย การสร้างตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized Linear Models)*. https://www.tgia.org/upload/file_group/3/download_861.pdf

อริญชย์ บุญมีคำ, วินัย โพธิ์สุวรรณ์, และ ธิดาพร ศุภภากร. (2562). ตัวแบบการถดถอยลอจิสติกสำหรับการวิเคราะห์การขาดอายุของกรมธรรม์ประกันชีวิต. *วารสารวิทยาศาสตร์บูรพา*, *24*(2), 754-767.

Alshamsi, A. S. (2014, November). Predicting Car Insurance Policies Using Random Forest. In *2014 10th International Conference on Innovations in Information Technology (IIT)* (pp. 128-132). IEEE.

Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, *4*(2), 145-154.

Chen, T., & Guestrin, C. (2016, August). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Dang, T. K., & Nguyen, H. H. X. (2022). A hybrid approach using decision tree and multiple linear regression for predicting students' performance based on learning progress and behavior. *SN Computer Science*, *3*(5), 393, https://doi.org/10.1007/s42979-022-01251-5.

Das, S., Nayak, J., Nayak, S., & Dey, S. (2022). Prediction of Life Insurance Premium during Pre- and Post-Covid-19: A Higher-Order Neural Network Approach. *Journal of The Institution of Engineers (India): Series B*, *103*(5), 1747-1773.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Fathurahman, M. (2022). Inverse Gaussian regression modeling and its application in neonatal mortality cases in Indonesia. *BAREKENG: Journal of Mathematics and Its Application*, *16*(4), 1197-1206.

Huadsri, S., Mekruksavanich, S., Jitpattanakul, A., & Phaphan, W. (2024, January). A Hybrid SARIMAX Model in Conjunction with Neural Networks for the Forecasting of Life Insurance Industry Growth in Thailand. In *2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical,*

Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON) (pp. 519-524). IEEE.

Huadsri, S., & Phaphan, W. (2024). The Development of Forecasting Models for Life Insurance Data by Employing Time-series Analysis and Machine Learning Technique. *WSEAS TRANSACTIONS on MATHEMATICS*, *23*, 196-205.

Kafková, S., & Křivánková, L. (2014). Generalized Linear Models in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *62*(2), 383-388.

Jain, P., Choudhury, A., Dutta, P., Kalita, K., & Barsocchi, P. (2021). Random Forest Regression-Based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes. *Processes*, *9*(11), 2095.

Lai, S. B. S., Shahri, N. H. N. B. M., Mohamad, M. B., Rahman, H. A. B. A., & Rambli, A. B. (2021). Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data. *Mathematics and Statistics*, *9*(3), 379-385.

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. *Front Genet*, 10:1077.

Lewis, C. D. (1982). *Industrial and Business Forecasting Methods: A Pracitical Guide to Exponential Smoothing and Curve Fitting*. Butterworth-Heinemann.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized Linear Models: with Applications in Engineering and the Sciences*. John Wiley & Sons.

Noor, M. R. B. M., & mat Isa, Z. (2014). Predicting Number of Purchasing Life Insurance. *Applied Mathematical Sciences*, *8*(82), 4087-4095.

Phaphan, W., Jitpattanakul, A., Huadsri, S., Budsaba, K., Phapan, W., & Mekruksavanich, S. (2024, March). Modeling Life Insurance Business Growth in Thailand using SARIMAX and Multilayer Perceptron. In *2024 16th International Conference on Computer and Automation Engineering (ICCAE)* (pp. 146-151). IEEE.

Seo, D. K., Kim, Y. H., Eo, Y. D., Park, W. Y., & Park, H. C. (2017). Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection. *Remote Sensing*, *9*(11), 1163.

Wu, J., Kong, L., Yi, M., Chen, Q., Cheng, Z., Zuo, H., & Yang, Y. (2022). Prediction and Screening Model for Products Based on Fusion Regression and XGBoost Classification. *Computational Intelligence and Neuroscience*, 4987639.