



การวิเคราะห์การถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของ
การเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์

A Multi-Output Regression Approach for Prediction Frequency and Severity of
Auto Insurance Claims

| | | | |
|----------------|-------------|--------------|-------------|
| นางสาวชฎารัตน์ | อิมสารพวงค์ | รหัสประจำตัว | 643021198-6 |
| นายปารเมศ | ศิริพรรณนท | รหัสประจำตัว | 643020446-8 |

รายงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

หลักสูตรสถิติและวิทยาการข้อมูล วิชาเอกสถิติศาสตร์ คณะวิทยาศาสตร์

มหาวิทยาลัยขอนแก่น

ปีการศึกษา 2567

การวิเคราะห์การถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของ
การเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์

A Multi-Output Regression Approach for Prediction Frequency and Severity of
Auto Insurance Claims

| | | | |
|----------------|-------------|--------------|-------------|
| นางสาวชฎารัตน์ | อิมสารพวงค์ | รหัสประจำตัว | 643021198-6 |
| นายปารเมศ | ศิริพรรณนที | รหัสประจำตัว | 643020446-8 |

รายงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

หลักสูตรสถิติและวิทยาการข้อมูล วิชาเอกสถิติศาสตร์ คณะวิทยาศาสตร์

มหาวิทยาลัยขอนแก่น

ปีการศึกษา 2567

หัวข้อโครงงานวิจัย

การวิเคราะห์การถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์

นักศึกษา

นางสาวชฎารัตน์ อิ่มสารพวงค์ รหัสประจำตัว 643021198-6

นายปารเมศ สิริพรรณนัท รหัสประจำตัว 643020446-8

อาจารย์ที่ปรึกษา

ผศ. ดร.ธิปไตย พงษ์ศาสตร์

ดร.พิชญ์ วิรัชโชติเสถียร

สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น อนุมัติให้รายงานฉบับนี้เป็นส่วนหนึ่งของ
การศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต (สถิติและวิทยาการข้อมูล)

.....อาจารย์ที่ปรึกษา

(ผศ. ดร.ธิปไตย พงษ์ศาสตร์)

วันที่ เดือน..... พ.ศ. 2568

.....อาจารย์ที่ปรึกษา

(ดร.พิชญ์ วิรัชโชติเสถียร)

วันที่ เดือน..... พ.ศ. 2568

.....หัวหน้าสาขาวิชาสถิติ

(ผศ. ดร.สุกัญญา เรืองสุวรรณ)

วันที่ เดือน..... พ.ศ. 2568

| | | |
|--------------------|---|--------------------------|
| หัวข้อโครงงานวิจัย | การวิเคราะห์การถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ | |
| นักศึกษา | นางสาวชฎารัตน์ อิมสารพวงค์ | รหัสประจำตัว 643021198-6 |
| | นายปารเมศ สิริพรรณนนท์ | รหัสประจำตัว 643020446-8 |
| อาจารย์ที่ปรึกษา | ผศ. ดร.ธิปไตย พงษ์ศาสตร์ | |
| | ดร.พิชญา วิรัชโชติเสถียร | |

บทคัดย่อ

การคาดการณ์ความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนจากผู้เอาประกันภัยมีความสำคัญต่อการบริหารจัดการความเสี่ยงและการกำหนดอัตราเบี้ยประกันภัยของบริษัทประกันภัยรถยนต์ หากการคาดการณ์ไม่มีความแม่นยำอาจส่งผลกระทบต่อเสถียรภาพทางการเงินของบริษัท งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองการถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนด้วยการเรียนรู้ของเครื่อง และเพื่อพัฒนาเว็บแอปพลิเคชันที่ช่วยในการทำนายความถี่และความรุนแรงในการเรียกร้องค่าสินไหมทดแทน โดยข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลกรมธรรม์ประกันภัยรถยนต์จากบริษัทประกันภัยแห่งหนึ่งในประเทศสเปน ตั้งแต่ปี พ.ศ. 2549 – 2558 จำนวน 80,924 ราย แบบจำลองการถดถอยพหุผลลัพธ์ที่ศึกษา ได้แก่ Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine และ Artificial Neural Network ทำการวัดประสิทธิภาพของแบบจำลองด้วยค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE), ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (SMAPE) ข้อมูลถูกแบ่งเป็นชุดฝึกและชุดทดสอบในอัตราส่วน 80:20 และฝึกแบบจำลองด้วย 10-fold Cross Validation (K = 10) นอกจากนี้ยังมีการปรับแต่งพารามิเตอร์ที่เหมาะสมสำหรับแต่ละแบบจำลองด้วยวิธี GridSearch

ผลการศึกษาพบว่า แบบจำลอง Multi-Output Random Forest Regression มีประสิทธิภาพสูงสุดในการทำนายความถี่ของการเรียกร้องค่าสินไหมทดแทน (MAE = 0.2427, RMSE = 0.7832, SMAPE = 28.8311) และความรุนแรงของการเรียกร้องค่าสินไหมทดแทน (MAE = 228.9800, RMSE = 864.6607, SMAPE = 29.2370) การวิเคราะห์ Shapley Additive Explanations (SHAP) แสดงให้เห็นว่าปัจจัยสำคัญที่มีผลต่อการทำนายมากที่สุดสามอันดับแรก คือ จำนวนกรมธรรม์ที่ผู้ถือกรมธรรม์มีอายุของใบอนุญาตขับขี่ และอายุของผู้ถือกรมธรรม์ นอกจากนี้ยังมีการเปรียบเทียบระหว่าง Multi-Output Regression และ Single-Output Regression พบว่าประสิทธิภาพโดยรวมใกล้เคียงกัน แต่ Multi-Output Regression มีข้อได้เปรียบด้านความสะดวก เนื่องจากสามารถทำนายหลายผลลัพธ์พร้อมกัน ลดขั้นตอนและ

เวลาในการคำนวณ เพื่อรองรับการใช้งานจริง นอกจากนี้ผู้วิจัยยังได้พัฒนาเว็บแอปพลิเคชันที่เชื่อมต่อกับ multi-output random forest regression ผ่าน Flask ช่วยให้บริษัทประกันภัยสามารถป้อนข้อมูลกรมธรรม์ และทราบผลการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนได้แบบเรียลไทม์

คำสำคัญ: การถดถอยพหุผลลัพธ์ จำนวนการเรียกร้องค่าสินไหมทดแทน ความรุนแรงของการเรียกร้องค่าสินไหมทดแทน การเรียนรู้ของเครื่อง การประกันภัยรถยนต์

สาขาวิชาสถิติ

ลายมือชื่อนักศึกษา.....

ปีการศึกษา 2567

(นางสาวชฎารัตน์ อิ่มสารพวงค์)

ลายมือชื่อนักศึกษา.....

(นายปารเมศ ศิริพรรณนท์)

ลายมือชื่ออาจารย์ที่ปรึกษา

(ผศ. ดร.ธิปไตย พงษ์ศาสตร์)

ลายมือชื่ออาจารย์ที่ปรึกษา

(ดร.พิชญา วิรัชโชติเสถียร)

| | | |
|------------------------|---|------------------------|
| Title | การวิเคราะห์การถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ | |
| Student | Miss. Chadarat Imsarapang | Student ID 643021198-6 |
| | Mister. Parames Siripathanon | Student ID 643020446-8 |
| Project Advisor | Asst. Prof. Dr. Tippatai Pongsart | |
| | Dr. Pitchaya Wiratchotisation | |

ABSTRACT

Predicting insurance claim frequency and severity is crucial for risk management and premium rate determination in automobile insurance. Inaccurate predictions may adversely affect the financial stability of insurance companies. The aims of the study is to develop and evaluate the performance of multi-output regression models for predicting the frequency and severity of insurance claims using machine learning techniques, to develop a web application to support real-time prediction of claim frequency and severity. The dataset used in this research consists of automobile insurance policy records from an insurance company in Spain, covering the period from 2006 to 2015, with a total of 80,924 records. The multi-output regression models examined include Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine, and Artificial Neural Network. Model performance is assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE). The dataset is divided into training and testing sets in an 80:20 ratio, with models trained using 10-fold cross-validation ($K = 10$). Additionally, the optimal parameters for each model were fine-tuned using GridSearch.

The results indicated that the Multi-Output Random Forest Regression model achieves the highest predictive performance for both claim frequency (MAE = 0.2427, RMSE = 0.7832, SMAPE = 28.8311) and claim severity (MAE = 228.9800, RMSE = 864.6607, SMAPE = 29.2370). The Shapley Additive Explanations (SHAP) analysis identified the three most influential factors in claim prediction: the number of policies held by the policyholder, the age of the driver's license, and the age of the policyholder. A comparison between Multi-Output Regression and Single-Output Regression demonstrated that while their overall

performance was similar, Multi-Output Regression offers a practical advantage by predicting multiple outputs simultaneously, thereby reducing computational steps and processing time. To support real-world applications, a Flask-based web application integrating the Multi-Output Random Forest Regression model was developed, enabling insurers to input policyholder data and obtain real-time claim predictions.

Keywords: Multi-Output Regression, Claims Frequency, Claims Severity, Machine Learning, Auto Insurance

Department of Statistics
Academic year 2024

Signature of student.....
(Miss. Chadarat Imsarapang)

Signature of student.....
(Mister. Parames Siripathanon)

Signature of project advisor.....
(Asst. Prof. Dr. Tippatai Pongsart)

Signature of project advisor.....
(Dr. Pitchaya Wiratchotisation)

กิตติกรรมประกาศ

การวิจัยเรื่อง “การวิเคราะห์การถดถอยพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์” ฉบับนี้ ดำเนินการวิจัยจนสำเร็จลุล่วงตามวัตถุประสงค์ของการศึกษา ทั้งนี้ผู้วิจัยขอขอบคุณอาจารย์ที่ปรึกษาการวิจัย ผู้ช่วยศาสตราจารย์ ดร.ธิปไตย พงษ์ศาสตร์ และ ดร.พิชญา วิรัชโชติเสถียร ที่ได้ให้คำปรึกษา แนะนำอย่างละเอียด รอบคอบ และช่วยแก้ไขปัญหามาในทุกระยะของการศึกษาด้วยความใส่ใจ ทำให้ผู้วิจัยสามารถดำเนินงานวิจัยได้อย่างมีประสิทธิภาพและบรรลุเป้าหมายที่วางไว้ ขอขอบคุณคณะกรรมการ รองศาสตราจารย์ ดร.วิชุดา ไชยศิริมงคล และ ดร.ธนพงศ์ อินทร์ ที่ได้ให้ข้อเสนอแนะอันเป็นประโยชน์ แลกเปลี่ยนความคิดเห็น และชี้แนะแนวทางในการปรับปรุงแก้ไข เพื่อให้งานวิจัยมีความสมบูรณ์และมีคุณค่ามากยิ่งขึ้น

ขอขอบคุณครอบครัวและเพื่อนอันเป็นที่รัก ที่คอยอยู่เคียงข้าง สนับสนุน และเป็นกำลังใจที่ช่วยให้ผู้วิจัยสามารถทำงานวิจัย ไม่ว่าจะเป็นคำพูดให้กำลังใจ การรับฟังเมื่อยามเหนื่อยล้า หรือการอยู่เคียงข้างในช่วงเวลาที่หนักหน่วง และขอขอบคุณตัวเองที่มุ่งมั่น ไม่ย่อท้อ อดทน และพยายามอย่างเต็มที่ในทุกขั้นตอนของการศึกษา แม้จะมีอุปสรรคหรือช่วงเวลาที่ยากลำบาก แต่ก็คอยให้กำลังใจตัวเองเสมอ จนสามารถดำเนินการวิจัยจนสำเร็จลุล่วง

สุดท้ายนี้ ผู้วิจัยขอแสดงความซาบซึ้งใจและขอบคุณทุกท่าน ทั้งที่ได้กล่าวนามและไม่ได้เอ่ยนาม ที่มีส่วนร่วมในการสนับสนุน ให้คำแนะนำ และมอบความเมตตาในรูปแบบต่าง ๆ ตลอดระยะเวลาการดำเนินงานวิจัย ผู้วิจัยหวังเป็นอย่างยิ่งว่างานวิจัยฉบับนี้จะเป็นประโยชน์ต่อผู้ที่ศึกษาเกี่ยวกับการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ หากมีข้อผิดพลาดประการใด คณะผู้วิจัยขออภัยไว้ ณ ที่นี้

คณะผู้วิจัย

มีนาคม 2568

สารบัญ

หน้า

| | |
|---|----|
| บทคัดย่อ | |
| กิตติกรรมประกาศ | |
| สารบัญ | ก |
| สารบัญตาราง | ง |
| สารบัญภาพ | จ |
| บทที่ 1 บทนำ | 1 |
| 1.1 ความเป็นมาและความสำคัญของปัญหา | 1 |
| 1.2 วัตถุประสงค์ของงานวิจัย | 2 |
| 1.3 สมมติฐานการวิจัย | 2 |
| 1.4 ขอบเขตการวิจัย | 2 |
| 1.4.1 ขอบเขตด้านข้อมูล | 2 |
| 1.4.2 ขอบเขตด้านแบบจำลองและตัวชี้วัดประสิทธิภาพของแบบจำลอง | 3 |
| 1.5 นิยามศัพท์เฉพาะ | 4 |
| 1.5.1 ศัพท์เฉพาะ | 4 |
| 1.5.1 คำย่อ | 4 |
| 1.6 ประโยชน์ที่คาดว่าจะได้รับการวิจัย | 4 |
| บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง | 5 |
| 2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง | 6 |
| 2.1.1 การประกันภัยรถยนต์ | 6 |
| 2.1.2 การเรียนรู้ของเครื่อง (Machine learning) | 7 |
| 2.1.3 การเรียนรู้เชิงลึก (Deep learning) | 9 |
| 2.1.4 การทำนายพหุผลลัพธ์ (Multi-Output Regression) | 10 |
| 2.1.5 แบบจำลองที่ใช้ในการวิเคราะห์การถดถอยพหุผลลัพธ์ | 11 |
| 2.1.5.1 การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้ (Multi-Output Random Forest Regression Regression) | 11 |
| 2.1.5.2 การถดถอยพหุผลลัพธ์แบบเอ็กซ์ตรีมเกรเดียนส์บูสติง (Multi-Output Extreme Gradient Boosting Regression) | 15 |
| 2.1.5.3 การถดถอยพหุผลลัพธ์แบบไลต์เกรเดียนส์บูสติง (Multi-Output Light Gradient Boosting Regression) | 20 |

สารบัญ (ต่อ)

| | หน้า |
|--|------|
| 2.1.5.4 การถดถอยพหุผลัฟร็โครงข่ายประสาทเทียม (Multi-Output Artificial Neural Network) | 21 |
| 2.1.6 การฝึกแบบจำลอง (Model train) | 23 |
| 2.1.7 การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) | 24 |
| 2.1.8 การประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลอง (Model evaluation and comparison) | 25 |
| 2.1.9 Shapley Additive Explanation (SHAP) | 26 |
| 2.2 งานวิจัยที่เกี่ยวข้อง | 27 |
| 2.3 กรอบแนวคิดการวิจัย | 31 |
| บทที่ 3 วิธีการดำเนินงานวิจัย | 32 |
| 3.1 การทำความเข้าใจธุรกิจ (Business Understanding) | 33 |
| 3.2 การศึกษาและทำความเข้าใจข้อมูล (Data Understanding) | 34 |
| 3.3 การเตรียมข้อมูล (Data Preparation) | 43 |
| 3.3.1 การทำความสะอาดข้อมูล (Data Cleansing) | 43 |
| 3.3.2 การแปลงข้อมูล (Data Transformation) | |
| 3.4 การสร้างแบบจำลองการถดถอยพหุผลัฟร็โดยใช้การเรียนรู้ของเครื่อง (Modeling) | 47 |
| 3.4.1 การแบ่งข้อมูล (Data partitioning) | 47 |
| 3.4.2 แบบจำลองการเรียนรู้ของเครื่อง (Model) | 48 |
| 3.5 การประเมินและเปรียบเทียบประสิทธิภาพแบบจำลอง (Model Evaluation and Comparison) | 54 |
| 3.6 การนำแบบจำลองไปใช้งานจริง (Deployment) | 55 |
| บทที่ 4 ผลการวิจัย | 57 |
| 4.1 ผลการศึกษาธุรกิจประกันภัยรถยนต์ | 58 |
| 4.2 ผลการศึกษาข้อมูลการเรียกร้องค่าสินไหมทดแทน | 59 |
| 4.3 ผลการเตรียมข้อมูลก่อนการวิเคราะห์ | 62 |

สารบัญ (ต่อ)

| | |
|--|----|
| 4.4 ผลการสร้างแบบจำลอง | 64 |
| 4.5 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง | |
| 4.6 ผลการพัฒนาเว็บแอปพลิเคชัน | |
| บทที่ 5 สรุปผลการวิจัย | 67 |
| 5.1 สรุปผลการวิจัย | 67 |
| 5.2 อภิปรายผลการวิจัย | 68 |
| 5.3 ประโยชน์ของสถิติ/สารสนเทศสถิติที่ใช้ในการวิจัย | 69 |
| 5.4 ข้อเสนอแนะ | 70 |
| เอกสารอ้างอิง | 72 |
| ภาคผนวก | 75 |

| | หน้า |
|--|------|
| ตารางที่ 1.1 ตัวแปรของข้อมูลประกันภัยรถยนต์ของบริษัทประกันภัย ประเทศสเปน | 3 |
| ตารางที่ 2.1 แบบจำลองและวิธีการที่ใช้ในงานวิจัยที่เกี่ยวข้อง | 30 |
| ตารางที่ 2.2 ตัวแปรที่ส่งผลต่อการทำนายในงานวิจัยที่เกี่ยวข้อง | 30 |
| ตารางที่ 3.1 รายละเอียดข้อมูลการประกันภัยรถยนต์ | 35 |
| ตารางที่ 3.2 ตัวอย่างข้อมูลที่ผิดพลาดที่จะลบออกจากชุดข้อมูล | 43 |
| ตารางที่ 3.3 ตัวอย่างข้อมูลการประกันภัยรถยนต์ สำหรับการวิเคราะห์ | 47 |
| ตารางที่ 3.4 ตัวอย่างการแปลงข้อมูล Has_claim | 47 |
| ตารางที่ 3.5 ร้อยละของข้อมูลชุดฝึกจำแนกตามการเรียกร้องค่าสินไหมทดแทน | 48 |
| ตารางที่ 3.6 ร้อยละข้อมูลชุดทดสอบจำแนกตามการเรียกร้องค่าสินไหมทดแทน | 48 |
| ตารางที่ 3.7 library ที่ใช้สำหรับวิเคราะห์ข้อมูล | 49 |
| ตารางที่ 3.8 ค่าเริ่มต้นของพารามิเตอร์สำหรับตัวแบบ M-RFR | 50 |
| ตารางที่ 3.9 ค่าเริ่มต้นของพารามิเตอร์สำหรับตัวแบบ M-XGBR | 50 |
| ตารางที่ 3.10 ค่าเริ่มต้นของพารามิเตอร์สำหรับตัวแบบ M-LGBMR | 51 |
| ตารางที่ 3.11 ค่าเริ่มต้นของพารามิเตอร์สำหรับตัวแบบ M-ANN | 51 |
| ตารางที่ 3.12 แสดงพารามิเตอร์สำหรับ GridSearchCV | 52 |
| ตารางที่ 3.13 ไฮเปอร์พารามิเตอร์สำหรับตัวแบบ M-RFR ที่กำหนดใน GridSearchCV | 52 |
| ตารางที่ 3.14 ไฮเปอร์พารามิเตอร์สำหรับตัวแบบ M-XGBR ที่กำหนดใน GridSearchCV | 53 |
| ตารางที่ 3.15 ไฮเปอร์พารามิเตอร์สำหรับตัวแบบ M-LGBMR ที่กำหนดใน GridSearchCV | 53 |
| ตารางที่ 3.16 ไฮเปอร์พารามิเตอร์สำหรับตัวแบบ M-ANN ที่กำหนดใน GridSearchCV | 53 |
| ตารางที่ 3.17 ไฮเปอร์พารามิเตอร์กำหนดใน GridSearchCV ของแต่ละแบบจำลอง | 54 |
| ตารางที่ 4.1 พารามิเตอร์ที่ดีที่สุดจากการค้นหาด้วย GridsearchCV ของแต่ละแบบจำลอง | 58 |
| ตารางที่ 4.2 ผลการทดสอบประสิทธิภาพการทำนายของแต่ละแบบจำลอง | 59 |
| ตารางที่ 4.3 แสดงข้อมูลบางส่วนของค่าจริงและค่าทำนายจากแบบจำลอง M-RFR | 60 |
| ตารางที่ 4.4 พารามิเตอร์และค่าที่ดีที่สุดจาก GridSearchCV ของ Single-Output Random Forest Regression | 61 |
| ตารางที่ 4.5 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองประเภทการถดถอยพหุผลลัพท์กับการถดถอยผลลัพท์เดียว | 62 |

สารบัญรูปภาพ

| | หน้า |
|---|------|
| ภาพที่ 2.1 ประเภทการเรียนรู้ของเครื่อง | 7 |
| ภาพที่ 2.2 ลักษณะการทำงานของการเรียนรู้แบบมีผู้สอน | 8 |
| ภาพที่ 2.3 ตัวอย่างประเภทแบบจำลองการเรียนรู้แบบมีผู้สอน | 9 |
| ภาพที่ 2.4 ตัวอย่างประเภทแบบจำลองการเรียนรู้แบบไม่มีผู้สอน | 9 |
| ภาพที่ 2.5 แผนภาพความสัมพันธ์ระหว่าง artificial intelligence, machine learning, deep learning | 10 |
| ภาพที่ 2.6 การทำนายพหุผลลัพธ์ | 10 |
| ภาพที่ 2.7 ลักษณะการทำงานของ Random Forest | 11 |
| ภาพที่ 2.8 ลักษณะการทำงานของ Decision Tree | 12 |
| ภาพที่ 2.9 ตัวอย่างการทำงานของ Decision Tree | 14 |
| ภาพที่ 2.10 แผนผังเซลล์ประสาทมนุษย์ | 21 |
| ภาพที่ 2.11 ตัวอย่างการทำงานแบบจำลอง ANN | 22 |
| ภาพที่ 2.12 กระบวนการทำงานของ 4-fold Cross-validation | 24 |
| ภาพที่ 2.13 กรอบแนวคิดการวิจัย | 31 |
| ภาพที่ 3.1 ขั้นตอนการดำเนินงานวิจัย | 33 |
| ภาพที่ 3.2 การตรวจสอบสถิติเบื้องต้นของข้อมูล | 36 |
| ภาพที่ 3.3 การกระจายตัวของจำนวนการเรียกร้อยค่าสินไหมทดแทนของข้อมูล | 38 |
| ภาพที่ 3.4 การกระจายตัวของความรุนแรงในการเรียกร้อยค่าสินไหมทดแทนทั้งหมด | 38 |
| ภาพที่ 3.5 การกระจายตัวของจำนวนกรรมธรรม์ประกันภัยรถยนต์ของข้อมูล | 39 |
| ภาพที่ 3.6 การกระจายตัวของอายุผู้ถือกรรมธรรม์ของข้อมูล | 39 |
| ภาพที่ 3.7 การกระจายตัวของอายุใบขับขี่ของข้อมูล | 40 |
| ภาพที่ 3.8 การกระจายตัวของจำนวนกรรมธรรม์ประกันภัยอื่นของข้อมูล | 40 |
| ภาพที่ 3.9 การตรวจสอบการกระจายของข้อมูลเชิงคุณภาพ | 41 |
| ภาพที่ 3.10 การตรวจสอบความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับเชิงคุณภาพ | 42 |
| ภาพที่ 3.11 แสดงผลการตรวจสอบค่าสูญหาย (Missing Value) | 44 |
| ภาพที่ 3.12 ลักษณะของ Box plot | 45 |
| ภาพที่ 3.13 ลักษณะของ Box plot | 45 |
| ภาพที่ 4.1 กราฟแท่งตัวชี้วัดประสิทธิภาพการทำนายความถี่ของการเรียกร้อยค่าสินไหมทดแทน | 59 |

สารบัญรูปภาพ (ต่อ)

| | |
|---|----|
| ภาพที่ 4.2 กราฟแท่งตัวชี้วัดประสิทธิภาพการทำนายความความรุนแรงของการเรียกร้องค่า สินไหม | 60 |
| ทดแทน | |
| ภาพที่ 4.3 ผล SHAP ของแบบจำลอง M-RFR ในการทำนายความถี่ของการเรียกร้องค่าสินไหม | 62 |
| ทดแทน | |
| ภาพที่ 4.4 ผล SHAP ของแบบจำลอง M-RFR ในการทำนายความรุนแรงของการเรียกร้องค่า สินไหม | 63 |
| ทดแทน | |
| ภาพที่ 4.5 หน้าต่างเว็บแอปพลิเคชัน | 64 |
| ภาพที่ 4.6 หน้าต่างเว็บแอปพลิเคชันสำหรับกรอกข้อมูลผู้เอาประกันภัย | 65 |
| ภาพที่ 4.7 หน้าต่างเว็บแอปพลิเคชันหลังจากกรอกข้อมูลผู้เอาประกัน | 65 |
| ภาพที่ 4.8 หน้าต่างเว็บแอปพลิเคชันที่แสดงผลการทำนาย | 66 |

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การประกันภัยรถยนต์ช่วยคุ้มครองความสูญเสียหรือความเสียหายที่เกิดจากการใช้รถยนต์แก่ผู้เอาประกันภัย ไม่ว่าจะเป็นชีวิต ร่างกาย ทรัพย์สินของบุคคลภายนอก รวมถึงบุคคลที่โดยสารในรถยนต์ โดยบริษัทประกันภัยจะประเมินความเสี่ยงด้วยปัจจัยต่างๆ เช่น ประวัติการขับขี่ อายุ ประสบการณ์การขับขี่และอื่น ๆ หากเกิดอุบัติเหตุทางรถยนต์และอยู่ในเงื่อนไขความคุ้มครองของกรมธรรม์ บริษัทประกันภัยรถยนต์จะจ่ายค่าสินไหมทดแทนตามเงื่อนไขในกรมธรรม์เมื่อเกิดความเสียหาย ซึ่งอาจรวมถึงค่ารักษาพยาบาล ค่าซ่อมที่ศูนย์บริการ อุบัติเหตุ หรือค่าธรรมเนียมอื่นๆ ทางกฎหมาย (บริษัทรู้ใจ จำกัด, 2566) หมายความว่าผู้เอาประกันภัยจะทำประกันภัยเพื่อถ่ายโอนความเสี่ยงของความสูญเสีย และในอีกทางกลับกันบริษัทประกันภัยจะต้องรับความเสี่ยงนั้นไว้

จากรายงานของสมาคมประกันวินาศภัยไทยพบว่า เบี้ยประกันภัยรับตรงสำหรับประกันภัยรถยนต์ในปี พ.ศ.2566 มีมูลค่ามากถึง 118,419 ล้านบาท จากผลประกอบการธุรกิจประกันวินาศภัยทั้งหมด 210,141 ล้านบาท คิดเป็น 56.35% มีอัตราการเติบโตอย่างต่อเนื่องโดยจากปีที่ผ่านมาอัตราการเติบโตเพิ่มขึ้น 6.00 % และ มีอัตราความเสียหายของการประกันภัยรถยนต์อยู่ที่ 59.4% ซึ่งหมายความว่าบริษัทประกันภัยจะจ่ายค่าสินไหมทดแทนประมาณ 59.4 บาทจากทุก ๆ 100 บาทที่ได้รับจากเบี้ยประกันภัย (สมาคมประกันวินาศภัยไทย, 2566) แสดงให้เห็นว่าประกันภัยรถยนต์มีความต้องการและเกี่ยวข้องกับคนจำนวนมาก นอกจากนี้จากข้อมูลของศูนย์อุบัติเหตุทางถนน พบว่ามีอัตราการเกิดอุบัติเหตุจราจรเพิ่มขึ้นจะทำให้มีการเรียกร้องค่าสินไหมทดแทนที่สูงขึ้น ส่งผลให้บริษัทประกันวินาศภัยต้องประเมินความเสี่ยงและปรับเบี้ยประกันภัยเพิ่มขึ้น (Inn Why, 2567) ดังนั้นบริษัทประกันภัยจำเป็นต้องมีการเตรียมตัวรับมือกับการเรียกร้องค่าสินไหมทดแทนในอนาคต เนื่องจากอัตราความเสียหายที่สูงอาจส่งผลกระทบต่อผลประกอบการของบริษัท

หลังจากวิกฤติโควิด-19 ในปี พ.ศ. 2565 พฤติกรรมการซื้อประกันภัยรถยนต์ในประเทศไทยเปลี่ยนแปลงไปเนื่องจากผู้เอาประกันภัยรู้สึกไม่คุ้มกับราคาเบี้ยประกันภัยที่ต้องจ่าย ทำให้ส่วนมากซื้อประกันภัยแบบตามความต้องการของผู้ใช้ (Personalize insurance) เพื่อตอบโจทย์และเหมาะสมกับแต่ละบุคคลอย่างแท้จริง (Amarin TV, 2565) การนำอัลกอริทึมมาช่วยในการวิเคราะห์ข้อมูลของลูกค้าแต่ละรายทำให้บริษัทสามารถเสนอแผนประกันภัยที่เหมาะสมกับพฤติกรรมการขับขี่หรือความต้องการด้านความคุ้มครอง และยังสร้างกลยุทธ์การตลาดที่มีประสิทธิภาพ (Binariks, 2024) โดยการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine learning) เพื่อการคาดการณ์ความถี่และความรุนแรงของการเรียกร้องมูลค่าสินไหมทดแทน ช่วยให้บริษัทสามารถประเมินความเสี่ยงที่สอดคล้องกับผู้เอาประกันภัยแต่ละราย และใช้ในการพิจารณาเบี้ยประกันภัยที่เหมาะสมให้กับผู้เอาประกันภัย ทำให้บริษัทสามารถบริหารจัดการความเสี่ยงได้ดีขึ้น

ดังนั้น เพื่อคาดการณ์ความถี่และมูลค่าความเสียหายของการเรียกร้องค่าสินไหมทดแทนจากผู้เอาประกันภัย ผู้วิจัยจะการเรียนรู้ของเครื่อง (Machine learning) ซึ่งเป็นองค์ประกอบหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence) ที่ช่วยให้คอมพิวเตอร์สามารถเรียนรู้และปรับปรุงประสิทธิภาพได้ด้วยตนเอง (Athiwat, 2019) มาใช้ในการทำนายโดยใช้การทำนายแบบพหุผลลัพธ์ (Multi-output Regression) ที่มีความสามารถในการทำนายค่าผลลัพธ์หลายตัวแปรพร้อมกัน และคำนึงถึงความสัมพันธ์ระหว่างตัวแปรตาม และการเปรียบเทียบประสิทธิภาพของแบบจำลองเพื่อทำการเลือกแบบจำลองที่ดีที่สุด

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อสร้างแบบจำลองทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ด้วยการเรียนรู้ของเครื่อง (Machine learning) แบบพหุผลลัพธ์

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง (Machine learning) ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทน

1.3 สมมติฐานการวิจัย

แบบจำลองการถดถอยพหุผลลัพธ์แบบเอ็กซ์ตรีมเกรเดียนต์บูสติง (Multi-Output Extreme Gradient Boosting Regression) สามารถทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ได้อย่างมีประสิทธิภาพสูงสุด เนื่องจากใช้การรวมต้นไม้หลายต้น (Decision Tree) โดยแต่ละต้นจะลดความคลาดเคลื่อนจากต้นก่อนหน้า และสามารถประมวลผลแบบขนาน (parallelization) ทำให้มีความเร็วในการประมวลผลสูงมาก ส่งผลให้สามารถจัดการกับข้อมูลที่ซับซ้อนได้ดี และยังลดความคลาดเคลื่อนในการทำนายได้อย่างมีประสิทธิภาพ

1.4 ขอบเขตการวิจัย

1.4.1 ขอบเขตด้านข้อมูล

ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลการประกันรถยนต์ของบริษัทประกันภัยแห่งหนึ่งในประเทศสเปน ตั้งแต่ปี ค.ศ. 2006 ถึง ค.ศ. 2015 เป็นระยะเวลา 10 ปี รวมทั้งสิ้น 80,924 ราย ประกอบด้วยตัวแปรอิสระ 8 ตัวแปร และตัวแปรตาม 2 ตัวแปร ได้มาจากการวิจัยของ Catalina Bolance and Raluca Vernic แสดงรายละเอียดดังตารางที่ 1.1

ตารางที่ 1.1 ข้อมูลประกันภัยรถยนต์ของบริษัทประกันภัยแห่งหนึ่งในประเทศสเปน

| ตัวแปรอิสระ | |
|----------------------------|--|
| ตัวแปร | คำอธิบายตัวแปร |
| 1. npol_auto (X_1) | จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี |
| 2. client_sex (X_2) | เพศของผู้ถือกรมธรรม์ (0=ชาย, 1=หญิง) |
| 3. client_age (X_3) | อายุของผู้ถือกรมธรรม์ |
| 4. lic_age (X_4) | อายุของใบอนุญาตขับขี่ |
| 5. city (X_5) | พื้นที่อยู่อาศัย (0=อื่นๆ, 1=ในเมืองใหญ่) |
| 6. north (X_6) | อาศัยอยู่ภาคเหนือหรือไม่ (0=ไม่ใช่, 1=ใช่) |
| 7. rest (X_7) | อาศัยอยู่พื้นที่อื่นๆของประเทศ (0=ไม่ใช่, 1=ใช่) |
| 8. client_nother (X_8) | จำนวนกรมธรรม์ประเภทอื่นในบริษัท |
| ตัวแปรตาม | |
| 1. nclaims_md (Y_1) | จำนวนการเรียกร้องค่าสินไหมทดแทนประกันภัยรถยนต์ที่มีความเสียหายต่อทรัพย์สิน |
| 2. cost_md (Y_2) | ความรุนแรงในการเรียกร้องค่าสินไหมทดแทนทั้งหมดสำหรับประกันภัยรถยนต์ที่มีความเสียหายต่อทรัพย์สิน (หน่วย: ยูโร) |

1.4.2 ขอบเขตด้านแบบจำลองและตัวชี้วัดประสิทธิภาพของแบบจำลอง

การศึกษาในงานวิจัยนี้ มีวัตถุประสงค์เพื่อสร้างแบบจำลองและเปรียบเทียบประสิทธิภาพการทำนายพหุผลลัพธ์ด้วยวิธีการเรียนรู้ของเครื่อง (Machine learning) ประกอบด้วยแบบจำลองดังนี้

1. การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้ (Multi-Output Random Forest Regression: M-RFF)
2. การถดถอยพหุผลลัพธ์แบบเอ็กซ์ตรีมเกรเดียนต์บูสติง (Multi-Output Extreme Gradient Boosting Regression: M-XGBR)
3. การถดถอยพหุผลลัพธ์แบบไลต์เกรเดียนต์บูสติง (Multi-Output Light Gradient Boosting Regression: M-LGBR)
3. การถดถอยพหุผลลัพธ์แบบโครงข่ายประสาทเทียม (Multi-Output Artificial Neural Network: M-ANN)

ทำการวัดประสิทธิภาพของแบบจำลอง จากค่าต่อไปนี้

1. ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE)
2. ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE)
3. ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (Symmetric Mean Absolute

Percentage Error: SMAPE)

1.5 นิยามศัพท์เฉพาะ.

1.5.1 ศัพท์เฉพาะ

ความถี่ในการเรียกร้องค่าสินไหมทดแทน คือ จำนวนการเรียกร้องค่าสินไหมทดแทนของประกันภัยรถยนต์ด้านทรัพย์สิน (property damage) จากผู้เอาประกันภัยแต่ละรายในช่วงปี ค.ศ. 2006 ถึง ค.ศ. 2015

ความรุนแรงในการเรียกร้องค่าสินไหมทดแทน คือ มูลค่าการเรียกร้องค่าสินไหมทดแทนของประกันภัยรถยนต์ด้านทรัพย์สิน (property damage) จากผู้เอาประกันภัยแต่ละรายในช่วงปี ค.ศ. 2006 ถึง ค.ศ. 2015

การเรียกร้องค่าสินไหมทดแทน คือ ความถี่หรือความรุนแรงในการเรียกร้องค่าสินไหมทดแทน

1.5.2 คำย่อ

| | |
|---------|---|
| RFR | Random Forest Regression |
| ANN | Artificial Neural Network |
| XGboost | Extreme Gradient Boosting |
| LGBR | Light Gradient Boosting Regression |
| M-RFR | Multi-Output Random Forest Regression |
| M-XGBR | Multi-Output Extreme Gradient Boosting Regression |
| M-LGBR | Multi-Output Light Gradient Boosting Regression |
| M-ANN | Multi-Output Artificial Neural Network |
| S-RFR | Single-Output Random Forest Regression |
| RSS | Residual Sum of Squares |
| MAE | Mean Absolute Error |
| RMSE | Root Mean Square Error |
| MAPE | Mean Absolute Percentage Error |
| SMAPE | Symmetric Mean Absolute Percentage Error |
| EDA | Exploratory Data Analysis |

1.6 ประโยชน์ที่คาดว่าจะได้รับการวิจัย

บริษัทประกันภัยรถยนต์สามารถนำแบบจำลองที่ดีที่สุดและคุณลักษณะที่มีผลกระทบต่อการทำนายความถี่และความรุนแรงในการเรียกร้องค่าสินไหมทดแทน มาประยุกต์ใช้กับข้อมูลของบริษัทประกันภัยเพื่อทำนายการเรียกร้องค่าสินไหมทดแทนและการกำหนดเบี้ยประกันภัยที่เหมาะสมให้กับผู้เอาประกันภัยได้อย่างมีประสิทธิภาพ

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการสร้างแบบจำลองพหุผลลัพธ์เพื่อทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ ได้อาศัยแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องดังนี้

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1.1 การประกันภัยรถยนต์

2.1.2 การเรียนรู้ของเครื่อง

2.1.3 การเรียนรู้เชิงลึก

2.1.4 การทำนายพหุผลลัพธ์

2.1.5 แบบจำลองที่ใช้ในการวิเคราะห์การถดถอยพหุผลลัพธ์

2.1.5.1 การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้

2.1.5.2 การถดถอยพหุผลลัพธ์แบบเอ็กตริมเกรเดียนต์บูตติง

2.1.5.3 การถดถอยพหุผลลัพธ์แบบโลทเกรเดียนต์บูตติง

2.1.5.4 การถดถอยพหุผลลัพธ์โครงข่ายประสาทเทียม

2.1.6 การฝึกแบบจำลอง

2.1.7 การปรับแต่งพารามิเตอร์

2.1.8 การประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลอง

2.1.9 Shapley Additive Explanation (SHAP)

2.2 งานวิจัยที่เกี่ยวข้อง

2.3 กรอบแนวคิดการวิจัย

โดยมีรายละเอียดแต่ละหัวข้อดังต่อไปนี้

2.1 แนวคิดทฤษฎีที่เกี่ยวข้อง

2.1.1 การประกันภัยรถยนต์

การประกันภัยรถยนต์คือคุ้มครองความสูญเสียหรือความเสียหายที่เกิดกับรถยนต์แก่ผู้เอาประกันภัยไม่ว่าจะเป็นชีวิต ร่างกาย ทรัพย์สินของบุคคลภายนอก รวมถึงบุคคลที่โดยสารในรถยนต์ โดยบริษัทประกันภัย จะรับผิดชอบค่าใช้จ่ายบางส่วนหรือทั้งหมดตามเงื่อนไขในกรมธรรม์ ด้วยจำนวนการใช้นานพาหนะที่เพิ่มขึ้นอย่างต่อเนื่อง การประกันภัยรถยนต์จึงมีความสำคัญอย่างยิ่งในสังคมปัจจุบัน ข้อมูลจากองค์การอนามัยโลก (WHO) รายงานว่าอุบัติเหตุจราจรทางถนนส่งผลให้มีผู้เสียชีวิตประมาณ 1.19 ล้านคนต่อปี และมีผู้ได้รับบาดเจ็บระหว่าง 20 ถึง 50 ล้านคน ซึ่งส่วนมากเป็นกลุ่มที่มีความเสี่ยง เช่น คนเดินถนน นักปั่นจักรยาน และผู้ขับขี่จักรยานยนต์ (WHO, 2024) ทำให้การประกันภัยรถยนต์เป็นสิ่งสำคัญที่ไม่เพียงช่วยลดภาระค่าใช้จ่ายจากความเสียหาย แต่ยังเสริมสร้างความมั่นใจแก่ผู้ใช้นานพาหนะในกรณีที่เกิดเหตุการณ์ไม่คาดคิด ผู้ขับขี่ที่มีการคุ้มครองจากประกันภัยสามารถเรียกร้องค่าสินไหมทดแทนเพื่อครอบคลุมความสูญเสียที่เกิดขึ้นได้

ปัจจัยหลักที่เกี่ยวข้องกับการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์มักมีความคล้ายคลึงกันในหลายประเทศ แม้จะมีความแตกต่างด้านกฎหมายหรือเงื่อนไขเฉพาะ แต่ปัจจัยพื้นฐาน เช่น ประวัติการเกิดอุบัติเหตุ อายุและเพศของผู้ขับขี่ ประวัติการเคลม ความถี่ในการขับขี่ พื้นที่ที่ใช้ขับขี่ และสภาพเศรษฐกิจ ล้วนเป็นปัจจัยสำคัญที่มีอิทธิพลต่อความเสี่ยง และสามารถนำมาใช้ในการพิจารณาการเรียกร้องค่าสินไหมทดแทนได้ อาทิ

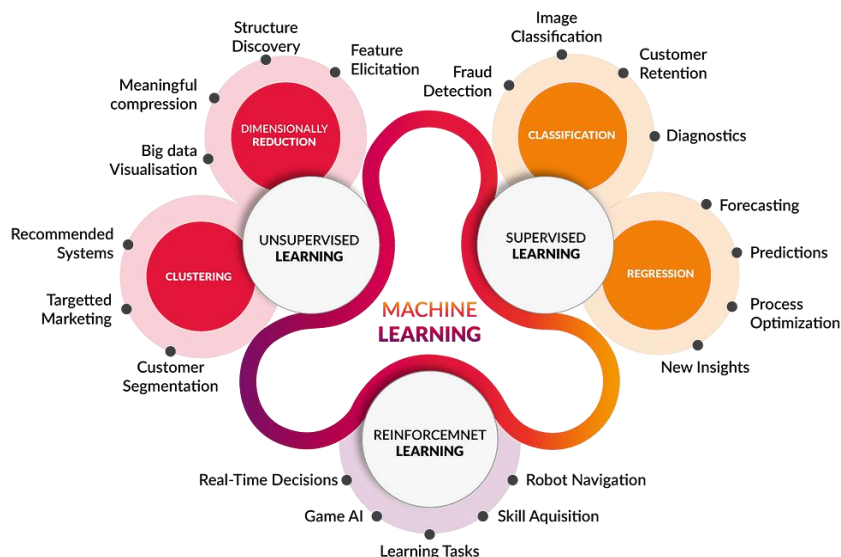
ประเทศสเปน พบว่าปัจจัยด้านเศรษฐกิจและสภาพภูมิอากาศเฉพาะของภูมิภาค ส่งผลต่อความถี่ในการเกิดอุบัติเหตุและความเสี่ยงที่อาจนำไปสู่การเรียกร้องค่าสินไหมทดแทน นอกจากนี้ยังมีการพิจารณา อายุของผู้ถือกรมธรรม์ ประสบการณ์การขับขี่ โดยคนขับที่อายุกว่า 25 ปี มักจะถือเป็นกลุ่มที่มีความเสี่ยงสูง ส่วนผู้ขับที่มีประสบการณ์และไม่มีประวัติการเกิดอุบัติเหตุมาก่อนจะได้รับการพิจารณาให้มีค่าเบี้ยประกันที่ต่ำกว่า ประเภทและอายุรถ (Johnson, 2024)

ประเทศไทย มีอัตราการเกิดอุบัติเหตุสูงที่สุดในเอเชียเนื่องด้วยมีพฤติกรรมการขับขี่ ถนนในบางพื้นที่ของประเทศไทยอาจไม่ได้รับการดูแลอย่างเหมาะสม ปัจจัยที่มีผลต่อการเรียกร้องค่าสินไหมทดแทนที่สำคัญคือ อายุของผู้ถือกรมธรรม์ อายุของรถยนต์ และเพศของผู้ขับขี่ ซึ่งล้วนเป็นข้อมูลที่ใช้ในการพิจารณาความเสี่ยงในการกำหนดเบี้ยประกัน (ปวริศา สุขเรื่อย และ สรรวม จงเจริญ, 2561)

เนื่องจากปัจจัยที่มีผลต่อการเรียกร้องค่าสินไหมทดแทนมีความคล้ายคลึงกันในหลายประเทศ การวิเคราะห์ข้อมูลจากประเทศหนึ่งจึงสามารถนำมาใช้เป็นข้อมูลอ้างอิงหรือแนวทางในการพัฒนาระบบการเรียกร้องค่าสินไหมทดแทนในประเทศอื่น ๆ ได้ ซึ่งจะช่วยเพิ่มประสิทธิภาพในการประเมินความเสี่ยงและการกำหนดเบี้ยประกันภัย

2.1.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง เป็นอัลกอริทึมที่ถูกพัฒนาขึ้นเพื่อจำลองการทำงานของสติปัญญามนุษย์ โดยอาศัยหลักการจากหลากหลายศาสตร์ เช่น ปัญญาประดิษฐ์ สถิติ ความน่าจะเป็น วิทยาการคอมพิวเตอร์ ทฤษฎีสารสนเทศ จิตวิทยา ทฤษฎีการควบคุม และปรัชญา อัลกอริทึมเหล่านี้สามารถทำนายผลลัพธ์ได้โดยไม่ต้องเขียนโปรแกรมคอมพิวเตอร์ที่ซับซ้อน โดยอาศัยการเรียนรู้จากชุดข้อมูลที่จัดเตรียมไว้ และประสบการณ์จากการทำซ้ำเพื่อทำนายผลลัพธ์ กระบวนการเรียนรู้นี้เรียกว่าการฝึกอบรม ซึ่งช่วยให้อัลกอริทึมสามารถปรับปรุงประสิทธิภาพได้ด้วยตนเองตลอดเวลา และยังสามารถเรียนรู้รูปแบบที่ซับซ้อนและละเอียดอ่อนได้มากกว่ามนุษย์ (Naqa et al., 2022) ประเภทของการเรียนรู้ของเครื่อง (Machine Learning) สามารถแบ่งออกเป็น 3 ประเภท ได้แก่ 1) การเรียนรู้แบบมีผู้สอน (Supervised Learning) 2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ 3) การเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) แสดงดังภาพที่ 2.1

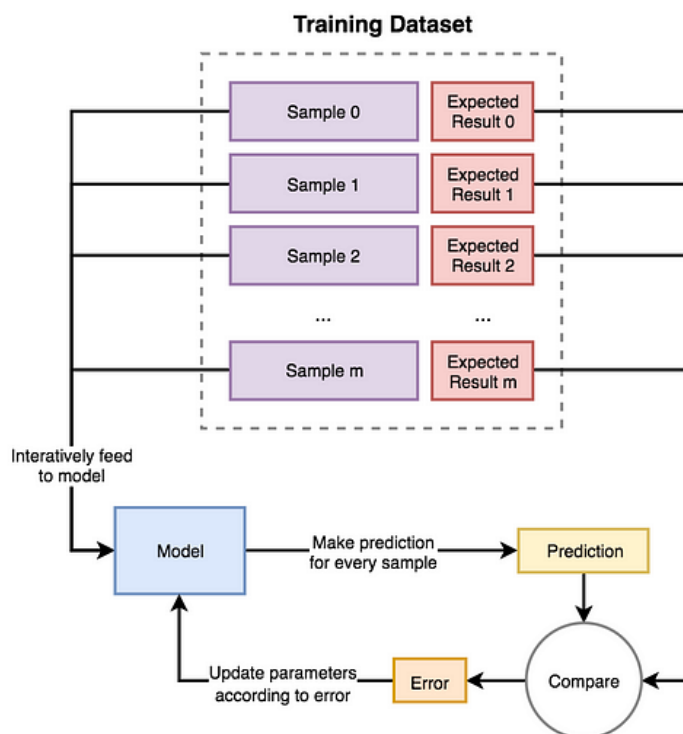


ภาพที่ 2.1 ประเภทการเรียนรู้ของเครื่อง

หมายเหตุ. จาก <https://resources.experfy.com/ai-ml/coding-deep-learning-for-beginners-types-of-machine-learning/>

1) การเรียนรู้แบบมีผู้สอน (Supervised Learning)

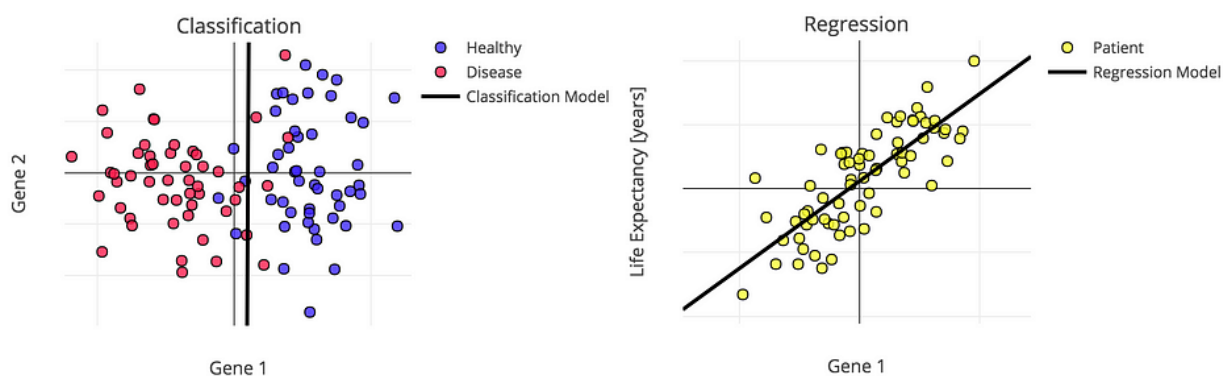
การเรียนรู้แบบมีผู้สอนเป็นเทคนิคที่อัลกอริทึมเรียนรู้จากชุดข้อมูลตัวอย่างที่ประกอบด้วยข้อมูลนำเข้า (input data) และผลลัพธ์ (label) ที่ถูกกำหนดไว้ล่วงหน้า แบบจำลองการเรียนรู้แบบมีผู้สอนจะหาค่าพารามิเตอร์ที่ทำนายผลลัพธ์ ในกระบวนการฝึกอบรมแบบจำลองจะทำนายค่าสำหรับแต่ละตัวอย่างและเปรียบเทียบกับผลลัพธ์ที่กำหนดไว้ หากมีความแตกต่างกัน (error) แบบจำลองจะเรียนรู้และปรับปรุงพารามิเตอร์เพื่อลดความผิดพลาดในการทำนายข้อมูลที่ไม่รู้จัก แสดงดังภาพที่ 2.2



ภาพที่ 2.2 ลักษณะการทำงานของการเรียนรู้แบบมีผู้สอน

หมายเหตุ. จาก <https://resources.experfy.com/ai-ml/coding-deep-learning-for-beginners-types-of-machine-learning/>

ประเภทของผลลัพธ์ในเทคนิคนี้แบ่งออกเป็น 2 ประเภท คือ การจำแนกประเภท (Classification) เป็นการกำหนดหมวดหมู่หรือประเภท เช่น การทำนายว่าเป็นโรคเบาหวานหรือไม่ การตรวจจับสนิทอโง และอีกประเภทหนึ่งคือ การถดถอย (Regression) ซึ่งผลการทำนายจะเป็นค่าต่อเนื่อง (Krzyk, 2023) แสดงดังภาพที่ 2.3

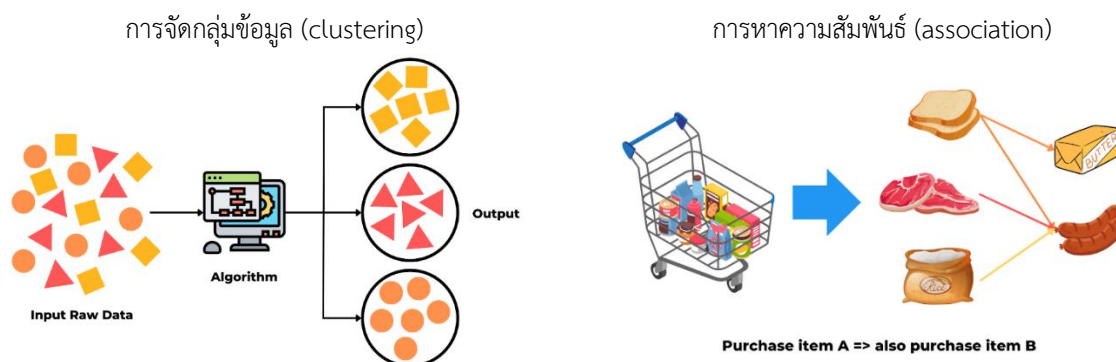


ภาพที่ 2.3 ตัวอย่างประเภทแบบจำลองการเรียนรู้แบบมีผู้สอน

หมายเหตุ. จาก <https://resources.experfy.com/ai-ml/coding-deep-learning-for-beginners-types-of-machine-learning/>

2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้แบบไม่มีผู้สอนเป็นเทคนิคที่อัลกอริทึมสามารถเรียนรู้และสำรวจชุดข้อมูล โดยที่ไม่มีการกำหนดผลลัพธ์ (label) ไว้ล่วงหน้า แบบจำลองจะตีความโครงสร้างและหาความสัมพันธ์ด้วยตนเอง การเรียนรู้แบบไม่มีผู้สอนถูกใช้ในงานต่างๆ เช่น สำหรับการจัดกลุ่มข้อมูล (clustering) การลดมิติข้อมูล (dimensionality reduction) การหาความสัมพันธ์ (association) (Eastgate Software, 2024) แสดงดังภาพที่ 2.4



ภาพที่ 2.4 ตัวอย่างประเภทแบบจำลองการเรียนรู้แบบไม่มีผู้สอน

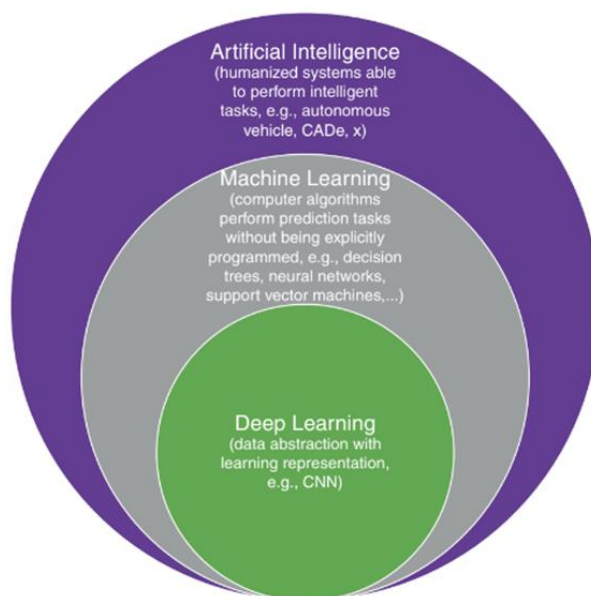
หมายเหตุ. จาก <https://eastgate-software.com/what-is-unsupervised-learning/>

3) การเรียนรู้แบบเสริมกำลัง(Reinforcement Learning)

การเรียนรู้แบบเสริมกำลังเป็นเทคนิคที่อัลกอริทึมเรียนรู้ผ่านการลองผิดลองถูกและการทำซ้ำ โดยจะได้รับรางวัล (Reward) ตามผลลัพธ์ของการกระทำ (Action) ที่ส่งผลต่อสิ่งแวดล้อม (Environment) ผู้กระทำจะเรียนรู้จากข้อผิดพลาดในอดีตเพื่อปรับปรุงและพัฒนาตนเองให้สามารถได้รับรางวัลที่ดีที่สุด ตัวอย่างการใช้งาน ได้แก่ ระบบขับเคลื่อนอัตโนมัติ (Self-driving car) หรือการซื้อขายหุ้นเพื่อให้ได้ผลตอบแทนสูงสุด (Stock Trading Optimization) (BDI, 2020)

2.1.3 การเรียนรู้เชิงลึก (Deep Learning)

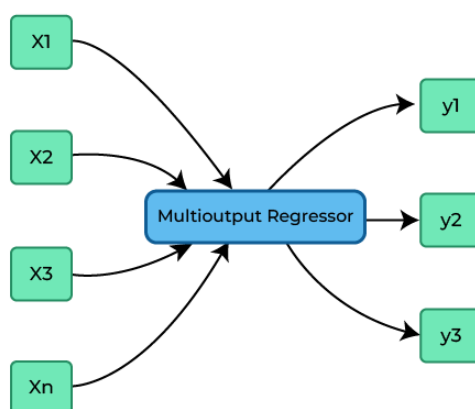
การเรียนรู้เชิงลึก เป็นสาขาย่อยของการเรียนรู้ของเครื่อง โดยมีจุดเด่นที่การเรียนรู้จากการแสดงข้อมูล (Representation Learning) จากข้อมูลพื้นฐานไปจนถึงข้อมูลที่ซับซ้อน ช่วยให้ระบบสามารถเรียนรู้จากข้อมูลดิบได้อย่างอัตโนมัติและสามารถจับรูปแบบเชิงลึกภายในข้อมูลได้โดยไม่ต้องมีการกำหนดคุณลักษณะด้วยมืออย่างชัดเจน ด้วยการใช้โครงข่ายประสาทเทียม (Neural Networks) (Naqa, Murphy and Li, 2022) สามารถนำมาใช้งานได้หลากหลาย เช่น การรู้จำภาพ (Image Recognition) การประมวลผลภาษาธรรมชาติ (Natural Language Processing) การรู้จำเสียง (Speech Recognition) และการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics) เป็นต้น มีแผนภาพแสดงดังภาพที่ 2.5



ภาพที่ 2.5 ความสัมพันธ์ระหว่าง artificial intelligence, machine learning และ deep learning
 หมายเหตุ. จาก What Are Machine and Deep Learning (p. 3-15), by Naqa et al., 2022

2.1.4 การทำนายพหุผลลัพธ์ (Multi-output regression)

การทำนายพหุผลลัพธ์ (Multi-output regression) หรือที่เรียกว่า multi-target regression หรือ multi-variate regression หรือ multi-response regression ซึ่งเป็นประเภทของการเรียนรู้แบบมีผู้สอน (Supervised Learning) จัดเป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องช่วยสร้างแบบจำลองสำหรับทำนายพหุผลลัพธ์โดยใช้ชุดข้อมูลตัวแปรทำนายเดียวกัน เทคนิคนี้สามารถแก้ปัญหาความซับซ้อนในโลกความเป็นจริงได้อย่างมีประสิทธิภาพมากกว่าการทำนายผลลัพธ์ที่ละตัว เนื่องจากแบบจำลองจะพิจารณาทั้งความสัมพันธ์ระหว่างคุณลักษณะ (Feature) กับตัวแปรตาม (Output) และความสัมพันธ์ระหว่างตัวแปรตามด้วยกันเอง ทำให้แบบจำลองที่สร้างขึ้นสามารถวิเคราะห์ความซับซ้อนได้ดีและมีประสิทธิภาพสูงในการทำนาย (Borchani et al., 2015) แสดงดังภาพที่ 2.6



ภาพที่ 2.6 การทำนายพหุผลลัพธ์

หมายเหตุ. จาก <https://www.geeksforgeeks.org/multioutput-regression-in-machine-learning/>

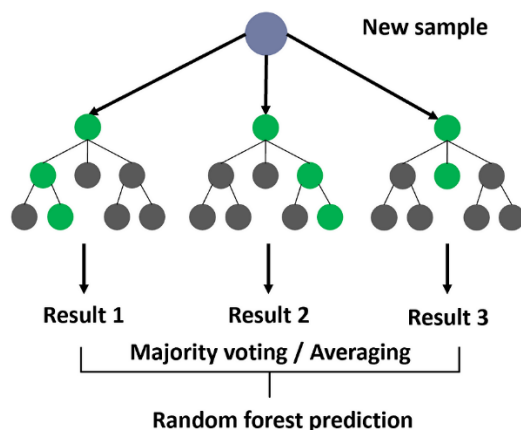
2.1.5 แบบจำลองที่ใช้ในการวิเคราะห์การถดถอยพหุผลลัพธ์ (Model)

แบบจำลองในงานวิจัยฉบับนี้ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) ประเภทการถดถอย (Regression) ที่มีผลลัพธ์ของการทำนายเป็นค่าที่ต่อเนื่อง แบบจำลองที่งานวิจัยใช้มีดังนี้

2.1.5.1 การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้ (Multi-Output Random Forest Regression: M-RFR)

การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้ เป็นวิธีการสร้างแบบจำลองที่สามารถจัดการกับข้อมูลที่มีหลายผลลัพธ์ในเวลาเดียวกัน โดยเฉพาะในกรณีที่ผลลัพธ์มีความสัมพันธ์กัน พัฒนามาจากการถดถอยแบบป่าสุ่ม (Random Forest Regressor) แบบผลลัพธ์เดียว สามารถใช้ได้ทั้งในงานที่เกี่ยวข้องกับการจำแนกประเภทและการถดถอย

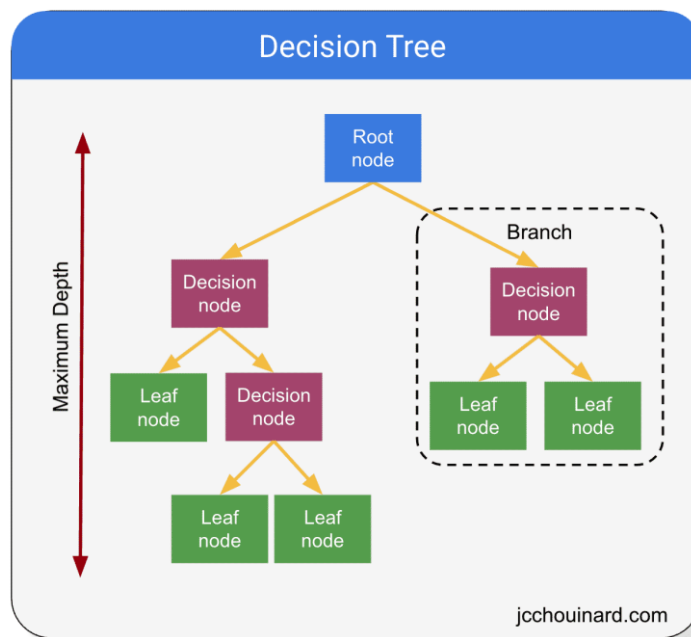
1) การถดถอยแบบป่าสุ่ม (Random Forest Regressor) เป็นวิธีการรวมแบบจำลองที่ใช้การถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression) หลายต้นในการทำนายผลลัพธ์ โดยใช้การสุ่มตัวอย่างแบบมีการคืนตัวอย่าง (bootstrap sampling) และการสุ่มเลือกคุณลักษณะเพื่อสร้างความหลากหลายของต้นไม้ในป่า ทำให้ข้อมูลเดียวกันสามารถถูกเลือกซ้ำได้ในแต่ละครั้ง วิธีนี้ช่วยสร้างความหลากหลายของชุดข้อมูลที่ใช้ในการฝึกแต่ละต้นไม้ ซึ่งส่งผลให้แบบจำลองมีความแม่นยำและความเสถียรภาพเพิ่มขึ้น นอกจากนี้การสุ่มเลือกคุณลักษณะและการรวมผลทำนายจากต้นไม้หลายต้นยังช่วยลดการเกิด overfitting ได้ ซึ่งทำให้แบบจำลองสามารถทำนายได้แม่นยำยิ่งขึ้นเมื่อถูกนำไปใช้กับข้อมูลชุดทดสอบหรือข้อมูลใหม่ ผลลัพธ์สุดท้ายของการทำนายจะได้จากการเฉลี่ยค่าทำนายจากต้นไม้แต่ละต้นในป่า (Abebe et al., 2020)



ภาพที่ 2.7 ลักษณะการทำงานของ Random Forest

หมายเหตุ. จาก <https://blog.devgenius.io/learning-random-forest-classification-using-iris-dataset-eeb930612e0e>

2) การถดถอยต้นไม้ตัดสินใจ (Decision Tree Regression) เป็นแบบจำลองการเรียนรู้ของเครื่องประเภทมีผู้สอน โดยมีเป้าหมายในการสร้างแบบจำลองเพื่อทำนายค่าผลลัพธ์ด้านการถดถอยหรือค่าที่ต่อเนื่อง การทำงานของ Decision Tree Regression จะใช้โครงสร้างต้นไม้ (tree structure) ในการสร้างแบบจำลองการทำนาย โดยการแบ่งข้อมูลออกเป็นกลุ่มย่อยตามลำดับของเงื่อนไขที่กำหนด โดยมีส่วนประกอบแสดงดังภาพที่ 2.8



ภาพที่ 2.8 ลักษณะการทำงานของ Decision Tree

หมายเหตุ. จาก <https://www.jcchouinard.com/decision-trees-in-machine-learning/>

จากภาพที่ 2.8 สามารถอธิบายส่วนประกอบได้ดังนี้

1. Root Node คือ โหนดเริ่มต้นของการตัดสินใจที่ประกอบด้วยข้อมูลทั้งหมด โหนดนี้จะทำการวิเคราะห์และแบ่งข้อมูลออกเป็นกลุ่มย่อย ซึ่งเมื่อทำการแบ่งแล้วโหนดรากจะสร้างโหนดย่อย (child nodes) เพื่อดำเนินการแบ่งข้อมูลในขั้นตอนถัดไป
2. Child nodes คือ โหนดย่อยที่เกิดจากการแบ่งของโหนดตัดสินใจ โดยแต่ละโหนดย่อยจะมีข้อมูลที่แตกต่างกันตามการแบ่งข้อมูลของโหนดก่อนหน้า
3. Decision node คือ โหนดที่ทำหน้าที่ในการแบ่งข้อมูลออกเป็นกลุ่มย่อยๆ เพื่อเพิ่มความแม่นยำในการทำนายค่าผลลัพธ์ โหนดนี้จะพิจารณาคุณลักษณะของข้อมูลและตัดสินใจว่าจะทำการแบ่งข้อมูลอย่างไรเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด
4. Leaf Node or Terminal Node คือ โหนดสุดท้ายที่ไม่สามารถแบ่งย่อยได้อีก ค่าในโหนดนี้จะเป็นค่าที่ใช้ในการทำนาย (prediction)

5. Branches or Sub-Tree คือ กิ่งของต้นไม้การตัดสินใจที่เชื่อมต่อกับโหนดตัดสินใจและโหนดย่อย โดยจำนวนชั้นของกิ่งที่มากที่สุดจากโหนดรากไปยังโหนดสุดท้ายเรียกว่า "Maximum Depth" ของต้นไม้ (Pathmind, n.d.)

การแบ่งโหนด (Splitting) กระบวนการในการแบ่งโหนด ทำได้ตามเกณฑ์หรือเงื่อนไขต่างๆ ซึ่งการถดถอยต้นไม้ตัดสินใจ จะเลือกคุณลักษณะ (feature) หนึ่งตัวเพื่อใช้ตัดสินใจในการแบ่งกลุ่มข้อมูลออกเป็นสองกลุ่มในลักษณะการแบ่งแบบไบนารีที่ซ้ำไปเรื่อย ๆ (recursive binary split) โดยจะเลือกจุดแบ่งที่ทำให้ค่าคลาดเคลื่อนกำลังสอง (Residual Sum of Squares : RSS) ต่ำที่สุด หลักการแบ่งข้อมูลในแต่ละโหนดสำหรับข้อมูลที่มี k ตัวแปร และ n ข้อมูล มีดังนี้

1. เลือกคุณลักษณะ (feature) คือ เลือก 1 ตัวแปรจากตัวแปรทั้งหมด k ตัวแปร แล้วเรียงลำดับค่าของข้อมูลตามคุณลักษณะที่เลือก
2. เลือกจุดแบ่ง (split point) คือ พิจารณาจุดแบ่งที่เป็นไปได้ทั้งหมดจากข้อมูล n ข้อมูล โดยจุดแบ่งจะเป็นตำแหน่งที่อยู่ระหว่างข้อมูลที่เรียงตามลำดับ จำนวนจุดแบ่งจะเท่ากับ $n - \text{min sample}$ จุด เมื่อ min sample คือจำนวนข้อมูลขั้นต่ำที่ต้องการสำหรับแต่ละโหนด
3. คำนวณค่า Residual Sum of Squares (RSS) คือ สำหรับแต่ละจุดแบ่งที่เป็นไปได้ จะคำนวณค่า RSS เพื่อวัดความแตกต่างระหว่างค่าจริงและค่าที่คาดการณ์ จุดแบ่งที่ดีที่สุดคือจุดที่ทำให้ค่า RSS ต่ำที่สุด เนื่องจากแสดงว่าการแบ่งข้อมูลที่ดีนั้นทำให้ข้อมูลในแต่ละกลุ่มมีความแตกต่างกันน้อยที่สุด มีสูตรคำนวณดังนี้

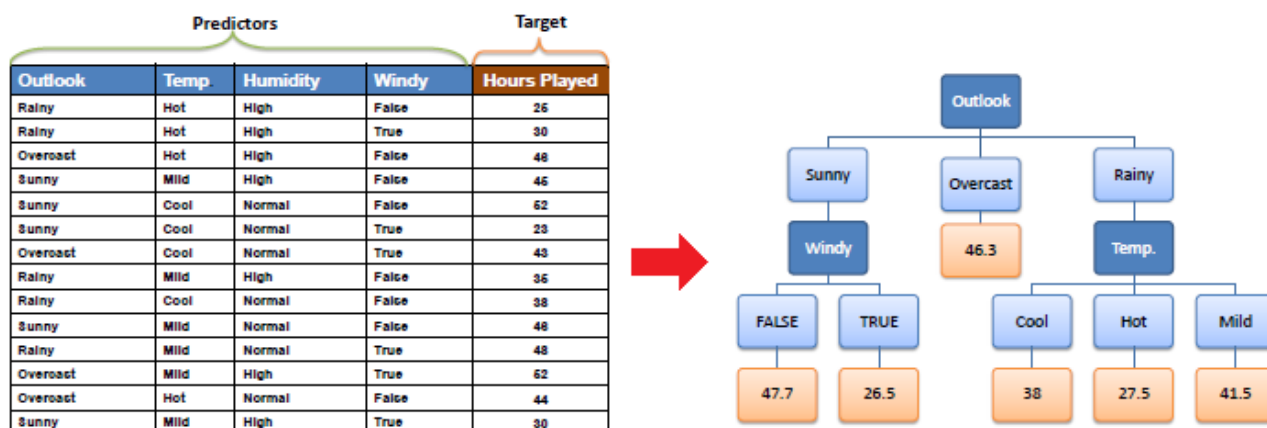
$$\sum_{j=1}^i \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

เมื่อ R_j คือ แต่ละกลุ่มของ observation ที่ถูกแบ่งออกมา ทั้งหมด j กลุ่ม

y_i คือ ค่าของตัวแปรตาม

\hat{y}_{R_j} คือ ค่าทำนายผลลัพธ์ในแต่ละกลุ่ม คำนวณจากค่าเฉลี่ยของตัวแปรตามในกลุ่มนั้นๆ

4. ทำซ้ำ คือ การตัดสินใจและการแบ่งข้อมูลจะดำเนินการไปเรื่อย ๆ จนกว่าจะถึงเงื่อนไขสิ้นสุด เช่น ความสูงของต้นไม้ (max depth) หรือจำนวนข้อมูลใน leaf node ต่ำกว่าเกณฑ์ที่กำหนด (min sample)
5. ทำนายค่า คือ เมื่อสร้างต้นไม้เสร็จแล้ว การทำนายค่าสำหรับข้อมูลใหม่จะทำได้โดยการคำนวณค่าเฉลี่ยของค่าจริงของตัวแปรผลลัพธ์ที่อยู่ใน leaf node นั้น ๆ (Daroontham, 2020)



ภาพที่ 2.9 ตัวอย่างการทำงานของ Decision Tree

หมายเหตุ. จาก https://www.saedsayad.com/decision_tree_reg.htm#:~:text=Decision%20tree%20builds%20regression%20or,decision%20nodes%20and%20leaf%20nodes

จากภาพที่ 2.9 เป็นการทำนาย Hours Played โดยเริ่มจากโหนดรากคือ Outlook แยกโหนดย่อยได้เป็น Sunny Overcast Rainy และในแต่ละรายการจะมีการแยกย่อยออกไป สุดท้ายผลการทำนายจะอยู่ที่โหนดสุดท้ายที่ไม่สามารถแยกย่อยไปอีก ซึ่งเป็นค่าทำนายเรียกว่า Leaf Node ตัวอย่างเช่น Outlook = Rainy แล้วมี Temp. = Hot ผลลัพธ์ของการทำนายจะได้ Hours Played = 27.5

3) การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้ (Multi-Output Random Forest

Regression) เป็นวิธีการสร้างแบบจำลองที่สามารถจัดการกับข้อมูลที่มีหลายผลลัพธ์ โดย Segal (1992) ได้เสนอการสร้างต้นไม้การถดถอยที่สามารถทำนายค่าผลลัพธ์หลายค่าได้พร้อมกันในแต่ละโหนด ซึ่งต้นไม้ถดถอยเหล่านี้มีพื้นฐานมาจากฟังก์ชันการแบ่ง (split function) แบบกำลังสองน้อยที่สุด (least squares) มีหลักการดังนี้

$$\phi(s, t) = SS(t) - SS(t_L) - SS(t_R)$$

เมื่อ $SS(t)$ คือ ผลรวมของค่าความคลาดเคลื่อนกำลังสองในโหนดที่ t

$$SS(t) = \sum_{i \in t} (y_i - \bar{y}(t))^2$$

หลักจากนั้นเพิ่มค่าถ่วงน้ำหนักความแปรปรวนเข้ากับค่าความคลาดเคลื่อนกำลังสอง เพื่อให้การสร้างโหนดมีความเป็นเนื้อเดียวกัน (homogeneous clusters) โดยพิจารณาจากชุดของผลลัพธ์ที่ต้องการทำนาย

$$SS(t) = \sum_{i \in t} (y_i - \bar{y}(t))' V^{-1}(t, \eta) (y_i - \bar{y}(t))$$

โดยที่ $V(t)$ เป็นเมทริกซ์ความแปรปรวนของโหนด t และ η เป็นพารามิเตอร์ที่กำหนดโครงสร้างการถ่วงน้ำหนักความแปรปรวน ดังนั้นการถดถอยพหุผลัฟต์แบบการสุ่มป่าไม้ สามารถจัดการกับการทำนายหลายตัวแปรได้โดยการเปลี่ยนต้นไม้ที่เป็นการทำนายแบบผลัฟต์เดียวให้เป็นหลายผลัฟต์ได้ (Linusson, 2013)

ผลัฟต์สุดท้ายจะเป็นค่าที่ทำนายสำหรับแต่ละตัวแปรผลัฟต์ในชุดข้อมูล มาจากค่าเฉลี่ยของข้อมูลใน Leaf node โดยในงานวิจัยนี้ใช้ Python library 'sklearn.ensemble' คำสั่ง MultiOutputRegressor(RandomForestRegressor) ของ scikit-learn ในการทำนายพหุผลัฟต์ สำหรับพารามิเตอร์หลักในการสร้างแบบจำลองมีดังนี้

- n_estimators : จำนวนต้นไม้ทั้งหมดที่ใช้ในการสร้างแบบจำลอง โดยทั่วไปหากมีค่ามาก ทำให้การทำนายแม่นยำมาก แต่จะใช้เวลาในสร้าง
- max_features : จำนวนของคุณลักษณะที่ถูกสุ่มมาสร้างต้นไม้ตัดสินใจแต่ละต้น
- max_depth : ความลึกสูงสุดของต้นไม้ในป่าแต่ละต้น
- min_samples_split : จำนวนตัวอย่างขั้นต่ำที่ต้องใช้ในการแบ่งที่โหนด
- min_samples_leaf : จำนวนตัวอย่างขั้นต่ำที่ต้องใช้ในโหนดใบ
- random_state : ค่าที่ใช้ในการสุ่มชุดข้อมูลในการฝึกสอนและชุดข้อมูลทดสอบ

2.1.5.2 การถดถอยพหุผลัฟต์แบบเอ็กซ์ตรีมเกรเดียนต์บูสติง (Multi-Output Extreme Gradient Boosting Regression: M-XGBR)

การถดถอยพหุผลัฟต์แบบเอ็กซ์ตรีมเกรเดียนต์บูสติง เป็นการการทำนายหลายผลัฟต์โดยใช้แบบจำลอง Extreme Gradient Boosting หรือ XGBoost ที่ได้รับการพัฒนาขึ้นจาก Gradient Boosting โดยเน้นการเพิ่มประสิทธิภาพและความสามารถในการประมวลผลของแบบจำลองที่ใช้ต้นไม้ตัดสินใจ (Decision Trees) แต่ละต้นไม้จะพยายามแก้ไขข้อผิดพลาดของต้นไม้ก่อนหน้า ทำให้แบบจำลองมีความแม่นยำมากขึ้น โดยใช้ Loss Function (Bentejac et al., 2019) มีขั้นตอนการทำงานของแบบจำลองการถดถอย XGBoost ดังนี้

1. กำหนดฟังก์ชันเริ่มต้น $f_0(x) = \arg \min_{\beta} \sum_{i=1}^N L(y_i, \beta)$ ซึ่งเป็นฟังก์ชันที่ทำให้การสูญเสีย $L(y_i, \beta)$ ต่ำที่สุดเท่าที่จะเป็นไปได้จากข้อมูลทั้งหมด

$$L_{xgb}(f_m) = \sum_{i=1}^N L(y_i, f_m(x_i))$$

| | | |
|-------|--------------------|--|
| เมื่อ | $L_{xgb}(f_m)$ | คือ ฟังก์ชันการสูญเสียที่วัดจากความต่างระหว่างค่าทำนายกับค่าจริงในรอบที่ m |
| | N | คือ จำนวนข้อมูลทั้งหมดในชุดฝึก |
| | $L(y_i, f_m(x_i))$ | คือ ค่าความแตกต่างระหว่างค่าจริง y_i กับค่าที่ทำนายจากแบบจำลองสำหรับข้อมูลลำดับที่ i |
| | β | คือ ค่าทำนาย |

โดยต้นไม้ต้นถัดไปจะพยายามแก้ไขข้อผิดพลาดจากต้นก่อนหน้า เมื่อตัวแปร m คือจำนวนรอบที่ได้ทำการปรับปรุงแบบจำลองไปแล้ว หรือเป็นลำดับของการเพิ่มต้นไม้การตัดสินใจ

2. เรียนรู้ซ้ำ สำหรับแต่ละรอบ $m = 1, \dots, M$:

$$L_{xgb}(\hat{f}_m) = \sum_{i=1}^N (L(y_i, \hat{f}_{m-1}(x_i)) + h_m(x_i))$$

เมื่อ $L(y_i, \hat{f}_{m-1}(x_i))$ คือ ค่าความแตกต่างระหว่างค่าจริง y_i กับค่าที่ทำนายจากแบบจำลองสำหรับข้อมูลลำดับที่ i ในรอบก่อนหน้า ($m - 1$)

$h_m(x_i)$ คือ ค่าการปรับปรุงที่แบบจำลองทำในรอบที่ m เพื่อลดข้อผิดพลาดที่เหลือจากรอบก่อนหน้า

ซึ่ง XGboost ใช้การขยายเทย์เลอร์ (Taylor expansion) ของฟังก์ชันการสูญเสียรอบ \hat{f}_m โดยพิจารณาพหุนามเทย์เลอร์ลำดับที่สองเพื่อช่วยประมาณค่าจุดต่ำที่สุด

$$L_{xgb}(\hat{f}_m) \approx \sum_{i=1}^N (L(y_i, \hat{f}_{m-1}(x_i)) + G_m(x_i) \cdot h_m(x_i) + \frac{1}{2} H_m(x_i) \cdot h_m^2(x_i))$$

2.1 คำนวณค่าอนุพันธ์อันดับหนึ่งและอนุพันธ์อันดับสองของฟังก์ชันการสูญเสีย

เมื่อ $G_m(x_i)$ คือ อนุพันธ์อันดับหนึ่งของฟังก์ชันการสูญเสีย และ $H_m(x_i)$ คือ อนุพันธ์อันดับสองของฟังก์ชันการสูญเสีย โดยสองค่านี้ช่วยให้ทราบถึงทิศทางและความโค้งของฟังก์ชันการสูญเสียที่ใช้ในการปรับปรุงการทำนาย มีสมการดังนี้

$$\text{Gradient} \quad G_m(x_i) = \left[\frac{\partial L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)} \right]_{\hat{f}(x_i) = \hat{f}_{m-1}(x_i)}$$

$$\text{Hessian} \quad H_m(x_i) = \left[\frac{\partial^2 L(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)^2} \right]_{\hat{f}(x_i) = \hat{f}_{m-1}(x_i)}$$

2.2 สร้างแบบจำลองการถดถอยใหม่ h_m โดยใช้ชุดข้อมูล x_i จาก Gradient และ Hessian ที่คำนวณได้ เพื่อให้ต้นไม้สามารถปรับปรุงฟังก์ชันการพยากรณ์ได้ดีขึ้น

$$\left\{ \left(x_i - \frac{\sum_{i \in I_j} G_m(x_i)}{\sum_{i \in I_j} H_m(x_i) + \lambda} \right) : i = 1, \dots, n \right\}$$

เมื่อ $\sum_{i \in I_j} G_m(x_i)$ คือ ผลรวมของ Gradient สำหรับข้อมูลทั้งหมดที่อยู่ในใบ j

$\sum_{i \in I_j} H_m(x_i)$ คือ ผลรวมของ Hessian สำหรับข้อมูลทั้งหมดที่อยู่ในใบ j

λ คือ ค่าพารามิเตอร์ regularization ที่ช่วยควบคุมไม่ให้ค่าน้ำหนักที่กำหนดให้กับแต่ละใบของต้นไม้ (Leaf Weights) ไม่ให้สูงเกินไป

j คือ จำนวนใบของต้นไม้

ในแต่ละโหนดของต้นไม้จะมีการเลือกจุดแบ่งที่เพิ่มความแม่นยำสูงสุด ซึ่ง Gain เป็นค่าที่ใช้วัดประสิทธิภาพของการแบ่งใบการแบ่ง หากจุดใดมีค่าสูงจะถือว่าความสามารถในการลดความสูญเสียมาก มีสมการดังนี้

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} G_m(x_i))^2}{\sum_{i \in I_L} H_m(x_i) + \lambda} + \frac{(\sum_{i \in I_R} G_m(x_i))^2}{\sum_{i \in I_R} H_m(x_i) + \lambda} - \frac{(\sum_{i \in I} G_m(x_i))^2}{\sum_{i \in I} H_m(x_i) + \lambda} \right] - \gamma$$

เมื่อ I_L และ I_R คือ กลุ่มของข้อมูลที่อยู่ในกลุ่มซ้ายและกลุ่มขวาหลังจากการแบ่งตามเกณฑ์ที่กำหนด

I คือ กลุ่มข้อมูลในใบก่อนการแบ่ง

γ คือ ค่าพารามิเตอร์ regularization ที่กำหนดค่า minimum Gain ที่ต้องการสำหรับการสร้างใบใหม่ ถ้า Gain ของการแบ่งมีค่าน้อยกว่าค่า γ แบบจำลองจะไม่สร้างใบใหม่ในส่วนนั้นเพื่อลดความซับซ้อนของต้นไม้

หลังจากนั้นจะทำการอัปเดตฟังก์ชันการทำนาย $\hat{f}_m = \hat{f}_{m-1} + \eta h_m$ เมื่อ η คือ ค่าพารามิเตอร์ควบคุมขนาดการเปลี่ยนแปลงที่เกิดจากต้นไม้ใหม่ หรือ learning rate เป็นค่าคงที่ที่ใช้ควบคุมอัตราการเรียนรู้ เพื่อให้การอัปเดตแต่ละรอบมีขนาดใหญ่มากเกินไป

3. เมื่อทำซ้ำทุกรอบ M แล้ว ผลลัพธ์คือฟังก์ชันการทำนาย \hat{f}_M ที่รวมการเรียนรู้จากทุกต้นไม้เข้าด้วยกัน (Schagen, 2023)

การถดถอยพหุผลลัพธ์แบบเอ็กตรีเมียร์เกรเดียนต์บูสติง สามารถทำได้โดยใช้แบบจำลองรวม (Combined Model) ที่ฝึกตัวแปรผลลัพธ์ทั้งหมดพร้อมกัน ซึ่งแบบจำลองจะพยายามประมาณฟังก์ชันที่มีหลายตัวแปร $f: \mathbb{R}^p \rightarrow \mathbb{R}^n$ ในการสร้างแบบจำลองรวม ต้นไม้การถดถอยจะมีใบที่เป็นเวกเตอร์แทนที่จะเป็นสเกลาร์ ซึ่งหมายความว่าแต่ละใบของต้นไม้สามารถให้ค่าทำนายสำหรับหลายผลลัพธ์ได้ในเวลาเดียวกัน มีหลักการดังนี้

1. กำหนดฟังก์ชันเริ่มต้น $\hat{f}_0(x) = \arg \min_{\beta} \sum_{i=1}^n L(y_i, \beta)$ ซึ่งเป็นฟังก์ชันที่ทำให้การสูญเสีย $L(y_i, \beta)$ ต่ำที่สุดเท่าที่จะเป็นไปได้จากข้อมูลทั้งหมด

ฟังก์ชันที่มีหลายตัวแปรได้ สำหรับฟังก์ชันการสูญเสียใด ๆ นั่นคือเมทริกซ์ที่มีหลายมิติ ใช้การคูณเทนเซอร์ (Tensor Product) เมื่อ U, V และ W เป็นเวกเตอร์สเปซ

$$\dim(W) = \dim(U) \cdot \dim(V)$$

ทำให้การขยายเทย์เลอร์ (Taylor expansion) สามารถการประมาณค่าการสูญเสียในบริบทของการทำนายหลายผลลัพธ์ได้

$$f(x) = f(x_0) + (Df(x_0))(x - x_0) + \dots + \frac{D^k f(x_0)}{k!} (x - x_0)^{\otimes k} + R(x_0)$$

เมื่อ $f(x)$ คือ ฟังก์ชันที่ต้องการประมาณค่า

$f(x_0)$ คือ ค่าของฟังก์ชันที่จุดเริ่มต้น x_0

$Df(x_0)$ คือ อนุพันธ์อันดับแรกของฟังก์ชันที่จุด x_0

$x - x_0$ คือ ความแตกต่างระหว่างค่าปัจจุบันกับค่าที่จุดเริ่มต้น

$\frac{D^k f(x_0)}{k!}$ คือ อนุพันธ์อันดับ k

$\otimes k$ คือ การคูณแบบ Tensor Product โดยทำซ้ำ k ครั้ง

หลังจากนั้นทำให้เป็นมาตรฐานโดยการใช้สองพจน์แรก (2-jet) ในการประมาณค่าฟังก์ชันเริ่มต้น มีสมการดังนี้

$$L(\hat{f}_m) = L(\hat{f}_{m-1}) + h_m \cdot DL(\hat{f}_{m-1}) + \frac{1}{2} h_m^{\otimes 2} \cdot D^2 L(\hat{f}_{m-1})$$

เมื่อ $L(\hat{f}_{m-1})$ คือ ฟังก์ชันการสูญเสียที่คำนวณจากค่าทำนายในรอบก่อนหน้า

h_m คือ ค่าการปรับปรุงที่แบบจำลองทำในรอบที่ m

$DL(\hat{f}_{m-1})$ คือ อนุพันธ์แรกของฟังก์ชันความสูญเสียที่จุด \hat{f}_{m-1}

$D^2 L(\hat{f}_{m-1})$ คือ อนุพันธ์อันดับสองของฟังก์ชันความสูญเสียที่จุด \hat{f}_{m-1}

2. เรียนรู้ซ้ำ สำหรับแต่ละรอบ $m = 1, \dots, M$:

ฟังก์ชันการสูญเสียในสำหรับของการทำนายหลายผลลัพธ์ โดยการรวมทุกมิติของใบ (leaves) ช่วยในการคำนวณฟังก์ชันสูญเสียของต้นไม้

$$\hat{L}(\hat{y}_m) = -\frac{1}{2} \sum_{i=1}^u \left(\frac{\sum_j ((G_m^i)_j)^2}{\sum_j ((\hat{H}_m^i)_j) + \lambda} \right) + \gamma T_m$$

เมื่อ $\hat{L}(\hat{y}_m)$ คือ ฟังก์ชันการสูญเสียที่คำนวณจากค่าทำนายในรอบที่ m

u คือ จำนวนของมิติของตัวแปรผลลัพธ์ หรือ จำนวนตัวแปรผลลัพธ์

T_m คือ จำนวนใบ (leaf nodes) ของต้นไม้ รอบที่ m

2.1 คำนวณค่าอนุพันธ์อันดับหนึ่งและอนุพันธ์อันดับสองของฟังก์ชันการสูญเสียสำหรับหลายผลลัพธ์

$$\begin{aligned} \text{Gradient} \quad G(\hat{y}) &= \left(\frac{\partial L}{\partial(\hat{y})_1}(\hat{y}), \dots, \frac{\partial L}{\partial(\hat{y})_u}(\hat{y}) \right) \\ \text{Hessian} \quad H(\hat{y}) &= \begin{pmatrix} \frac{\partial^2 L}{\partial(\hat{y})_1^2}(\hat{y}) & \cdots & \frac{\partial^2 L}{\partial(\hat{y})_1 \partial(\hat{y})_u}(\hat{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L}{\partial(\hat{y})_u \partial(\hat{y})_1}(\hat{y}) & \cdots & \frac{\partial^2 L}{\partial(\hat{y})_u^2}(\hat{y}) \end{pmatrix} \end{aligned}$$

2.2 สร้างแบบจำลองการถดถอยใหม่ h_m โดยใช้ชุดข้อมูล x_i จาก Gradient และ Hessian ที่คำนวณได้ เพื่อให้ต้นไม้สามารถปรับปรุงฟังก์ชันการพยากรณ์ได้ดีขึ้น

$$\left\{ \left(x_i, -\frac{\sum_j (G_m^i)_j}{\sum_j (\hat{H}_m^i)_j + \lambda} \right) : i = 1, \dots, n \right\}$$

ในแต่ละโหนดของต้นไม้จะมีการเลือกจุดแบ่งที่เพิ่มความแม่นยำสูงสุด ซึ่ง Gain เป็นค่าที่ใช้วัดประสิทธิภาพของการแบ่งใบการแบ่ง หากจุดใดมีค่าสูงจะถือว่าความสามารถในการลดความสูญเสียมาก มีสมการดังนี้

$$\text{Gain} = \frac{1}{2} \left[\sum_{i=1}^u \left(\frac{\sum_{j \in l_l} ((G_m^i)_j)^2}{\sum_{j \in l_l} ((\hat{H}_m^i)_j) + \lambda} \right) + \sum_{i=1}^u \left(\frac{\sum_{j \in l_r} ((G_m^i)_j)^2}{\sum_{j \in l_r} ((\hat{H}_m^i)_j) + \lambda} \right) - \sum_{i=1}^u \left(\frac{\sum_{j \in l} ((G_m^i)_j)^2}{\sum_{j \in l} ((\hat{H}_m^i)_j) + \lambda} \right) \right] - \gamma$$

หลังจากนั้นจะทำการอัปเดตฟังก์ชันการทำนาย $\hat{f}_m = \hat{f}_{m-1} + \eta h_m$ เมื่อ η คือค่าพารามิเตอร์ควบคุมขนาดการเปลี่ยนแปลงที่เกิดจากต้นไม้ใหม่ หรือ learning rate เป็นค่าคงที่ที่ใช้ควบคุมอัตราการเรียนรู้ เพื่อไม่ให้เกิดแต่ละรอบมีขนาดใหญ่เกินไป

3. เมื่อทำซ้ำหุ้รอบ M หรือจนกว่าจะถึงเงื่อนไขหยุด เช่น หรือจนกว่าจะถึงเงื่อนไขการหยุดฟังก์ชันการสูญเสียลดลงหรือครบเงื่อนไขของพารามิเตอร์ที่กำหนด ดังนั้นผลลัพธ์คือฟังก์ชันการทำนาย \hat{f}_M ที่รวมการเรียนรู้จากทุกต้นไม้เข้าด้วยกัน

การถดถอยพหุผลลัพธ์แบบเอ็กตริมเกรเดียนต์บูสติง มีความสามารถในการจับความสัมพันธ์ระหว่างตัวแปรผลลัพธ์ได้ดี เนื่องจากใช้โมเดลเดียวในการพยากรณ์หลายผลลัพธ์พร้อมกัน การใช้ฟังก์ชันการสูญเสียที่เหมาะสมร่วมกับ L2 Regularization ช่วยลดความซับซ้อนของต้นไม้และลดโอกาสในการเกิด overfitting นอกจากนี้ การคำนวณอนุพันธ์อันดับแรกและอันดับสองทำให้สามารถหาจุดแบ่งได้อย่างรวดเร็วและแม่นยำ ในขณะที่การอัปเดตพารามิเตอร์ในแต่ละรอบช่วยเพิ่มประสิทธิภาพในการเรียนรู้ ทำให้กระบวนการสร้างแบบจำลองรวดเร็วขึ้น (Schagen, 2023) โดยในงานวิจัยนี้ใช้ Python library ‘xgboost’ ร่วมกับ MultiOutputRegressor คำสั่ง MultiOutputRegressor(xgb) ของ scikit-learn ในการทำนายพหุผลลัพธ์ สำหรับพารามิเตอร์หลักในการสร้างแบบจำลองมีดังนี้

- n_estimators : จำนวนต้นไม้ที่ใช้ในการสร้างแบบจำลอง

- eta : ควบคุมความเร็วในการเรียนรู้ของแบบจำลอง
- max_depth : ความลึกสูงสุดของต้นไม้แต่ละต้น
- min_child_weight : ค่าขั้นต่ำของน้ำหนักที่ใช้ในการแบ่งโหนด
- subsample : อัตราการสุ่มข้อมูลตัวอย่างในแต่ละรอบการสร้างต้นไม้
- seed : คือค่าที่ใช้ในการตั้งค่า random_state เพื่อให้การสุ่มค่าต่างๆ คงที่

2.1.5.3 การถดถอยพหุผลลัพธ์แบบไลต์เกรเดียนต์บูสติง (Multi-Output Light Gradient Boosting Regression : LGBMR)

แบบจำลอง LGBR เป็นแบบจำลองที่มีโครงสร้างเป็นแบบต้นไม้หลายๆ ต้น โดยต้นไม้เหล่านี้จะถูกสร้างขึ้นจากข้อมูลฝึก คือ แบบจำลองจะทำการค้นหาตัวแปรอิสระที่ส่งผลต่อตัวแปรตามโดยจะเปรียบเทียบความสามารถในการแบ่งกลุ่มค่าผลลัพธ์ของตัวแปรตาม ตามค่าตัวแปรอิสระที่เปลี่ยนแปลงไป จากนั้นจะสร้างเงื่อนไขว่า ถ้าตัวแปรอิสระดังกล่าวมีค่าอยู่ในช่วงแต่ละช่วง ตัวแปรตามควรมีค่าเท่าไร โดยจะใช้ค่าเฉลี่ยของตัวแปรตามที่ถูกจัดกลุ่มอยู่ในกลุ่มเดียวกัน ต่อมาแบบจำลองจะตรวจสอบการแบ่งกลุ่มว่าจะสามารถแยกค่าตัวแปรตามที่มีค่าแตกต่างกันออกจากกันและจับกลุ่มค่าตัวแปรตามไว้ให้อยู่ใกล้เคียงกัน จะดีกว่าการไม่แบ่งกลุ่มหรือไม่ ถ้าหากไม่ดีกว่าจะไม่ทำการแบ่งกลุ่มนั้น เมื่อได้ตัวแปรอิสระที่ใช้ในการแบ่งกลุ่มแล้ว แบบจำลองจะเริ่มสร้างต้นไม้โดยใช้ตัวแปรอิสระดังกล่าวแตกกิ่งค่าความเป็นไปได้ของตัวแปรตามออกมา จากนั้นแบบจำลองจะทำซ้ำตามที่กล่าวมากับตัวแปรอิสระอื่น ๆ ที่ยังไม่ถูกเลือกจนได้ต้นไม้ที่ประกอบด้วยกิ่งจำนวนมาก ซึ่งจะมีประสิทธิภาพในการทำนายค่าตัวแปรตาม โดยแบบจำลอง LGBR มีความสามารถในการจัดการข้อมูลขนาดใหญ่ และทำงานได้เร็วมากเมื่อเทียบกับ Gradient Boosting (FINNOMENA, 2565) และคุณสมบัติที่โดดเด่นของ LGBR คือความสามารถในการฝึกโดยอิงจากฟังก์ชันการสูญเสีย (loss function) ที่แตกต่างกัน

โดยหลักการหลักของ LGBR คือใช้กลยุทธ์การเติบโตของต้นไม้แบบ Leaf-wise ซึ่งต่างจาก Gradient Boosting แบบเดิมที่ใช้ Level-wise โดย Leaf-wise จะเลือกขยายใบที่ลดค่าฟังก์ชันการสูญเสียได้มากที่สุดก่อน จะทำให้มีการเรียนรู้ที่รวดเร็วและมีประสิทธิภาพมากขึ้น และ LGBR ใช้เทคนิค Histogram-based Algorithm ที่ลดการคำนวณที่จำเป็นในการหาจุดแบ่ง (split points)

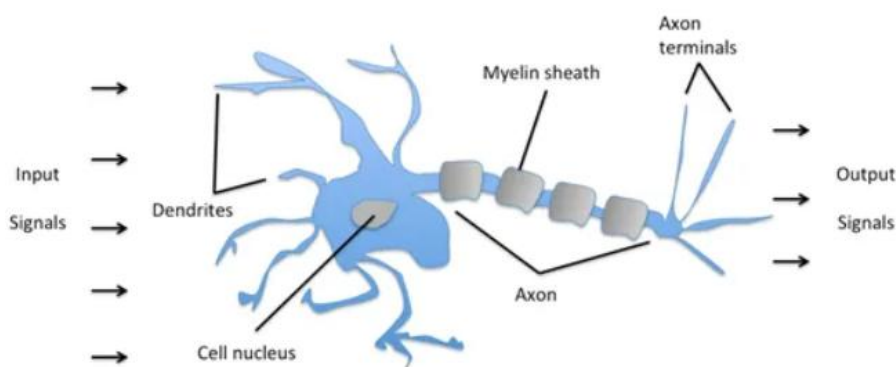
โดยในงานวิจัยนี้ใช้ Python library 'lightgbm' ร่วมกับ MultiOutputRegressor คำสั่ง MultiOutputRegressor(lightgbm) ของ scikit-learn ในการทำนายพหุผลลัพธ์ สำหรับพารามิเตอร์หลักในการสร้างแบบจำลองมีดังนี้

- objective : ประเภทของปัญหาที่ต้องการจัดการ
- learning_rate : อัตราการเรียนรู้ที่ใช้ในการปรับค่าถ่วงน้ำหนักของแบบจำลอง
- n_estimators : จำนวนต้นไม้ที่ใช้ในการสร้างแบบจำลอง
- max_depth : ความลึกสูงสุดของต้นไม้แต่ละต้น

- subsample : อัตราการสุ่มข้อมูลตัวอย่างในแต่ละรอบการสร้างต้นไม้
- min_child_weight : ค่าขั้นต่ำของน้ำหนักที่ใช้ในการแบ่งโหนด
- num_leaves : จำนวนใบสูงสุดในแต่ละต้นไม้
- min_data_in_leaf : จำนวนข้อมูลขั้นต่ำในแต่ละใบ

2.1.5.4 การถดถอยพหุผลลัพธ์โครงข่ายประสาทเทียม (Multi-Output artificial neural network: M-ANN)

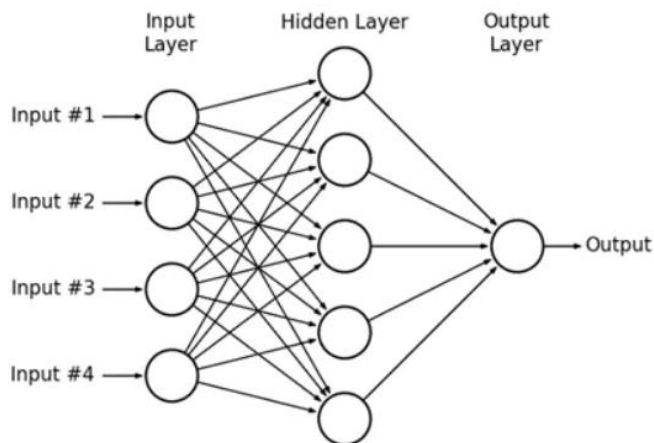
แบบจำลอง Artificial Neural Network (ANN) เป็นหนึ่งในวิธีการของการเรียนรู้เชิงลึกซึ่งอยู่ภายใต้แนวทางของการเรียนรู้แบบมีผู้สอนที่สามารถนำไปประยุกต์ใช้กับงานได้หลายด้าน เช่น การทำนาย การจัดกลุ่ม เป็นต้น โดยแนวคิดหลักของ ANN ได้รับแรงบันดาลใจจากโครงสร้างและการทำงานของสมองมนุษย์ ซึ่งพยายามเลียนแบบการทำงานของเซลล์ประสาทเพื่อประมวลผลข้อมูล แสดงดังภาพที่ 2.12



ภาพที่ 2.10 แผนผังเซลล์ประสาทมนุษย์

หมายเหตุ. จาก https://www.researchgate.net/publication/343921333_Modeling_and_Predicting_of_Motor_Insurance_Claim_Amount_using_Artificial_Neural_Network

ANN ประกอบไปด้วย โหนด หลายโหนด ซึ่งคล้ายกับเซลล์ประสาทของมนุษย์ที่เชื่อมต่อกันด้วยลิงค์และโต้ตอบกัน มีส่วนประกอบสำคัญ คือ ข้อมูลป้อนเข้า (Input layer) ชั้นซ่อน (Hidden Layer) และข้อมูลส่งออก (Output layer) ซึ่งมีหลักการทำงานดังภาพที่ 2.11



ภาพที่ 2.11 ตัวอย่างการทำงานแบบจำลอง ANN

หมายเหตุ. จาก https://www.researchgate.net/publication/343921333_Modeling_and_Predicting_of_Motor_Insurance_Claim_Amount_using_Artificial_Neural_Network

จากภาพที่ 2.11 โครงสร้างพื้นฐานของ ANN จะประกอบไปด้วย 3 ส่วนหลัก

ชั้นข้อมูลป้อนเข้า (Input layer) เป็นชั้นที่จะทำการป้อนข้อมูลเข้า จำนวนโหนดในชั้นนี้จะขึ้นอยู่กับข้อมูลที่นำเข้าแบบจำลองหรือตัวแปรอิสระที่จะนำเข้าสู่แบบจำลอง

ชั้นซ่อน (Hidden Layer) คือชั้นที่อยู่ระหว่างชั้นข้อมูลนำเข้าและชั้นข้อมูลส่งออก ซึ่งมีบทบาทสำคัญในการเรียนรู้ของแบบจำลอง ในชั้นซ่อนสามารถมีหลายชั้นและแต่ละชั้นสามารถมีจำนวนโหนด เพอร์เซพตรอน (Perceptron) ได้ตามต้องการ เพอร์เซพตรอนเป็นหน่วยที่รับข้อมูลจากชั้นก่อนหน้า แล้วคำนวณและส่งผลลัพธ์ผ่านฟังก์ชันกระตุ้น (Activation Function)

ชั้นข้อมูลส่งออก (Output layer) เป็นชั้นที่จะนำข้อมูลจากการคำนวณชั้นที่ผ่านมาไปใช้ จำนวนโหนดในชั้นนี้จะขึ้นอยู่กับตัวแปรตาม ซึ่งสามารถมีมากกว่า 1 โหนดได้ (อัครพล พรหมพิริยะพงษ์, 2566)

ซึ่ง ANN สามารถทำนายแบบพหุผลลัพธ์ได้โดยใช้ MLPRegressor (Multi-Layer Perceptron Regressor) ใน scikit-learn ซึ่งเป็นแบบจำลองปัญญาประดิษฐ์ประเภท ANN ออกแบบมาเพื่อใช้ในการแก้ปัญหาการถดถอย ซึ่งคาดการณ์ตัวแปรตามแบบต่อเนื่องและสามารถทำนายผลลัพธ์ได้หลายตัวแปรโดยไม่ต้องใช้ร่วมกับ Multi-Output Regression ดังนั้น MLPRegressor เป็นทางเลือกที่ดีสำหรับการแก้ปัญหาทำนายค่าตัวเลขที่ต้องการแบบจำลองที่มีความยืดหยุ่นสูงและสามารถจัดการกับความซับซ้อนของข้อมูลได้ซึ่งในงานวิจัยจะใช้พารามิเตอร์หลักดังนี้

- hidden_layer_sizes: การกำหนดจำนวนชั้นและจำนวนหน่วยประสาทในแต่ละ hidden layer เช่น (4,2) หมายความว่าในชั้นซ่อนจะมีทั้งหมด 2 ชั้น โดยชั้นแรกจะมี 4 โหนด และชั้นที่สองจะมี 2 โหนด

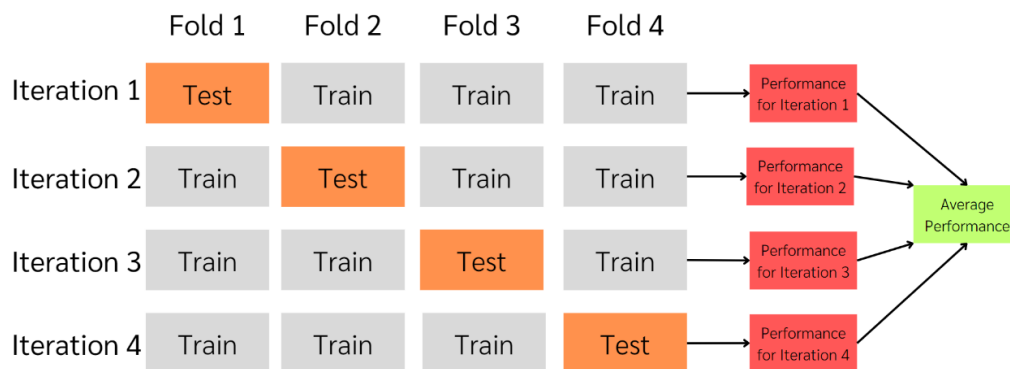
- solver: การปรับน้ำหนักของแบบจำลอง เช่น lbfgf ใช้สำหรับปัญหา optimization และ adam ใช้สำหรับการฝึกโครงข่ายประสาท
- activation: ฟังก์ชันที่ใช้ในการแปลงข้อมูลจากแต่ละชั้นของโครงข่ายประสาทเทียม ก่อนส่งไปยังชั้นถัดไป จะสามารถกำหนดเป็น 'identity' เป็นการส่งไปไหนต่อไ้โดยชั้นฟังก์ชันเอกลักษณ์ $f(x)=x$ และ 'relu' เป็นการส่งไปยังไหนต่อไ้โดยใช้ฟังก์ชัน Rectified Linear Unit คือ $f(x)=\max(0,x)$
- learning_rate: อัตราการเรียนรู้ที่ใช้ในการปรับพารามิเตอร์

2.1.6 การฝึกแบบจำลอง (Model train)

การฝึกสอนแบบจำลองการเรียนรู้ของเครื่องหมายถึงกระบวนการที่ใช้ข้อมูลการฝึกเพื่อปรับพารามิเตอร์ต่างๆของแบบจำลองให้สามารถทำนายผลลัพธ์ได้อย่างแม่นยำ การฝึกแบบจำลองจะทำงานกับข้อมูลที่มีป้ายกำกับ ซึ่งประกอบด้วยคุณลักษณะและค่าผลลัพธ์ที่ต้องการ โดยผ่านขั้นตอนการเรียนรู้หลายรอบเพื่อค้นหาพารามิเตอร์ที่เหมาะสมที่สุด ซึ่งสามารถลดความผิดพลาดระหว่างผลลัพธ์ที่คาดการณ์และผลลัพธ์จริง ในงานวิจัยนี้จะใช้เทคนิค K-Fold Cross-Validation โดยใช้ชุดข้อมูลฝึกอบรมที่แบ่งมา 80% สำหรับการฝึกและทดสอบแบบจำลอง เพื่อตรวจสอบความแม่นยำของแบบจำลองและปรับปรุงประสิทธิภาพ

K-fold Cross-validation เป็นเทคนิคการเรียนรู้ของเครื่องที่ใช้เพื่อประเมินประสิทธิภาพของแบบจำลอง โดยเฉพาะเมื่อข้อมูลมีจำกัด เทคนิคนี้จะทำการแบ่งข้อมูลออกเป็น K ส่วน ซึ่งเรียกว่า folds โดยในแต่ละรอบการประเมินหนึ่ง fold จะถูกใช้เป็นชุดทดสอบ (Test Set) หรือชุดตรวจสอบ (Validation Set) ขณะที่ folds ที่เหลือจะรวมกันเป็นชุดฝึก (Training Set) แบบจำลองจะได้รับการฝึกและทดสอบ K ครั้ง โดยการเปลี่ยน fold ที่ใช้เป็นชุดทดสอบในแต่ละครั้ง วิธีการนี้ช่วยให้การประเมินแบบจำลองมีความแม่นยำและสม่ำเสมอมากขึ้น แสดงดังภาพที่ 2.14

หลังจากเสร็จสิ้นกระบวนการ K-fold Cross-validation แบบจำลองจะถูกทดสอบเพิ่มเติมกับข้อมูลที่ไม่ได้ใช้ในการฝึกหรือในการ Cross-validation ซึ่งเรียกว่า test set เพื่อประเมินความแม่นยำที่แท้จริงของแบบจำลองในสภาพแวดล้อมที่ไม่เคยเห็นมาก่อน



ภาพที่ 2.12 กระบวนการทำงานของ 4-fold Cross-validation

2.1.7 การปรับไฮเปอร์พารามิเตอร์ (Hyperparameter Tuning)

การปรับไฮเปอร์พารามิเตอร์เป็นกระบวนการสำคัญในการพัฒนาแบบจำลองการเรียนรู้ของเครื่อง เนื่องจากไฮเปอร์พารามิเตอร์ที่เลือกใช้นั้นมีผลต่อประสิทธิภาพของแบบจำลองอย่างมาก ไม่น้อยไปกว่าพารามิเตอร์ที่เรียนรู้จากข้อมูล การกำหนดค่าไฮเปอร์พารามิเตอร์จะทำล่วงหน้าก่อนการฝึกแบบจำลอง ซึ่งหากไฮเปอร์พารามิเตอร์ถูกตั้งค่าอย่างเหมาะสม จะช่วยเพิ่มประสิทธิภาพการทำนายของแบบจำลองได้แม่นยำยิ่งขึ้น

ไฮเปอร์พารามิเตอร์ คือค่าที่ผู้ใช้กำหนดเองก่อนที่จะเริ่มการเรียนรู้ เช่น สำหรับ Random Forest การตั้งค่า `n_estimators` เป็นจำนวนต้นไม้ที่ใช้ในการสร้างแบบจำลอง การกำหนดค่าไฮเปอร์พารามิเตอร์อย่างเหมาะสมมีความสำคัญเพื่อควบคุมการฝึกแบบจำลองและให้ได้ผลลัพธ์ที่ดีที่สุด โดยวิธีค้นหาไฮเปอร์พารามิเตอร์ วิธีการค้นหาไฮเปอร์พารามิเตอร์ที่เหมาะสมมีหลายวิธี เช่น Manual Search Grid Search และ Random Search มีรายละเอียดดังนี้

1) Manual Search คือวิธีการที่เลือกค่าไฮเปอร์พารามิเตอร์จากประสบการณ์และความคิดเห็นส่วนตัว โดยสร้างแบบจำลองจากค่าที่เลือกและทำการวัดความแม่นยำของแบบจำลองนั้น ๆ เป็นระยะ ๆ จนกว่าจะได้ค่าความแม่นยำที่พึงพอใจ วิธีนี้ขึ้นอยู่กับความรู้และประสบการณ์ของผู้พัฒนาเพื่อกำหนดค่าไฮเปอร์พารามิเตอร์ที่ดีที่สุดสำหรับแบบจำลอง

2) Grid Search คือ เทคนิคการค้นหาแบบกริดที่ใช้ในการค้นหาค่าไฮเปอร์พารามิเตอร์ที่เหมาะสมอย่างเป็นระบบ โดยจะสร้างตารางของค่าที่กำหนดไว้ล่วงหน้าและทำการทดสอบแต่ละชุดพารามิเตอร์อย่างครบถ้วน จากนั้นจะประเมินประสิทธิภาพของแบบจำลองที่สร้างขึ้นจากชุดพารามิเตอร์เหล่านั้น เมื่อการทดสอบครบทุกชุดแล้ว แบบจำลองที่มีชุดพารามิเตอร์ที่ให้ผลลัพธ์แม่นยำที่สุดจะถือว่าดีที่สุด แม้ว่าวิธีนี้จะเข้าใจง่ายและตรงไปตรงมา แต่การใช้ Grid Search อาจใช้เวลาค่อนข้างมาก โดยเฉพาะเมื่อจำนวนชุดพารามิเตอร์และความละเอียดของกริดสูง (BDI, 2024)

3) Random Search เป็นเทคนิคที่ใช้ในการค้นหาค่าพารามิเตอร์ที่ดีที่สุดสำหรับแบบจำลองการเรียนรู้ของเครื่อง โดยการสุ่มเลือกค่าพารามิเตอร์จากช่วงที่กำหนดไว้ ซึ่งแตกต่างจาก Grid Search ที่จะทดสอบทุกค่าที่เป็นไปได้ในช่วงที่กำหนดไว้ Random Search จะเลือกทดสอบค่าบางส่วนที่สุ่มมาเท่านั้น ทำให้ใช้เวลาน้อยกว่าและสามารถค้นหาพารามิเตอร์ที่เหมาะสมได้อย่างรวดเร็ว (Bergstra & Bengio, 2012)

2.1.8 การประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลอง (Model Evaluation and Comparison)

เพื่อประเมินและเปรียบเทียบประสิทธิภาพแบบจำลองการถดถอยพหุผลลัพธ์โดยทั่วไป จะคำนวณค่าเฉลี่ยของความคลาดเคลื่อนในการทำนายตัวแปรตามแต่ละตัวแปรแยกกัน จากนั้นจึงนำมาหาค่าเฉลี่ย เพื่อประเมินประสิทธิภาพของแบบจำลอง โดยมักจะวัดจากค่าดังนี้ 1. ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) 2. ค่ารากที่สองของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (Root Mean Squared Error : RMSE) เนื่องจากค่า MSE เป็นค่าวัดความถูกต้องของการทำนายที่วัดจากการยกกำลังสองของค่าความคลาดเคลื่อน โดยจะให้ความสำคัญกับค่าทำนายที่แตกต่างไปจากค่าจริงมาก จึงทำให้อ่อนไหวต่อค่าความคลาดเคลื่อนที่มีขนาดใหญ่ หากมีค่าความคลาดเคลื่อนสูงเมื่อยกกำลังจะทำให้มีค่ามาก จึงนิยมใช้ RMSE ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง

1) ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) มีสมการดังนี้

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

2) ค่ารากที่สองของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (Root Mean Squared Error: RMSE) มีสมการดังนี้

$$RMSE = \sqrt{MSE}$$

การเปรียบเทียบประสิทธิภาพของทั้งค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) และค่ารากที่สองของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (RMSE) หากค่าของการคำนวณมีค่าน้อยแสดงว่าแบบจำลองมีประสิทธิภาพมาก (Borchani et al., 2015)

3) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) เป็นการประเมินความแม่นยำของแบบจำลองการถดถอยโดยการวัดผลต่างสัมบูรณ์โดยเฉลี่ยระหว่างค่าที่คาดการณ์ไว้กับค่าเป้าหมาย ข้อดีของมาตรวัดนี้คือมีความทนทานต่อค่าผิดปกติมากกว่าวิธีอื่น โดยให้การแสดงข้อผิดพลาดที่สมดุล และตีความได้ง่ายเนื่องจากไม่สนใจทิศทางของความคลาดเคลื่อน (Ahmed, 2023) โดยมีสมการดังนี้

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

โดย n คือ จำนวนข้อมูลทดสอบ
 y_i คือ ค่าจริงของข้อมูลลำดับที่ i เมื่อ $i = 1, 2, 3, \dots, n$
 \hat{y}_i คือ ค่าทำนายของข้อมูลลำดับที่ i เมื่อ $i = 1, 2, 3, \dots, n$

4) ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error: MAPE)

ซึ่งใช้การรายงานผลเป็นร้อยละความคลาดเคลื่อนระหว่างค่าที่ทำนายกับค่าผลลัพธ์จริง ทำให้สามารถทำความเข้าใจถึงระดับความแม่นยำได้ แต่ยังมีข้อจำกัดอยู่คือตัวหารของสมการคือค่าจริง หากค่าจริงเป็น 0 จะไม่สามารถหาค่าได้ หรือหากใกล้เคียง 0 จะทำให้ค่าความคลาดเคลื่อนมีค่าผิดปกติ โดยมีสมการดังนี้

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i|} \times 100$$

ค่าความคลาดเคลื่อนเฉลี่ยเปอร์เซ็นต์ (MAPE) ไม่สามารถคำนวณได้เมื่อค่าจริงเท่ากับศูนย์ เนื่องจากการหารด้วยค่าจริง จึงได้มีการพัฒนาตัวชี้วัดประสิทธิภาพใหม่ที่เรียกว่า ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร

5) ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (Symmetric Mean

Absolute Percentage Error: SMAPE) เพื่อแก้ไขปัญหาดังกล่าว SMAPE เป็นตัววัดความผิดพลาดที่ปรับปรุงจาก MAPE โดยเปลี่ยนตัวหารจากค่าจริงเพียงอย่างเดียว เป็นค่าเฉลี่ยระหว่างค่าจริงและค่าที่ทำนายในรูปแบบค่าสัมบูรณ์ ซึ่งช่วยลดผลกระทบของค่าจริงที่เป็นศูนย์หรือใกล้ศูนย์ในการคำนวณ (Hmong.in.th., n.d.)

$$\text{SMAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}} \times 100$$

y_i คือ ค่าจริงของข้อมูลลำดับที่ i เมื่อ $i = 1, 2, 3, \dots, n$

\hat{y}_i คือ ค่าทำนายของข้อมูลลำดับที่ i เมื่อ $i = 1, 2, 3, \dots, n$

2.1.9 Shapley Additive Explanation (SHAP)

Shapley Additive Explanation เป็นเครื่องมือแสดงภาพที่สามารถใช้ในการอธิบายการทำนายของแบบจำลองการเรียนรู้ของเครื่อง SHAP ใช้แนวคิดพื้นฐานจาก ค่า Shapley (Shapley Value) ที่มาจาก ทฤษฎีเกม (Game Theory) มีหลักการคือ ตัวแปรในแบบจำลองจะถูกมองว่าเป็น "ผู้เล่น" และการคำนวณค่า Shapley เพื่อค่าประเมินผลกระทบของตัวแปรแต่ละตัวในแบบจำลอง กล่าวคือ ผลกระทบจากการร่วมมือของผู้เล่นแต่ละคนเป็นเท่าไร โดยเฉลี่ย การเปลี่ยนแปลงของการคาดการณ์เมื่อมีตัวแปรเข้าร่วมชุดข้อมูลแต่ละชุด หากมีค่า Shapley Value สูง แสดงว่าทำให้ผลการทำนายมีความแม่นยำขึ้นมาก โดยมีสมการและแนวคิดของ Shapley Value ดังนี้

ขนาดของผลกระทบ ϕ_j ที่มีค่าของตัวแปรตัวที่ j ส่งผลให้การทำนายของจุดข้อมูลที่พิจารณาต่างจากค่าเฉลี่ยของทั้งชุดข้อมูล เช่น หากแบบจำลองที่ใช้เป็นแบบจำลองเชิงเส้น

$$\hat{f}(x) = \beta_0 + \beta_1 + \dots + \beta_p x_p$$

ค่าผลกระทบของ ϕ_j ที่ค่าของตัวแปรที่ j จะเขียนได้เป็น

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

โดย $\beta_j x_j$ คือการประเมินผลกระทบของตัวแปร x_j ณ จุดข้อมูลนั้น และ $E(\beta_j X_j)$ คือค่าเฉลี่ยของผลกระทบของตัวแปรนี้จากจุดข้อมูลทั้งหมด ซึ่งค่า ϕ_j คือค่าผลต่างของการทำนายเมื่อมีตัวแปรนั้นเทียบกับค่าเฉลี่ยของการทำนาย เมื่อนำค่าผลกระทบจากตัวแปรทุกตัวมารวมกัน จะได้ผลลัพธ์ของการทำนายเท่ากับผลต่างของค่าทำนายจุดนั้นกับค่าเฉลี่ยการทำนายของแบบจำลอง ดังสมการ

$$\begin{aligned} \sum_{j=1}^p \phi_j(\hat{f}) &= \sum_{j=1}^p (\beta_j x_j - E(\beta_j X_j)) \\ &= (\beta_0 + \sum_{j=1}^p \beta_j x_j) + (\beta_0 + \sum_{j=1}^p E(\beta_j X_j)) \\ &= \hat{f}(x) - E(\hat{f}(x)) \end{aligned}$$

(ปฏิภาณ ประเสริฐสม และ พีรต สามะศิริ, 2566)

2.2 งานวิจัยที่เกี่ยวข้อง

Poufinas et al.(2023) ได้ศึกษาการทำนายมูลค่าการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ เนื่องจากมีผลต่อกระแสเงินสด การตั้งราคาเบี้ยประกันภัย และการจัดการความเสี่ยงของบริษัทประกันภัย โดยมีการนำเสนอตัวแปรที่ไม่ใหม่ในการเรียกร้องค่าสินไหมทดแทน คือ สภาพอากาศและยอดขายทำนายด้วยแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ดังนี้ 1. Support Vector Machines (SVM) 2. Decision Trees 3. Random Forests 4. Xgboost เพื่อทำนายมูลค่าการเรียกร้องค่าสินไหมทดแทนเฉลี่ยต่อรถยนต์ที่ทำประกันในแต่ละไตรมาส ใช้ข้อมูลการเรียกร้องค่าสินไหมทดแทนจากพอร์ตโฟลิโอประกันภัยรถยนต์ในกรุงเอเธนส์ ประเทศกรีซ ตั้งแต่ปี ค.ศ. 2008 ถึง ค.ศ. 2020 ผลการวิจัยแสดงให้เห็นว่าตัวแปรที่มีอิทธิพลมากที่สุด 3 ตัว ได้แก่ ยอดขายรถยนต์ใหม่ที่มีการชะลอข้อมูล 3 ไตรมาสและ 1 ไตรมาส และอุณหภูมิต่ำสุดของสถานีอากาศ Elefsina ที่ชะลอข้อมูล 3 ไตรมาส และผลการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง พบว่าแบบจำลองที่ดีที่สุดคือ Random Forests และรองลงมาคือ XGBoost

Kumar et al. (2020) ได้ศึกษาการทำนายจำนวนมูลค่าการเรียกร้องค่าสินไหมทดแทนจากการประกันภัยรถยนต์ในประเทศอินเดีย ซึ่งใช้ข้อมูลการเรียกร้องค่าสินไหมทดแทนประกันภัยรถยนต์ต่อทรัพย์สินตั้งแต่ปี ค.ศ.1981 ถึงปี ค.ศ. 2016 สำหรับวิเคราะห์ โดยใช้แบบจำลองหลายประเภทคือ Generalized

Linear Model (GLM) สำหรับการวิเคราะห์ข้อมูลที่มีการแจกแจงแบบไม่ปกติ ARIMA (Autoregressive Integrated Moving Average) ใช้ข้อมูลในอดีตเพื่อทำนายอนาคต และ Artificial Neural Network (ANN) ทำนายข้อมูลที่มีโครงสร้างไม่เชิงเส้น โดยมีความสามารถในการปรับตัวเองตามข้อมูลที่ได้รับ และได้ทำการเปรียบเทียบ ผลการทำนายกับข้อมูลจริงของมูลค่าการเรียกร้อยค่าสินไหมทดแทนในช่วง 36 ปี ผลลัพธ์การวิจัยพบว่าแบบจำลอง ANN มีประสิทธิภาพของการทำนายดีที่สุด โดยมีค่า RMSE เท่ากับ 0.17601 ในขณะที่แบบจำลอง GLM ได้ค่า RMSE เท่ากับ 1.183 และ ARIMA ได้ค่า RMSE เท่ากับ 1.3748 ดังนั้นแบบจำลอง ANN มีความแม่นยำสูงกว่าแบบจำลองอื่นๆในการทำนายมูลค่าการเรียกร้อยค่าสินไหมทดแทนประกันภัยรถยนต์ต่อทรัพย์สิน ซึ่งจะช่วยให้บริษัทประกันภัยสามารถคาดการณ์การเรียกร้อยค่าสินไหมทดแทนในอนาคตได้แม่นยำยิ่งขึ้น

Chen et al. (2020) ศึกษาการสร้างแบบจำลอง LightGBM (Light Gradient Boosting Machine) เพื่อทำนายความถี่ในการเรียกร้อยค่าสินไหมทดแทนในประกันภัยรถยนต์ ข้อมูลที่ใช้มาจากบริษัทประกันภัยรถยนต์ของสหรัฐอเมริกา มีทั้งหมด 10,305 แถว และ 24 ตัวแปร แบ่งออกเป็นสามกลุ่มได้แก่ ตัวแปรที่เกี่ยวกับเจ้าของรถ ตัวแปรที่เกี่ยวกับยานพาหนะ ตัวแปรที่เกี่ยวกับกรรมสิทธิ์ เพื่อใช้ในการทำนายความถี่ในการเรียกร้อยค่าสินไหมทดแทนในประกันภัยรถยนต์ โดยมีการเปรียบเทียบวิธี LightGBM กับวิธีต่างๆ ได้แก่ 1. Gradient Decision Boosting Tree 2. Artificial Neural Network 3. Support Vector Machine 4. Generalized Linear Models ซึ่งพบว่า LightGBM สามารถจัดการกับข้อมูลที่ไม่สมดุลและมีความสามารถในการทำนายที่ดีที่สุดเมื่อเทียบกับแบบจำลองอื่น มีประสิทธิภาพดังนี้ Accuracy เท่ากับ 0.835 และค่า AUC เท่ากับ 0.907

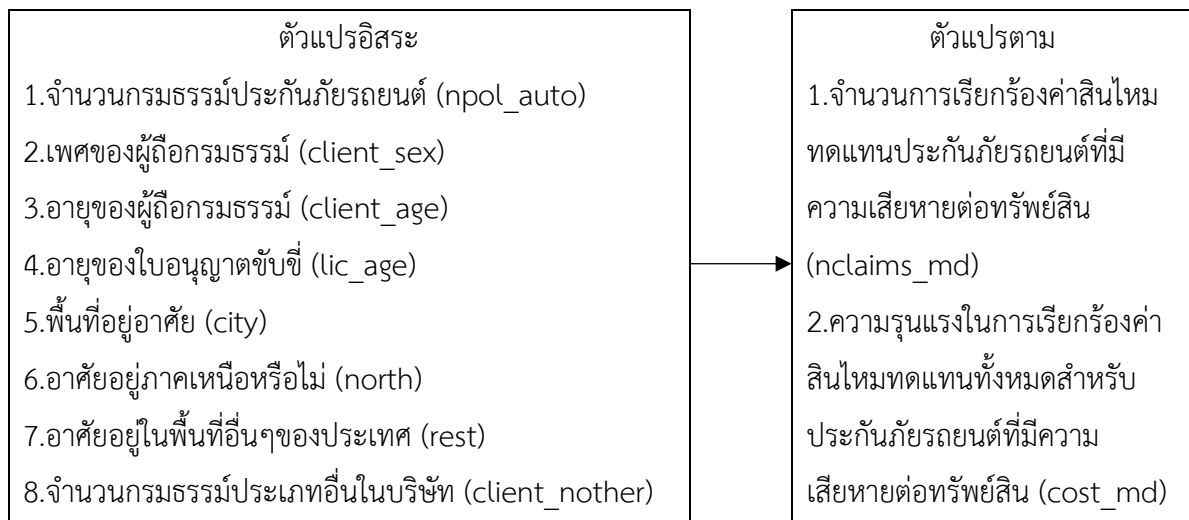
Jin (2021) ศึกษาการทำนายความรุนแรงในการเรียกร้อยค่าสินไหมทดแทนรายบุคคล โดยพิจารณามูลค่าการเรียกร้อยค่าสินไหมทดแทนที่มีการรายงานแต่ยังไม่ได้รับการชำระ (Reported But Not Settled) สำหรับประกันวินาศภัยแบบจำลองการเรียนรู้ของเครื่องประเภท ensemble ML models ทั้งหมด 3 แบบจำลอง ได้แก่ 1. XGBoost 2. Random Forest 3. Extra Trees มีข้อมูลสำหรับการวิจัยทั้งหมด 500,914 ข้อมูล และ 32 ตัวแปร โดยใช้วิธีการ Grid Search เพื่อค้นหาค่าพารามิเตอร์ที่ดีที่สุดสำหรับแต่ละแบบจำลอง และนำมาทดสอบกับชุดข้อมูลทดสอบ ผลการวิจัยพบว่า XGBoost ให้ค่า RMSE ที่ต่ำที่สุดและมีเวลาในการคำนวณที่เร็วที่สุดเมื่อเปรียบเทียบกับวิธี ensemble ML models อื่น ๆ นอกจากนี้ XGBoost ยังช่วยให้สามารถวิเคราะห์และตีความปัจจัยที่มีผลต่อการทำนายมูลค่าการเรียกร้อยค่าสินไหมทดแทน โดยใช้เทคนิค Tree SHAP พบว่าตัวแปรที่มีอิทธิพลต่อการทำนายมากที่สุดได้แก่ อายุของผู้เอาประกันภัย ไตรมาสที่เกิดอุบัติเหตุ จำนวนปีของการพัฒนาระหว่างชุดฝึกและชุดทดสอบ และส่วนของร่างกายที่ได้รับบาดเจ็บ ตามลำดับ

กิตติศักดิ์ และคณะ (2564) ได้ศึกษาการตรวจจับการเรียกร้องค่าสินไหมทดแทนประกันภัยรถยนต์ ซึ่งมีวัตถุประสงค์ในงานวิจัยคือ 1) หาปัจจัยที่มีอิทธิพลต่อการเรียกร้องค่าสินไหมทดแทน 2) เปรียบเทียบประสิทธิภาพการทำนายของวิธีการคัดเลือกตัวแปรอิสระ 3 วิธี คือ ไม่มีการคัดเลือกคุณลักษณะ วิธีการถดถอยลอจิสติกทีละขั้น และวิธีต้นไม้ตัดสินใจ 3) เปรียบเทียบประสิทธิภาพการทำนายของ 3 อัลกอริทึม คือ วิธีนาอิวเบย์ วิธีสุ่มป่าไม้ และวิธีบูสต์ติงปรับได้ และ 4) เปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมร่วมกับวิธีการคัดเลือกคุณลักษณะ จากการวิจัยพบว่าคุณลักษณะที่มีอิทธิพลต่อการเรียกร้องค่าสินไหมทดแทนคือ ลักษณะการเกิดเหตุ ภูมิภาคที่เกิดเหตุ อายุผู้เอาประกันภัย ประเภทการซ่อม ทุนประกันภัย และจำนวนเงินที่จ่ายรวมทั้งการเรียกร้องค่าสินไหมทดแทน

Sun et al. (2024) ได้ศึกษาการประเมินความเสี่ยงการประกันภัยรถยนต์ซึ่งต้องการหาปัจจัยที่มีผลต่อประกันภัยรถยนต์และเปรียบเทียบประสิทธิภาพของแบบจำลองโดยใช้แบบจำลอง Actuarial Transformer (AT) ร่วมกับแบบจำลองแบบ Tree-Based คือ XGBoost, LightGBM, CatBoost เพื่อเพิ่มความแม่นยำของการทำนาย วัดประสิทธิภาพโดยใช้ค่า Poisson Deviance และ Improvement Index และยังใช้ SHAP เพื่อหาคุณลักษณะที่มีผลต่อการทำนาย จากการศึกษพบว่าแบบจำลองที่ดีที่สุดคือ AT(CatBoost) ซึ่งอธิบายได้ว่าการนำ AT เข้ามาสามารถช่วยความแม่นยำของแบบจำลองเพิ่มขึ้นได้ และตัวแปรที่ส่งผลต่อการทำนายคือ โบนัสสมาลัส (BonusMalus), อายุของผู้ขับขี่ (DriveAge), ความหนาแน่นของประชากรในเมืองที่ผู้ขับขี่อาศัยอยู่ (Density) และ อายุของรถ (VehAge)

ผู้วิจัยได้สรุปแบบจำลองและวิธีการที่ใช้ในการสร้างแบบจำลอง และสรุปตัวแปรที่ส่งผลต่อการทำนายการเรียกร้องค่าสินไหมทดแทนที่ได้จากการทบทวนวรรณกรรม ดังตารางที่ 2.1 และตารางที่ 2.2 ตามลำดับ

2.3 กรอบแนวคิดการวิจัย



ภาพที่ 2.13 กรอบแนวคิดการวิจัย

บทที่ 3

วิธีการดำเนินงานวิจัย

งานวิจัยเรื่อง การวิเคราะห์การถดถอยพหุผลลัพ์ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ มีวัตถุประสงค์เพื่อสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทน สำหรับการเตรียมและวิเคราะห์ข้อมูล ผู้วิจัยใช้ภาษา Python ผ่าน Google Colaboratory เวอร์ชัน 3.10.12 โดยมีขั้นตอนการดำเนินการวิจัย ดังนี้

3.1 การทำความเข้าใจธุรกิจ

3.2 การศึกษาและทำความเข้าใจข้อมูล

3.3 การเตรียมข้อมูล

3.3.1 การทำความสะอาดข้อมูล

3.3.2 การแปลงข้อมูล

3.4 การสร้างแบบจำลองการถดถอยพหุผลลัพ์โดยใช้การเรียนรู้ของเครื่อง

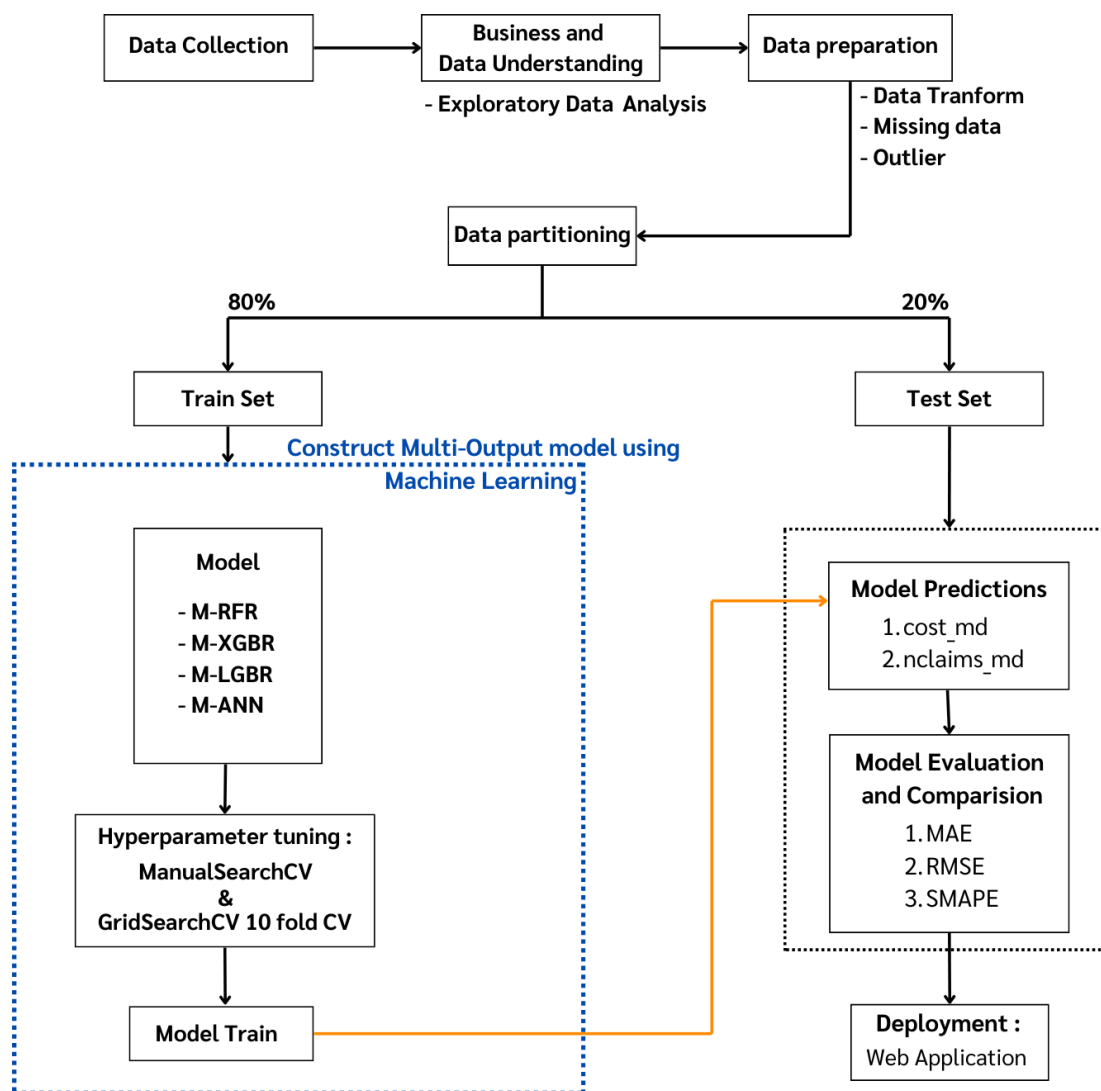
3.4.1 การแบ่งข้อมูล

3.4.2 การสร้างแบบจำลองการเรียนรู้ของเครื่อง

3.5 การประเมินและเปรียบเทียบประสิทธิภาพแบบจำลอง

3.6 การนำแบบจำลองไปใช้งานจริง

โดยมีรายละเอียดขั้นตอนของการดำเนินงานวิจัย ดังภาพที่ 3.1



ภาพที่ 3.1 ขั้นตอนการดำเนินงานวิจัย

3.1 การทำความเข้าใจธุรกิจ (Business Understanding)

ในขั้นตอนแรก จะทำการศึกษาว่าธุรกิจประกันภัยรถยนต์คืออะไร มีความสำคัญต่อภาคเศรษฐกิจอย่างไร มีความเสี่ยงด้านใดบ้างที่ทำให้ธุรกิจประสบความล้มเหลว รวมถึงศึกษาว่ามีปัจจัยใดบ้างที่ส่งผลกระทบต่อความเสี่ยง

3.2 การศึกษาและทำความเข้าใจข้อมูล (Data Understanding)

ในขั้นตอนนี้จะทำความเข้าใจข้อมูล จัดทำ data dictionary และใช้ Exploratory Data Analysis: EDA ในการพรรณนาข้อมูล ได้แก่ การหาค่าเฉลี่ย ร้อยละ ส่วนเบี่ยงเบนมาตรฐาน ค่าสูงสุดต่ำสุด การกระจายตัวของข้อมูล และ scatter plot ดูความสัมพันธ์ของข้อมูล

ข้อมูลที่ใช้ในการศึกษา คือข้อมูลการเรียกร้องค่าสินไหมทดแทนซึ่งเป็นข้อมูลทุติยภูมิ เป็นข้อมูลของบริษัทประกันภัยแห่งหนึ่งในประเทศสเปน ตั้งแต่วันที่ 1 มกราคม ค.ศ. 2006 ถึง 31 ธันวาคม ค.ศ. 2015 ระยะเวลารวม 10 ปี มีทั้งหมด 10 ตัวแปร จำนวน 80,924 แถว ได้มาจากการวิจัยของ Catalina Bolance and Raluca Vernic

3.3 การเตรียมข้อมูลก่อนการวิเคราะห์ (Data Preparation)

วัตถุประสงค์ของขั้นตอนนี้ คือการจัดการกับข้อมูลให้อยู่ในรูปแบบที่เหมาะสมก่อนการวิเคราะห์ โดยใช้โปรแกรม Google Collaboratory ในการเตรียมข้อมูล มีรายละเอียดการเตรียมข้อมูลดังนี้

3.3.1 การทำความสะอาดข้อมูล (Data Cleansing)

การทำความสะอาดข้อมูลเป็นการตรวจสอบข้อมูลและปรับปรุงข้อมูลให้มีคุณภาพมากขึ้น โดยหาจุดผิดของข้อมูลและหาวิธีการปรับปรุงข้อมูลนั้น ในงานวิจัยนี้ประกอบด้วย

1) การตรวจสอบความถูกต้องและความแนบเนียนของข้อมูล

เป็นการตรวจสอบข้อมูลว่า มีการบันทึกมาถูกต้องแนบเนียนหรือไม่ จะทำการตรวจสอบความแนบเนียนภายนอก (External consistency) ในแต่ละตัวแปร จะพิจารณาค่าที่เป็นไปได้ของข้อมูล หากพบว่าในแถวใด มีค่าของข้อมูลที่ไม่ถูกต้องแม้เพียงหนึ่งตัว จะทำการตัดข้อมูลทั้งแถวนั้นออก

2) การจัดการกับค่าสูญหาย (Missing Value)

ทำการตรวจสอบทุกแถว หากพบว่ามีข้อมูลสูญหายในแถวใด จะทำการตัดทั้งแถวนั้นออก

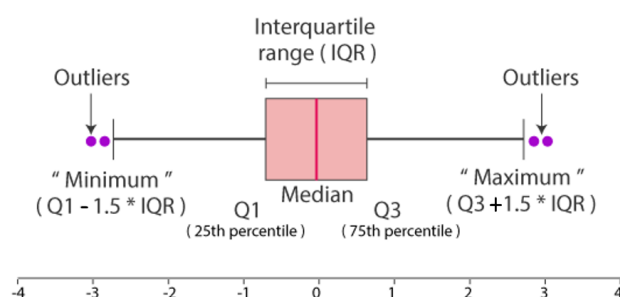
3) การจัดการกับค่านอกเกณฑ์ (Outliers)

โดยทั่วไปหากค่านอกเกณฑ์มีจำนวนน้อยกว่า 5% ของข้อมูลทั้งหมด จะคงค่านอกเกณฑ์ไว้เพื่อนำมาวิเคราะห์ต่อไป สำหรับข้อมูลเชิงคุณภาพที่มีค่าเป็น 0 และ 1 ไม่จำเป็นต้องจัดการกับค่านอกเกณฑ์ แต่สามารถใช้วิธีการถ่วงน้ำหนักเพื่อให้แบบจำลองรับรู้ถึงความไม่สมดุลของข้อมูล (Kutner et al., 2004) มีหลักการในการหาค่านอกเกณฑ์ดังนี้ ค่านอกเกณฑ์คือค่าที่มากกว่า $Q3 + 1.5 * IQR$ หรือ $Q1 - 1.5 * IQR$

เมื่อ Q1 คือ ค่าจากเปอร์เซ็นต์ไทล์ที่ 25 ของข้อมูล

Q3 คือ ค่าจากเปอร์เซ็นต์ไทล์ที่ 75 ของข้อมูล

IQR (Interquartile Range) คือ $Q3 - Q1$ แสดงรายละเอียดดังภาพที่ 3.3



ภาพที่ 3.3 ลักษณะของ Box plot

อย่างไรก็ตาม ข้อมูลทางการประกันภัยรถยนต์มีความเฉพาะเจาะจง โดยปกติในรอบปีของการประกันภัยรถยนต์ ผู้เอาประกันภัยส่วนใหญ่ไม่ได้มีการเรียกร้องค่าสินไหมทดแทน เมื่อเกิดการเรียกร้องค่าสินไหมทดแทนขึ้นค่าเหล่านี้จึงเป็นค่านอกเกณฑ์ และมูลค่าของการเรียกร้องค่าสินไหมทดแทนนั้นอาจสูงซึ่งผลต่อการดำเนินธุรกิจประกันภัย การตัดค่านอกเกณฑ์นี้ออกอาจทำให้การวิเคราะห์ไม่สะท้อนความเป็นจริง นอกจากนี้ค่านอกเกณฑ์เป็นส่วนสำคัญของการวิเคราะห์ความเสี่ยงของบริษัทประกันภัย และบริษัทประกันภัยจำเป็นต้องใช้ข้อมูลที่ครอบคลุมทุกสถานการณ์เพื่อให้สามารถกำหนดเบี้ยประกันภัยที่เหมาะสม ดังนั้นผู้วิจัยจะทำการวิเคราะห์เบื้องต้นว่า ค่านอกเกณฑ์มีจำนวนร้อยละเท่าใด แต่จะไม่ทำการตัดค่านอกเกณฑ์ออก โดยจะนำข้อมูลทั้งหมดที่ผ่านการเตรียมข้อมูลแล้วไปสร้างแบบจำลอง

3.3.2 การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูล เพื่อให้ข้อมูลพร้อมที่จะนำเข้าในการสร้างแบบจำลองทำนายการเรียกร้องค่าสินไหมทดแทน การแปลงข้อมูลนอกจากจะช่วยป้องกันปัญหาค่าทำนายที่ไม่เหมาะสมแล้ว ยังช่วยลดความแปรปรวนในข้อมูลและเพิ่มประสิทธิภาพการทำนายด้วยแบบจำลองการเรียนรู้ของเครื่อง ในงานวิจัยนี้ประกอบด้วย

- 1) แปลงข้อมูลเชิงคุณภาพให้อยู่ในรูปของตัวเลข
- 2) แปลงข้อมูลของตัวแปรตาม ได้แก่ ความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทน (Y_1 และ Y_2 ตามลำดับ) เนื่องจากค่าตัวแปรตามทั้งสองไม่สามารถติดลบได้

ในการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับตัวแปรตามหรือตัวแปรผลลัพธ์ด้วยแบบจำลองการเรียนรู้ของเครื่อง อาจเกิดปัญหาค่าทำนายของแบบจำลองที่มีค่าติดลบได้ ส่งผลให้ผลลัพธ์ไม่สอดคล้องกับบริบทตัวแปรผลลัพธ์ เพื่อแก้ไขปัญหานี้จึงมีการแปลงข้อมูลของตัวแปรโดยใช้ฟังก์ชันลอการิทึมธรรมชาติ (Logarithmic Transformation) ในการแปลงค่าตัวแปรผลลัพธ์ก่อนนำไปใช้ในการฝึกสอนแบบจำลอง (Model Training) มีการปรับค่า $\ln(y)$ โดยการเพิ่มค่าคงที่ 10^{-2} เพื่อหลีกเลี่ยงปัญหาค่าศูนย์ $\ln(0)$ ที่ไม่สามารถหาค่าได้ การเพิ่มค่าคงที่ 10^{-2} ช่วยป้องกันปัญหาค่าทำนายที่เป็นลบ โดยสมการที่ใช้ในการแปลงข้อมูลมีดังนี้

$$y' = \ln(y + 10^{-2})$$

หลังจากการทำนายโดยแบบจำลองเสร็จสิ้นแล้ว ค่าทำนายจะถูกแปลงกลับมาเป็นค่าตัวแปรเดิม (y) เพื่อให้ผลลัพธ์สอดคล้องกับข้อมูลจริงและเหมาะสมต่อการนำไปใช้ในเชิงปฏิบัติ โดยสมการที่ใช้ในการแปลงกลับคือ

$$y = \exp(y') - 10^{-2}$$

3.4 การสร้างแบบจำลองการถดถอยพหุผลลัพธ์โดยใช้การเรียนรู้ของเครื่อง (Modeling)

ในงานวิจัยนี้ทำการการสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อทำนายจำนวน(ความถี่) และความรุนแรงในการเรียกร้องค่าสินไหมทดแทน จำนวน 4 แบบจำลอง ได้แก่

- 1) การถดถอยพหุผลลัพธ์แบบการสุ่มป่าไม้ (Multi-Output Random Forest Regression: M-RFR)
- 2) การถดถอยพหุผลลัพธ์แบบเอ็กซ์ตรีมเกรเดียนต์บูสติง (Multi-Output XGBoost Regression: M-XGBR)
- 3) การถดถอยพหุผลลัพธ์แบบไลต์เกรเดียนต์บูสติง (Multi-Output Light Gradient Boosting Regression: M-LGBR)
- 4) การถดถอยพหุผลลัพธ์แบบโครงข่ายประสาทเทียม (Multi-Output Artificial Neural Network: M-ANN)

ก่อนที่จะทำการสร้างแบบจำลองจะทำการแบ่งข้อมูลที่ผ่านมาขั้นตอนการเตรียมข้อมูลมาแล้ว จากนั้นทำการปรับแต่งพารามิเตอร์เพื่อใช้ในการสร้างแบบจำลอง สำหรับการคัดเลือกตัวแปรอิสระที่จะนำมาใช้สร้างแบบจำลอง พบว่าข้อมูลที่ได้มีจำนวนตัวแปรอิสระไม่มาก กล่าวคือ มีจำนวน 8 ตัวแปร ดังนั้นผู้วิจัยจะนำตัวแปรอิสระทั้งหมดมาใช้ในการสร้างแบบจำลอง

3.4.1 การแบ่งข้อมูล (Data partitioning)

ข้อมูลที่ผ่านมาการเตรียมมาแล้วจะถูกแบ่งออกเป็นสองส่วน คือชุดข้อมูลฝึก (Train Set) คิดเป็น 80% จากข้อมูลทั้งหมด และชุดข้อมูลทดสอบ (Test Set) คิดเป็น 20% จากข้อมูลทั้งหมด หากข้อมูลของตัวแปรตามเป็นข้อมูลที่ไม่สมดุล (Imbalance data) จะทำการแบ่งข้อมูลด้วยวิธี stratify เพื่อให้สัดส่วนการแบ่งกลุ่มของข้อมูลทั้งสองชุดยังคงใกล้เคียงกัน

การฝึกแบบจำลอง งานวิจัยนี้ใช้วิธี K-fold Cross Validation โดยกำหนดค่า K เท่ากับ 10 หมายความว่าชุดข้อมูลฝึกจะถูกแบ่งเป็น 10 ส่วนย่อยที่มีขนาดเท่ากัน ในแต่ละรอบของการฝึกแบบจำลอง จะเลือกใช้ข้อมูล 9 ส่วนสำหรับฝึก และอีก 1 ส่วนที่เหลือใช้สำหรับทดสอบ แบบจำลองจะฝึกและทดสอบซ้ำกันทั้งหมด 10 รอบ โดยเปลี่ยนส่วนข้อมูลที่ใช้ทดสอบในแต่ละรอบ ทำให้ข้อมูลทั้งหมดถูกใช้ทั้งในการฝึกและทดสอบแบบสลับกัน กระบวนการนี้ช่วยเพิ่มความแม่นยำและความเสถียรของแบบจำลอง

3.4.2 แบบจำลองการเรียนรู้ของเครื่อง (Model)

การสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อทำนายความถี่และความรุนแรงในการเรียกร้องค่าสินไหมทดแทน ใช้ข้อมูลชุดฝึกในการพัฒนาและปรับแต่งไฮเปอร์พารามิเตอร์ของแบบจำลอง ซึ่งจำเป็นต้องกำหนดค่าพารามิเตอร์ที่เหมาะสมสำหรับแต่ละแบบจำลอง ในงานวิจัยนี้มีวิธีการปรับแต่งค่าพารามิเตอร์ดังนี้

ขั้นที่ 1 ปรับแต่งไฮเปอร์พารามิเตอร์ด้วยวิธี Manual Search เพื่อสำรวจค่าพารามิเตอร์เบื้องต้น

ขั้นที่ 2 นำค่าไฮเปอร์พารามิเตอร์จากขั้นตอนที่ 1. ไปกำหนดขอบเขตสำหรับการปรับแต่งไฮเปอร์พารามิเตอร์เพิ่มเติมด้วยวิธี GridSearchCV เพื่อทดสอบทุกค่าที่เป็นไปได้ในช่วงพารามิเตอร์ที่กำหนด

ผู้วิจัยใช้ Google Colaboratory ในการสร้างแบบจำลอง มี library ที่เกี่ยวข้องกับงานวิจัยแสดงในตารางที่ 3.1

ตารางที่ 3.1 library ที่ใช้สำหรับวิเคราะห์ข้อมูล

| library | Description |
|-------------------------|---|
| google.colab | ใช้สำหรับการเข้าถึงและทำงานร่วมกับ Google Drive |
| pandas | ใช้สำหรับการจัดการและวิเคราะห์ข้อมูลในรูปแบบ DataFrame เช่น CSV, Excel เป็นต้น |
| numpy | ใช้สำหรับการคำนวณเชิงตัวเลข เช่น คำนวณเกอเมตริก |
| pyarrow.parquet | ใช้สำหรับการอ่านและเขียนไฟล์ Parquet ซึ่งเป็นไฟล์ที่มีการบีบอัดและจัดเก็บข้อมูลแบบ Columnar |
| scipy.stats | ใช้สำหรับการทดสอบทางสถิติต่าง ๆ เช่น Pearson correlation |
| sklearn.datasets | ใช้สำหรับการสร้างชุดข้อมูลตัวอย่างสำหรับการทดสอบหรือฝึกอบรมแบบจำลอง |
| sklearn.ensemble | ใช้สำหรับการสร้างแบบจำลองการเรียนรู้แบบรวม (ensemble learning) เช่น ExtraTreesRegressor |
| sklearn.multioutput | ใช้สำหรับการสร้างแบบจำลองการทำนายหลายผลลัพธ์ |
| sklearn.model_selection | ใช้สำหรับการแบ่งชุดข้อมูล การทำ cross-validation (train_test_split) และการค้นหาค่าพารามิเตอร์ที่ดีที่สุดด้วย GridSearchCV |
| sklearn.metrics | ใช้ในการวัดผลของแบบจำลอง เช่น MAE, RMSE |
| matplotlib.pyplot | ใช้สำหรับการสร้างกราฟและแผนภูมิเพื่อแสดงผลข้อมูลที่เป็นภาพ |
| seaborn | ใช้สำหรับการสร้างกราฟที่มีความซับซ้อนและสวยงามมากขึ้น |
| time | ใช้สำหรับการจัดการและวัดเวลาในการประมวลผลของโค้ด |
| sklearn.neural_network | ใช้สำหรับการสร้างแบบจำลอง Artificial Neural Network (ANN) |
| lightgbm | ใช้สำหรับการสร้างแบบจำลอง LGBM |
| xgboost | สำหรับการสร้างแบบจำลอง Extreme Gradient Boosting |

ขั้นที่ 1 การปรับแต่งไฮเปอร์พารามิเตอร์ด้วยวิธี Manual Search มีขั้นตอน ดังนี้

1. ทดสอบด้วยค่าเริ่มต้น (default) ของแบบจำลอง เพื่อให้ทราบประสิทธิภาพเบื้องต้นของแบบจำลอง

2. เลือกไฮเปอร์พารามิเตอร์ที่ต้องการทดสอบ โดยทดสอบทีละพารามิเตอร์เพื่อตรวจสอบว่าการปรับค่าแต่ละค่าให้ประสิทธิภาพที่ดีขึ้นหรือไม่ เช่น Random Forest Regression ที่ทำการทดสอบ `n_estimators` ทดสอบที่ค่า 100, 200, 300, 400, 500 และ 1000

3. ทดสอบแต่ละพารามิเตอร์ในแบบจำลอง จากนั้นพิจารณาประสิทธิภาพของแบบจำลองที่ได้ เมื่อพบว่าการปรับค่าพารามิเตอร์ไปในทิศทางใด ๆ ทำให้ประสิทธิภาพของแบบจำลองลดลงหรือไม่ดีขึ้น จะถือว่าเป็นจุดที่เหมาะสมในการหยุดการทดสอบและนำค่าไฮเปอร์พารามิเตอร์ไปกำหนดขอบเขตของค่าไฮเปอร์พารามิเตอร์ สำหรับทำ GridSearchCV

ค่าเริ่มต้นของพารามิเตอร์ในแต่ละแบบจำลอง แสดงดังตารางที่ 3.2- 3.5 ตามลำดับ

ตารางที่ 3.2 ค่าเริ่มต้นของพารามิเตอร์สำหรับแบบจำลอง M-RFR

| ไฮเปอร์พารามิเตอร์ | ค่าเริ่มต้น |
|--------------------------------|-------------|
| <code>n_estimators</code> | 100 |
| <code>max_depth</code> | None |
| <code>min_samples_split</code> | 2 |
| <code>min_sample_leaf</code> | 1 |

`max_depth = None` หมายถึงไม่มีการกำหนดความลึกสูงสุดของต้นไม้ โดยโหนดจะขยายออกจนกว่าโหนดสุดท้ายของต้นไม้จะไม่สามารถแบ่งได้อีก หรือจำนวนตัวอย่างในแต่ละใบจะน้อยกว่า `min_samples_split`

ตารางที่ 3.3 ค่าเริ่มต้นของพารามิเตอร์สำหรับแบบจำลอง M-XGBR

| ไฮเปอร์พารามิเตอร์ | ค่าเริ่มต้น |
|-------------------------------|-------------|
| <code>n_estimators</code> | 100 |
| <code>eta</code> | 0.3 |
| <code>max_depth</code> | 6 |
| <code>min_child_weight</code> | 1 |
| <code>subsample</code> | 1 |

ตารางที่ 3.4 ค่าเริ่มต้นของพารามิเตอร์สำหรับแบบจำลอง M-LGBR

| ไฮเปอร์พารามิเตอร์ | ค่าเริ่มต้น |
|----------------------------|-------------|
| <code>learning_rate</code> | 0.1 |
| <code>n_estimators</code> | 100 |
| <code>max_depth</code> | -1 |
| <code>subsample</code> | 1 |

| | |
|------------------|-------|
| min_child_weight | 0.001 |
| num_leaves | 31 |
| min_data_in_leaf | 20 |

ตารางที่ 3.5 ค่าเริ่มต้นของพารามิเตอร์สำหรับแบบจำลอง M-ANN

| ไฮเปอร์พารามิเตอร์ | ค่าเริ่มต้น |
|--------------------|-------------|
| Hidden_layer_sizes | (100,) |
| Solver | adam |
| activation | relu |
| Learning_rate | constant |

Solver = adam หมายถึง แบบจำลองจะปรับค่าอัตราการเรียนรู้และน้ำหนักในโมเดลให้มีประสิทธิภาพสูงสุด

Activation = relu หมายถึง ฟังก์ชันในโหนดซ่อน ให้ค่าเป็น 0 ถ้าอินพุตติดลบ และคงค่าเดิมถ้าอินพุตเป็นบวก

learning_rate = constant หมายถึง ใช้อัตราการเรียนรู้คงที่ตลอดการฝึกแบบจำลอง

ขั้นที่ 2 การปรับแต่งไฮเปอร์พารามิเตอร์เพิ่มเติมด้วยวิธี GridSearchCV

หลังจากได้ขอบเขตไฮเปอร์พารามิเตอร์สำหรับแต่ละแบบจำลองแล้ว จะทำการทดสอบทุกที่เป็นไปได้ในช่วงพารามิเตอร์ที่กำหนดด้วยวิธี GridSearchCV โดย Google Colaboratory มี library GridSearchCV ซึ่งมีค่าพารามิเตอร์ แสดงดังตารางที่ 3.6

ตารางที่ 3.6 พารามิเตอร์สำหรับ GridSearchCV

| พารามิเตอร์ | ความหมายและค่าที่กำหนด |
|--------------|---|
| estimator | แบบจำลองที่ต้องการ เช่น RandomForest, SVM ฯลฯ |
| param_grid | กำหนดค่าพารามิเตอร์ต่าง ๆ ที่ทดสอบจากแบบจำลอง เช่น แบบจำลอง Random Forest จะกำหนดพารามิเตอร์ดังนี้ 'estimator__n_estimators': [100, 200, 300], 'estimator__max_depth': [10, 20, 30] เป็นต้น |
| Scoring | ค่าที่ใช้ในการประเมินผลการทำงานของแบบจำลอง ในงานวิจัยนี้กำหนดเป็น MAE |
| n_jobs | กำหนดเป็น -1 หมายความว่าใช้ทุก CPU core ที่มีอยู่เพื่อเร่งกระบวนการคำนวณ |
| refit | กำหนดเป็น True แบบจำลองจะถูกฝึกซ้ำด้วยค่าพารามิเตอร์ที่ดีที่สุด |
| cv | Cross Validation ในงานวิจัยนี้จะใช้ K-fold Cross validation ซึ่งกำหนด K=10 |
| pre_dispatch | จำนวนงานที่ถูกเตรียมไว้ล่วงหน้าก่อนที่การทำงานจะเริ่ม โดยกำหนดให้เท่ากับ 2*n_jobs (ค่าเริ่มต้น) นั่นคือเตรียมงานล่วงหน้าเป็นสองเท่าของจำนวนการทำงาน |

| | |
|--------------------|--|
| error_score | ค่าที่จะใช้แทนค่าเมื่อแบบจำลองไม่สามารถฝึกได้หรือเกิดข้อผิดพลาดในการประเมิน กำหนดเป็น nan |
| return_train_score | คะแนนของการฝึกซ้อม (training scores) กำหนดเป็น False หมายความว่าคะแนนค่าเฉพาะคะแนนในการทดสอบ (test scores) |
| verbose | ความละเอียดของข้อความที่แสดงเมื่อกำลังประมวลผล กำหนดค่าเป็น 0 หมายความว่าจะไม่มีการพิมพ์ข้อความ |

การสร้างตัวแบบในงานวิจัยนี้จะคงค่าไฮเปอร์พารามิเตอร์ของ GridSearchCV ไว้ในทุก ๆ แบบจำลอง แต่จะมีการปรับเปลี่ยน estimator และ ค่า param_grid ของแต่ละแบบจำลอง เพื่อให้มีความเหมาะสมกับแบบจำลอง โดยแต่ละแบบจำลองมีขอบเขตของไฮเปอร์พารามิเตอร์ แสดงดังตารางที่ 3.7

ตารางที่ 3.7 ขอบเขตของไฮเปอร์พารามิเตอร์กำหนดใน GridSearchCV ของแต่ละแบบจำลอง

| แบบจำลอง | ไฮเปอร์พารามิเตอร์ | ค่าที่กำหนด |
|---|--|--|
| Multi-Output Random Forest Regression | n_estimators max_depth min_samples_split min_samples_leaf | [1100, 1200, 1300] [6, 7, 8, 9] [100, 110, 120, 130, 140] [40, 50, 60] |
| Multi-Output Extreme Gradient Boosting Regression | n_estimators max_depth learning_rate min_child_weight subsample | [100, 200, 300] [5, 10, 15, 20] [0.05, 0.1, 0.2] [10, 15, 20] [0.6, 0.7, 0.8] |
| Multi-Output Light Gradient Boosting Regression | n_estimators max_depth learning_rate num_leaves min_data_in_leaf | [100, 200, 300] [-1, 5, 15, 25] [0.05, 0.1, 0.2] [30, 35, 40, 45] [60, 65, 70] |
| Multi-Output Artificial Neural Network | hidden_layer_sizes solver activation learning_rate | [(64,),(128,),(128,64), (256,128), (512,256)] ['adam'] ['identity', 'relu'] ['constant', 'adaptive'] |

หลังจากที่ได้พารามิเตอร์ที่ดีที่สุดสำหรับแต่ละแบบจำลองจากวิธี GridSearchCV แล้ว จะนำพารามิเตอร์เหล่านั้นมาทดสอบกับชุดข้อมูลทดสอบที่แบบจำลองไม่เคยพบมาก่อน เพื่อประเมินและเปรียบเทียบประสิทธิภาพของแบบจำลองว่าแบบจำลองใดมีประสิทธิภาพดีที่สุด

3.5 การประเมินและเปรียบเทียบประสิทธิภาพแบบจำลอง (Evaluation)

3.5.1 การประเมินและเปรียบเทียบประสิทธิภาพแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลองในงานวิจัยนี้จะดำเนินการโดยนำแบบจำลองที่ผ่านการปรับแต่งพารามิเตอร์แล้ว มาทดสอบกับข้อมูลชุดทดสอบเพื่อวัดความสามารถในการทำนาย ซึ่งในงานวิจัยมีการทำนายสองผลลัพธ์ แต่ละผลลัพธ์จะถูกประเมินด้วยมาตรวัดดังนี้

1. ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) โดยยิ่งค่า MAE ต่ำ คือเข้าใกล้ศูนย์ยิ่งแสดงถึงความแม่นยำในการทำนาย
2. ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error: RMSE) โดยยิ่งค่า RMSE ต่ำ คือเข้าใกล้ศูนย์ยิ่งแสดงถึงความแม่นยำในการทำนาย
3. ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (Symmetric Mean Absolute Percentage Error: SMAPE) เป็นค่าที่ใช้วัดขนาดของความคลาดเคลื่อนของการทำนายกับค่าจริงโดยไม่คำนึงถึงทิศทางและถูกแปลงให้อยู่ในรูปแบบของร้อยละ ยิ่งค่า MAPE ต่ำ คือเข้าใกล้ศูนย์ยิ่งแสดงถึงความแม่นยำในการทำนาย

ในงานวิจัยนี้ การเปรียบเทียบประสิทธิภาพของแบบจำลองจะให้น้ำหนักกับค่า MAE มากที่สุด โดยพิจารณาแบบจำลองที่มีค่า MAE ต่ำที่สุดว่าเป็นแบบจำลองที่มีความสามารถในการทำนายมากที่สุด เนื่องจาก MAE เหมาะสมกับการพิจารณาสำหรับข้อมูลมีค่านอกเกณฑ์ (outliers)

3.5.2 การเปรียบเทียบประสิทธิภาพของแบบจำลองพหุผลลัพธ์ที่ดีที่สุดกับแบบจำลองผลลัพธ์เดียว

เพื่อประเมินว่าแบบจำลองประเภทการถดถอยพหุผลลัพธ์ที่ได้จากขั้นตอน 3.5.1 ว่ามีประสิทธิภาพเหนือกว่าแบบจำลองเดียวกันที่เป็นประเภทการถดถอยผลลัพธ์เดียว (Single-Output Regression) หรือไม่ จะทำการสร้างแบบจำลองการถดถอยผลลัพธ์เดียว โดยทำนายเป็นแยกผลลัพธ์ ได้แก่ 1. ทำนายความถี่ของการเรียกร้อยค่าสินไหมทดแทน และ 2. ทำนายความรุนแรงของการเรียกร้อยค่าสินไหมทดแทน โดยจะทำการปรับแต่งพารามิเตอร์ให้เหมาะสมในการสร้างแบบจำลอง ใช้ข้อมูลชุดสร้างและทดสอบแบบจำลองเช่นเดียวกันกับแบบจำลองพหุผลลัพธ์ ทำการประเมินประสิทธิภาพด้วยค่า MAE RMSE และ SMAPE

3.5.3 Shapley Additive Explanation (SHAP)

หลังจากได้แบบจำลองพหุผลลัพ์ที่ดีที่สุดในขั้นตอน 3.5.1 จะทำการวิเคราะห์ SHAP ซึ่งเป็นเครื่องมือที่ช่วยในการอธิบายผลของการทำนายจากแบบจำลอง เพื่อทำความเข้าใจคุณลักษณะที่มีผลกระทบต่อการตัดสินใจของแบบจำลองการทำนาย หรือเพื่อประเมินผลกระทบของแต่ละตัวแปรต่อผลลัพธ์การทำนาย ค่าของ Shapley value เป็นค่าที่บ่งบอกว่าตัวแปรแต่ละตัวส่งผลต่อการตัดสินใจของแบบจำลองในลักษณะใด ถ้าค่าเป็นบวกหมายความว่าตัวแปรนั้นช่วยผลักดันค่าการทำนายให้สูงขึ้น และถ้าค่าเป็นลบหมายความว่าตัวแปรนั้นทำให้ค่าการทำนายลดลง ตัวแปรที่มีค่า Shapley Value สูงสุดจะถูกพิจารณาว่าเป็นปัจจัยสำคัญที่สุดในการทำนายผลลัพ์ การวิเคราะห์นี้ช่วยเพิ่มความโปร่งใสและความน่าเชื่อถือของแบบจำลอง เนื่องจากผู้ใช้งานสามารถเข้าใจกลไกการตัดสินใจของแบบจำลองได้ชัดเจนยิ่งขึ้น

ผู้วิจัยจะนำเสนอแบบรูปภาพด้วย SHAP summary plot เป็นกราฟที่แสดงค่าผลกระทบของตัวแปรต่าง ๆ ในแบบจำลอง โดยจุดสีแดงจะแสดงผลในเชิงบวก (ค่าการทำนายเพิ่ม) และจุดสีฟ้าในเชิงลบ (ค่าการทำนายลด) การนำเสนอผลลัพ์ในรูปแบบนี้ช่วยให้นักวิจัยและผู้ใช้งานสามารถตรวจสอบความสมเหตุสมผลของผลการทำนายได้อย่างชัดเจนและเข้าใจง่าย อีกทั้งยังสนับสนุนการตัดสินใจที่แม่นยำมากขึ้นในการนำแบบจำลองไปประยุกต์ใช้ในสถานการณ์จริง

3.6 การนำแบบจำลองไปใช้งานจริง (Deployment)

การนำแบบจำลองไปใช้งานจริง ผู้วิจัยจะนำแบบจำลองที่มีประสิทธิภาพดีที่สุดจากการทดสอบมาพัฒนาเป็นเว็บแอปพลิเคชัน เพื่อให้บริษัทประกันภัยรถยนต์สามารถนำไปใช้งานทำนายผลลัพ์ได้สะดวกและรวดเร็ว โดยใช้โปรแกรม Visual Studio Code (VS Code) และใช้ Flask ซึ่งเป็น Web API เชื่อมต่อแบบจำลอง Machine Learning เข้ากับระบบเว็บ โดยหน้าต่างของเว็บแอปพลิเคชัน (User Interface) พัฒนาโดยใช้ภาษา HTML, CSS, JavaScript มีขั้นตอนการทำงานของเว็บแอปพลิเคชัน ดังนี้

1. ผู้ใช้ป้อนข้อมูลผ่านเว็บแอปพลิเคชัน ดังนี้

- 1) จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี
- 2) เพศของผู้ถือกรมธรรม์
- 3) อายุของผู้ถือกรมธรรม์
- 4) อายุของใบอนุญาตขับขี่
- 5) พื้นที่อยู่อาศัย
- 6) การอาศัยอยู่ภาคเหนือหรือไม่
- 7) อาศัยอยู่พื้นที่อื่น ๆ ของประเทศ
- 8) จำนวนกรมธรรม์ประเภทอื่นในบริษัทเดียวกัน

2. ประมวลผลข้อมูลและทำนายผลลัพธ์ เว็บแอปพลิเคชันใช้ Flask เป็นตัวกลางในการรับข้อมูลที่ใช้ป้อนเข้ามา จากนั้นทำการประมวลผลข้อมูล เช่น การแปลงให้อยู่ในรูปแบบที่สามารถนำไปใช้งานกับแบบจำลองได้ ข้อมูลที่ผ่านการประมวลผลแล้วจะถูกส่งเข้าสู่แบบจำลองเพื่อทำนายผลลัพธ์ และ Flask จะรับค่าทำนายกลับมาเพื่อนำไปแสดงผลแก่ผู้ใช้

3. แสดงผลลัพธ์ ผลลัพธ์จากการทำนายจะถูกส่งกลับไปยังผู้ใช้ในรูปแบบตัวเลข ดังนี้

- 1) ความรุนแรงในการเรียกร้องค่าสินไหมทดแทนทั้งหมดสำหรับประกันภัยรถยนต์
- 2) จำนวนการเรียกร้องค่าสินไหมทดแทนสำหรับประกันภัยรถยนต์

สำหรับการนำเว็บแอปพลิเคชันไปใช้งานจริง ผู้วิจัยมีแผนจะเผยแพร่ผ่านแพลตฟอร์ม Google Cloud / AWS / Azure เพื่อให้ผู้ใช้สามารถเข้าถึงเว็บแอปพลิเคชันผ่านเว็บเบราว์เซอร์

บทที่ 4

ผลการวิจัย

4.1 ผลการศึกษาธุรกิจประกันภัยรถยนต์

ธุรกิจประกันภัยรถยนต์ จัดเป็นประกันภัยเบ็ดเตล็ด (casualty insurance) ที่มีความสำคัญต่อภาคเศรษฐกิจของโลกตลาดประกันภัยมีมูลค่าประมาณ 5.8 ล้านล้านเหรียญสหรัฐ โดยการลงทุนในเทคโนโลยีด้านประกันภัย (InsurTech) มีมูลค่าประมาณ 7.2 พันล้านเหรียญสหรัฐในประเทศไทย จากรายงานของ คปภ./สมาคมประกันวินาศภัย พบว่ามีเบี้ยประกันภัยรับโดยตรงรวมประมาณ 285,028 ล้านบาท ซึ่งการประกันภัยรถยนต์มีสัดส่วนสูงที่สุด คิดเป็น 57% ของเบี้ยประกันวินาศภัยทั้งหมด หรือประมาณ 161,340 ล้านบาทโดยปกติจะมีระยะเวลาความคุ้มครองเป็นปีต่อปี โดยบริษัทจะเก็บเบี้ยประกันเพียงครั้งเดียวก่อนระยะเวลาเอาประกันภัย ความเสี่ยงทางการเงิน จัดว่าเป็นความเสี่ยงที่มีความสำคัญอย่างยิ่งต่อธุรกิจประกันภัยรถยนต์ โดยเฉพาะอย่างยิ่งที่มาจาก การเรียกร้องค่าสินไหมทดแทนของผู้เอาประกันภัย โดยในปี พ.ศ.2566 พบว่า มีมูลค่าสูงถึง 918,067 ล้านบาท (สมาคมประกันวินาศภัย ,2566)

ในปัจจุบัน ความถี่และ ความรุนแรงของการเรียกร้องค่าสินไหมทดแทนมีความผันผวนสูง ส่งผลโดยตรงต่อการกำหนดเบี้ยประกันภัย ความมั่นคงของธุรกิจ และการรักษาความยั่งยืนในระยะยาว ดังนั้นบริษัทประกันภัยจำเป็นต้องมีการคาดการณ์ที่แม่นยำ เพื่อให้สามารถตั้งราคาประกันภัยได้เหมาะสม ลดความเสี่ยงในการตั้งราคาผิดพลาด ซึ่งอาจนำไปสู่ การสูญเสียรายได้หรือความไม่พึงพอใจของลูกค้า ดังนั้นการวิจัยครั้งนี้จึงมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการถดถอยพหุผลลัพธ์ (Multi-output Regression) สำหรับทำนายทั้งความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนพร้อมกัน ทั้งนี้การเรียกร้องค่าสินไหมทดแทนจากประกันภัยรถยนต์ได้รับอิทธิพลจากปัจจัยเฉพาะของผู้เอาประกันภัย เช่น อายุของผู้ถือกรมธรรม์ ซึ่งเป็นข้อมูลที่ถูกนำไปศึกษาและวิเคราะห์ในหลายประเทศ (Chen et al., 2020) นอกจากนี้ การคัดเลือกแบบจำลองอาจส่งผลต่อความถูกต้องของการพยากรณ์ค่าสินไหมทดแทน เนื่องจากการเลือกแบบจำลองอาจทำให้ช่วงพยากรณ์ผิดพลาดและความแปรปรวนต่ำเกินจริง (Selection Effect) จึงควรใช้แบบ Full Model ซึ่งรวมตัวแปรทั้งหมดที่มีอยู่ในข้อมูลโดยไม่ทำการคัดเลือกตัวแปร จะทำให้ผลลัพธ์ที่น่าเชื่อถือกว่า (Hong, Kuffner, & Martin, 2018)

ABCDEFGHIJK

4.2 ผลการศึกษาข้อมูลการเรียกร้องค่าสินไหมทดแทน

4.2.1 รายละเอียดของข้อมูล

จากการศึกษาข้อมูลการเรียกร้องค่าสินไหมทดแทนของบริษัทประกันภัยแห่งหนึ่งในประเทศสเปน พบว่า มีทั้งหมด 10 ตัวแปร จำนวน 80,924 แถว มีข้อมูลเกี่ยวกับลักษณะของผู้เอาประกันภัย ประกอบด้วย 1) ความรุนแรงในการเรียกร้องค่าสินไหมทดแทนทั้งหมดสำหรับประกันภัยรถยนต์ 2) จำนวนการเรียกร้องค่าสินไหมทดแทนสำหรับประกันภัยรถยนต์ 3) จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี 4) เพศของผู้ถือกรมธรรม์ 5) อายุของผู้ถือกรมธรรม์ 6) อายุของใบอนุญาตขับขี่ 7) พื้นที่อยู่อาศัย 8) การอาศัยอยู่ภาคเหนือหรือไม่ 9) อาศัยอยู่พื้นที่อื่นๆของประเทศ 10) จำนวนกรมธรรม์ประเภทอื่นในบริษัทเดียวกัน สำหรับประกันภัยรถยนต์ มีรายละเอียดข้อมูลแสดงดังตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดข้อมูลการประกันภัยรถยนต์

| ตัวแปร | ชื่อตัวแปร | ความหมายของตัวแปร | ประเภทของข้อมูล (มาตรวัด) | ช่วงของข้อมูล/ ค่าที่เป็นไปได้ |
|--------|------------|--|------------------------------|-----------------------------------|
| Y_1 | nclaims_md | จำนวนการเรียกร้องค่าสินไหมทดแทน ประกันภัยรถยนต์ที่มีความเสียหายต่อ ทรัพย์สิน | ปริมาณ (อัตราส่วน) | 0 - 40 |
| Y_2 | cost_md | ความรุนแรงในการเรียกร้องค่าสินไหม ทดแทนทั้งหมดสำหรับประกันภัย รถยนต์ที่มีความเสียหายต่อทรัพย์สิน (มูลค่า) | ปริมาณ (อัตราส่วน) | 0 - 65875.22 |
| X_1 | npol_auto | จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ ถือกรมธรรม์มี | ปริมาณ (อัตราส่วน) | 1 - 35 |
| X_2 | client_sex | เพศของผู้ถือกรมธรรม์ | คุณภาพ (นามบัญญัติ) | Man Woman |
| X_3 | client_age | อายุของผู้ถือกรมธรรม์ | ปริมาณ (อันตรภาค) | 20 - 90 |
| X_4 | lic_age | อายุของใบอนุญาตขับขี่ | ปริมาณ (อันตรภาค) | 1 - 70 |
| X_5 | city | พื้นที่อยู่อาศัย | คุณภาพ (นามบัญญัติ) | 0 = อื่นๆ 1 = อยู่ในเมืองใหญ่ |
| X_6 | north | อาศัยอยู่ภาคเหนือหรือไม่ | คุณภาพ (นามบัญญัติ) | 0 = ไม่ 1 = ใช่ |
| X_7 | rest | อาศัยอยู่พื้นที่อื่นๆของประเทศ | คุณภาพ | 0 = ไม่ใช่ |

| | | | | |
|-------|---------------|--|-----------------------|---------|
| | | | (นามบัญญัติ) | 1 = ใช่ |
| X_8 | client_nother | จำนวนกรรมธรรม์ประเภทอื่นในบริษัท เดียวกัน | ปริมาณ (อัตราส่วน) | 0 - 23 |

ในกรณีตัวแปร lic_age (X_4) คือ อายุของใบอนุญาตขับขี่ เป็นระยะเวลาที่ผู้ขับขี่ได้รับใบอนุญาตขับขี่ โดยคำนวณจากปีที่ได้รับใบอนุญาตครั้งแรกจนถึงปีปัจจุบัน เช่น ถ้าลูกค้าได้รับใบอนุญาตขับขี่ตอนอายุ 20 ปี และปัจจุบันลูกค้าอายุ 30 ปี "lic_age" จะเท่ากับ 10 ปี

rest (X_7) คืออาศัยอยู่พื้นที่อื่นๆของประเทศ มีค่าที่เป็นไปได้สองค่า ได้แก่ 1 = ใช่ ($X_5 = 0$ และ $X_6 = 0$) และมีค่าเป็น 0 = ไม่ใช่ (อยู่ในพื้นที่ใดพื้นที่หนึ่ง ยกเว้น $X_5 = 1$ และ $X_6 = 1$)

client_nother (X_8) คือ จำนวนกรรมธรรม์ประเภทอื่นในบริษัท ที่ไม่ใช่ประกันภัยรถยนต์และประกันภัยบ้าน (เช่น ประกันภัยอุบัติเหตุ ประกันชีวิตแบบตลอดชีพ ประกันชีวิตแบบบำนาญ)

4.2.2 ผลการวิเคราะห์ข้อมูลเชิงสำรวจ

ผลการวิเคราะห์สถิติเบื้องต้นของข้อมูลผู้ถือกรรมธรรม์ แสดงดังภาพที่ 4.2

1) ผลการวิเคราะห์ข้อมูลเชิงสำรวจ

| | nclaims_md | cost_md | npol_auto | client_sex | client_age | lic_age | client_nother | city | north | rest |
|--------|--------------|--------------|--------------|------------|--------------|--------------|---------------|---------|---------|---------|
| count | 80924.000000 | 80924.000000 | 80924.000000 | 80924 | 80924.000000 | 80924.000000 | 80924.000000 | 80924.0 | 80924.0 | 80924.0 |
| unique | NaN | NaN | NaN | 2 | NaN | NaN | NaN | 2.0 | 2.0 | 2.0 |
| top | NaN | NaN | NaN | Man | NaN | NaN | NaN | 0.0 | 0.0 | 1.0 |
| freq | NaN | NaN | NaN | 61731 | NaN | NaN | NaN | 64994.0 | 57566.0 | 41636.0 |
| mean | 0.255511 | 234.202818 | 2.44679 | NaN | 53.24177 | 29.937682 | 0.219020 | NaN | NaN | NaN |
| std | 0.822942 | 916.829434 | 2.10436 | NaN | 13.11958 | 11.359151 | 0.655378 | NaN | NaN | NaN |
| min | 0.000000 | 0.000000 | 1.00000 | NaN | 18.00000 | 1.000000 | 0.000000 | NaN | NaN | NaN |
| 25% | 0.000000 | 0.000000 | 1.00000 | NaN | 43.00000 | 21.000000 | 0.000000 | NaN | NaN | NaN |
| 50% | 0.000000 | 0.000000 | 2.00000 | NaN | 53.00000 | 30.000000 | 0.000000 | NaN | NaN | NaN |
| 75% | 0.000000 | 0.000000 | 3.00000 | NaN | 63.00000 | 38.000000 | 0.000000 | NaN | NaN | NaN |
| max | 40.000000 | 65875.220000 | 35.00000 | NaN | 90.00000 | 70.000000 | 23.000000 | NaN | NaN | NaN |

ภาพที่ 4.2 ผลการวิเคราะห์สถิติเบื้องต้นของข้อมูล

จากภาพที่ 4.2 ข้อมูลผู้ถือกรรมธรรม์ประกอบด้วยตัวแปรเชิงปริมาณและเชิงคุณภาพ สำหรับตัวแปรเชิงปริมาณจะปรากฏค่า count mean std min 25% 50% 75% max และสำหรับตัวแปรเชิงคุณภาพจะปรากฏค่า count unique top freq ในกรณีที่ปรากฏค่า NaN หมายความว่าไม่สามารถหาสำหรับตัวแปรประเภทนั้นได้ โดยพบว่า ทุกตัวแปรที่มีจำนวนข้อมูลทั้งหมด 80,924 แถว

จำนวนการเรียกร้องค่าสินไหมทดแทน (nclaims_md) มีค่าเฉลี่ยเท่ากับ 0.2555 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.8229 และค่ามัธยฐานเท่ากับ 0 ซึ่งพบว่าข้อมูลส่วนใหญ่ไม่มีการเรียกร้องค่าสินไหมทดแทน และค่าสูงสุดของข้อมูลเท่ากับ 40 แสดงว่ามีค่าห่างจากข้อมูลส่วนใหญ่หรือค่านอกเกณฑ์อยู่

ความรุนแรงในการเรียกร้องค่าสินไหมทดแทน (cost_md) มีค่าเฉลี่ยเท่ากับ 234.2028 ยูโร ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 916.8294 ค่าต่ำสุดของข้อมูลเท่ากับ 0 ค่าที่อยู่ตำแหน่ง 25% ของข้อมูลเท่ากับ 0 ค่ามัธยฐานเท่ากับ 0 ค่าที่อยู่ตำแหน่ง 75% ของข้อมูลเท่ากับ 0 และค่าสูงสุดของข้อมูลเท่ากับ 65875.2200 แสดงว่ามีค่าห่างจากข้อมูลส่วนใหญ่หรือค่านอกเกณฑ์อยู่

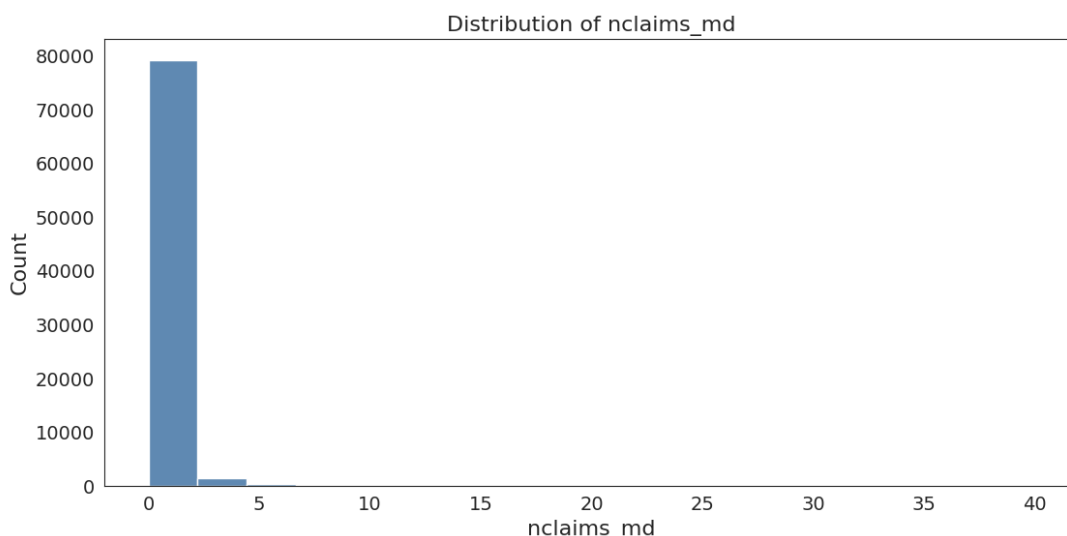
จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี (npol_auto) มีค่าเฉลี่ยเท่ากับ 2.4467 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 2.1043 ค่าต่ำสุดของข้อมูลเท่ากับ 1 ค่าที่อยู่ตำแหน่ง 25% ของข้อมูลเท่ากับ 1 ค่ามัธยฐานข้อมูลเท่ากับ 2 ค่าที่อยู่ตำแหน่ง 75% ของข้อมูลเท่ากับ 3 และค่าสูงสุดของข้อมูลเท่ากับ 35 ซึ่งมีค่าห่างจากค่ากลางค่อนข้างมาก

อายุของผู้ถือกรมธรรม์ (client_age) มีค่าเฉลี่ยเท่ากับ 53.2417 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 13.1195 นั่นคือผู้ถือกรมธรรม์ส่วนใหญ่อยู่ในช่วงวัยกลางคนถึงผู้สูงอายุ ค่าต่ำสุดของข้อมูลเท่ากับ 18 ค่าที่อยู่ตำแหน่ง 25% ของข้อมูลเท่ากับ 43 ค่ามัธยฐานเท่ากับ 53 ค่าที่อยู่ตำแหน่ง 75% ของข้อมูลเท่ากับ 63 และค่าสูงสุดของข้อมูลเท่ากับ 90

อายุของใบอนุญาตขับขี่ (lic_age) มีค่าเฉลี่ยเท่ากับ 29.9376 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 11.3591 ค่าต่ำสุดของข้อมูลเท่ากับ 1 ค่าที่อยู่ตำแหน่ง 25% ของข้อมูลเท่ากับ 21 ค่ามัธยฐานเท่ากับ 30 ค่าที่อยู่ตำแหน่ง 75% ของข้อมูลเท่ากับ 38 และค่าสูงสุดของข้อมูลเท่ากับ 70

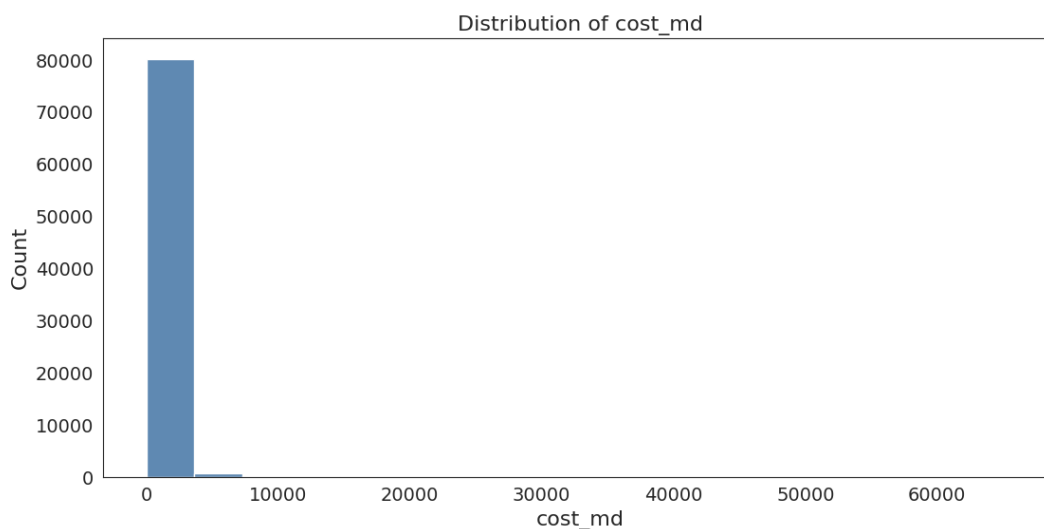
จำนวนกรมธรรม์ประเภทอื่นในบริษัท (client_nother) มีค่าเฉลี่ยเท่ากับ 0.2190 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.6553 ค่าต่ำสุดของข้อมูลเท่ากับ 0 ค่าที่อยู่ตำแหน่ง 25% ของข้อมูลเท่ากับ 0 ค่ามัธยฐานเท่ากับ 0 ค่าที่อยู่ตำแหน่ง 75% ของข้อมูลเท่ากับ 0 พบว่าส่วนใหญ่ไม่มีกรมธรรม์ประเภทอื่นในบริษัทเดียวกันและค่าสูงสุดของข้อมูลเท่ากับ 23

สำหรับตัวแปรที่เป็นข้อมูลเชิงคุณภาพ มีจำนวนค่าที่ไม่ซ้ำกันในตัวแปร (unique) เท่ากับ 2 และค่าอื่น ๆ ดังนี้ 1. เพศของผู้ถือกรมธรรม์ (client_sex) ส่วนใหญ่เป็นเพศชาย 2. พื้นที่อยู่อาศัย (city) ผู้ถือกรมธรรม์ส่วนใหญ่ไม่ได้อยู่ในเมืองใหญ่ โดยมีจำนวน 64,994 ราย 3. อาศัยอยู่ภาคเหนือหรือไม่ (north) 4. ผู้ถือกรมธรรม์ส่วนใหญ่ไม่ได้อาศัยพื้นที่ภาคเหนือ อาศัยอยู่พื้นที่อื่น ๆ ของประเทศ (rest) พบว่าส่วนใหญ่อาศัยอยู่ในพื้นที่อื่นของประเทศ คือไม่ได้อยู่ทั้งภาคเหนือและในเมืองใหญ่ จำนวน 41,636 ราย



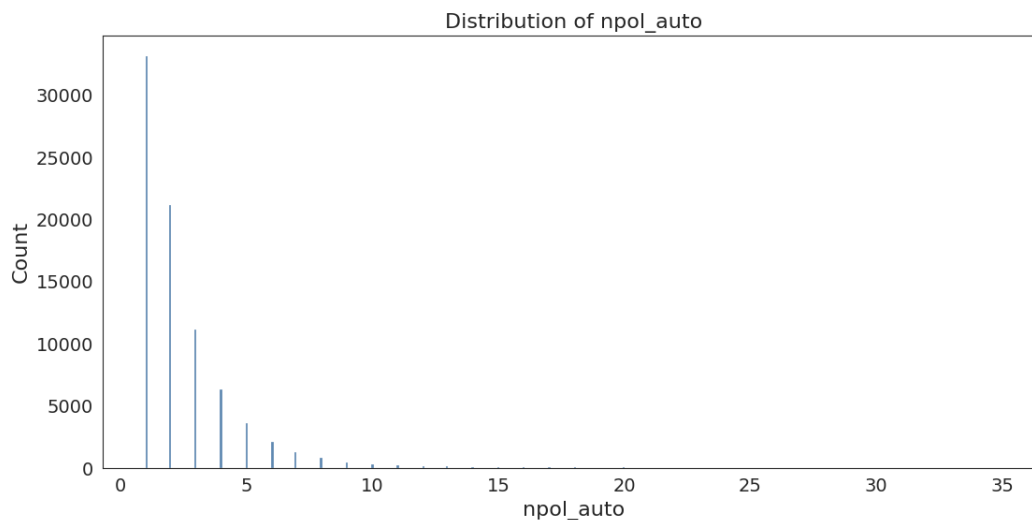
ภาพที่ 4.3 การกระจายตัวของข้อมูลความถี่ของการเรียกร้องค่าสินไหมทดแทน

จากภาพที่ 4.3 พบว่าจำนวนการเรียกร้องค่าสินไหมทดแทนประกันภัยรถยนต์ที่มีความเสียหายต่อทรัพย์สินมีการกระจายแบบเบ้ขวา ส่วนใหญ่มีจำนวนการเรียกร้องค่าสินไหมทดแทนต่ำมาก อยู่ในช่วง 0 ถึง 2 ครั้ง ส่งผลให้เกิดความไม่สมดุลในข้อมูล ซึ่งต้องพิจารณาปรับสมดุลข้อมูลเพื่อให้การสร้างแบบจำลองมีความแม่นยำและเสถียรมากขึ้น



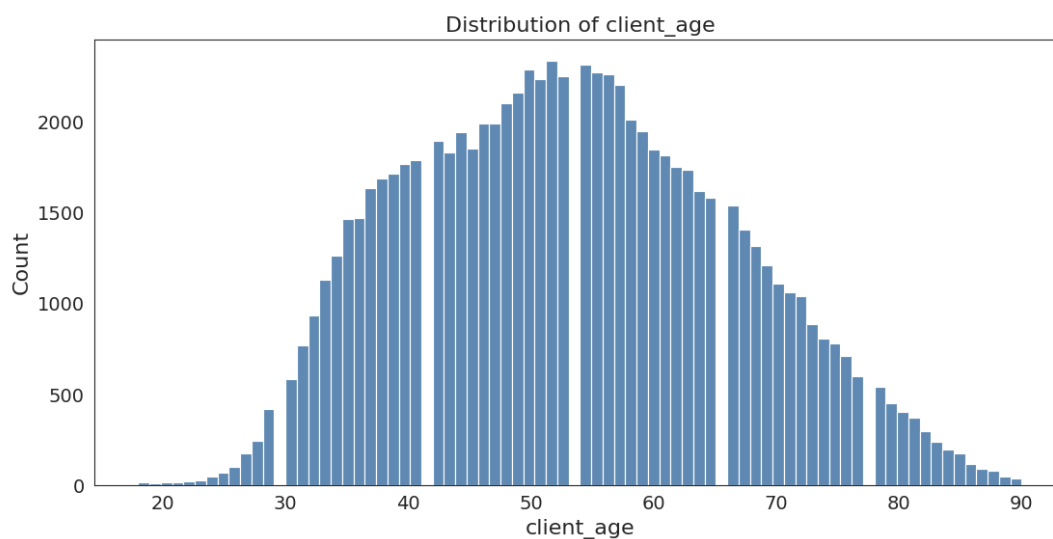
ภาพที่ 4.4 การกระจายตัวของข้อมูลความรุนแรงในการเรียกร้องค่าสินไหมทดแทน

จากภาพที่ 4.4 พบว่าความรุนแรงในการเรียกร้องค่าสินไหมทดแทนมีการกระจายแบบเบ้ขวา ส่วนใหญ่มีค่าต่ำกว่า 10,000 การที่เป็น 0 มากหมายความว่าไม่เคยมีการเรียกร้องค่าสินไหมทดแทนทำให้ไม่มีมูลค่าการเรียกร้องค่าสินไหมทดแทนทั้งหมดเช่นกัน



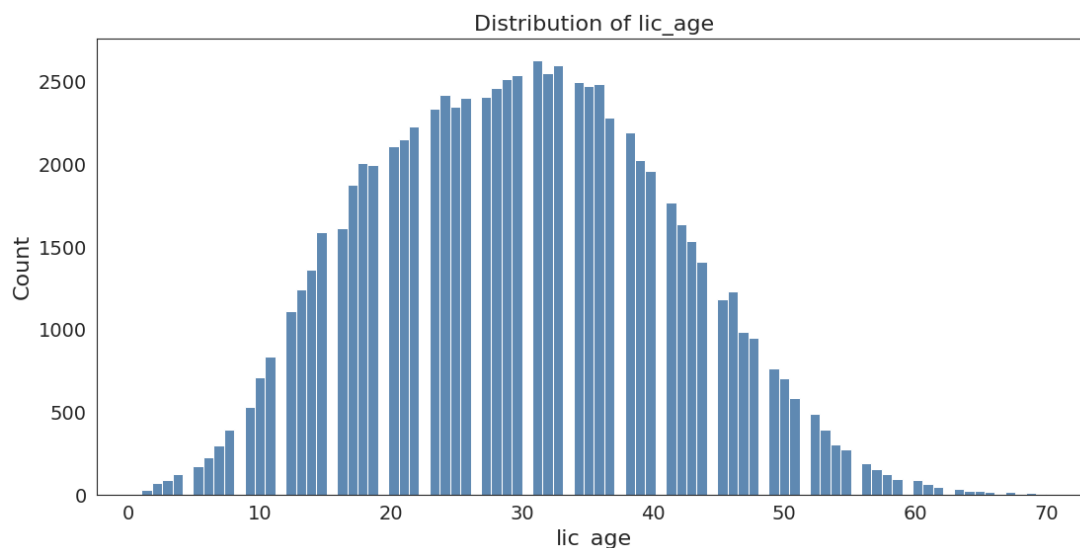
ภาพที่ 4.5 การกระจายตัวของข้อมูลจำนวนกรรมธรรม์ประกันภัยรถยนต์

จากภาพที่ 4.5 พบว่าจำนวนกรรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรรมธรรม์มีการกระจายแบบเบ้ขวา จำนวนกรรมธรรม์ที่รถยนต์ที่มีต่อบุคคลส่วนใหญ่อยู่ที่ 1 ถึง 3 เล่มกรรมธรรม์



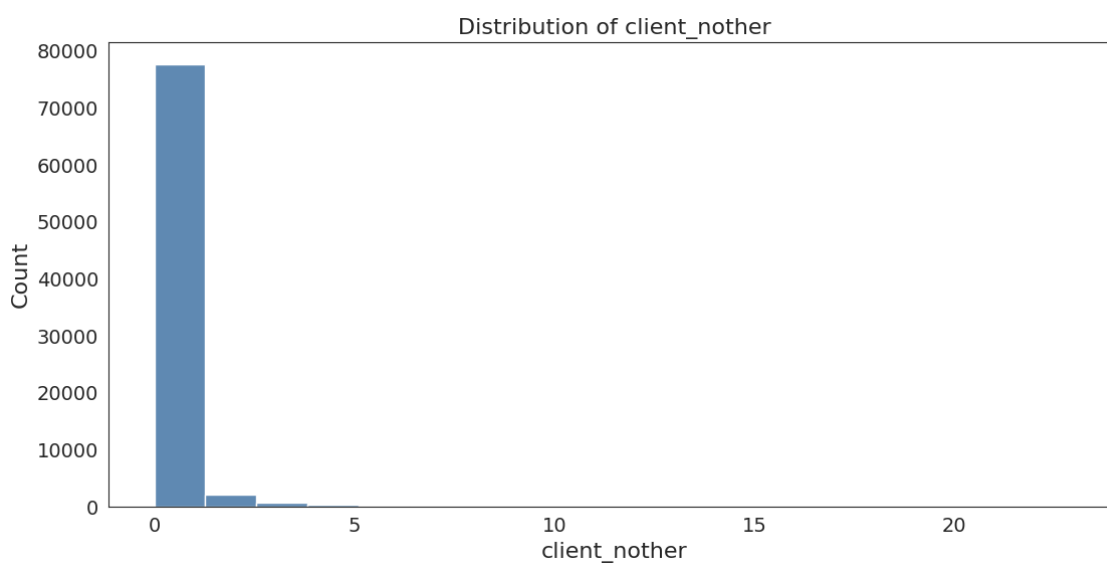
ภาพที่ 4.6 การกระจายตัวของข้อมูลอายุผู้ถือกรรมธรรม์

จากภาพที่ 4.6 พบว่าอายุของผู้ถือกรรมธรรม์ มีการกระจายตัวค่อนข้างสมมาตร อายุของผู้ใช้บริการส่วนใหญ่จะอยู่ในช่วงวัยกลางคนถึงวัยผู้ใหญ่ตอนปลาย คือช่วง 30 ถึง 70 ปี



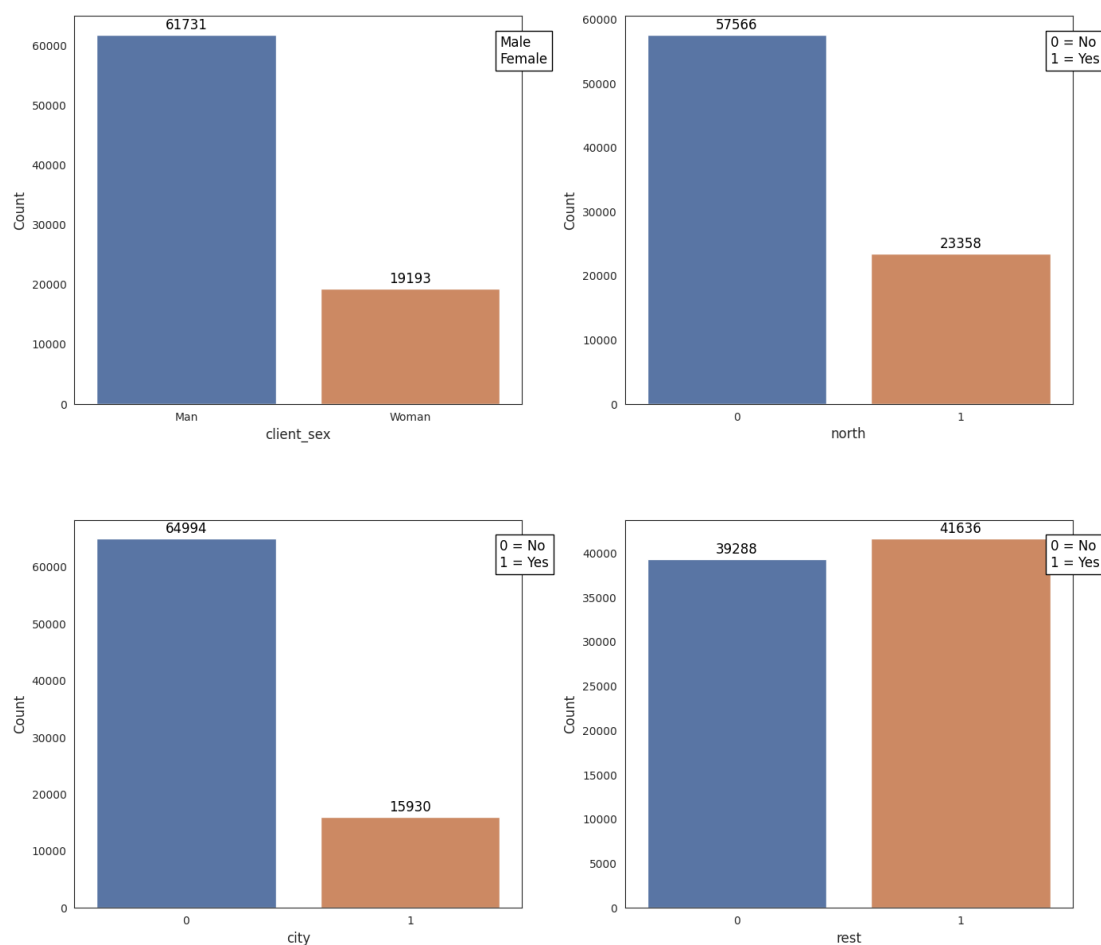
ภาพที่ 4.7 การกระจายตัวของข้อมูลอายุใบขับขี่

จากภาพที่ 4.7 พบว่าอายุของใบอนุญาตขับขี่ มีการกระจายตัวค่อนข้างสมมาตร ส่วนใหญ่มีใบอนุญาตขับขี่มานานประมาณ 20 ถึง 40 ปี



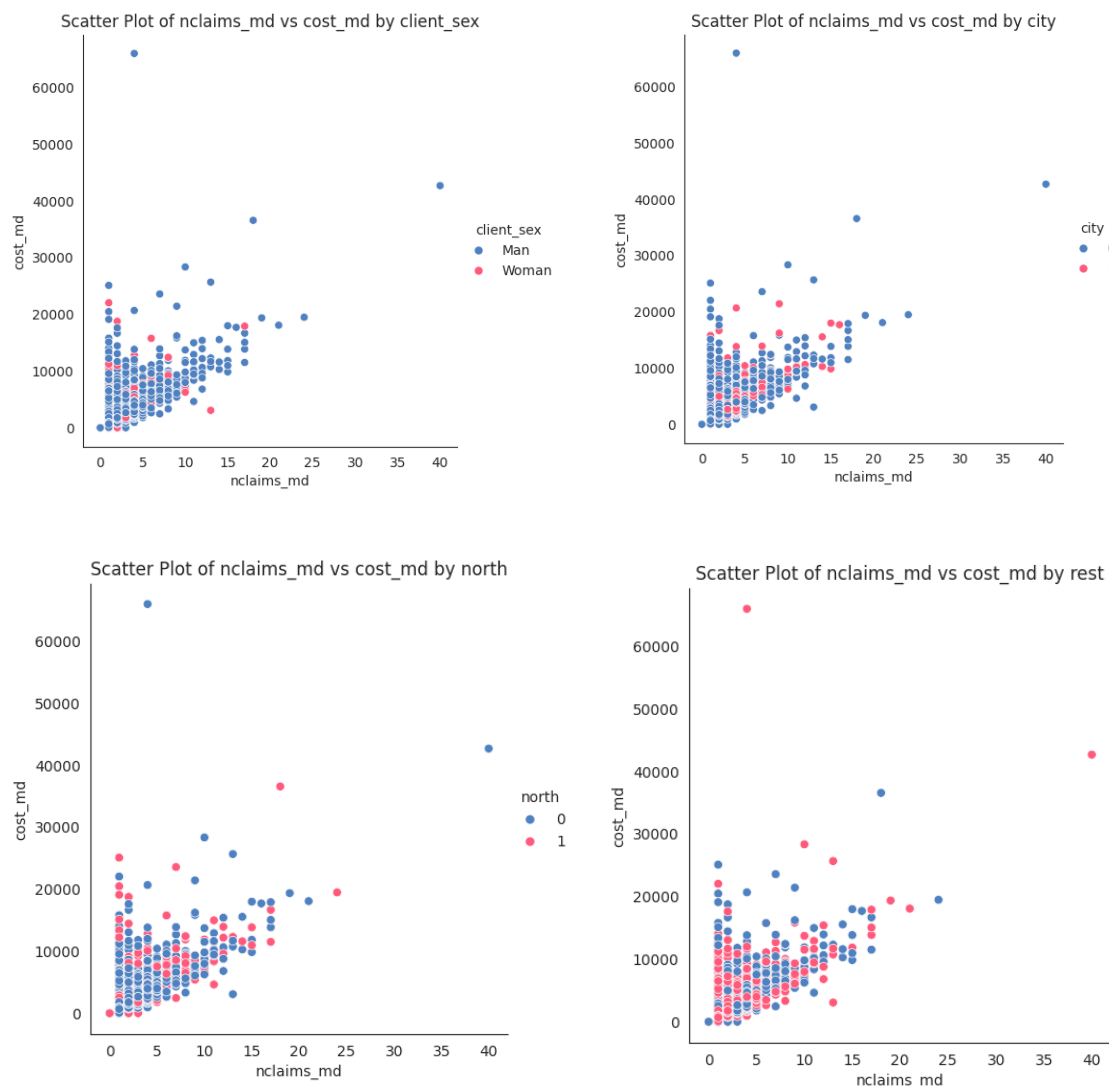
ภาพที่ 4.8 การกระจายตัวของข้อมูลจำนวนกรรมธรรม์ประกันภัยอื่น

จากภาพที่ 4.8 พบว่าจำนวนกรรมธรรม์ประเภทอื่นในบริษัท มีการกระจายแบบเบ้ขวา มีการถือกรรมธรรม์อื่นที่ไม่ใช่รถยนต์หรือบ้านค่อนข้างน้อยหรือไม่มีเลยคือ 0 ถึง 1 เล่มกรรมธรรม์



ภาพที่ 4.9 การตรวจสอบการกระจายของข้อมูลเชิงคุณภาพ

จากภาพที่ 4.9 สามารถอธิบายได้ดังนี้ เพศของผู้ถือกรรมธรรม์ (client_sex) พบว่าส่วนใหญ่ผู้ถือกรรมธรรม์เป็นเพศชาย พื้นที่อยู่อาศัย (city) พบว่าผู้ถือกรรมธรรม์ส่วนใหญ่ไม่ได้อาศัยอยู่ในเมืองใหญ่ (บาเซโลน่า หรือ มาดริด) อาศัยอยู่ภาคเหนือหรือไม่ (north) พบว่าผู้ถือกรรมธรรม์ส่วนใหญ่อยู่พื้นที่ที่ไม่ใช่ภาคเหนือ และการอาศัยอยู่พื้นที่อื่นๆของประเทศ (rest) พบว่าจำนวนคนที่อาศัยในเขตพื้นที่ทั้งในเมืองและภาคเหนือ มีความใกล้เคียงกับคนที่อาศัยในพื้นที่อื่น ๆ



ภาพที่ 4.10 Scatter plot แสดงความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณกับเชิงคุณภาพ

จากภาพที่ 4.10 สามารถอธิบายได้ว่าส่วนใหญ่ข้อมูลมีการกระจายตัวอยู่ของจำนวนการเรียกร้องค่าสินไหมทดแทนมีค่าน้อยกว่า 10 และความรุนแรงในการเรียกร้องค่าสินไหมทดแทนอยู่ในช่วงไม่เกิน 20,000 ยูโร นอกจากนี้ยังชี้ให้เห็นถึงความสัมพันธ์ระหว่างจำนวนและความรุนแรงในการเรียกร้องค่าสินไหมทดแทนที่เป็นไปในเชิงบวก กล่าวคือเมื่อจำนวนการเรียกร้องค่าสินไหมทดแทนเพิ่มขึ้น ความรุนแรงในการเรียกร้องค่าสินไหมทดแทนก็มีแนวโน้มที่จะเพิ่มขึ้นตามไปด้วยและสามารถอธิบายแต่ละกราฟได้ดังนี้

Scatter Plot of nclaims_md vs cost_md by Client Sex ผู้ถือกรรมธรรม์ผู้ชายมีการกระจายตัวของจำนวนการเรียกร้องค่าสินไหมทดแทนในระดับสูงกว่า และบางกรณีก็มีความรุนแรงในการเรียกร้องค่าสินไหมทดแทนสูงกว่าผู้หญิง

Scatter Plot of nclaims_md vs cost_md by city ผู้ถือกรรมธรรม์ที่อาศัยอยู่ในเมืองใหญ่ และผู้ที่ไม่ได้อาศัยอยู่ในเมืองใหญ่มีรูปแบบการกระจายตัวของการจำนวนและความรุนแรงในการเรียกร้องค่าสินไหมทดแทนคล้ายคลึงกัน แต่ในบางกรณีกลุ่มที่อาศัยในเมืองใหญ่มีแนวโน้มมากกว่า

Scatter Plot of nclaims_md vs cost_md by north ผู้ถือกรมธรรม์ที่อาศัยอยู่ในภาคเหนือและผู้ที่ไม่ได้อาศัยอยู่ในภาคเหนือมีรูปแบบการกระจายตัวของการจำนวนและความรุนแรงในการเรียกร้องค่าสินไหมทดแทนคล้ายคลึงกัน

Scatter Plot of nclaims_md vs cost_md by rest ผู้ถือกรมธรรม์อาศัยพื้นที่อื่นๆของประเทศการกระจายตัวของความรุนแรงในการเรียกร้องค่าสินไหมทดแทนที่สูงกว่า หากเทียบกับจำนวนการเรียกร้องค่าสินไหมที่เท่ากัน

4.3 ผลการเตรียมข้อมูลก่อนการวิเคราะห์

4.3.1 ผลการทำความสะอาดข้อมูล

1) ผลการตรวจสอบความถูกต้องและความแนบเนียนของข้อมูล

เมื่อพิจารณาตัวแปร $client_age$ (X_3) คืออายุของผู้ถือกรมธรรม์ และ lic_age (X_4) คืออายุของใบอนุญาตขับขี่ ดังนั้น " lic_age " จะต้องมีย่านน้อยกว่า " $client_age$ " เสมอ เพราะลูกค้าจะได้รับใบอนุญาตขับขี่หลังจากที่เกิดมาแล้ว และสำหรับอายุขั้นต่ำที่สามารถทำใบขับขี่ได้ในประเทศสเปน (ช่วงปี ค.ศ. 2006 ถึง ค.ศ. 2015) คืออายุ 18

ผู้วิจัยจึงคำนวณค่าความแตกต่างระหว่าง $client_age$ และ lic_age หากผลลัพธ์น้อยกว่า 18 ปี จะถือว่าข้อมูลนั้นผิดพลาดและจะถูกลบออกจากข้อมูล ดังตัวอย่างในตารางที่ 4.2 หลังตรวจสอบแล้วพบว่าข้อมูลที่น้อยกว่า 18 อยู่ 2,734 แถว จาก 80,924 แถว เมื่อลบออกแล้วเหลือข้อมูลอยู่ 78,190 แถว

ตารางที่ 4.2 ตัวอย่างข้อมูลที่มีการบันทึกผิดพลาด

| $client_age$ | lic_age | $client_age - lic_age$ |
|---------------|------------|--------------------------|
| 25 | 10 | 15 |
| 30 | 22 | 8 |

พิจารณาตัวแปร $nclaims_md$ (Y_1) คือ จำนวนการเรียกร้องค่าสินไหมทดแทนประกันภัยรถยนต์ที่มีความเสียหายต่อทรัพย์สิน และ $cost_md$ (Y_2) คือ ความรุนแรงในการเรียกร้องค่าสินไหมทดแทนทั้งหมดสำหรับประกันภัยรถยนต์ที่มีความเสียหายต่อทรัพย์สิน (มูลค่า) เนื่องจากมีข้อมูลที่ $nclaims_md$ มีค่า แต่ $cost_md$ เป็น 0 หมายความว่า จำนวนการเรียกร้องค่าสินไหมขึ้นแต่ไม่มีการจ่ายค่าสินไหมทดแทน ซึ่งอาจมีสาเหตุมาจากการเรียกร้องถูกปฏิเสธเนื่องจากไม่อยู่ในเงื่อนไขความคุ้มครองของกรมธรรม์ หรือ ค่าเสียหายต่ำกว่า Deductible (ค่าเสียหายส่วนแรกที่ผู้เอาประกันภัยต้องรับผิดชอบเอง) ทำให้บริษัทประกันภัยไม่ต้องจ่ายเงิน จึงพิจารณาลบข้อมูลนี้ออกก่อนการวิเคราะห์ ซึ่งมีจำนวน 467 แถว จาก 78,190 แถว เมื่อลบแล้วเหลือข้อมูลอยู่ 77,723 แถว

2) ผลการจัดการกับค่าสูญหาย

จากการตรวจสอบข้อมูลในงานวิจัยนี้พบว่า ไม่มีค่าค่าสูญหาย แสดงผลดังภาพที่ 4.11

| Number of missing values | |
|--------------------------|---|
| nclaims_md | 0 |
| cost_md | 0 |
| npol_auto | 0 |
| client_sex | 0 |
| client_age | 0 |
| lic_age | 0 |
| client_nother | 0 |
| city | 0 |
| north | 0 |
| rest | 0 |

ภาพที่ 4.11 ผลการตรวจสอบค่าสูญหาย (Missing Value)

3) ผลการตรวจสอบค่านอกเกณฑ์

จากการวิเคราะห์พบว่ามีตัวแปรที่มีค่านอกเกณฑ์มากกว่า 5% ของข้อมูลทั้งหมดดังนี้

cost_md (Y_1) มีค่านอกเกณฑ์จำนวน 11,132 แถว (คิดเป็น 14.58%)

nclaims_md (Y_2) มีค่านอกเกณฑ์จำนวน 11,132 แถว (คิดเป็น 14.58%)

client_nother (X_8) มีค่านอกเกณฑ์จำนวน 11,598 แถว (คิดเป็น 14.92%)

มีรายละเอียดข้อมูลแสดงดังภาพที่ 4.12

Column: nclaims_md
Lower Bound: 0.0
Upper Bound: 0.0
Number of outliers: 11332
Percentage of outliers: 14.58%

Column: cost_md
Lower Bound: 0.0
Upper Bound: 0.0
Number of outliers: 11332
Percentage of outliers: 14.58%

Column: npol_auto
Lower Bound: -2.0
Upper Bound: 6.0
Number of outliers: 3385
Percentage of outliers: 4.36%

Column: client_age
Lower Bound: 15.5
Upper Bound: 91.5
Number of outliers: 0
Percentage of outliers: 0.00%

Column: lic_age
Lower Bound: -4.5
Upper Bound: 63.5
Number of outliers: 86
Percentage of outliers: 0.11%

Column: client_nother
Lower Bound: 0.0
Upper Bound: 0.0
Number of outliers: 11598
Percentage of outliers: 14.92%

ภาพที่ 4.12 ผลการตรวจสอบค่านอกเกณฑ์

4.3.2 ผลการแปลงข้อมูล

1) จากตัวแปรทั้งหมด มีตัวแปร เพศของผู้ถือกรมธรรม์ เป็นข้อมูลเชิงคุณภาพ มีค่าที่เป็นไปได้คือ Man, Woman เพื่อให้สามารถนำไปวิเคราะห์ในโปรแกรมได้ จึงทำการแปลงข้อมูลให้อยู่ในรูปของตัวเลข โดยกำหนดให้ 0 = Man, 1 = Woman

2) หลังจากการแปลงข้อมูลของตัวแปรตามแล้ว มีผลแสดงตัวอย่างข้อมูล ดังตารางที่ 4.3

ตารางที่ 4.3 ตัวอย่างข้อมูลการประกันภัยรถยนต์ที่ทำการแปลงเรียบร้อยแล้ว

| | คนที่ 1 | คนที่ 2 | คนที่ 3 | คนที่ 4 | คนที่ 5 |
|----------------------------------|---------|---------|---------|---------|---------|
| nclaims_md ก่อนการแปลงข้อมูล | 0 | 2 | 1 | 0 | 1 |
| nclaims_md_Log หลังการแปลงข้อมูล | -4.6051 | 0.6981 | 0.0099 | -4.6051 | 0.0099 |
| cost_md ก่อนการแปลงข้อมูล | 0 | 1764 | 882 | 0 | 4967.47 |
| cost_md_Log หลังการแปลงข้อมูล | -4.6051 | 7.4753 | 6.7822 | -4.6051 | 8.5107 |
| npol_auto | 2 | 4 | 6 | 3 | 8 |
| client_sex | 0 | 0 | 0 | 1 | 1 |
| client_age | 42 | 81 | 59 | 68 | 48 |
| lic_age | 8 | 34 | 32 | 45 | 24 |
| city | 0 | 0 | 0 | 1 | 1 |
| north | 0 | 0 | 1 | 0 | 0 |
| rest | 1 | 1 | 0 | 0 | 0 |
| client_nother | 1 | 0 | 2 | 0 | 0 |

4.4 ผลการสร้างแบบจำลอง

4.4.1 ผลการแบ่งข้อมูล

จากการตรวจสอบพบว่า ตัวแปรความถี่ในการเรียกร้องค่าสินไหมทดแทนเป็นข้อมูลไม่สมดุล ผู้วิจัยจึงแบ่งข้อมูลด้วยวิธี stratify ใช้ข้อมูล “Has_claim” ที่มาจากการแปลงข้อมูล nclaims_md ซึ่งตัวแปรนี้จะไม่ถูกนำไปสร้างแบบจำลอง แต่ใช้ในการสร้างความสมดุลในการแบ่งข้อมูล ค่าของข้อมูลที่เป็นไปได้คือ

0 หมายถึง ไม่มีการเรียกร้องค่าสินไหมทดแทน

1 หมายถึง มีการเรียกร้องค่าสินไหมทดแทน ตั้งแต่ 1 ครั้งขึ้นไป

มีตัวอย่างการแปลงข้อมูลดังตารางที่ 4.4

ในงานวิจัยนี้จะใช้ library ที่ชื่อว่า `model_selection` คำสั่ง `train_test_split(X, y, test_size=0.2, random_state=42, stratify=data_claims['Has_claim'])`

ตารางที่ 4.4 ตัวอย่างการแปลงข้อมูล Has_claim

| nclaims_md | Has_claim |
|------------|-----------|
| 0 | 0 |
| 1 | 1 |
| 4 | 1 |
| 0 | 0 |

หลังจากทำการแบ่งข้อมูลแล้ว สัดส่วนของกลุ่มข้อมูลทั้งสองชุดยังคงใกล้เคียงกัน โดยร้อยละของข้อมูลจำแนกตามการเรียกร้องค่าสินไหมทดแทนของข้อมูลชุดฝึกแสดงดังตารางที่ 4.5 และข้อมูลชุดทดสอบแสดงดังตารางที่ 4.6

ตารางที่ 4.5 ร้อยละของข้อมูลชุดฝึกจำแนกตามการเรียกร้องค่าสินไหมทดแทน

| มีการเรียกร้องค่าสินไหมทดแทนใช่หรือไม่ | จำนวน | ร้อยละ |
|--|--------|--------|
| ไม่ใช่ | 53,112 | 85.42 |
| ใช่ | 9,066 | 14.58 |
| รวม | 62,178 | 100.00 |

ตารางที่ 4.6 ร้อยละข้อมูลชุดทดสอบจำแนกตามการเรียกร้องค่าสินไหมทดแทน

| มีการเรียกร้องค่าสินไหมทดแทนใช่หรือไม่ | จำนวน | ร้อยละ |
|--|--------|--------|
| ไม่ใช่ | 13,279 | 85.42 |
| ใช่ | 2,266 | 14.58 |
| รวม | 15,545 | 100.00 |

4.4.2 ผลการปรับแต่งพารามิเตอร์ของแต่ละแบบจำลอง

ในขั้นตอนการปรับแต่งค่าพารามิเตอร์ด้วยวิธี Manual Search สำหรับการทดสอบ `n_estimators` ทดสอบที่ค่า 100, 200, 300, 400, 500, 1000 พบว่ามีประสิทธิภาพดีขึ้นเมื่อจำนวนต้นไม้สูงขึ้น และยังพบว่าค่าประสิทธิภาพของแบบจำลองลดลงเมื่อมีจำนวนต้นไม้มากกว่า 500 ต้น

ในขั้นตอนการปรับแต่งค่าพารามิเตอร์ของแบบจำลองดำเนินการโดยใช้วิธี GridSearchCV ซึ่งเป็นกระบวนการที่ทดลองค่าพารามิเตอร์ทุกค่าที่เป็นไปได้ภายในช่วงที่กำหนด พร้อมกับการประเมินผลแบบ 10-

fold Cross Validation (CV) เพื่อค้นหาชุดค่าพารามิเตอร์ที่ให้ผลลัพธ์ดีที่สุดสำหรับแต่ละแบบจำลอง ใน ผลการวิเคราะห์จากการปรับแต่งพารามิเตอร์ แสดงไว้ในตารางที่ 4.7

ตารางที่ 4.7 พารามิเตอร์ที่ดีที่สุดจากการค้นหาด้วย GridsearchCV ของแต่ละแบบจำลอง

| Model Multi-Output | Tuning Parameters | Best Parameters |
|---|--|---|
| Random Forest Regression | n_estimators : [1100, 1200, 1300] max_depth : [6, 7, 8, 9] min_samples_split : [100, 110, 120, 130, 140] min_samples_leaf : [40, 50, 60] | 1300 9 130 60 |
| Extreme Gradient Boosting Regression | n_estimators : [100, 200, 300] max_depth : [5, 10, 15, 20] learning_rate : [0.05, 0.1, 0.2] min_child_weight : [10, 15, 20] subsample : [0.6, 0.7, 0.8] | 200 5 0.05 10 0.6 |
| Light Gradient Boosting Regression | n_estimators : [100, 200, 300] max_depth : [-1, 5, 15, 25] learning_rate : [0.05, 0.1, 0.2] num_leaves : [30, 35, 40, 45] min_data_in_leaf : [60, 65, 70] | 200 15 0.05 30 60 |
| Artificial Neural Network | hidden_layer_sizes : [(64,),(128,),(128,64) (256,128), (512,256)] solver : ['adam'] activation : ['identity', 'relu'] learning_rate : ['constant', 'adaptive'] | (128, 64) Adam relu constant |

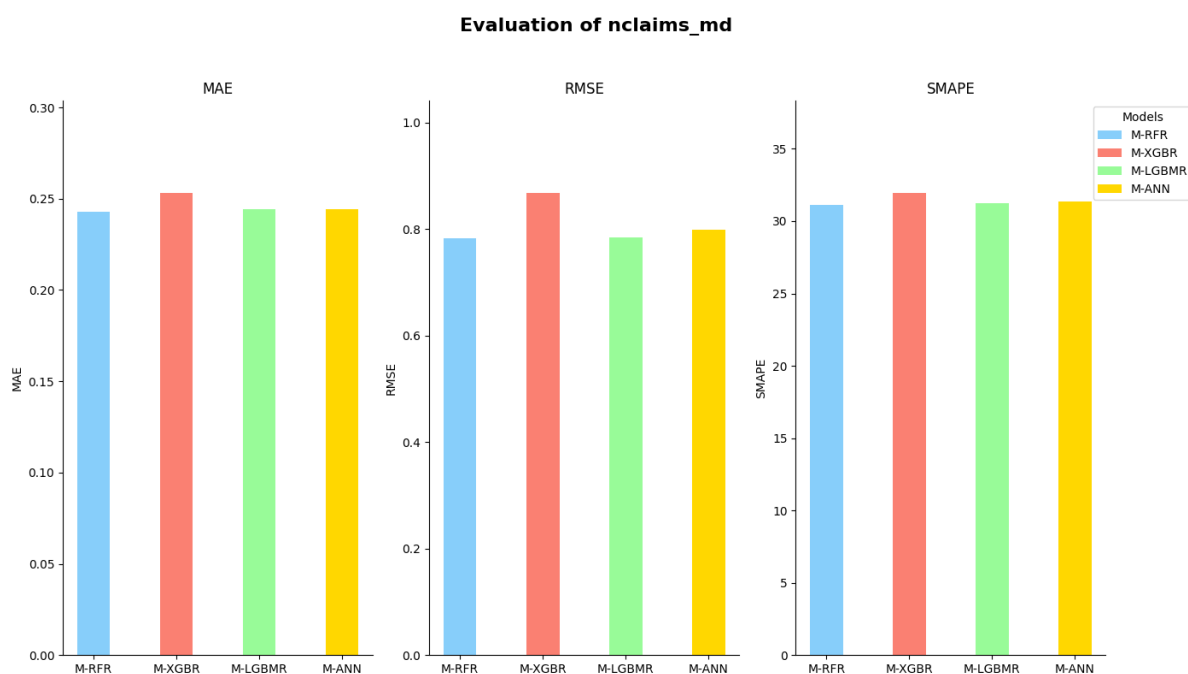
4.5 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลอง

4.5.1 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองประเภทการถดถอยพหุผลลัพธ์

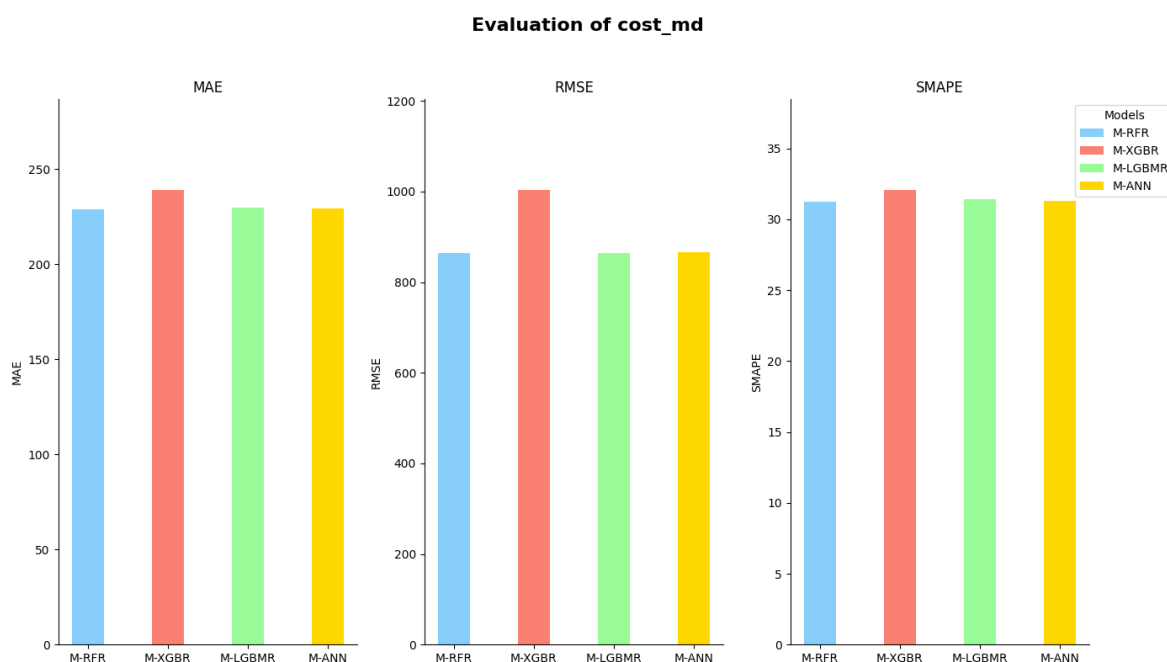
พารามิเตอร์ที่ดีที่สุดของแต่ละแบบจำลองจะถูกทดสอบกับชุดข้อมูลทดสอบเพื่อประเมินและเปรียบเทียบประสิทธิภาพในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ ทั้ง 4 แบบจำลอง โดยมีผลตัวเลขแสดงประสิทธิภาพดังในตารางที่ 4.8 และกราฟเปรียบเทียบประสิทธิภาพในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนดังภาพที่ 4.13 และ 4.14 ตามลำดับ

ตารางที่ 4.8 ผลการทดสอบประสิทธิภาพการทำนายของแต่ละแบบจำลอง

| Model | nclaims_md | | | cost_md | | | Duration (seconds) |
|---------|------------|--------|---------|----------|-----------|---------|-----------------------|
| | MAE | RMSE | SMAPE | MAE | RMSE | SMAPE | |
| M-RFR | 0.2427 | 0.7832 | 28.8311 | 228.9800 | 864.6607 | 29.2370 | 1658 |
| M-XGBR | 0.2520 | 0.8621 | 29.0140 | 238.9562 | 1003.0973 | 29.4286 | 24 |
| M-LGBMR | 0.2440 | 0.7863 | 28.9626 | 229.7467 | 865.2349 | 29.2945 | 35 |
| M-ANN | 0.2443 | 0.7998 | 28.8917 | 229.3862 | 866.5528 | 29.2581 | 1937 |



ภาพที่ 4.13 ประสิทธิภาพการทำนายความถี่ของการเรียกร้องค่าสินไหมทดแทน



ภาพที่ 4.13 ประสิทธิภาพการทำนายความความรุนแรงของการเรียกร้องค่าสินไหมทดแทน

จากตารางที่ 4.8 พบว่าแบบจำลองที่สามารถทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ได้ดีที่สุดคือ Multi-Output Random Forest Regression โดยในการทำนายความถี่ของการเรียกร้องค่าสินไหมทดแทน มีค่า MAE เท่ากับ 0.2427 ค่า RMSE เท่ากับ 0.7832 ค่า SMAPE เท่ากับ 31.1281 ในการทำนายความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ มีค่า MAE เท่ากับ 228.9800 ค่า RMSE เท่ากับ 864.6607 และค่า SMAPE เท่ากับ 31.2500 และทำการทำนายโดยใช้ข้อมูลชุดทดสอบแสดงดังในตารางที่ 4.9

ตารางที่ 4.9 ค่าจริงและค่าจากการทำนายโดยใช้แบบจำลอง M-RFR จากข้อมูลชุดทดสอบ

| ตัวแปร | คนที่ 1 | คนที่ 2 | คนที่ 3 | คนที่ 4 | คนที่ 5 |
|--------------------|----------|---------|----------|----------|---------|
| client_sex | 0 | 1 | 0 | 1 | 0 |
| client_age | 54 | 56 | 53 | 53 | 53 |
| lic_age | 32 | 37 | 33 | 28 | 35 |
| city | 0 | 1 | 0 | 0 | 0 |
| north | 1 | 0 | 1 | 0 | 0 |
| rest | 0 | 0 | 0 | 1 | 1 |
| client_nother | 0 | 0 | 0 | 5 | 0 |
| npol_auto | 13 | 2 | 12 | 24 | 8 |
| cost_md | 0 | 2170.76 | 358 | 3505.01 | 0 |
| cost_md Predict | 250.6056 | 0 | 146.5017 | 508.5177 | 0 |
| nclaims_md | 0 | 5 | 1 | 5 | 0 |
| nclaims_md Predict | 1 | 0 | 1 | 2 | 0 |

4.5.2 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองประเภทการถดถอยพหุผลลัพธ์กับการถดถอยผลลัพธ์เดียว

จากผลการวิจัยในข้อ 4.5.1 พบว่าแบบจำลองที่สามารถทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนได้ดีที่สุดคือ การถดถอยพหุผลลัพธ์แบบสุ่มป่า (Multi-Output Random Forest Regression) จึงสร้างแบบจำลองการถดถอยพหุผลลัพธ์แบบสุ่มป่าประเภทการถดถอยผลลัพธ์เดียว (Single-Output Regression) โดยทำนายเป็นแยกผลลัพธ์ ได้แก่ 1.ทำนายความถี่ของการเรียกร้องค่าสินไหมทดแทน และ 2.ทำนายความรุนแรงของการเรียกร้องค่าสินไหมทดแทน โดยมีการปรับแต่งพารามิเตอร์และค่าที่ดีที่สุด แสดงดังในตารางที่ 4.10

ตารางที่ 4.10 ค่าพารามิเตอร์จาก GridSearchCV ของ Single-Output Random Forest Regression

| Random Forest Model | Tuning Parameters | Best Parameters |
|---------------------|--|-----------------|
| cost_md | n_estimators : [100, 200, 300] | 100 |
| | max_depth : [20, 22, 24, 26] | 26 |
| | min_samples_split : [90, 100, 110, 130, 150] | 150 |
| | min_samples_leaf : [50, 55, 60] | 50 |
| nclaims_md | n_estimators : [100, 200, 300] | 100 |
| | max_depth : [18, 20, 22, 24, 26] | 18 |
| | min_samples_split : [80, 100, 120, 140] | 80 |
| | min_samples_leaf : [40, 50, 60] | 50 |

นำพารามิเตอร์ที่ดีที่สุดของแบบจำลองการถดถอยผลลัพธ์เดียวมาทดสอบกับข้อมูลชุดทดสอบ และเปรียบเทียบประสิทธิภาพของแบบจำลองการถดถอยผลลัพธ์เดียวกับการทำนายจากแบบจำลองการถดถอยพหุผลลัพธ์ ผลการทดสอบดังแสดงในตารางที่ 4.11

ตารางที่ 4.11 ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองประเภทการถดถอยพหุผลลัพธ์กับการถดถอยผลลัพธ์เดียว

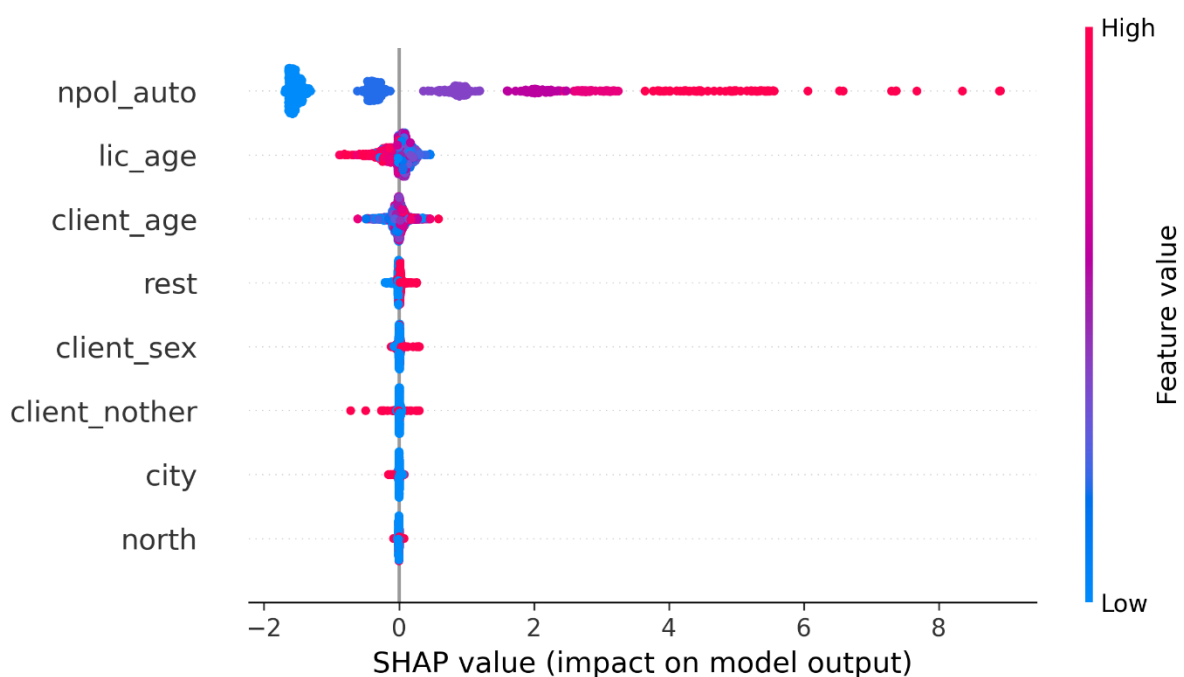
| Random Forest Model | MAE | RMSE | SMAPE |
|---------------------|-----------------|-----------------|----------------|
| S - nclaims_md | 0.2482 | 0.8074 | 29.8663 |
| M - nclaims_md | 0.2427 | 0.7832 | 31.1281 |
| S - cost_md | 225.5343 | 849.2932 | 198.1965 |
| M - cost_md | 228.9800 | 864.6607 | 31.2500 |

จากตารางที่ 4.11 พบว่าแบบจำลองประเภทการถดถอยพหุผลลัพธ์กับการถดถอยผลลัพธ์เดียวมีประสิทธิภาพใกล้เคียงกัน โดย

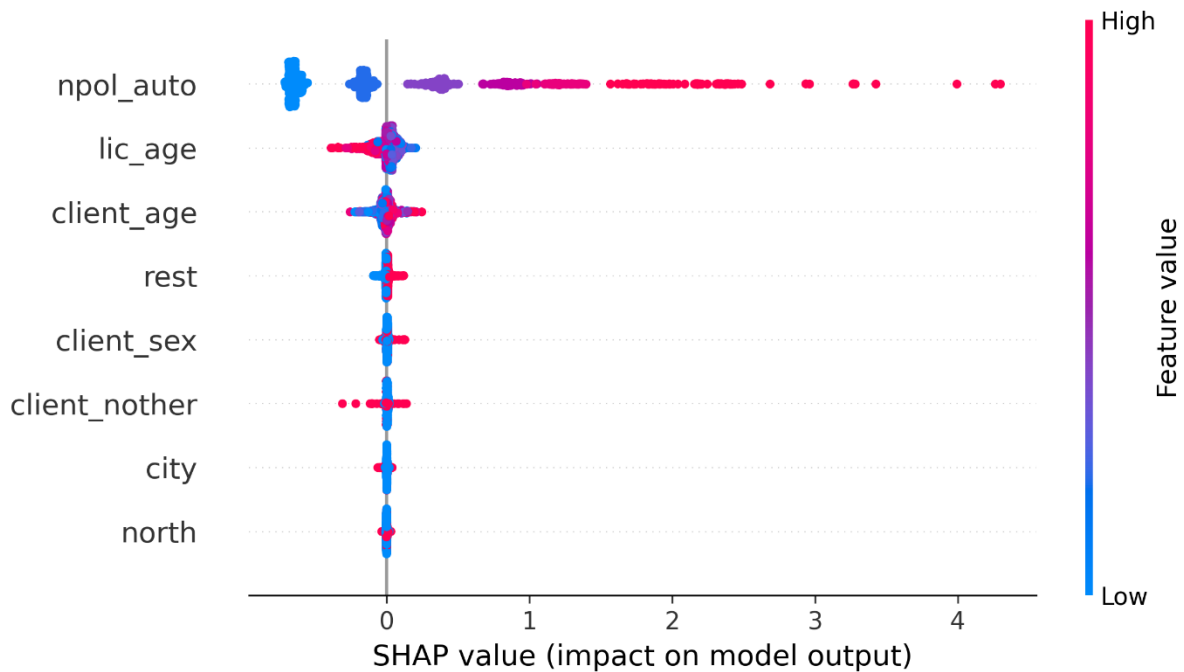
- การทำนายความถี่ในการเรียกรถเช่าใหม่ทดแทนแบบจำลองที่สามารถทำนายได้ดีคือ Multi-Output Random Forest Regression มีค่า MAE เท่ากับ 0.2427 ค่า RMSE เท่ากับ 0.7832 และค่า SMAPE เท่ากับ 31.1281
- การทำนายความรุนแรงในการเรียกรถเช่าใหม่ทดแทนแบบจำลองที่สามารถทำนายได้ดีคือ Single-Output Random Forest Regression มีค่า MAE เท่ากับ 225.5343 ค่า RMSE เท่ากับ 849.2932 และค่า SMAPE เท่ากับ 198.1965

4.5.3 ผลการอธิบายแบบจำลองด้วยเทคนิค SHAP

ผลการทำนายจาก Multi-Output Random Forest Regression สามารถอธิบายเพิ่มเติมได้โดยใช้เทคนิค Shapley Additive Explanation (SHAP) เพื่อให้เข้าใจถึงคุณลักษณะที่ส่งผลต่อการทำนายของแบบจำลองทั้งในเชิงบวกและเชิงลบ ผลการวิเคราะห์ SHAP แสดงในภาพที่ 4.3 และ 4.4



ภาพที่ 4.3 ผล SHAP ของแบบจำลอง M-RFR ในการทำนายความถี่ของการเรียกรถเช่าใหม่ทดแทน



ภาพที่ 4.4 ผล SHAP ของแบบจำลอง M-RFR ในการทำนายความรุนแรงของการเรียกร้องค่าสินไหมทดแทน จากภาพที่ 4.3 และ 4.4 แสดงผล SHAP ที่สะท้อนถึงความสำคัญของคุณลักษณะต่อการทำนาย ความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนของแบบจำลอง M-RFR พบว่าคุณลักษณะที่ ผลกระทบมากต่อการทำนาย 3 อันดับแรก ได้แก่

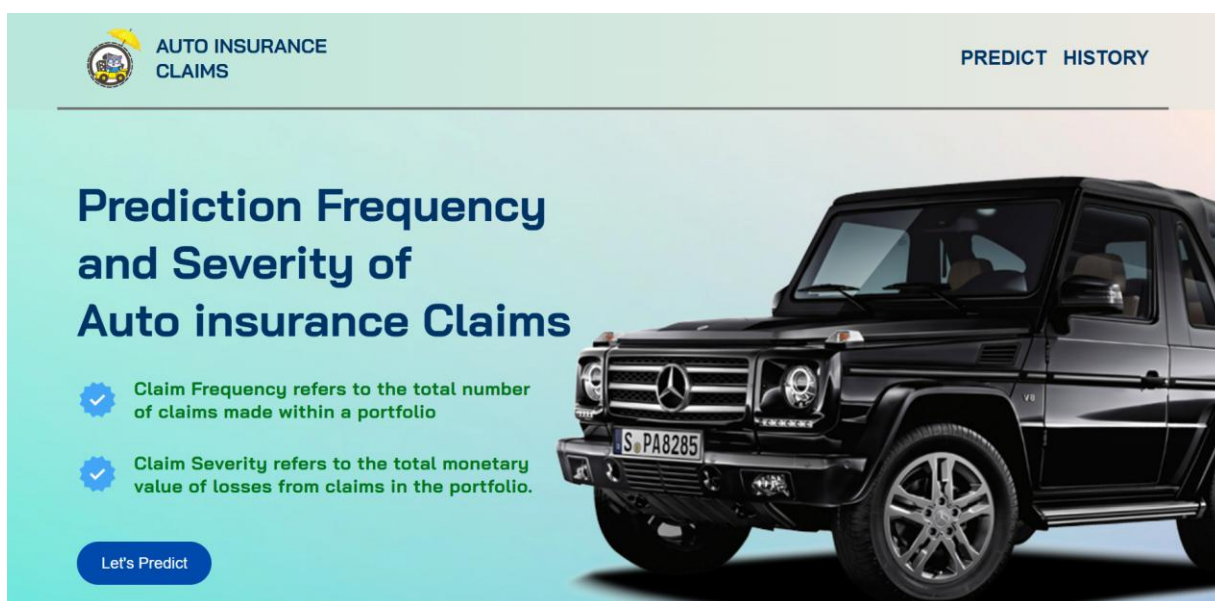
- จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี (npol_auto) มีผลกระทบเชิงบวกกับการทำนาย หายความว่าผู้ที่ถือกรมธรรม์ประกันภัยรถยนต์จำนวนมากมีแนวโน้มที่จะเรียกร้องค่าสินไหม ทั้งความถี่และความรุนแรงมาก
- อายุของใบอนุญาตขับขี่ (lic_age) มีผลกระทบทั้งเชิงลบและเชิงบวก สามารถพิจารณาหลากหลาย ดังนี้ 1. หากมีประสบการณ์ในการขับขี่มากจะมีแนวโน้มการร้องค่าสินไหมทดแทนน้อยลง 2. หาก ประสบการณ์ในการขับขี่มากและอายุมาก อาจมีความเสี่ยงเพิ่มขึ้นจากข้อจำกัดด้านสมรรถภาพทาง ร่างกาย ส่งผลให้ความรุนแรงของการเรียกร้องค่าสินไหมสูง และ 3. หากผู้ถือกรมธรรม์มีประสบการณ์ ในการขับขี่น้อยลง มีแนวโน้มที่จะเรียกร้องค่าสินไหมทดแทนที่สูงขึ้น
- อายุของผู้ถือกรมธรรม์ (client_age) มีผลกระทบทั้งเชิงลบและเชิงบวก ซึ่งสัมพันธ์กับพฤติกรรมการ ขับขี่ สามารถพิจารณาหลากหลายดังนี้ 1. ผู้ถือกรมธรรม์ที่มีอายุในช่วงวัยกลางคนมักเป็นกลุ่มที่มี ความรับผิดชอบสูงและมีพฤติกรรมการขับขี่ที่ระมัดระวัง ทำให้ความถี่และความรุนแรงของการ เรียกร้องค่าสินไหมลดลง 2. ผู้ถือกรมธรรม์ที่อายุน้อยบางรายอาจขาดประสบการณ์ในการขับขี่ ส่งผล ให้ความเสี่ยงต่อการเกิดอุบัติเหตุเพิ่มขึ้น และอาจทำให้การเรียกร้องค่าสินไหมมีความถี่หรือรุนแรง มากขึ้น 3. ผู้ถือกรมธรรม์ที่มีอายุเยอะมากอาจเผชิญข้อจำกัดด้านสมรรถภาพทางร่างกาย เช่น การ

มองเห็นหรือการตอบสนองที่ลดลง ซึ่งเพิ่มความเสี่ยงต่ออุบัติเหตุและส่งผลให้การเรียกร้องค่าสินไหมรุนแรงขึ้น

คุณลักษณะที่มีผลกระทบต่อการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทน ได้แก่ การอาศัยอยู่ในพื้นที่อื่น ๆ ของประเทศ (rest) เพศของผู้ถือกรมธรรม์ (client_sex) จำนวนกรมธรรม์ประเภทอื่นในบริษัทเดียวกัน (client_nothor) พื้นที่อยู่อาศัย (city) การอาศัยอยู่ในภาคเหนือหรือไม่ (north) คุณลักษณะเหล่านี้มีผลกระทบต่อการทำนายในระดับต่ำ ซึ่งอาจพิจารณาตัดออกเพื่อลดความซับซ้อนของแบบจำลองและเพิ่มประสิทธิภาพในอนาคต

4.6 ผลการพัฒนาเว็บแอปพลิเคชัน

ผลการพัฒนาเว็บแอปพลิเคชันสำหรับการทำนายความถี่และมูลค่าการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์ มีหน้าต่างแสดงดังภาพที่ 4.16 การกรอกข้อมูลดังภาพที่ 4.17, 4.18 และผลการทำนายดังภาพที่ 4.19



ภาพที่ 4.16 หน้าต่างเว็บแอปพลิเคชัน

Enter Individual Policyholder Information

| | |
|----------------|---|
| n_pol | Number of motor insurance policies held by the client. |
| gender | Gender of the client |
| age | Age of the client. |
| license age | Number of years since the client obtained their driving license. |
| client_norther | The number of other types of insurance policies held by the client within the company ▼ |
| city | The client resides in a major city ▼ |
| north | The client resides in the northern region. ▼ |
| rest | The client resides in other areas of the country ▼ |

*Rest is automatically generated based on other variables; no manual input is required.

ภาพที่ 4.17 หน้าต่างเว็บแอปพลิเคชันสำหรับกรอกข้อมูลผู้เอาประกันภัย (1)

Enter Individual Policyholder Information

| | |
|------|---|
| 16 | Number of motor insurance policies held by the client. |
| Male | Gender of the client |
| 26 | Age of the client. |
| 11 | Number of years since the client obtained their driving license. |
| 13 | The number of other types of insurance policies held by the client within the company ▼ |
| Yes | The client resides in a major city ▼ |
| No | The client resides in the northern region. ▼ |
| No | The client resides in other areas of the country ▼ |


*North is automatically populated based on City variables. No manual input is required.

*Rest is automatically generated based on other variables; no manual input is required.

ภาพที่ 4.18 หน้าต่างเว็บแอปพลิเคชันหลังจากกรอกข้อมูลผู้เอาประกันภัย (2)

Enter Individual Policyholder Information

| | |
|--|---|
| <input type="text" value="16"/> | Number of motor insurance policies held by the client. |
| <input type="text" value="Male"/> | Gender of the client |
| <input type="text" value="26"/> | Age of the client. |
| <input type="text" value="11"/> | Number of years since the client obtained their driving license. |
| <input type="text" value="13"/> | The number of other types of insurance policies held by the client within the company ▼ |
| <input type="text" value="Yes"/> | The client resides in a major city ▼ |
| <input type="text" value="No"/> | The client resides in the northern region. ▼ |
| <i>*North is automatically populated based on City variables. No manual input is required.</i> | |
| <input type="text" value="No"/> | The client resides in other areas of the country ▼ |
| <i>*Rest is automatically generated based on other variables; no manual input is required.</i> | |



| | |
|--|--|
| Claim Frequency <input type="text" value="1"/> | Claim Severity <input type="text" value="220.3943"/> |
|--|--|

ภาพที่ 4.19 หน้าต่างเว็บแอปพลิเคชันที่แสดงผลการทำนาย

บทที่ 5

สรุปผลการวิจัย

การวิเคราะห์การถดถอยพหุผลัฟฟ์ในการทำนายความถี่และความรุนแรงของการเรียกร้อยค่าสินไหมทดแทนในการประกันภัยรถยนต์ มีวัตถุประสงค์เพื่อสร้างแบบจำลองทำนายความถี่และความรุนแรงของการเรียกร้อยค่าสินไหมทดแทนในการประกันภัยรถยนต์ด้วยการเรียนรู้ของเครื่อง และเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องในการทำนายความถี่และความรุนแรงของการเรียกร้อยค่าสินไหมทดแทน ข้อมูลที่ใช้ในการวิจัยครั้งนี้คือข้อมูลการประกันภัยรถยนต์ของบริษัทแห่งหนึ่งในประเทศสเปน ตั้งแต่ปี ค.ศ. 2006 ถึง ค.ศ. 2015 เป็นระยะเวลา 10 ปี รวมทั้งสิ้น 80,924 ราย สามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้ใช้แบบจำลองการถดถอยพหุผลัฟฟ์ในการทำนายความถี่และความรุนแรงในการเรียกร้อยค่าสินไหมทดแทนทั้งหมด 4 แบบจำลอง ได้แก่ Random Forest, Extreme Gradient Boosting, Light Gradient Boosting และ Artificial Neural Network พบว่า Random Forest มีประสิทธิภาพในการทำนายมากที่สุด โดยมีผลคือ

1. ในการทำนายความถี่ของการเรียกร้อยค่าสินไหมทดแทนในการประกันภัยรถยนต์มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เท่ากับ 0.2427 ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เท่ากับ 0.7832 ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (SMAPE) เท่ากับ 28.8311
2. ในการทำนายความรุนแรงของการเรียกร้อยค่าสินไหมทดแทนในการประกันภัยรถยนต์ มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เท่ากับ 228.9800 ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เท่ากับ 864.6607 ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (SMAPE) เท่ากับ 29.2370

ผลการอธิบายแบบจำลองด้วย Shapley Additive Explanation (SHAP) พบว่าตัวแปรที่มีผลต่อการทำนายทั้งความถี่และความรุนแรงในการเรียกร้อยค่าสินไหมทดแทน ได้แก่ จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี อายุของใบอนุญาตขับขี่ และอายุของผู้ถือกรมธรรม์

จากผลการวิจัยพบว่าแบบจำลองพหุผลัฟฟ์โดยใช้ Random Forest มีประสิทธิภาพในการทำนายสูงสุด จึงได้สร้างแบบจำลองการถดถอยพหุผลัฟฟ์เดียวโดยใช้ Random Forest เหมือนกัน เพื่อเปรียบเทียบประสิทธิภาพกับแบบจำลองการถดถอยพหุผลัฟฟ์ พบว่าแบบจำลองการถดถอยพหุผลัฟฟ์มีความแม่นยำใน

การทำนายความถี่ของการเรียกร้อยค่าสินไหมทดแทนดีกว่า ขณะที่แบบจำลองการถดถอยผลลัพธ์เดียวมีความแม่นยำในการทำนายความรุนแรงของการเรียกร้อยค่าสินไหมทดแทนสูงกว่า ซึ่งทั้งสองแบบจำลองมีประสิทธิภาพในการทำนายโดยรวมใกล้เคียงกัน โดยผลการวิเคราะห์แบบจำลองผลลัพธ์เดียว พบว่าการทำนายความถี่ในการเรียกร้อยค่าสินไหมทดแทน มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เท่ากับ 0.2482 ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เท่ากับ 0.8074 ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (SMAPE) เท่ากับ 29.8663 และการทำนายความรุนแรงในการเรียกร้อยค่าสินไหมทดแทน มีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) เท่ากับ 225.5343 ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เท่ากับ 849.2932 ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (SMAPE) เท่ากับ 198.1965

ผู้วิจัยได้พัฒนาเว็บแอปพลิเคชันเพื่อให้บริษัทประกันภัยรถยนต์สามารถเข้าถึงและทำนายความถี่และความรุนแรงของการเรียกร้อยค่าสินไหมทดแทนได้สะดวกและรวดเร็ว การทำงานของระบบเริ่มจากผู้ใช้กรอกข้อมูลของผู้เอาประกันภัยที่หน้าเว็บแอปพลิเคชัน จากนั้นกดทำนายเพื่อให้ระบบประมวลผลด้วยแบบจำลอง Multi-Output Random Forest Regression และผลลัพธ์จะแสดง 2 ค่า คือ ค่าความถี่และความรุนแรงของการเรียกร้อยค่าสินไหมทดแทนในการประกันภัยรถยนต์

5.2 อภิปรายผลการวิจัย

การสร้างแบบจำลองพหุผลลัพธ์ในการทำนายความถี่และความรุนแรงในการเรียกร้อยค่าสินไหมทดแทนในการประกันภัยรถยนต์ โดยใช้แบบจำลองกา

รเรียนรู้ของเครื่อง จำนวน 4 แบบจำลอง ได้แก่ Random Forest, Extreme Gradient Boosting (XGboost), Light Gradient Boosting (LightGBM), Artificial Neural Network (ANN) ผลการวิจัยพบว่า Random Forest มีประสิทธิภาพในการทำนายสูงสุดเมื่อเปรียบเทียบกับแบบจำลอง และเพื่อทำความเข้าใจตัวแปรที่มีผลต่อการทำนาย ผู้วิจัยใช้ Shapley Additive Explanation (SHAP) ในการวิเคราะห์ความสำคัญของตัวแปร พบว่าตัวแปรที่ส่งผลต่อการทำนายทั้งความถี่และความรุนแรงของการเรียกร้อยค่าสินไหมทดแทน 3 อันดับแรก ได้แก่ จำนวนกรมธรรม์ของผู้ถือกรมธรรม์ อายุของผู้ถือกรมธรรม์ และอายุของใบอนุญาตขับขี่ ซึ่งอายุของผู้ถือกรมธรรม์ สอดคล้องกับงานวิจัยของ Jin (2021) ที่ระบุว่า

การฝึกแบบจำลอง XGBoost ใช้เวลาสร้างแบบจำลองน้อยที่สุดที่ 24 วินาที ตามด้วย LightGBM ที่ 35 วินาที เนื่องจากใช้เทคนิค Gradient Boosting ที่มีประสิทธิภาพสูง ส่วน Random Forest ใช้เวลาที่ 27 นาที 38 วินาที เนื่องจากต้องสร้างต้นไม้จำนวนมาก ขณะที่ ANN ใช้เวลามากที่สุดที่ 32 นาที 17 วินาที เนื่องจากต้องอัปเดตพารามิเตอร์หลายรอบผ่าน Backpropagation ปัจจัยที่ส่งผลกระทบต่อระยะเวลาการสร้างแบบจำลอง ได้แก่ ขนาดข้อมูล จำนวนพารามิเตอร์ และโครงสร้างแบบจำลอง โดย ANN ใช้เวลาเพิ่มขึ้นแบบ

เชิงเส้นเมื่อขนาดข้อมูลเพิ่มขึ้น ในขณะที่ LightGBM และ XGBoost ยังคงใช้เวลาน้อยกว่า แสดงถึงประสิทธิภาพในการจัดการข้อมูลขนาดใหญ่

การเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองการถดถอยพหุผลลัพธ์ (Multi-Output Regression) กับแบบจำลองการถดถอยผลลัพธ์เดียว (Single-Output Regression) พบว่าประสิทธิภาพของทั้งสองแบบจำลองมีความใกล้เคียงกันในภาพรวม แต่แบบจำลอง Multi-Output Regression มีข้อได้เปรียบในด้านความสะดวก เนื่องจากสามารถทำนายหลายผลลัพธ์ได้พร้อมกัน在一次เดียว ช่วยลดขั้นตอนและเวลาในการคำนวณ ดังนั้นในการประยุกต์ใช้งานจริง ผู้วิจัยได้พัฒนาเว็บแอปพลิเคชันที่เชื่อมต่อกับแบบจำลอง Multi-Output Random Forest Regression ผ่าน Flask ซึ่งช่วยให้บริษัทประกันภัยรถยนต์สามารถใช้งานแบบจำลองได้อย่างสะดวกและรวดเร็ว โดยผู้ใช้กรอกข้อมูลของผู้เอาประกันภัยลงในระบบ และกดทำนาย จากนั้นระบบจะประมวลผลและแสดงผลลัพธ์ประกอบด้วยค่าทำนายสองค่าคือ ความถี่และความรุนแรงของการเรียกร้องค่าสินไหมในการประกันภัยรถยนต์

5.3 ประโยชน์ของสถิติ/สารสนเทศสถิติที่ใช้ในการวิจัย

5.3.1 การแปลงข้อมูลก่อนการวิเคราะห์

การแปลงข้อมูลด้วยฟังก์ชันลอการิทึมธรรมชาติ (Logarithmic Transformation) ช่วยหลีกเลี่ยงค่าทำนายที่ติดลบ ลดการกระจายตัวของข้อมูลที่มีการเบี่ยงเบนมาก (skewness) และผลกระทบของค่าผิดปกติ (outliers) ทำให้มีการกระจายตัวของข้อมูลมีความสมดุลยิ่งขึ้น ส่งผลให้ผลลัพธ์สอดคล้องกับข้อมูลจริงและการทำนายมีความแม่นยำมากขึ้น

5.3.2 การเรียนรู้ของเครื่อง

การสร้างแบบจำลองทำนายด้วยการเรียนรู้ของเครื่อง (Machine Learning) สามารถทำนายหลายผลลัพธ์พร้อมกันได้ และมีการปรับปรุงประสิทธิภาพของแบบจำลองหลายวิธี เช่น การปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) เพื่อหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับแต่ละแบบจำลอง และการใช้เทคนิค Cross-Validation (CV) เพื่อประเมินความสามารถของแบบจำลองในชุดข้อมูลที่แตกต่างกัน วิธีดังกล่าวช่วยเพิ่มความแม่นยำในการทำนายและทำให้แบบจำลองทำนายได้อย่างมีประสิทธิภาพในสถานการณ์ที่หลากหลาย

5.3.3 การพัฒนาเว็บแอปพลิเคชัน

การสร้างเว็บแอปพลิเคชันด้วยภาษา Python, HTML, CSS, JavaScript และเชื่อมต่อกับแบบจำลองผ่าน Flask ช่วยให้สามารถนำแบบจำลองไปใช้งานได้ในรูปแบบที่ผู้ใช้ (Users) สามารถเข้าถึงได้ง่ายผ่านเบราว์เซอร์ โดยไม่จำเป็นต้องติดตั้งซอฟต์แวร์เพิ่มเติม

5.4 ข้อเสนอแนะ

5.4.1 ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1) บริษัทประกันภัยสามารถนำผลการวิจัยไปใช้ในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนพร้อมกัน ที่มาจากผู้เอาประกันรายเดียวกัน ซึ่งมีข้อมูลต่างๆ สอดคล้องกับงานวิจัยนี้ โดยเฉพาะอย่างยิ่งในประเทศสเปน หรือภูมิภาคยุโรปที่มีความเสี่ยงที่คล้ายกัน สำหรับบริษัทที่มีข้อมูลของตัวแปรอิสระแตกต่างจากงานวิจัยนี้ อาจพิจารณาสร้างแบบจำลอง Random Forest โดยใช้ตัวแปรอิสระที่มีมาใช้ในการทำนายการเรียกร้องค่าสินไหมทดแทน ซึ่งสามารถประยุกต์ใช้วิธีการเตรียมข้อมูล และขั้นตอนการสร้างแบบจำลองจากงานวิจัยนี้ ซึ่งจะช่วยให้บริษัทประกันภัยสามารถคาดการณ์แนวโน้มของการเรียกร้องค่าสินไหมทดแทนในอนาคตได้อย่างแม่นยำและเตรียมการรับมือได้อย่างเหมาะสม โดยแบบจำลอง Random Forest จะใช้เวลาในการประมวลผลค่อนข้างนานหากมีข้อมูลจำนวนมาก บริษัทประกันภัยอาจใช้พิจารณาใช้แบบจำลองที่มีระยะเวลาในการฝึกน้อยกว่าแต่ประสิทธิภาพใกล้เคียงกับ Random Forest ได้แก่ แบบจำลอง LightGBM และ XGBoost

นอกจากนี้ การวิเคราะห์ผลกระทบของปัจจัยต่าง ๆ ด้วยเทคนิค Shapley Additive Explanation (SHAP) พบว่า จำนวนกรมธรรม์ประกันภัยรถยนต์ที่ผู้ถือกรมธรรม์มี อายุของใบอนุญาตขับขี่ และอายุของผู้ถือกรมธรรม์ เป็นปัจจัยสำคัญที่ส่งผลต่อการเรียกร้องค่าสินไหมทดแทน เป็นข้อมูลช่วยสนับสนุนการตัดสินใจในการกำหนดเบี้ยประกันภัยของผู้เอาประกันภัยแต่ละรายอย่างเหมาะสม รวมถึงนำไปใช้ในการกำหนดกลยุทธ์และวางแผนบริหารจัดการความเสี่ยงอื่นๆ ในลำดับต่อไป

2) เนื่องจากข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลในอดีต ซึ่งอาจไม่สอดคล้องกับบริบทสภาพแวดล้อม หรือพฤติกรรมของผู้เอาประกันภัยเมื่อเวลาเปลี่ยนแปลงไป ดังนั้นควรมีการอัปเดตแบบจำลองเมื่อเวลาผ่านไประยะหนึ่ง นอกจากนี้อาจพิจารณานำปัจจัยอื่น ๆ เพิ่มเติมมาใช้ในการสร้างแบบจำลอง เช่น การเปลี่ยนแปลงของปัจจัยทางเศรษฐกิจ สังคม และพฤติกรรมของผู้เอาประกันภัยในช่วงเวลาต่าง ๆ เพื่อให้สอดคล้องกับสถานการณ์ปัจจุบัน ซึ่งอาจส่งผลต่อความถี่และมูลค่าของการเรียกร้องค่าสินไหมทดแทน

5.4.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

1) ข้อมูลที่ใช้ในการสร้างแบบจำลองเป็นข้อมูลทางด้านประชากรศาสตร์ ซึ่งอาจส่งผลต่อความแม่นยำ เนื่องจากไม่สะท้อนถึงการเปลี่ยนแปลงของตลาดหรือพฤติกรรมของผู้เอาประกันภัยเมื่อเวลาเปลี่ยนแปลงไป ในการทำวิจัยครั้งต่อไปอาจทำการเพิ่มจำนวนตัวแปร และปัจจัยทางด้านอื่นๆ เช่น เศรษฐกิจ สังคม มาใช้ในการสร้างแบบจำลอง

2) ในการวิจัยครั้งนี้ได้ดำเนินการทดลองเฉพาะแบบจำลองต้นไม้ประเภท Ensemble ได้แก่ Random Forest, XGBoost และ LightGBM และโครงข่ายประสาทเทียม (ANN) เท่านั้น ในการวิจัยครั้งต่อไป อาจพิจารณานำแบบจำลองประเภทอื่น ๆ เช่น แบบจำลองการถดถอยเชิงเส้น (Linear Regression), การถดถอยด้วยเวกเตอร์สนับสนุน (SVR) และแบบจำลอง K-Nearest Neighbors (KNN) เพื่อขยายขอบเขตของการศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองที่หลากหลาย

3) ในการพิจารณาจัดการข้อมูลก่อนนำมาสร้างแบบจำลอง เนื่องจากในการวิจัยครั้งนี้ได้ใช้ข้อมูลจริงโดยไม่ได้ดำเนินการตัดหรือจัดการค่าที่อยู่นอกเกณฑ์ (outliers) ซึ่งอาจส่งผลกระทบต่อความแม่นยำของผลลัพธ์ อย่างไรก็ตาม จากการทดลองเบื้องต้นพบว่า เมื่อมีการตัดค่าที่อยู่นอกเกณฑ์ออกแล้ว ค่าร้อยละความคลาดเคลื่อนสัมบูรณ์เฉลี่ยแบบสมมาตร (SMAPE) ลดลงจาก 30 เหลือ 20 หมายความว่า การจัดการค่าที่อยู่นอกเกณฑ์อาจช่วยเพิ่มความแม่นยำของแบบจำลองได้

4) การพัฒนาแบบจำลองโดยใช้เทคนิคการเลือกคุณสมบัติ (feature selection) อาจช่วยเพิ่มประสิทธิภาพของแบบจำลองได้ โดยการคัดเลือกเฉพาะคุณสมบัติที่มีความสำคัญต่อการทำนายจะช่วยลดความซับซ้อนของแบบจำลอง และอาจส่งผลให้มีความแม่นยำสูงขึ้น รวมถึงลดการเกิดปัญหาการ overfitting ซึ่งช่วยให้แบบจำลองสามารถทำงานได้ดีขึ้นในสภาพแวดล้อมที่ไม่เคยเห็นข้อมูลมาก่อน

เอกสารอ้างอิง

กิตติศักดิ์ จังพานิช, ศุภเจษฎา สิวันนา, วรรณพร เซวณชวานิล, กุลภัสสรณ์ ชีวลักษณะณาสีทธิ์, และธีระวัฒน์

สีมากันทร. (2564). การประยุกต์ใช้การเรียนรู้ของเครื่องในการตรวจจับการเรียกร้องค่าสินไหม

ทดแทนประกันภัยรถยนต์ที่ไม่มีการออกสำรวจภัย. *วารสารวิทยาศาสตร์ มหาวิทยาลัยหัวเฉียวเฉลิม*

พระเกียรติ, 27(2), 50-63.

บริษัท รูจี้ใจ จำกัด. (2566). **ประกันรถยนต์คืออะไร**. ค้นเมื่อ 1 กรกฎาคม 2567, จาก <https://www.roojai.com/insurance-glossary/what-is-car-insurance/>

ปวีศา สุขเรื่อย และ สำรวม จงเจริญ. (2561). ตัวแบบการถดถอยที่มีผลกระทบจากค่าศูนย์ ประยุกต์ใช้กับ จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนในประกันภัยรถยนต์ภาคสมัครใจ. *วารสาร มหาวิทยาลัยศรีนครินทรวิโรฒ (สาขาวิทยาศาสตร์และเทคโนโลยี) ปีที่ 10 ฉบับที่ 20 กรกฎาคม - ธันวาคม 2561.*

ปฏิภาณ ประเสริฐสม และ พีรตล สามะศิริ. (2566). **ตีความโมเดล Machine Learning: ตัวอย่างและการตีความ Shapley value**. ค้นเมื่อ 10 ตุลาคม 2567, จาก <https://bdi.or.th/big-data-101/shapley-value-example/>

สมาคมประกันวินาศภัยไทย. (2023). **คาดผลประกอบการธุรกิจประกันวินาศภัย ปี 66**. The General Insurance Association of Thailand. สืบค้นเมื่อ 8 กรกฎาคม 2567. https://www.tgia.org/newsandevents-detail-EN_1344_1

อัศรพล พรหมพิริยะพงษ์. (2566). **การวิเคราะห์คุณลักษณะที่สามารถคาดการณ์ปริมาณการเพาะปลูกแบบเกษตรแม่นยำ: กรณีศึกษาผลผลิตข้าว**.วิทยานิพนธ์ปริญญาโท สาขาวิชาการแปรรูปธุรกิจทางดิจิทัล คณะ วิทยาศาสตร์มหาบัณฑิต มหาวิทยาลัยธรรมศาสตร์

Athiwat. (2019). **Machine Learning คืออะไร?**. Medium. Retrieved July 1, 2024. from <https://bit.ly/3Un9MhJ>

Ahmed, M. W. (2023). **Understanding mean absolute error (MAE) in regression: A practical guide**. Retrieved July 5, 2024, from <https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df>

Amarin TV. (2567). **สินเชื่อเงินด่วน สมัครง่าย ไม่ต้องคำประกัน เช็กได้ที่นี่**. Amarin TV. สืบค้นเมื่อ 8 กรกฎาคม พ.ศ. 2567. <https://www.amarintv.com/spotlight/money-product/detail/37444>

Abebe, M., Shin, Y., Noh, Y., Lee, S. & Lee, I. (2020). Machine Learning Approaches for Ship Speed Prediction toward Energy Efficient Shipping. *Applied Sciences, 10(7).*

- BDI. (2020). **Introduction to reinforcement learning**. Retrieved September 1, 2024, from <https://bdi.or.th/big-data-101/introduction-to-reinforcement-learning/>
- Binariks. (2024). **The Role of Big Data in Personalizing Insurance**. Retrieved November 8, 2024. https://binariks.com/blog/big-data-in-insurance-personalization/?fbclid=IwY2xjawGanHRleHRuA2FlbQlXMAABHdU2ecPwChD6Wka4xMOZYVW8_GpdUENV18cCf8rqCc7dvqwVtmGVex6wPw_aem_sHYGaj8WVrWHM3a6F5IW-w
- Bentejac, C., Csorgo, A., Munoz, M. G. (preprint). **A Comparative Analysis of XGBoost**. <https://doi.org/10.48550/arXiv.1911.01914>
- Borchani, H., Varando, G., Bielza, C., & Larrañaga, P. (2015). A survey on multi-output regression. **WIREs Data Mining and Knowledge Discovery**, 5(5), 216-233
- Chen, Y., Hu, M., Xie, Y., Qiu, R. (2020). Claim frequency predicting based on LightGBM. **Journal of Nonlinear and Analysis**, 22(8), 1757-1770.
- Daroontham, W. (2563). **เจาะลึก Random Forest (Part 2 of รู้จัก Decision Tree, Random Forest และ XGBoost)**. ค้นเมื่อ 5 กรกฎาคม 2567, จาก <https://bit.ly/40hcc5f>
- Daroontham, W. (2563). **รู้จัก Decision Tree, Random Forrest และ XGBoost (Part 1)**. ค้นเมื่อ 9 สิงหาคม 2567, จาก <https://bit.ly/4eTTsgM>
- Eastgate Software. (2024). **What is unsupervised learning?**. Retrieved July 25, 2024, from <https://eastgate-software.com/what-is-unsupervised-learning/>
- FINNOMENA. (2565). **รู้จัก Light Gradient Boosting Machine (LightGBM) โมเดลสุดล้ำสำหรับงานด้านการเงิน**. สืบค้นเมื่อ 27 สิงหาคม, 2567. <https://www.finnomena.com/finnomena-ic/light-gradient-boosting-machine-model/>
- GeeksforGeeks. (2023). **Multioutput regression in machine learning**. Retrieved August 20, 2024, from <https://www.geeksforgeeks.org/multioutput-regression-in-machine-learning/>
- Hmong.in.th. (n.d.). **Symmetric mean absolute percentage error**. Retrieved January 28, 2025, from https://hmong.in.th/wiki/Symmetric_mean_absolute_percentage_error
- Hong, L., Kuffner, T., Martin, R. (2023). **On Prediction of Future Insurance Claims When the Model Is Uncertain**, 12(1), 90-99.
- Inn Why. (2567). **Insure World : จับตา!วินาศภัยไทย เบียดทะลุ3.7แสนล้าน! ประกันทรัพย์สิน/รถ/พีเอส-สุขภาพดันเติบโต**. ค้นเมื่อ 8 พฤศจิกายน 2567. <https://shorturl.at/YLUmX>
- Jin, F. F. (2021). Using decision tree ensemble methods for the estimation of individual claims

- reserving (Master's thesis, Erasmus School of Economics). Erasmus University.
- Johnson, A. (2024). **Car & Moto Insurance In Spain >> Get The Right Cover At The Right Price.** Moving TO SPAIN. Retrieved November 7, 2024. https://movingtospain.com/car-insurance-in-spain/?fbclid=IwY2xjawGanF5leHRuA2FlbQIxMAABHa-v4lZFq4OjvzTr5r0FxeR10J59VwSG1qDdvMfTJJJug284wjqeXOQurA_aem_Q7bB-2P6whnKV_DiVXvNQ
- Krzyk, K. (2023). **Coding deep learning for beginners – Types of machine learning.** Retrieved July 3, 2024, from <https://resources.experfy.com/ai-ml/coding-deep-learning-for-beginners-types-of-machine-learning/>
- Kutner, H. M., Nachtsheim, J. C., Neter, J., Li, W. (2004). **Applied Linear Statistical Models.** 5 th ed. [n.p.]: McGraw-Hill.
- Kumar, V. S., Satpathi, D. K., Kumar, P. T. V. P., Haragopal, V. V. (2020). Modeling and Predicting of Motor Insurance Claim Amount using Artificial Neural Network. **International Journal of Recent Technology and Engineering** ISSN:2277-3878
- Naqa, E. I., Murphy, M. J., & Li, H. (2022). **What Are Machine and Deep Learning.** In Machine and Deep Learning in Oncology, Medical Physics and Radiology (pp. 3-15). https://doi.org/10.1007/978-3-030-83047-2_1
- Pathmind. (n.d.). **Decision tree.** Pathmind Wiki. Retrieved July 27, 2024. <https://wiki.pathmind.com/decision-tree>
- Poufinas, T., Gogas, P., Papadimitriou, T., Zaganidis, E. (2023). Machine Learning in Forecasting Motor Insurance Claims. **Risks** **2023**, **11**(9), 164. <https://doi.org/10.3390/risks11090164>
- Schagen, V. S. (2023). **Multi Target XGBoost Cash Flow Prediction An Efficient Machine Learning Algorithm For Future Liability Projections** (Master's thesis, Master of Science). The Delft University of Technology.
- Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. *IEEE Access*, PP(99), 1–1
- WHO. (2024). Road traffic injuries. Retrieved November 7, 2024. <https://shorturl.asia/H1C0n>
- XGBoost. (n.d.). Multi-output. Retrieved August 27, 2024. <https://xgboost.readthedocs.io/en/stable/tutorials/multioutput.html>

ภาคผนวก

ภาคผนวก ก

ตารางการดำเนินโครงการ

ตารางที่ ก.1 ตารางการดำเนินโครงการ

[illegible]

ภาคผนวก ข

Implementation Code

ตัวอย่างโค้ดที่ใช้ในการพัฒนาแบบจำลองในการทำนายความถี่และความรุนแรงของการเรียกร้องค่าสินไหมทดแทนในการประกันภัยรถยนต์

ข.1 การแบ่งข้อมูลสำหรับพัฒนาแบบจำลอง

```
# สมมุติว่า 'Has_claim' เป็นคอลัมน์ที่บอกว่าการเคลมประกันหรือไม่
X = data_claims_cleaned[['npol_auto', 'client_sex', 'client_age', 'lic_age',
'client_nother', 'cities2', 'north', 'rest']]
y = data_claims_cleaned[['nclaims_md', 'cost_md']]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42,
stratify=data_claims_cleaned['Has_claim'])
# สร้าง DataFrame สำหรับชุดฝึก
train_data = pd.concat([X_train, y_train], axis=1)

# สร้าง DataFrame สำหรับชุดทดสอบ
test_data = pd.concat([X_test, y_test], axis=1)
```

ข.2 ตัวอย่างการฝึกแบบจำลอง Multi-output Random Forest

```
# กำหนดพารามิเตอร์สำหรับ Grid Search
param_grid = {
    "estimator__n_estimators": [1400,1500],
    "estimator__max_depth": [10],
    "estimator__min_samples_split": [130],
    "estimator__min_samples_leaf": [60]
}
kf = KFold(n_splits=10, shuffle=True, random_state=42)

# ฟังก์ชันสำหรับการ Fit GridSearchCV และบันทึกผลลัพธ์
def run_multioutput_grid_search_optimized(X_train, y_train, param_grid):
    xgb = MultiOutputRegressor(RandomForestRegressor(random_state=42))
    grid_search = GridSearchCV(
        estimator=xgb,
```



```

    param_grid=param_grid,
    scoring="neg_mean_absolute_error",
    cv=kf,
    return_train_score=True,
    verbose=1,
    n_jobs=-1,
)

# รัน GridSearch โดยใช้ฟังก์ชัน fit_with_sample_weights
start_time = time.time()
grid_search.fit(X_train, y_train) # ส่ง sample_weight ตรงๆ ไปที่ fit()
total_time = time.time() - start_time

# สรุปผล
results_df = pd.DataFrame(grid_search.cv_results_)
results_df["total_time"] = total_time
results_df["avg_fold_time"] = total_time / kf.get_n_splits()

return results_df, grid_search.best_params_

# รัน Multi-Output GridSearchCV
print("Running Multi-Output GridSearchCV...")
results, best_params = run_multioutput_grid_search_optimized(
    X_train, y_train, param_grid
)
print(f"Best parameters: {best_params}")

# บันทึกเฉพาะพารามิเตอร์ที่ดีที่สุดและค่าเฉลี่ยใน result_summary
result_summary = {
    "Best_Parameters": [str(best_params)], # Best hyperparameters
    "Average_Train_Score": [results.filter(like="mean_train_score").mean(axis=0).iloc[0]], #
    "Average training score"
}

```

```

"Average_Validation_Score":
[results.filter(like="mean_test_score").mean(axis=0).iloc[0]], # Average validation score
    "Average_Fold_Time": [results["avg_fold_time"].mean()] # Average fold time
}

print("Detailed results saved to 'multioutput_results.csv' and 'detailed_results.csv'.")

# ส่วนที่เพิ่ม: คำนวณ RMSE, MAE, และ SMAPE

def calculate_smape(y_true, y_pred):
    return 100 * np.mean(2 * np.abs(y_pred - y_true) / (np.abs(y_true) + np.abs(y_pred)))

# ใช้พารามิเตอร์ที่ดีที่สุดจาก GridSearchCV
cleaned_best_params = {key.replace("estimator__", ""): value for key, value in
best_params.items()}

# ใช้พารามิเตอร์ที่ดีที่สุดจาก GridSearchCV
xgb_best = MultiOutputRegressor(RandomForestRegressor(random_state=42,
**cleaned_best_params))

# ฝึกโมเดลด้วยข้อมูล train
xgb_best.fit(X_train, y_train)

# ทำนายค่าจากข้อมูล test
y_pred = xgb_best.predict(X_test)

# คำนวณค่าตัวชี้วัดแยกตาม y1 และ y2
metrics = {}
for i, target_name in enumerate(y.columns):
    y_true = y_test.iloc[:, i]
    y_pred_target = y_pred[:, i]

```

```

metrics[f"MAE_{target_name}"] = mean_absolute_error(y_true, y_pred_target)
metrics[f"RMSE_{target_name}"] = mean_squared_error(y_true, y_pred_target) ** 0.5
metrics[f"SMAPE_{target_name}"] = calculate_smape(y_true.values, y_pred_target)

# สร้าง DataFrame สำหรับผลลัพธ์ทั้งหมดในแถวเดียว
result_row = {"Best_Parameters": str(best_params)}
result_row.update(metrics)

df_metrics = pd.DataFrame([result_row])

print("Test metrics saved to 'model_test_metrics.csv'.")

y_pred_df = pd.DataFrame(y_pred, columns=["Predicted_y1", "Predicted_y2"],
index=y_test.index)

# สร้างคอลัมน์ใหม่สำหรับค่าที่ปรับตามเงื่อนไข
adjust_predict_y1 = y_pred_df["Predicted_y1"].copy()
adjust_predict_y2 = y_pred_df["Predicted_y2"].copy()

# เงื่อนไข: ถ้า adjust_predict_y1 < 0 หรือ adjust_predict_y2 < 0 ให้ทั้งคู่เป็น 0
adjust_predict_y1 = np.where((adjust_predict_y1 < 0) | (adjust_predict_y2 < 0), 0,
adjust_predict_y1)
adjust_predict_y2 = np.where((adjust_predict_y1 < 0) | (adjust_predict_y2 < 0), 0,
adjust_predict_y2)

# ปัดทศนิยมของ adjust_predict_y1: ถ้ามากกว่า 0.5 ให้ปัดขึ้น ถ้าน้อยกว่า 0.5 ให้ปัดลง
adjust_predict_y1 = np.where(adjust_predict_y1 > 0.5, np.ceil(adjust_predict_y1),
np.floor(adjust_predict_y1))

# เงื่อนไข: ถ้า adjust_predict_y1 == 0 ให้ adjust_predict_y2 = 0
adjust_predict_y2 = np.where(adjust_predict_y1 == 0, 0, adjust_predict_y2)

```

```

# เงื่อนไข: ถ้า adjust_predict_y2 == 0 ให้ adjust_predict_y1 = 0
adjust_predict_y1 = np.where(adjust_predict_y2 == 0, 0, adjust_predict_y1)

# เพิ่มคอลัมน์ใหม่เข้าไปใน DataFrame
y_pred_df["adjust_predict_y1"] = adjust_predict_y1
y_pred_df["adjust_predict_y2"] = adjust_predict_y2
# รวมข้อมูล X_test, y_test และ y_pred
y_test_with_pred = pd.concat([X_test_df, y_test.reset_index(drop=True),
y_pred_df.reset_index(drop=True)], axis=1)
# เปลี่ยนชื่อคอลัมน์
y_test_with_pred.columns = list(X_test_df.columns) + ["nclaims_md", "cost_md",
"Predicted_y1", "Predicted_y2", "adjust_predict_y1", "adjust_predict_y2"]
# บันทึกไฟล์ CSV
y_test_with_pred.to_csv("A_nes_data_best_parameter_ของแบบจำลอง_XGBoost.csv",
index=False)

```

ข.3 Shapley Additive Explanation (SHAP)

```

import os
import matplotlib.pyplot as plt

# สร้างโฟลเดอร์สำหรับบันทึกผล
save_path = '/content/drive/MyDrive/โปรเจกจบ กับอ.จีปไตย/บทที่ 4/SHAP/'
os.makedirs(save_path, exist_ok=True) # สร้างโฟลเดอร์หากยังไม่มี

#save_path = "/home/workspace/chanoknan/claim_predict/"
#os.makedirs(save_path, exist_ok=True) # สร้างโฟลเดอร์หากยังไม่มี

# สร้าง SHAP Explainer
explainer = shap.Explainer(xgb_best.predict, X_train, feature_names=X_train.columns)

# คำนวณ SHAP values

```

```

shap_values = explainer(X_test)

# แสดงค่า SHAP value ของแต่ละตัวแปรแยกตามเป้าหมาย
for i, target_name in enumerate(y.columns): # y.columns คือชื่อ Target เช่น Target1, Target2
    print(f"\nSHAP values summary for {target_name}:")

    # สร้าง DataFrame สำหรับ SHAP values
    shap_df = pd.DataFrame(shap_values[:, i].values, columns=X_train.columns,
index=X_test.index)

    # คำนวณค่าเฉลี่ย (Mean Absolute SHAP Value) ของแต่ละตัวแปร
    mean_shap_values = shap_df.abs().mean().sort_values(ascending=False)
    print(mean_shap_values) # แสดงค่า Mean Absolute SHAP Value ของแต่ละตัวแปร

    # แสดง SHAP summary plot และบันทึกผล
    plt.figure()
    shap.summary_plot(shap_values[:, i], X_test, feature_names=X_train.columns,
show=False) # show=False เพื่อบันทึกก่อนแสดง
    plt.title(f"SHAP Summary Plot for {target_name}")

    # บันทึกผลลงใน Google Drive
    file_path = f"{save_path}shap_summary_{target_name}.png"
    plt.savefig(file_path, dpi=300, bbox_inches='tight') # DPI 300 เพื่อความคมชัด
    print(f"Saved {file_path}")
    plt.close() # ปิดกราฟเพื่อประหยัดหน่วยความจำ

    # บันทึกผลลงใน Google Drive
    file_path = f"{save_path}shap_summary_{target_name}.png"
    plt.savefig(file_path, dpi=300, bbox_inches='tight') # DPI 300 เพื่อความคมชัด
    print(f"Saved {file_path}")
    plt.close() # ปิดกราฟเพื่อประหยัดหน่วยความจำ

```

ข.4 GitHub ที่รวบรวมโค้ดสำหรับ Web Page & Web Application

ภาคผนวก ค

วิธีการใช้งาน Web Application

ภาคผนวก ง

หน้าต่าง Web Page ที่ผู้วิจัยพัฒนาสำหรับงานวิจัย

