การตรวจจับการฉ้อโกงประกันภัยรถยนต์โดยใช้การวิเคราะห์ข้อความและการเรียนรู้ของเครื่อง

ภูริต อำนวยชัย[1*], ศุภร คนธภักดี[2]

บทคัดย่อ

วัตถุประสงค์ของงานวิจัยเพื่อศึกษาวิเคราะห์ข้อมูลจากข้อความร่วมกันกับการใช้คุณลักษณะอื่นๆมาประกอบร่วมกันนำมาประยุกต์ใช้กับเทคนิคการเรียนรู้ของเครื่อง(Machine Learning) เพื่อสร้างแบบจำลองเพื่อทำนายการคาดการความน่าจะเป็นว่าเคลมจะเกิดการทุจริต และเปรียบเทียบประสิทธิภาพของแบบจำลองการแยกประเภท(Classification) ร่วมกับการทดลองกับการจัดการความไม่สมดุลกันของข้อมูล โดยใช้ชุดข้อมูลการเคลมสินไหมรถยนต์ของบริษัทเอเชียประกันภัย1950 จำกัด(มหาชน) ที่เกิดเคลมในช่วง ม.ค. 2563 ถึง ธ.ค. 2563 โดยรวบรวมข้อมูลการทุจริตเคลมในช่วง ม.ค. 2563 ถึง เม.ย. 2564 จำนวนข้อมูลทั้งหมด 58,579 แถว โดยได้ทำการทดลองด้วย 4 วิธีหลักดังนี้ 1. สร้างแบบจำลองทดลองกับข้อมูลที่มีความไม่สมดุล 2. สร้างแบบจำลองทดลองกับข้อมูลที่จัดการกับความไม่สมดุลด้วยวิธี Random Oversampling 3. สร้างแบบจำลองทดลองกับข้อมูลที่จัดการกับความไม่สมดุลด้วยวิธี SMOTE 4. นำแบบจำลองและวิธีการจัดการความไม่สมดุลของข้อมูลที่เลือกมาทำการปรับจูนพารามิเตอร์ ผู้วิจัยได้ทำการทดลองโดยเปรียบเทียบจากค่า Accuracy, Precision, Recall และ F1-Score ในแต่ละวิธีการที่ทำการวิจัย ซึ่งแบบจำลองที่ให้ค่าผลลัพธ์ที่ดีที่สุดคือ Random Forest และวิธีการจัดการกับความไม่สมดุลกันของข้อมูลคือ SMOTE โดยให้ค่า Accuracy=0.99, Precision=0.803, Recall=0.241, F1-Score=0.371 โดยใช้เวลาเทรนแบบจำลองเพียง 12นาที จากการทดลองแบบจำลอง Random Forest ร่วมกับการทำ SMOTE สามารถให้ผลลัพธ์ที่ดีกว่าและใช้เวลาในการเทรนที่ไม่มาก ในแง่ของการใช้คุณลักษณะข้อความกับคุณลักษณะที่ไม่ใช่ข้อความพบว่าแบบจำลองยังให้ความสำคัญกับคุณลักษณะที่ไม่ใช่ข้อความมากกว่า

คำสำคัญ : ทุจริตเคลมรถยนต์, การวิเคราะห์ข้อความ, การเรียนรู้ของเครื่อง, ความไม่สมดุลกันของข้อมูล, เทคนิคป่าแบบสุ่ม

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 091-7895398 E-mail address: phurit.anc@g.swu.ac.th

Motor Insurance Fraud Detection Using Text Analysis and Machine Learning

Phurit Amnuaychai[1*], Subhorn Khonthapagdee [2]

Abstract

The purpose of this research was to analyze the data from the text attributes and categorical attributes, in order to generate a model using machine learning techniques. We use the dataset from motor insurance claims of Asia Insurance Company 1950 (Public) that originated in the period from Jan. 2020 to Dec. 2020 and fraudulent claims data from Jan. 2020 to Apr. 2021, which is a total of 58,579. The machine learning (ML) algorithms such as Naive Bayes classifier, Logistic regression, Random Forest and support vector machine were applied to the dataset. To handle an imbalanced dataset, in this study, we compare two methods: random oversampling and SMOTE. These models were evaluated using Accuracy, Precision, Recall and F1-Score. We found that Random Forest using SMOTE achieved the best result, with the values Accuracy=0.99, Precision=0.803, Recall=0.241, F1-Score=0.371.

Keywords : Motor Claim Fraud, Text Analytics, Machine Learning, Imbalance Data, Random Forest Technique

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 091-7895398 E-mail address: phurit.anc@g.swu.ac.th

บทนำ

ในอุตสาหกรรมประกันภัยรถยนต์มีการแข่งขันสูง เนื่องจากจำนวนรถที่จดทะเบียนและต่อภาษีกับกรมการขนส่งทางบก ทั้งรถเก่าและรถใหม่มีอัตราเพิ่มขึ้นในทุกๆปี โดยจำนวนรถสะสม ทั่วประเทศ ณ วันที่ 30 ก.ย. 2559 ถึงปี 2563 มีอัตราเพิ่มขึ้น ร้อยละ 2.06 หรือประมาณ 1ล้านคัน/ปี [17] ส่งผลให้ปริมาณรถยนต์ที่ใช้งานอยู่บนท้องถนนมีการทำประกันภัยรถยนต์เพิ่มขึ้น [2] จากการที่บริษัทประกันภัยรับความเสี่ยงในการรับประกันภัยไว้เป็นปริมาณมากขึ้นเท่าไหร่ ยิ่งส่งผลให้มีโอกาสเกิดการทุจริต/ฉ้อโกงการเคลมประกันมากขึ้นอย่างหลีกเลี่ยงไม่ได้

การฉ้อโกงและทุจริตการเคลมประกันอาจทำได้ในหลายวิธีการกล่าวคือ อาจมีการทุจริตจากตัวผู้เอาประกันภัยเอง , ผู้เอาประกันภัยร่วมมือกันกับคู่กรณีร่วมกันทำให้เกิดการทุจริต และการทุจริตอีกวิธีหนึ่งคือเจ้าหน้าที่สำรวจภัยร่วมมือกับผู้เอาประกันภัยดำเนินการทุจริตการเคลม สาเหตุทั้งหมดที่กล่าวมาเป็นปัจจัยที่ทำให้บริษัทประกันภัยต้องชดใช้สินไหมเกินความเป็นจริง ทำให้บริษัทประกันภัยต้องสูญเสียจำนวนเงินในการเคลมอย่างมหาศาล ส่งผลกระทบให้ อัตราส่วนค่าสินไหมทดแทน(Loss Ratio) และอัตราส่วนค่าใช้จ่ายในการจัดการค่าสินไหมทดแทน (Loss Adjustment Expense Ratio) เพิ่มขึ้น ซึ่งทำให้มีผลกระทบต่อการคำนวนเบี้ยประกันของแต่ละผลิตภัณฑ์รถยนต์ที่จะขายในอนาคต และอาจทำให้สูญเสียลูกค้าในกลุ่มลูกค้าที่ไม่มีการทุจริตอีกด้วย

ด้วยสาเหตุนี้การตรวจสอบการฉ้อโกง/การทุจริตในการเคลมสินไหมจำเป็นอย่างยิ่ง ในปัจจุบันตรวจสอบโดยเจ้าหน้าที่สินไหม ซึ่งต้องใช้เวลานานและประสบการณ์จากผู้เชี่ยวชาญเป็นอย่างมาก อาจส่งผลให้อาจเกิดการพิจารณาผิดพลาด และส่งผลให้การตรวจสอบการเคลมที่ไม่ทุจริตเกิดการล่าช้า ทำให้ลูกค้าอาจจะร้องเรียนกับทาง คปภ.ได้

จากทั้งหมดที่กล่าวมานี้ นำไปสู่การแก้ไขปัญหาที่เกิดขึ้นโดยอาศัยเทคโนโลยีและวิธีการทางด้านการวิเคราะห์ข้อมูลเข้ามาช่วยสนับสนุน เทคนิค Machine Learning เป็นวิธีการหนึ่งที่จะช่วยให้การตรวจสอบการทุจริต/ฉ้อโกง ให้เป็นไปได้อย่างรวดเร็วและถูกต้องมากกว่าการทำงานโดยใช้มนุษย์เป็นผู้กระทำ

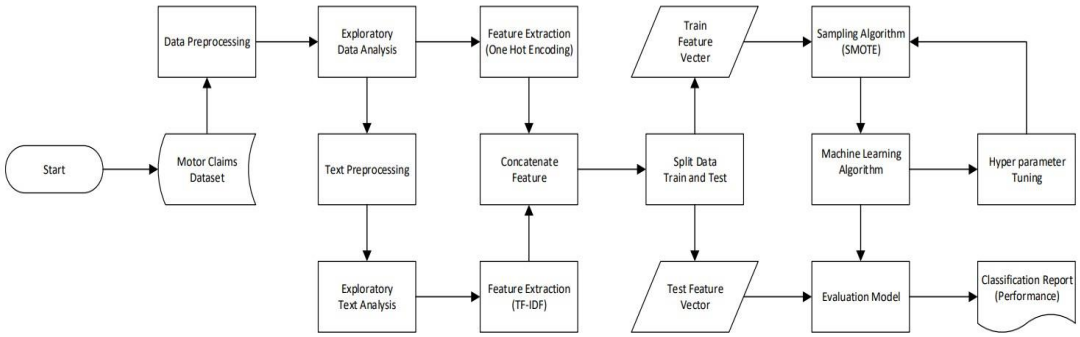
ในงานวิจัยนี้เนื่องจากเป็นข้อมูลที่ไม่สมดุลกันมาก จึงต้องใช้วิธีการสุ่มข้อมูลเกิน (Oversampling) เป็นการเพิ่มจำนวนข้อมูลที่อยู่ในกลุ่มส่วนน้อย หลังจากนั้นจะนำข้อความจากรายงานสำรวจภัยมาวิเคราะห์ มาประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing Techniques) ในการสกัดคุณลักษณะ (Features Extraction) จากข้อความร่วมกับการสร้าง แบบจำลองสำหรับการจำแนกประเภท (Classification Model) โดยเลือกใช้อลักอริทึมการเรียนรู้ของ เครื่อง (Machine Learning Algorithms)

บทความวิจัยเรื่อง Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique [11] ได้สาธิตแนวทางในการสร้างเครื่องตรวจจับการฉ้อโกงประกันภัยรถยนต์ที่ใช้การเรียนรู้ด้วยเครื่อง ซึ่งจะคาดการณ์การเรียกร้องค่าสินไหมทดแทนจากการประกันภัยที่ฉ้อฉลจากชุดข้อมูลบันทึกการเคลมรถยนต์กว่า 15,420 รายการ แบบจำลองที่เสนอสร้างขึ้นโดยใช้เทคนิคการสุ่มตัวอย่างเกินจริงของชนกลุ่มน้อยสังเคราะห์ (SMOTE) ซึ่งขจัดความไม่สมดุลของคลาสของชุดข้อมูล และใช้แบบจำลองการแยกประเภท Random Forest ความแม่นยำของแบบจำลองที่เสนอด้วยค่าเฉลี่ยถ่วงน้ำหนักที่สอดคล้องกัน (Weighted Averages) มีค่า Sensitivity 99.9% โดยใช้เวลา 1.43 วินาทีในการเทรนแบบจำลอง

บทความวิจัยเรื่อง Leveraging deep learning with LDA-base text analytics to detect automobile insurance fraud [15] บทความนี้ขอเสนอรูปแบบการเรียนรู้เชิงลึกแบบใหม่สำหรับรถยนต์การตรวจจับการฉ้อโกงประกันภัยที่ใช้การวิเคราะห์ข้อความโดยใช้ Latent Dirichlet Allocation(LDA) เพื่อแยกคุณสมบัติข้อความที่ซ่อนอยู่ในคำอธิบายข้อความของอุบัติเหตุที่ปรากฏในคำกล่าวอ้างและใช้โครงข่ายประสาทเทียมในการเทรนกับชุดข้อมูลซึ่งรวมถึงข้อความฟีเจอร์และฟีเจอร์ตัวเลขสำหรับตรวจจับการ

อ้างสิทธิ์ที่เป็นการฉ้อโกงผลการทดลองเปิดเผยว่ากรอบการทำงานตามการวิเคราะห์ข้อความที่นำเสนอมีประสิทธิภาพดีกว่าแบบดั้งเดิม นอกจากนี้ผลการทดลองแสดงให้เห็นว่าโครงข่ายประสาทเทียมมีประสิทธิภาพดีกว่าโมเดลแมชชีนเลิร์นนิงที่ใช้กันอย่างแพร่หลายเช่น Random Forest และSupport Vector Machine ดังนั้นโครงข่ายประสาทเทียมและ LDA จึงมีศักยภาพที่เหมาะสมสำหรับการตรวจจับการฉ้อโกงประกันภัยรถยนต์

วิธีดำเนินการ



ภาพประกอบ 1 แสดง Flowchart กระบวนการสร้างแบบจำลอง

ขั้นตอนที่ 1 : การเก็บรวบรวมข้อมูล

ผู้วิจัยนำข้อมูลการเคลมประกันรถยนต์ทั้งที่มีการทุจริตเคลมทั้งหมดในช่วงปี 2562 – เม.ย. 2564 และไม่ทุจริตเคลมในปี 2563 โดยได้รับความอนุเคราะห์ข้อมูลจากบริษัทเอเชียประกันภัย1950 จำกัด(มหาชน) การระบุประเภทว่าเป็นการทุจริตและไม่ทุจริตเคลมนั้นระบุโดยผู้เชี่ยวชาญของฝ่ายสินไหมรถยนต์ ประกอบไปด้วยคุณลักษณะทั้งหมด 17 คุณลักษณะ(16 คุณลักษณะที่ไม่ใช่ข้อความ กับ 1 คุณลักษณะที่เป็นข้อความ) พร้อมกับ 1 คุณลักษณะที่ติดป้ายกำกับว่าเป็นการทุจริตหรือไม่ทุจริตเคลม

ขั้นตอนที่ 2 : การนำเข้าข้อมูล การเตรียมข้อมูล การสำรวจข้อมูล และแปลงคุณลักษณะของข้อมูล ที่ไม่ใช่ข้อความ

จะใช้ภาษาไพทอนในการวิเคราะห์ข้อมูลและการเรียนรู้ของเครื่อง เริ่มจากการนำเข้าชุดข้อมูลที่เป็นไฟน์ csv จากนั้นจะทำการเตรียมข้อมูลในกระบวนการทำ Data Preprocessing [10] เปลี่ยนประเภทของข้อมูล, จัดการกับ Missing value, จัดการกับข้อมูลที่เป็น Outliner และจัดการกับข้อมูลที่เป็น Null จากนั้นจะทำการสำรวจข้อมูลเบื้องต้น เพื่อหาข้อมูลเชิงลึก โดยจะใช้ไลบรารี่ Numpy, Pandas, Matplotlip, Sklearn, Seaborn และทำการเพิ่มคุณลักษณะที่คำนวณเพิ่ม เช่น ระยะเวลาเกิดเคลมหลังจากเริ่มคุ้มครอง หลังจากที่ได้ผ่านกระบวนการทำการเตรียมข้อมูลมาแล้วจะทำการแปลงข้อมูลให้อยู่ในรูปแบบตัวเลขเพื่อให้สามารถสร้างแบบจำลองได้ โดยการทำ One-hot Encoding โดยใช้วิธีการ Dummy Variable Encoding

ขั้นตอนที่ 3 : การเตรียมข้อมูล การสำรวจข้อมูล และสร้างคุณลักษณะของข้อมูล ที่เป็นข้อความ

การเตรียมข้อมูลของข้อความเราจะใช้ไลบรารี่ของ PythaiNLP และNLTK(Natural Language Toolkit) [8] มาช่วยจัดการกับข้อมูลที่เป็นข้อความโดยนำข้อความทั้งหมดมารวมกันเพื่อสร้าง Bag of word(Dictionary) ของทุกๆ รายงานสำรวจภัย (comment) จำนวนทั้งสิ้น 56,495 รายการข้อความ หลังจากนั้นจะทำการทำความสะอาดข้อมูลโดยจะมีขั้นตอนดังนี้ 1.ใช้ Corpus thai_stopwords() ของ PythaiNLP มาทำการเพิ่มเติม stopwords คำบางส่วนเข้าไป 2. ทำการลบข้อมูลตัวอักษรที่เป็นภาษาอังกฤษ , ตัวเลข , อักขระต่างๆ จนเหลือแต่ตัวอักษรที่เป็นคำภาษาไทย 3.ทำการจัดการกับการพิมพ์ข้อความที่เรียงผิดหรือใช้ผิดอักษร(Normalize) 4.ทำการตัดคำด้วย word_tokenize engine = “newmm” 5.ทำ Stopwords Removal เพื่อลดคำที่ไม่ได้ใช้ออกไป 6.ทำการเลือกคำที่ตัดออกมาแล้วมีตัวอักษรที่มากกว่า 2 ตัวอักษรมาดำเนินการ

หลังจากที่ผ่านการทำ Pre-processing ของข้อความทั้งหมดแล้วจะมานับคำทั้งหมดโดยคำที่ผ่านการทำการ Unique ของคำมาแล้วได้จำนวนคำทั้งสิ้น 21,182 คำหลังจากนั้นจะทำการสำรวจคำดูว่าคำไหนที่พบเจอและใช้มากที่สุด รวมถึงทำ Word Cloud เพื่อจับกลุ่มคำโดยเรียงจากคำที่มีมากที่สุดไปน้อยที่สุดเพื่อให้มองเห็นคำที่ถูกใช้มากที่สุดได้ง่าย

จากนั้นเราจะทำการสร้างคุณลักษณะของข้อความที่ตัดคำมาแล้วให้เป็นตัวเลขโดยเราจะใช้วิธีการ TFIDF [14] เพื่อจะเตรียมข้อมูลก่อนจะนำไปเข้าแบบจำลองเพื่อฝึกสอน โดยเราจะกำหนดค่าพารามิเตอร์ TfidfVectorizer เพื่อทดสอบประสิทธิภาพของแบบจำลองโดยเราใช้พารามิเตอร์ use_idf = true , norm = l2 , max_features=1000

หลักการที่ใช้ในการพิจารณาเลือกค่าพารามิเตอร์ max_feature = 1000 นั้นจะพิจารณาจากค่า TFIDF ของคำที่มีความสำคัญที่มีค่ามากที่สุดเป็นจำนวณ 1000 คำนำมาใช้เป็นคุณลักษณะในส่วนของข้อความ โดยที่เกณฑ์(threshold) คือเลือกผลรวมค่า TFIDF ของคำแต่ละคำที่มีค่ามากกว่าเท่ากับ 60

ขั้นตอนที่ 4 : รวมคุณลักษณะที่ไม่ใช่ข้อความ และคุณลักษณะของข้อความเข้าด้วยกัน

หลังจากที่ทำ Pre-processing ทั้งคุณลักษณะที่ไม่ใช่ข้อความ และคุณลักษณะที่เป็นข้อความ และผ่านการทำ One-hot encoding และ word vector เรียบร้อยแล้วเราจะนำคุณลักษณะทั้งหมดมาทำการรวมกันโดยการนำมาต่อกัน (Concatenate)

ขั้นตอนที่ 5 : การแบ่งข้อมูลสำหรับการเทรนและทดสอบ

จากนั้นจะทำการแบ่งข้อมูลเพื่อใช้สำหรับการเทรนแบบจำลองและทดสอบแบบจำลองโดยจะแบ่งข้อมูลเทรนออกเป็น 70% และทดสอบออกเป็น 30%

ขั้นตอนที่ 6 : ทำการ Scale ข้อมูล

เราต้องการทำให้ค่าในแต่ละคุณลักษณะอยู่ใน Scale มาตรฐานเดียวกัน จะส่งผลให้แบบจำลองได้ค่าความแม่นยำที่ดีขึ้น ในงานวิจัยนี้เราใช้สูตร StandardScaler() ของ Scikit-Learn โดยเราจะทำการ Scale ข้อมูลที่เป็นคุณลักษณะที่ไม่ใช่ Class Label โดยข้อมูลคุณลักษณะ(Feature) ทั้งหมดที่นำมาใช้จะเหลือเพียง 1,089 คุณลักษณะโดยจะทำการ Scale ข้อมูลแยกกันระหว่างข้อมูลที่ใช้ฝึกฝนกับข้อมูลที่ใช้ในการทดสอบ

ขั้นตอนที่ 7 : การสร้างแบบจำลองในการทำนาย ร่วมกันกับการทดลองการแก้ปัญหาการไม่สมดุลกันของข้อมูล

การทดลองกับโมเดลแบบจำลองที่เลือกมาทำวิจัย ร่วมกับการทำ 10-Folds Cross Validation [12] โดยแบบจำลองการจำแนกประเภททั้ง 4 แบบจำลองที่ใช้ในงานวิจัยคือ 1. Naïve Bayes 2. Logistic Regression 3. Support Vector Machine 4. Random Forest โดยจะใช้ค่ามาตรฐาน (Default) ของแบบจำลองแต่ละประเภทใช้ในการทดลอง หลังจากนั้นจะนำมาทดสอบกับข้อมูลที่ใช้ทดสอบกับแบบจำลองการจำแนกประเภททั้ง 4 และนำมาเปรียบเทียบประสิทธิภาพของแต่ละวิธีการโดยวิธีการที่

ทดลองคือ 1. ทดลองกับข้อมูลที่ไม่สมดุลกัน(Imbalance Data) 2.ทดลองกับข้อมูลที่แก้ปัญหาความไม่สมดุลกันด้วยวิธี Random Oversampling 3.ทดลองกับข้อมูลที่แก้ปัญหาความไม่สมดุลกันด้วยวิธี SMOTE [9]

ขั้นตอนที่ 8 : การปรับจูนพารามิเตอร์กับแบบจำลองและวิธีการที่เลือก

หลังจากที่ได้ทำการทดลองกับแบบจำลองจำแนกประเภทร่วมกับวิธีการจัดการกับข้อมูลที่ไม่สมดุลกันเป็นที่เรียบร้อยแล้ว ทางผู้วิจัยจะเลือกแบบจำลองที่มีประสิทธิภาพโดยการดูค่า Accuracy, Recall , Precision และF1-Score [6] เพื่อพิจารณาเลือกแบบจำลองนำมาทดลองเพิ่มเติมโดยการปรับจูนพารามิเตอร์ของแบบจำลองที่เลือกมาโดยใช้ GridSearchCV เป็นตัวหาพารามิเตอร์ที่ดีเหมาะสมที่สุดเพื่อนำมาใช้งาน แบบจำลองที่ผู้วิจัยเลือกมาทดลองเพิ่มเติมนั้นคือ Random Forest นำมาทดลองปรับจูนพารามิเตอร์โดยใช้ GridSearchCV กับข้อมูลที่ได้ผ่านการแก้ปัญหาการไม่สมดุลกันของข้อมูลโดยใช้เทคนิค SMOTE พารามิเตอร์ที่ทางผู้วิจัยใช้ในการปรับจูนในครั้งนี้คือ n_estimators และ max_features โดยค่าพารามิเตอร์ที่ปรับจูนโดยใช้ GridSearchCV ในการหาพารามิเตอร์ที่ดีที่สุดผู้วิจัยจะใช้ค่าดังนี้ 'n_estimators': [50, 100, 200, 300, 400, 500, 600, 700, 800] และ 'max_features': ['auto', 'sqrt', 'log2'] โดยจะใช้รวมกับการทำ 10-Fold Cross Validation

ผลการวิจัยและอภิปรายผลการวิจัย

การทดลองโดยการศึกษาตามขบวนการและขั้นตอนตลอดจนการวัดประสิทธิภาพ เพื่อให้บรรลุจุดประสงค์ของการวิจัยที่ได้กำหนดไว้ ประกอบด้วย 3 วิธีดังนี้

วิธีที่ 1 ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ไม่สมดุลกัน

ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลมีความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับไม่เป็นการทุจริตเคลม จากผลลัพธ์จะเห็นได้ว่าภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy=0.991, Precision=1.0 , Recall=0.212, F1-Score=0.35 โดย Naïve Bayes ให้ค่า Recall=0.833, Accuracy=0.401, Precision=0.017, F1-Score=0.033 โดยที่Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น จากภาพรวมของการวัดประสิทธิภาพโดยพิจารณาจากค่า F1-Score(weighted average ระหว่าง Precision และ Recall) ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 7:11 นาที ดังตารางที่ 1

ตารางที่ 1 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ไม่สมดุลกัน

Imbalance Data				
Model	NB	LR	RF	SVM
Accuracy	0.401	0.986	0.991	0.984
Precision	0.017	0.367	1	0.304

Recall	0.833	0.291	0.212	0.321
F1-Score	0.033	0.325	0.35	0.312
Train Duration	0:00:38	0:01:04	0:07:11	2:59:56

วิธีที่ 2 ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ทำ Random Oversampling

ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้ถูกแก้ไขความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับไม่เป็นการทุจริตเคลม โดยใช้วิธีการ Random Oversampling จากผลลัพธ์จะเห็นได้ว่าภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy=0.991, Precision=0.977, Recall=0.207, F1-Score=0.342 โดย Naïve Bayes ให้ค่า Recall=0.823, Accuracy=0.391, Precision=0.016, F1-Score=0.032 โดยที่Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น จากภาพรวมของการวัดประสิทธิภาพโดยพิจารณาจากค่า F1-Score(weighted average ระหว่าง Precision และ Recall) ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 21:01 นาที ดังตารางที่ 2

ตารางที่ 2 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ Random Oversampling

Random Oversampling				
Model	NB	LR	RF	SVM
Accuracy	0.391	0.957	0.991	0.964
Precision	0.016	0.138	0.977	0.159
Recall	0.823	0.493	0.207	0.468
F1-Score	0.032	0.216	0.342	0.237
Train Duration	0:01:18	0:03:18	0:21:01	5 days, 12:04:14

วิธีที่ 3 ผลลัพธ์ของการจำแนกประเภทกับข้อมูลที่ทำ SMOTE

ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้ถูกแก้ไขความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับไม่เป็นการทุจริตเคลม โดยใช้วิธีการ SMOTE จากผลลัพธ์จะเห็นได้ว่าภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy=0.99, Precision=0.803, Recall=0.241, F1-Score=0.371371 โดย Naïve Bayes ให้ค่า Recall=0.7, Accuracy=0.652, Precision=0.024, F1-Score=0.0046 โดยที่Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น จากภาพรวมของการวัดประสิทธิภาพโดยพิจารณาจากค่า F1-

Score(weighted average ระหว่าง Precision และ Recall) ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 12:00 นาที ดังตารางที่ 3

ตารางที่ 3 ผลลัพธ์ที่ได้จากการทดลองแบบจำลองการแยกประเภทร่วมกับข้อมูลที่ทำ SMOTE

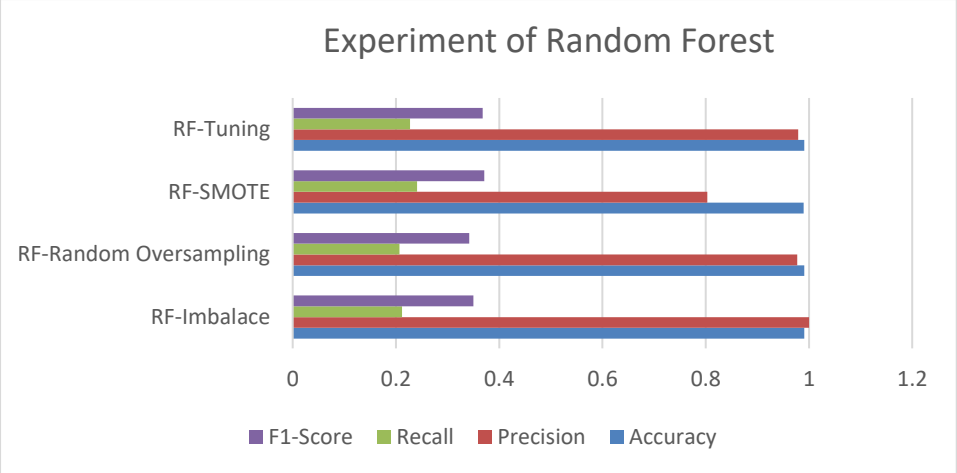
SMOTE				
Model	NB	LR	RF	SVM
Accuracy	0.652	0.977	0.99	0.977
Precision	0.024	0.211	0.803	0.207
Recall	0.7	0.36	0.241	0.335
F1-Score	0.046	0.266	0.371	0.256
Train Duration	0:02:54	0:04:34	0:12:00	1 day, 11:12:16

การเปรียบเทียบผลของการทดลองของแบบจำลองที่เลือก Random Forest

จากการทดลองทั้งหมดผู้วิจัยนำแบบจำลองที่เราเลือกมาทดลองคือ Random Forest ซึ่งผู้วิจัยได้นำการทดลองแต่ละวิธีการมาเปรียบเทียบเพื่อดูประสิทธิภาพของแบบจำลอง Random Forest เพื่อให้เห็นว่าวิธีการใดเหมาะสม โดยวิธีการแก้ปัญหาความไม่สมดุลกันของข้อมูลด้วยวิธีการ SMOTE ให้ค่า Precision น้อยกว่าวิธีการอื่นแต่ค่า Recall และ F1-Score จะให้ผลลัพธ์ที่ดีกว่าเมื่อเทียบกับวิธีการอื่น ดังนั้นจึงเลือกแบบจำลอง Random Forest กับวิธีการจัดการความไม่สมดุลของข้อมูล SMOTE มาทำการปรับพารามิเตอร์เพิ่ม โดยมีรายละเอียดดังตารางเปรียบเทียบตาม ตารางที่ 4

ตารางที่ 4 เปรียบเทียบ Random Forest กับวิธีการทดลองแบบต่างๆ

	RF-Imbalance	RF-Random Oversampling	RF-SMOTE	RF-Tuning
Accuracy	0.991	0.991	0.99	0.991
Precision	1	0.977	0.803	0.979
Recall	0.212	0.207	0.241	0.227
F1-Score	0.35	0.342	0.371	0.368
Train Duration	0:07:11	0:21:01	0:12:00	0:23:37



ภาพประกอบ 2 ผลลัพธ์ที่ได้จากการทดลองแบบจำลอง Random Forest ในแต่ละวิธีการทดลอง

การทดลองปรับจูนพารามิเตอร์ของแบบจำลองและวิธีการที่เลือกและผลลัพธ์

การทดลองในการปรับจูนพารามิเตอร์โดยใช้ GridSearchCV และ 10 Fold Cross-Validation ในการเลือกพารามิเตอร์ที่ดีที่สุดของแบบจำลอง Random Forest ที่ทดลองกับข้อมูลที่ผ่านการทำ SMOTE ผลลัพธ์ที่ได้ค่าของพารามิเตอร์ที่เลือกออกมา คือ 'max_features': 'log2', 'n_estimators': 600 โดยใช้เวลาในการทำ GridSearchCV ทั้งสิ้น 4 ชั่วโมง 47 นาทีในการหาพารามิเตอร์ที่ดีที่สุด และเมื่อนำค่าพารามิเตอร์ที่ได้มาทดสอบเพื่อดูความแม่นยำและระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองจะได้ผลลัพธ์ดังนี้ ค่า Accuracy=0.991, Precision=0.979, Recall=0.227, F1-Score=0.368 โดยใช้เวลาการฝึกฝนแบบจำลอง 23:37 นาที เมื่อเปรียบเทียบกับวิธีการที่ใช้พารามิเตอร์มาตรฐาน (Default) ของแบบจำลองแล้ว พารามิเตอร์ที่เป็นมาตรฐานสามารถให้ค่า Recall และF1-Score ที่ดีกว่า ดังตารางที่ 5

ตารางที่ 5 ผลลัพธ์ของการทดลองปรับจูนพารามิเตอร์ของแบบจำลองที่เลือก

	RF+SMOTE (Tuning)	RF+SMOTE (Default)
Accuracy	0.991	0.99
Precision	0.979	0.803
Recall	0.227	0.241
F1-Score	0.368	0.371
Train Duration	0:23:37	0:12:00

สรุปผลการวิจัย

ในการวิจัยครั้งนี้เป็นการวิจัยเพื่อศึกษากระบวนการวิเคราะห์ข้อความร่วมกันกับคุณลักษณะที่ไม่ใช่ข้อความเพื่อนำไปสู่การตรวจจับการทุจริตเคลม โดยใช้เทคนิคการเรียนรู้ของเครื่อง(Machine Learning) นำมาสร้างแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในแต่ละการทดลอง หลังจากนั้นได้เลือกแบบจำลองที่มีประสิทธิภาพที่สุดกับชุดข้อมูลในงานวิจัยนี้นำมาปรับจูนพารามิเตอร์เพื่อทำการทดลองเพื่อได้ผลลัพธ์ออกมา และทำการทดสอบข้อมูลใหม่กับแบบจำลองที่เราคิดว่าให้ประสิทธิภาพที่ดี

เนื่องจากการวิจัยนี้ผู้วิจัยได้มุ่งเน้นไปที่การนำคุณลักษณะที่ไม่ใช่ข้อความ มาใช้ร่วมกันกับคุณลักษณะที่เป็นข้อความ ในขั้นตอน Preprocess ทั้งคุณลักษณะที่ไม่ใช่ข้อความและคุณลักษณะที่เป็นข้อความค่อนข้างมีความสำคัญเป็นอย่างมาก ถ้าเราทำความสะอาดข้อมูลรวมถึงการตัดคำของคุณลักษณะของข้อความออกมาไม่ดีส่งผลให้จะไม่มีคุณค่าเท่าที่ควรในส่วนของคุณลักษณะที่ไม่ใช่ข้อความเราสามารถจัดกลุ่มรวมประเภทของคุณลักษณะเดียวกันให้ลดน้อยลงได้เพื่อเพิ่มความสำคัญของประเภทย่อยที่อยู่ในคุณลักษระเดียวกันมากขึ้น และทำให้คุณลักษณะที่นำมาใช้ลดน้อยลงไปด้วย ทำให้การฝึกฝนแบบจำลองทำได้เร็วมากขึ้น ในส่วนของคุณลักษณะที่เป็นข้อความการทำ Feature Selection และ Feature Importance ค่อนข้างมีความสำคัญต่อการเลือกคำมาใช้ เนื่องจากถ้าเราไม่เลือกคำที่มีความสำคัญมากๆมาใช้คุณลักษณะข้อความจะมีจำนวนมากส่งผลให้การฝึกฝนแบบจำลองใช้เวลาค่อนข้างมากและใช้ทรัพยากรหน่วยความจำ(Memory) มากทำให้ไม่สามารถฝึกฝนแบบจำลองได้สำเร็จเพราะเกิดปัญหาหน่วยความจำไม่เพียงพอ ในงานวิจัยนี้จึงใช้เวลาในขั้นตอน Preprocess ค่อนข้างมาก ผู้วิจัยได้ทำการทดลองตรวจสอบคุณลักษณะที่สำคัญ(Feature Importance) ของคุณลักษณะทั้งหมดร่วมกับแบบจำลอง Random Forest ใน 100 อันดับแรกผลปรากฏว่าคุณลักษณะที่มีความสำคัญที่แบบจำลองเลือกใช้ส่วนใหญ่เป็นคุณลักษณะที่ไม่ใช่ข้อความเช่น ‘เวลาที่เกิดเหตุ(ชั่วโมง)’,’อายุรถ’,’วันที่เกิดเหตุหลังจากวันที่กรมธรรม์เริ่มคุ้มครอง’,’สาเหตุของการเกิดอุบัติเหตุ(เช่น รถประกันเสียหลัก,เฉี่ยวชนคู่กรณี เป็นต้น)’,’ลักษณะของการใช้รถยนต์(เช่น ส่วนบุคคล,เพื่อการพาณิชย์,ใช้รับจ้างสาธารณะ เป็นต้น)’,’ประเภทของตัวรถยนต์(เช่น เก๋ง2ตอน,รถแท็กซี่,กระบะแวน เป็นต้น)’,’ผลคดี(เช่นฝ่ายถูก,ฝ่ายผิด,ประมาทร่วม เป็นต้น)’ ส่วนคุณลักษณะที่เป็นข้อความแบบจำลองที่ได้เลือกมาใช้เป็นคำที่ไม่ได้สื่อความหมายไปในทางที่จะสื่อว่าเป็นการทุจริตเคลม เช่น ‘นุ่น’,’นุ่ง’,’นุ่ม’,’นุด’,’นี้’ แต่มีความสำคัญที่น้อยกว่าคุณลักษณะที่ไม่ใช่ข้อความ

การทดลองส่วนที่เกี่ยวข้องกับความไม่สมดุลกันของข้อมูล ผู้วิจัยได้ลองทดสอบแบบจำลองแต่ละแบบจำลองกับข้อมูลทั้งที่ไม่สมมดุลกัน และข้อมูลที่ได้จัดการให้สมมดุลกัน พบว่าประสิทธิภาพของแบบจำลองที่ได้ทดลองกับข้อมูลที่ผ่านการทำ Oversampling ในวิธีการ SMOTE นั้นมีประสิทธิภาพที่ดีกว่าวิธีการ Random Oversampling สำหรับในชุดข้อมูลที่เราใช้ในงานวิจัยในครั้งนี้

ในการทดลองโดยสร้างแบบจำลองการจำแนกประเภทโดยใช้ข้อมูลที่ได้ถูกแก้ไขความไม่สมดุลกันของคลาสที่เป็นการทุจริตเคลมกับไม่เป็นการทุจริตเคลม โดยใช้วิธีการ SMOTE จากผลลัพธ์จะเห็นได้ว่าภาพรวมของการวัดประสิทธิภาพของแบบจำลอง Random Forest มีประสิทธิภาพและความแม่นยำในการทำนายได้ดีกว่าแบบจำลองประเภทอื่นๆ โดยที่ค่า Accuracy=0.99, Precision=0.803, Recall=0.241, F1-Score=0.371 โดย Naïve Bayes ให้ค่า Recall=0.7, Accuracy=0.652, Precision=0.024, F1-Score=0.0046 โดยที่Random Forest มีประสิทธิภาพที่ดีกว่าแบบจำลองการแยกประเภทชนิดอื่น จากภาพรวมของการวัดประสิทธิภาพโดยพิจารณาจากค่า F1-Score(weighted average ระหว่าง Precision และ Recall) ระยะเวลาที่ใช้ในการฝึกฝนแบบจำลองของ Random Forest ก็ไม่ได้ใช้เวลามากเกินไปซึ่งใช้เวลาเพียง 12:00 นาที

ปัญหาที่ทางผู้วิจัยพบกับข้อมูลในชุดนี้คือปัญหา Overfitting ซึ่งแบบจำลองที่นำมาใช้ในการทดลองกับการจัดการข้อมูลทั้ง 3 วิธีการก็ยังให้ผลลัพธ์ที่แบบจำลอง Overfitting กับข้อมูลในคลาสที่มีข้อมูลส่วนมาก(คลาสที่ไม่ทุจริตเคลม) อยู่ดีโดยผู้วิจัยได้

พิจารณาการวัดประสิทธิภาพจากค่า Accuracy, Precision, Recall, F1-Score ร่วมกับ Confusion Matrix โดยค่า Accuracy ในแต่ละแบบจำลองให้ผลลัพธ์ที่ดีมาก แต่ค่า F1-Score ให้ผลลัพธ์ที่ไม่ค่อยดี เมื่อได้มาดูค่า Confusion Matrix แล้วพบว่าค่า TN(True Negative) ออกมาค่อนข้างแม่นยำ แต่ค่า TP(True Positive) ออกมาค่อนข้างไม่แม่นยำเท่าที่ควร แบบจำลองเลยเอนเอียงไปในทางทำนายคลาสที่ไม่ทุจริตเคลม(คลาส 0) ได้ดีกว่าคลาสที่ทุจริตเคลม(คลาส 1) ที่เราสนใจ

งานวิจัยในอนาคต จากปัญหาที่เราพบในการดำเนินการวิจัยในครั้งนี้ทั้งหมด ผู้วิจัยได้เล็งเห็นว่าในขั้นตอนกระบวนการทำ Preprocess และ Text Preprocess มีความสำคัญมาก ในงานวิจัยนี้สามารถที่จะจัดการกับข้อมูลในขั้นตอนนี้เพิ่มเติมได้อีกและเพื่อเพิ่มประสิทธิภาพของคุณลักษณะที่เป็นข้อความผู้วิจัยคิดว่าจำเป็นต้องสร้าง Dictionary คำที่มีความทุจริตเพื่อนำมาใช้ในการให้น้ำหนักของคำในข้อความด้วย ในส่วนของการทำ Feature Importance สามารถช่วยให้เราเลือกคุณลักษณะที่มีความสำคัญในงานวิจัยออกมาได้อย่างดีซึ่งอาจจะพิจารณาทำเพิ่มเติมกับคุณลักษณะที่ไม่ใช่ข้อความ ในส่วนของการลดขนาดของ Feature ลงทางผู้วิจัยได้พิจารณาแล้วว่างานวิจัยสามารถต่อยอดนำเทคนิคการจัดกลุ่มของประโยคและกลุ่มของคำโดยใช้วิธีการ LDA(Latent Dirichlet Allocation) ทดลองเพิ่มเติมก่อนที่จะลองทดสอบกับแบบจำลองทั้ง 4 แบบจำลองที่เราได้ใช้ในการดำเนินการวิจัยในครั้งนี้ ในอีกวิธีการที่ผู้วิจัยคาดว่าจะนำมาทดลองกับชุดข้อมูลนี้คือการทำ Under sampling กับการจัดการข้อมูลที่ไม่สมดุลกัน และนำวิธีการ Deep Learning เข้ามาร่วมในการทดลองกับการทำ LDA เพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพที่มากขึ้นและเพื่อทดสอบการแก้ปัญหาการ Overfitting กับข้อมูลชุดนี้ในอนาคตต่อไป

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับความอนุเคราะห์ข้อมูลที่ใช้ในการดำเนินการวิจัยจากบริษัทเอเชียประกันภัย1950 จำกัด(มหาชน) และได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] (คปภ.), ส. (2020). กรอบการลงทุนตามความเสี่ยง(Risk-Base Capital Framework). สืบค้นจาก https://www.oic.or.th/sites/default/files/3029-9267-2.pdf

[2] (คปภ.), ส. (2563). รายงานภาวะธุรกิจประกันภัยไทย ประจำปี 2563. สืบค้นจาก https://www.oic.or.th/th/industry/statistic/data/39/2

[3] Aninditya, A., Hasibuan, M. A., และ Sutoyo, E. (2019, 5-7 Nov. 2019). Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy. Paper presented at the 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS).

[4] Arreerard, R., และ Senivongse, T. (2018, 12-13 July 2018). Thai Defamatory Text Classification on Social Media. Paper presented at the 2018 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD).

[5] Brownlee, J. (2014). An Introduction to Feature Selection. Retrieved from https://machinelearningmastery.com/an-introduction-to-feature-selection/

[6] chengz. (2019). วัดประสิทธิภาพ Model จาก Confusion Matrix. สืบค้นจาก https://medium.com/@cheng3374/วัดประสิทธิภาพ-model-จาก-confusion-matrix-69d391bcd48

[7] Das, A. (2019). Oversampling to remove class imbalance using SMOTE. Transportation Research Part C: Emerging Technologies. Retrieved from https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35

[8] Ganesan, K. (2019). All you need to know about text preprocessing for NLP and Machine Learning. Retrieved from https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

[9] Goswami, D. S. (2020). Class Imbalance, SMOTE, borderline SMOTE, ADASYN. Retrieved from https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804

[10] Goyal, K. (2021). Data Preprocessing in Machine Learning: 7 Easy Steps To Follow. Retrieve from https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/

[11] Harjai, S., Khatri, S. K., และ Singh, G. (2019, 21-22 Nov. 2019). Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique. Paper presented at the 2019 4th International Conference on Information Systems and Computer Networks (ISCON).

[12] Kumar, A. (2020). K-Fold Cross Validation – Python Example. Retrieved from https://vitalflux.com/k-fold-cross-validation-python-example/

[13] Prasasti, I. M. N., Dhini, A., และ Laoh, E. (2020, 17-18 Oct. 2020). Automobile Insurance Fraud Detection using Supervised Classifiers. Paper presented at the 2020 International Workshop on Big Data and Information Security (IWBIS).

[14] Prasertsom, P. (2020). สกัดใจความสำคัญของข้อความด้วยเทคนิคการประมวลผลทางภาษาเบื้องต้น: TF-IDF. สืบค้นจาก https://bigdata.go.th/big-data-101/tf-idf-1/

[15] Wang, Y., และ Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems, 105, 87-95.

[16] Yuenyong, S., และ Sinthupinyo, S. (2020). Gender classification of thai facebook usernames. International Journal of Machine Learning and Computing, 10(5).

[17] กองแผนงานกรมการขนส่งทางบก, ก. (2559 - 2563). รายงานสถิติการขนส่ง ปีงบประมาณ 2559 – 2563. สืบค้นจาก https://web.dlt.go.th/statistics/