

การพยากรณ์ราคารถยนต์มือสองโดยใช้เทคนิคการสกัดข้อมูลจากเว็บไซต์ และแบบจำลองการเรียนรู้ของเครื่อง

ภูริทรัพย์ เดชพิพัฒนประชา¹ วรรัตน์ จงไกรจักร^{2*} อัครชัย อินทนิล³ วิมลณัฐ สุขพล⁴

^{1,3,4} คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏภูเก็ต, ภูเก็ต, ประเทศไทย

^{2*} คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏภูเก็ต, ภูเก็ต, ประเทศไทย

*อีเมลผู้ประพันธ์บทความ: worrarat.j@pkru.ac.th

รับต้นฉบับ: 7 สิงหาคม 2568; รับบทความฉบับแก้ไข: 8 ตุลาคม 2568; ตอรับบทความ: 5 กุมภาพันธ์ 2569

เผยแพร่ออนไลน์: 27 เมษายน 2569

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการพยากรณ์ราคารถยนต์มือสองในประเทศไทย โดยใช้เทคนิคการสกัดข้อมูลจากเว็บไซต์และแบบจำลองการเรียนรู้ของเครื่อง ข้อมูลถูกรวบรวมจากเว็บไซต์ Kaidee.com, One2Car.com และ Chobrod.com จำนวน 55,989 รายการ หลังการทำความสะอาดเหลือ 42,823 รายการ ซึ่งได้คุณลักษณะพื้นฐาน 13 รายการสำหรับสร้างแบบจำลอง ผู้วิจัยได้ดำเนินการทดลองสองลักษณะ ได้แก่ 1) การใช้เฉพาะคุณลักษณะพื้นฐาน และ 2) การเพิ่มคุณลักษณะใหม่ 6 รายการ ได้แก่ อายุรถยนต์ อัตราการใช้งานต่อปี กำลังสองของระยะทางสะสม กำลังสองของขนาดเครื่องยนต์ ภาระสะสมการใช้งาน และภาระเชิงเวลา เพื่อเพิ่มศักยภาพของแบบจำลองในการเรียนรู้เชิงลึก โดยเปรียบเทียบประสิทธิภาพของแบบจำลอง XGBoost, Random Forest, LightGBM, CatBoost และ Gradient Boosting ด้วยตัวชี้วัด MAE, RMSE, MAPE, R² และ Accuracy ผลการทดลองพบว่า XGBoost ให้ผลดีที่สุด โดยเมื่อเพิ่มคุณลักษณะใหม่ ค่า R² เพิ่มขึ้นจาก 0.9262 เป็น 0.9419 และ Accuracy จากร้อยละ 89.42 เป็นร้อยละ 92.38 พร้อมลดค่า MAPE จากร้อยละ 10.58 เหลือร้อยละ 7.62 แสดงว่าการสร้างคุณลักษณะใหม่ช่วยเพิ่มความแม่นยำของแบบจำลองอย่างมีนัยสำคัญ การวิเคราะห์ความสำคัญของคุณลักษณะระบุว่าปัจจัยหลักคือ ประเภทเชื้อเพลิง ประเภทรถ ขนาดเครื่องยนต์ ยี่ห้อ และค่ากำลังสองของขนาดเครื่องยนต์ ผลการวิจัยยืนยันว่าการผสมผสานเทคนิคการเรียนรู้ของเครื่องร่วมกับการสร้างคุณลักษณะใหม่ช่วยเพิ่มประสิทธิภาพการพยากรณ์ และสนับสนุนการตัดสินใจอย่างโปร่งใสในตลาดรถยนต์มือสองของประเทศไทย

คำสำคัญ: การวิเคราะห์ข้อมูล การเรียนรู้ของเครื่อง การพยากรณ์ราคา รถยนต์มือสอง การสกัดข้อมูล



Used Car Price Prediction Using Web Scraping and Machine Learning Models

Bhurisub Dejjipatpracha¹ Worrarat Jongkrajak^{2*} Akarachai Inthanil³ Wimonnat Sukpol⁴

^{1,3,4}*Faculty of Science and Technology, Phuket Rajabhat University, Phuket, Thailand*

^{2*}*Faculty of Management Sciences, Phuket Rajabhat University, Phuket, Thailand*

*Corresponding author e-mail: worrarat.j@pkru.ac.th

Received: 7 August 2025; Revised: 8 October 2025; Accepted: 5 February 2026

Published online: 27 April 2026

Abstract

This study aimed to develop a predictive model for used car prices in Thailand using web scraping techniques and machine learning algorithms. Data were collected from Kaidee.com, One2Car.com, and Chobrod.com, totaling 55,989 records. After data cleaning, 42,823 valid records remained, containing 13 fundamental attributes for model construction. Two experiments were conducted: (1) using only the basic features and (2) incorporating six newly engineered features—car age, annual usage rate, squared mileage, squared engine size, cumulative usage load, and temporal load—to enhance the model's learning capability. The performance of five models, including XGBoost, Random Forest, LightGBM, CatBoost, and Gradient Boosting, was compared using MAE, RMSE, MAPE, R^2 , and accuracy. The results showed that XGBoost achieved the best prediction performance. With the additional features, the R^2 value improved from 0.9262 to 0.9419, accuracy increased from 89.42% to 92.38%, and MAPE decreased from 10.58% to 7.62%, indicating that feature engineering significantly enhanced model accuracy. Feature importance analysis revealed that the most influential factors affecting used car prices were fuel type, car type, engine size, brand, and squared engine size. The findings confirm that integrating machine learning with feature engineering substantially improves predictive performance and can serve as a decision-support tool for buyers, sellers, and financial institutions to promote transparency and fairness in Thailand's used car market.

Keywords: Data analysis, Machine learning, Price prediction, Used car, Web scraping

บทนำ

ปัจจุบันตลาดรถยนต์มือสองมีบทบาทสำคัญต่ออุตสาหกรรมยานยนต์ทั่วโลก รวมถึงในประเทศไทย ซึ่งรถยนต์มือสองเป็นทางเลือกที่ได้รับความนิยมจากผู้บริโภคที่ต้องการเป็นเจ้าของรถยนต์ในราคาที่ประหยัดกว่ารถยนต์ใหม่ รถยนต์มือสองสามารถตอบสนองความต้องการของผู้ใช้ ครอบคลุมผู้ใช้ทั่วไปและผู้ที่ต้องการรถยนต์สำหรับใช้งานเฉพาะทาง ในราคาที่เข้าถึงได้ ส่งผลให้ธุรกิจซื้อขายรถยนต์มือสองมีการเติบโตอย่างรวดเร็ว ทั้งในรูปแบบของโชว์รูม การซื้อขายแบบดั้งเดิม และแพลตฟอร์มการตลาดออนไลน์ที่เพิ่มความสะดวกให้แก่ผู้ซื้อและผู้ขายให้สามารถเข้าถึงการทำธุรกรรมได้อย่างรวดเร็ว อย่างไรก็ตาม การกำหนดราคาของรถยนต์มือสองในตลาดยังคงขาดมาตรฐานราคากลาง โดยราคาขึ้นอยู่กับหลายปัจจัย เช่น สภาพรถยนต์ ปีที่ผลิต เลขไมล์ ยี่ห้อรุ่นของรถยนต์ ประวัติการใช้งานและกลยุทธ์ทางการตลาด [1], [2] อีกทั้งราคาที่ตั้งขึ้นโดยร้านค้าหรือผู้ขายรายย่อยนั้น ไม่ได้เป็นการการันตีราคาที่แท้จริงของรถยนต์ ส่งผลให้ผู้ซื้อที่มีความเสี่ยงที่จะซื้อในราคาเกินราคาปกติ หรือซื้อรถที่มีคุณภาพต่ำกว่าที่ควรจะเป็น แม้ว่าผู้ซื้อจะสามารถเปรียบเทียบราคาได้จากหลากหลายแหล่งข้อมูล เช่น เว็บไซต์ซื้อขายรถยนต์มือสอง หรือแพลตฟอร์มออนไลน์ แต่ต้องใช้เวลาและความชำนาญ อีกทั้งยังขาดเครื่องมือที่สามารถช่วยพยากรณ์ราคาได้อย่างเป็นระบบและแม่นยำ [3], [4]

การพยากรณ์ราคามีหลายวิธี แต่วิธีที่ได้รับความนิยมคือ การเรียนรู้ของเครื่อง (machine learning) ที่ใช้การสร้างแบบจำลองพยากรณ์ราคา โดยใช้ชุดข้อมูลเพื่อเป็นข้อมูลสำหรับนำเข้าแบบจำลอง [5] ดังนั้นงานวิจัยนี้จึงมีเป้าหมายในการรวบรวมข้อมูลรถยนต์มือสองเพื่อนำมาเป็นข้อมูลสำหรับการตัดสินใจของแบบจำลองและเปรียบเทียบแบบจำลอง เพื่อหาแบบจำลองที่เหมาะสมกับชุดข้อมูลนี้ โดยเก็บรวบรวมข้อมูลด้วยการสกัดข้อมูลจากเว็บไซต์ (web scraping) โดยเลือกเว็บไซต์ที่ได้รับความนิยมในการซื้อขายรถยนต์มือสองในประเทศไทย ได้แก่ Kaidee.com, One2Car.com และ Chobrod.com ที่สะท้อนราคาจริงในตลาดประเทศไทย หลังจากเปรียบเทียบและได้แบบจำลองที่เหมาะสมที่สุดแล้ว การใช้แบบจำลองจะช่วยให้ผู้ซื้อสามารถประเมินราคาที่เหมาะสมของรถยนต์แต่ละคันได้อย่างสะดวกและรวดเร็ว ลดความเสี่ยงจากการซื้อรถในราคาที่ไม่สมเหตุสมผลและส่งเสริมให้ตลาดรถยนต์มือสองมีความโปร่งใส ซึ่งงานวิจัยนี้มีวัตถุประสงค์คือ รวบรวมข้อมูลรถยนต์มือสองจากเว็บไซต์ประกาศขายออนไลน์โดยใช้เทคนิคการสกัดข้อมูลจากเว็บไซต์ และเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องที่เหมาะสมกับการพยากรณ์ราคาของรถยนต์มือสอง ทั้งนี้ผลลัพธ์จากงานวิจัยยังสามารถนำไปประยุกต์ใช้ในเชิงพาณิชย์ โดยเป็นเครื่องมือสนับสนุนการตัดสินใจในการซื้อขายรถยนต์มือสองทั้งในระดับผู้บริโภคทั่วไปและธุรกิจที่เกี่ยวข้อง

ทบทวนวรรณกรรม

การพยากรณ์ราคาของรถยนต์มือสองถือเป็นปัญหาวิจัยที่มีความสำคัญต่อผู้ที่ต้องการซื้อรถยนต์ เนื่องจากชุดข้อมูลมีลักษณะซับซ้อนและหลากหลาย ซึ่งเป็นปัจจัยที่มีผลต่อการกำหนดราคาในตลาด เช่น สภาพทางกายภาพของรถยนต์มือสอง พฤติกรรมและรูปแบบการใช้งาน รวมถึงความผันผวนและแนวโน้มของกลไกตลาด ทั้งนี้ผู้วิจัยได้รวบรวมงานวิจัยที่เกี่ยวข้องสามารถจัดแบ่งได้เป็น 4 กลุ่ม ประกอบด้วย 1) การสกัดข้อมูลจากเว็บไซต์ 2) แบบจำลองการเรียนรู้ของเครื่อง 3) การประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง และ 4) งานวิจัยที่เกี่ยวข้องกับการพยากรณ์ราคาของรถยนต์มือสอง โดยมีรายละเอียดดังต่อไปนี้

1. การสกัดข้อมูลจากเว็บไซต์

ในปัจจุบันข้อมูลที่ปรากฏบนเว็บไซต์นับเป็นแหล่งข้อมูลที่สามารถนำไปใช้ในการวิเคราะห์ วิจัย และพัฒนาระบบอัจฉริยะต่าง ๆ ได้ อย่างไรก็ตาม เว็บไซต์ส่วนใหญ่มักแสดงข้อมูลในรูปแบบของ HTML ซึ่งไม่อยู่ในโครงสร้างที่เหมาะสมต่อการนำมาใช้งานโดยตรง ดังนั้นจึงต้องใช้เทคนิคการสกัดข้อมูลจากเว็บไซต์เพื่อดึงข้อมูลที่ต้องการออกจากโครงสร้างหน้าเว็บไซต์และแปลงให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ ซึ่งจากงานวิจัยที่ [6] ได้อธิบายข้อมูลเบื้องต้นเกี่ยวกับเทคนิคการสกัดข้อมูลจากเว็บไซต์ เช่น HTTP programming, DOM parsing และ HTML parsers รวมถึงแนะนำซอฟต์แวร์และเครื่องมือยอดนิยมสำหรับการดึงข้อมูลจากเว็บไซต์ เช่น BeautifulSoup และ Scrapy การสกัดข้อมูลจากเว็บไซต์มีจุดเด่นในเรื่องการช่วยดึงข้อมูลที่อยู่บนเว็บเพจมาใช้อย่างรวดเร็วและสะดวก ข้อมูลสามารถนำไปจัดเก็บ วิเคราะห์ หรือแปลงเป็นรูปแบบตามที่ต้องการได้ทันที โดยอาจจะมีข้อจำกัดในเรื่องนโยบายหรือข้อกำหนดการใช้งานบนบางเว็บไซต์ และยังคงจัดการกับการเปลี่ยนแปลงโครงสร้างเว็บไซต์ที่เกิดขึ้นบ่อยครั้ง

2. แบบจำลองการเรียนรู้ของเครื่อง

แบบจำลองการเรียนรู้ของเครื่อง เป็นวิธีการวิเคราะห์ข้อมูลที่ใช้แบบจำลองเรียนรู้จากชุดข้อมูล ระบุรูปแบบ และตัดสินใจได้อัตโนมัติ ซึ่งแบบจำลองการเรียนรู้โดยทั่วไปแบ่งออกเป็น 4 ประเภทหลัก คือ 1) การเรียนรู้แบบมีผู้สอน (supervised learning) เป็นกระบวนการฝึกฝนแบบจำลองด้วยชุดข้อมูลที่มีป้ายกำกับ ซึ่งประกอบด้วยข้อมูลนำเข้าและผลลัพธ์ เพื่อเรียนรู้ความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม ตัวอย่างเช่น การจำแนกประเภท (classification) และการถดถอย (regression) ดังงานวิจัยที่ [7] 2) การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) เป็นกระบวนการวิเคราะห์ข้อมูลที่ไม่มีป้ายกำกับ เพื่อหารูปแบบของข้อมูล ด้วยเทคนิคการจัดกลุ่ม (clustering) การค้นหากฎความสัมพันธ์ (association rule mining) และการลดมิติข้อมูล (dimensionality reduction) ดังงานวิจัยที่ [8] 3) การเรียนรู้แบบกึ่งมีผู้สอน (semi-supervised learning) เป็นวิธีการที่ประยุกต์ข้อมูลที่มีป้ายกำกับและไม่มีป้ายกำกับในการฝึกฝนแบบจำลอง เพื่อเพิ่มประสิทธิภาพการทำนายในสถานการณ์ที่ข้อมูลที่มีป้ายกำกับมีจำนวนจำกัด ดังงานวิจัยที่ [9] 4) การเรียนรู้เชิงเสริมแรง (reinforcement learning) เป็นการเรียนรู้ของเครื่องที่ใช้สัญญาณป้อนกลับในรูปของการกำหนดรางวัลหรือบทลงโทษเพื่อนำข้อมูลมาพัฒนาแบบจำลองดังงานวิจัยที่ [10]

3. การประเมินแบบจำลอง

การประเมินแบบจำลองต้องใช้เกณฑ์ที่เหมาะสมกับลักษณะและประเภทของปัญหา เนื่องจากแต่ละปัญหามีเป้าหมายและข้อกำหนดในการวัดที่แตกต่างกัน สำหรับการเรียนรู้แบบจัดกลุ่มสามารถประเมินได้ทั้งกรณีที่ไม่มี Ground truth โดยพิจารณาจากความหนาแน่นและการแยกของข้อมูลโดยมีตัวชี้วัด เช่น Silhouette score และ Davies–Bouldin index และกรณีที่มี Ground truth ใช้ตัวชี้วัด เช่น Adjusted rand index หรือ Normalized mutual information ในการเปรียบเทียบผลลัพธ์กับกลุ่มข้อมูลจริง ส่วนการเรียนรู้แบบจำแนกประเภทใช้ตัวชี้วัดสำหรับการประเมินได้แก่ Accuracy, Precision, Recall, F1-score และ Receiver-operator curve ดังงานวิจัยที่ [11], [12], [13] นอกจากนี้ยังมีตัวชี้วัด Matthews correlation coefficient และ Cohen’s kappa ที่มีความน่าเชื่อถือในกรณีข้อมูลไม่สมดุล และในการเรียนรู้แบบถดถอยใช้กับการทำนายตัวแปรต่อเนื่อง โดยใช้ตัวชี้วัดสำหรับการประเมินได้แก่ รากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) ค่าความคลาดเคลื่อนสัมบูรณ์ร้อยละเฉลี่ย (MAPE) และค่าสัมประสิทธิ์การตัดสินใจ (R^2) และค่าความแม่นยำของการพยากรณ์ (Forecast Accuracy: FA) ดังงานวิจัยที่ [18], [19] ซึ่งแต่ละตัวชี้วัดมีข้อจำกัดที่ต่างกันไป ดังนั้นการเลือกตัวแปรเพื่อการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้จึงขึ้นอยู่กับวัตถุประสงค์ของการวิจัยและลักษณะของชุดข้อมูลเป็นสำคัญ

4. งานวิจัยที่เกี่ยวข้องกับการพยากรณ์ราคารถยนต์มือสอง

ผู้วิจัยได้รวบรวมและวิเคราะห์งานวิจัยที่เกี่ยวข้องกับการพยากรณ์ราคารถยนต์มือสอง โดยพิจารณาจากวิธีการหรืออัลกอริทึมที่ใช้ชุดข้อมูล การประเมินประสิทธิภาพ และผลการวิจัย เพื่อนำมาประกอบการเลือกแนวทางที่เหมาะสมสำหรับการเปรียบเทียบในงานวิจัยฉบับนี้ ทั้งนี้สาระสำคัญของแต่ละงานวิจัยสรุปไว้ดังตารางที่ 1

ตารางที่ 1: สรุปสาระสำคัญของการทบทวนวรรณกรรมที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัย	วิธีการ/อัลกอริทึมที่ใช้	ชุดข้อมูล	การประเมินประสิทธิภาพ	ผลการวิจัย
1	Forecasting the Used Car Price Index Using Time Series Forecasting Methods [4]	- วิธีแยกส่วนประกอบ - วิธีการปรับเรียบด้วยเส้นโค้งเลขชี้กำลัง	ดัชนีราคารถยนต์มือสอง (ม.ค. 2562 – ธ.ค. 2566, 60 เดือน)	- Mean Absolute Deviation (MAD) - MAPE	วิธีการปรับเรียบด้วยเส้นโค้งเลขชี้กำลังแบบวินเทอร์มีความเหมาะสมที่สุด MAD = 3.9535 และ MAPE = 4.0722

ตารางที่ 1: สรุปสาระสำคัญของการทบทวนวรรณกรรมที่เกี่ยวข้อง (ต่อ)

ลำดับ	ชื่องานวิจัย	วิธีการ/อัลกอริทึมที่ใช้	ชุดข้อมูล	การประเมินประสิทธิภาพ	ผลการวิจัย
2	Comparative Analysis of Supervised Learning Algorithms for Valuating Used Car Prices [11]	- Multiple Linear Regression - Decision Tree Regression - Random Forest Regression	ชุดข้อมูล Car Dekho (13 คุณลักษณะ, 19974 รายการ)	- R^2 - RMSE - MAE	Random Forest Regression ความแม่นยำสูงสุดร้อยละ 94.10 เทียบกับ Decision Tree ที่ร้อยละ 92.45 และ Multiple Linear Regression ที่ร้อยละ 89.85
3	Comparative Analysis of XGBoost and Random Forest for Used Car Price Prediction [12]	- XGBoost - Random Forest	ชุดข้อมูลจาก OLX Indonesia (40,596 แถว, 7 คุณลักษณะหลัก ขยายเป็น 11 คุณลักษณะ)	- R^2 - RMSE - MAE	สำหรับแบบจำลองทั่วไป Random Forest มีประสิทธิภาพดีกว่า ($R^2 = 0.9485$) ด้วยความเสถียรและความสอดคล้องกันที่ดีกว่า สำหรับแบบจำลองเฉพาะแบรนด์ XGBoost ประสิทธิภาพดีอย่างสม่ำเสมอ สำหรับแบรนด์ส่วนใหญ่ เช่น Toyota โดยมี R^2 อยู่ที่ 0.96
4	Predicting the Selling Price of Cars Using Business Intelligence with the Feed-forward Backpropagation Algorithms [13]	Feed-forward Backpropagation	ชุดข้อมูล BMW จาก Kaggle (15 คุณลักษณะ, 325 กรณี, ปี ค.ศ. 1995–2017)	Mean Error	ราคาที่ทำนายใกล้เคียงกับราคาจริงด้วยค่าเฉลี่ยความผิดพลาดร้อยละ 11.46
5	Predicting Used Car Prices in Metro Manila Using Artificial Neural Networks on Web-Scraped Data [14]	Artificial Neural Network (ANN) - Multilayer Feedforward Neural Network (MFNN) with Back Propagation Algorithm	สกัดข้อมูลจากเว็บ “cars.trovit.ph” ใน Metro Manila	- MAE - MSE - RMSE - R^2 - Adjusted R^2	- หลังการปรับจูน Hyperparameter: MSE = 31,905,395,405.08, RMSE = 179,402.63, MAE = 125,546.52, R^2 Score = 0.7031 - ค่า RMSE และ MAE สูงเกินไปในบริบทของราคารถยนต์มือสอง
6	Prediction of Used Car Price Based on Supervised Learning Algorithm [15]	- Extra Trees Regressor - Random Forest Regressor (RF) - Ridge Regression - Decision Tree Regressor (DT)	ชุดข้อมูลราคาจาก Ali Tian Chi (ปี ค.ศ. 1996–2019, 6019 รายการ, 12 คุณลักษณะ)	- R^2 - MAE - MSE - RMSE - RMSLE - MAPE	- ผลการวิจัยพบว่า Extra Trees Regressor (ET) ให้ประสิทธิภาพดีที่สุด ($R^2 = 0.9807$) หลังการปรับปรุง ซึ่งสูงกว่า RF (0.8448), Ridge (0.8417), DT (0.8001) - คุณลักษณะ ระบบเกียร์อัตโนมัติ (automatic transmission) มีความสำคัญสูงสุด
7	Predictive Analytics on Used Car Prices Using Business Intelligence of Bayesian Networks for Sales Risk Reduction [16]	Bayesian Networks (BN) - Probabilistic Modelling	ชุดข้อมูลจาก Craigslist (18,319 รายการ: 2,884 รถหุ, 15,435 รถไม่หุ)	Accuracy	- แบบจำลองเฉพาะสำหรับรถหุ มีความแม่นยำสูงสุด (ร้อยละ 83.5) - BN สามารถทำนายราคาภายใต้สภาวะที่ไม่แน่นอนได้

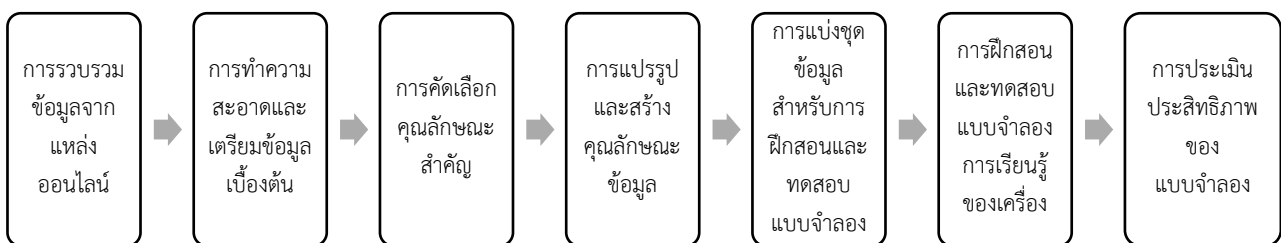
ตารางที่ 1: สรุปสาระสำคัญของการทบทวนวรรณกรรมที่เกี่ยวข้อง

ลำดับ	ชื่องานวิจัย	วิธีการ/อัลกอริทึมที่ใช้	ชุดข้อมูล	การประเมินประสิทธิภาพ	ผลการวิจัย
8	Residual Value of Used Car Analysis and Prediction [17]	Neural Network, CatBoost, XGBoost, LightGBM, Random Forest, Decision Tree, LASSO regression	บันทึกการขายต่อจริงของรถยนต์เช่าซื้อ (กว่า 7,800 คัน) จากบริษัทในนิวซีแลนด์	- MAE - MAPE	- แบบจำลอง CatBoost ให้ประสิทธิภาพสูงสุด ในการทำนายมูลค่างเฉลี่ย (MAE = 1195.29, MAPE = ร้อยละ 5.88) - ประสิทธิภาพดีกว่าวิธีเดิมของบริษัทประมาณร้อยละ 50 - การรวมตัวแปรเศรษฐกิจมหภาคช่วยเพิ่มความแม่นยำ

จากการทบทวนวรรณกรรมพบว่า การพยากรณ์ราคารถยนต์มือสองมีการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องอย่างหลากหลาย โดยเฉพาะการเรียนรู้ของเครื่องแบบมีผู้สอน เช่น Random Forest, XGBoost, Extra Trees และ Neural Network แสดงผลลัพธ์ที่แม่นยำ เมื่อนำไปใช้กับชุดข้อมูลจริงจากเว็บไซต์หรือธุรกิจรถยนต์มือสอง ตัวแปรในการประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง ได้แก่ R^2 , RMSE, MAE และ MAPE ซึ่งแบบจำลองที่มีค่า R^2 สูงและค่าความคลาดเคลื่อนต่ำจะถูกพิจารณาว่ามีความเหมาะสมที่สุด การวิจัยนี้จึงมุ่งเน้นการเปรียบเทียบแบบจำลอง โดยใช้คุณลักษณะสำคัญจากข้อมูลจริง และประเมินผลด้วยเกณฑ์มาตรฐานเพื่อให้สามารถพยากรณ์ราคาที่ใกล้เคียงกับความเป็นจริงมากที่สุดในตลาดรถยนต์มือสองของไทย

ระเบียบวิธีวิจัย

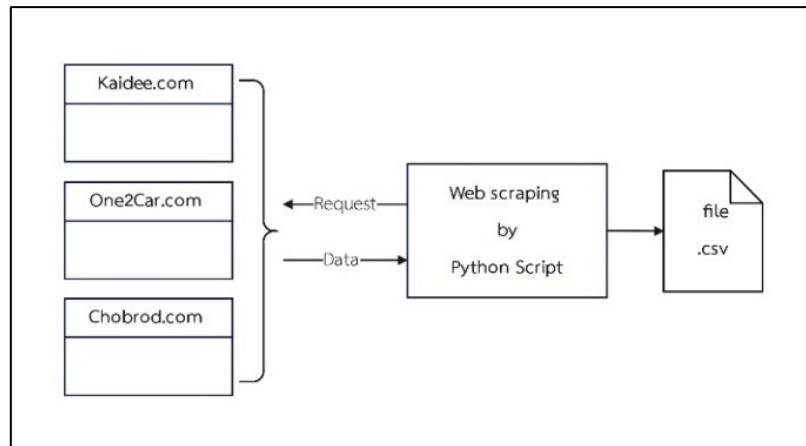
งานวิจัยนี้มีขั้นตอนการดำเนินวิจัยทั้งหมด 6 ขั้นตอน ดังแสดงในรูปที่ 1



รูปที่ 1: ขั้นตอนการดำเนินวิจัย

1. การรวบรวมข้อมูลจากแหล่งออนไลน์

ผู้วิจัยทำการรวบรวมข้อมูลรถยนต์มือสองจากเว็บไซต์ประกาศขายออนไลน์ ซึ่งประกอบด้วย Kaidee.com One2Car.com และ Chobrod.com โดยใช้เทคนิคการสกัดข้อมูลจากเว็บไซต์ ผ่าน Requests Library และ BeautifulSoup ด้วยภาษา Python เพื่อให้ได้ข้อมูลที่เกี่ยวข้อง ดังแสดงได้ตามขั้นตอนในรูปที่ 2 ตัวอย่างเช่น ยี่ห้อและรุ่นของรถยนต์ ปีที่ผลิต เลขไมล์ ราคาขาย ระบบเกียร์ ประเภทเชื้อเพลิง ประเภทของระบบขับเคลื่อน และจังหวัดที่ประกาศขาย เป็นต้น ข้อมูลในขั้นตอนนี้อยู่ในรูปแบบไฟล์ .csv อย่างไรก็ตาม ข้อมูลที่ได้จากแหล่งต่าง ๆ นั้นจะไม่สามารถนำมาใช้งานได้โดยตรง เนื่องจากความไม่เหมือนกันของข้อมูล เช่น บางแหล่งข้อมูลไม่มีข้อมูลเกี่ยวกับเครื่องยนต์ หรือข้อมูลที่สกัดได้ไม่ครบถ้วนสมบูรณ์ ดังนั้นต้องนำข้อมูลเหล่านี้ไปผ่านกระบวนการทำความสะอาดข้อมูลและเตรียมข้อมูลเบื้องต้นต่อไป



รูปที่ 2: ขั้นตอนการรวบรวมข้อมูลจากแหล่งออนไลน์

2. การทำความสะอาดและเตรียมข้อมูลเบื้องต้น

ข้อมูลที่ได้จากการรวบรวมจะผ่านการทำความสะอาด และปรับข้อมูลให้พร้อมใช้งานได้อย่างถูกต้อง ซึ่งมีกระบวนการดังต่อไปนี้

2.1 การตรวจสอบคุณลักษณะของข้อมูล: เพื่อระบุข้อมูลที่ขาดหายไป เนื่องจากรูปแบบการกรอกข้อมูลในเว็บไซต์จากผู้ขายแต่ละแหล่งข้อมูลไม่เป็นมาตรฐานเดียวกัน รวมถึงไม่บังคับให้ผู้ลงขายกรอกข้อมูลที่ครบถ้วน โดยข้อมูลลักษณะนี้จะถูกลบออกไป

2.2 การตรวจสอบค่าผิดปกติของข้อมูล: ข้อมูลที่สัปดาห์นั้นประกอบด้วยข้อมูลราคาของรถยนต์ที่แพงเกินความเป็นจริง ดังนั้นงานวิจัยนี้จึงใช้ค่าของข้อมูล ณ จุดที่ 75 เป็นค่าอ้างอิง เพื่อกำหนดกรอบของราคารถยนต์มือสองที่ไม่ผิดปกติ ดังนั้นข้อมูลที่ไม่เข้าเงื่อนไขนี้จะถูกลบออกไป

2.3 การตรวจสอบข้อมูลที่มีความซ้ำซ้อนกัน: เนื่องจากในแต่ละเว็บไซต์ไม่มีข้อจำกัดสำหรับการลงขายสินค้าขึ้นเดียวกันกับเว็บไซต์อื่น ดังนั้นข้อมูลบางชุดที่ผู้ขายลงจะซ้ำซ้อนกับเว็บไซต์อื่น เพราะผู้ขายต้องการให้ผู้ซื้อมีการมองเห็นการขายรถยนต์เพิ่มขึ้น ดังนั้นขั้นตอนนี้จะมีการลบข้อมูลที่ซ้ำซ้อนออกจากข้อมูลดิบ

2.4 การตรวจสอบรูปแบบข้อมูลว่าตรงกันหรือไม่: เช่น ข้อมูลวันที่และเวลา ซึ่งแต่ละแหล่งข้อมูลไม่เหมือนกัน ในกระบวนการนี้ใช้วิธีการแปลงข้อมูลเวลาให้อยู่ในรูปแบบของ วัน/เดือน/ปี รวมถึงปรับรูปแบบข้อมูลอื่นให้อยู่ประเภทเดียวกัน เช่น เลขทศนิยม

2.5 การแปลงข้อมูลตัวอักษรให้อยู่ในรูปแบบเดียวกัน: โดยปรับให้เป็นตัวอักษรลักษณะตัวพิมพ์เล็ก เพื่อให้คำเดียวกันมีความหมายตรงกัน เช่น “TOYOTA” “Toyota” และ “toyota”

2.6 การนำข้อมูลไปใช้ในแบบจำลอง: งานวิจัยนี้ได้เข้ารหัสข้อมูลที่เป็นข้อความให้เป็นตัวเลขและแยกเป็นคอลัมน์แต่ละลักษณะ เช่นคำว่า “toyota” หรือ “honda” ที่ถูกจัดรูปให้เป็นตัวพิมพ์เล็กจะถูกเข้ารหัสด้วยตัวเลข เช่น 1 และ 0 ซึ่งจะช่วยให้สามารถนำข้อมูลไปวิเคราะห์ได้

2.7 การตรวจสอบข้อมูลที่ผิดพลาด: เช่น ข้อมูลรุ่น Giulia และ Giulis ซึ่งจะเห็นได้ว่าความหมายของข้อมูลคือ Giulia ดังนั้นข้อมูลที่ผิดจะถูกแก้ไขให้ถูกต้อง

เมื่อข้อมูลผ่านกระบวนการทำความสะอาดและเตรียมข้อมูลเบื้องต้นแล้วจะได้ข้อมูลที่พร้อมใช้งานเพื่อที่จะดำเนินการในขั้นตอนการแปรรูปและสร้างคุณลักษณะต่อไป

3. การแปรรูปและสร้างคุณลักษณะ

ในขั้นตอนนี้ผู้วิจัยได้นำคุณลักษณะของชุดข้อมูลบางคุณลักษณะไปคำนวณและสร้างเป็นคุณลักษณะใหม่ เพื่อช่วยให้แบบจำลองการเรียนรู้ของเครื่องพยากรณ์ราคารถยนต์มือสองได้แม่นยำมากขึ้น ซึ่งประกอบด้วยคุณลักษณะของข้อมูลต่อไปนี้

3.1 อายุรถยนต์: การคำนวณอายุของรถยนต์มือสองสามารถหาได้จากสมการดังต่อไปนี้

$$A_v = Y_p - Y_M \quad (1)$$

เมื่อ A_v คือ อายุรถยนต์

Y_p คือ ปี พ.ศ. ปัจจุบัน

Y_M คือ ปี พ.ศ. ที่รถยนต์ผลิต

3.2 อัตราการใช้งานต่อปี: ผู้วิจัยสร้างคุณลักษณะนี้ขึ้นมาเพื่อนำมาประเมินสภาพรถยนต์มือสอง ซึ่งคำนวณได้ตามสมการต่อไปนี้

$$U_A = \frac{M_v}{A_v} \quad (2)$$

เมื่อ U_A คือ อัตราการใช้งานต่อปี (กิโลเมตร/ปี)

M_v คือ ระยะทางสะสม (กิโลเมตร)

3.3 กำลังสองของระยะทางสะสม: คุณลักษณะนี้ใช้เพื่อประเมินความสัมพันธ์ที่ไม่เป็นเชิงเส้นระหว่างระยะทางที่รถยนต์วิ่งและราคาของรถยนต์มือสอง โดยแสดงได้ดังสมการต่อไปนี้

$$M_s = M_v^2 \quad (3)$$

เมื่อ M_s คือ กำลังสองของระยะทางสะสม

3.4 กำลังสองของขนาดเครื่องยนต์: คุณลักษณะนี้มีวัตถุประสงค์เช่นเดียวกับกำลังสองของระยะทางสะสม โดยแสดงได้สมการดังต่อไปนี้

$$E_s = E_v^2 \quad (4)$$

เมื่อ E_s คือ กำลังสองของขนาดเครื่องยนต์

E_v คือ ขนาดของเครื่องยนต์

3.5 ภาระการใช้งานสะสม: การสร้างคุณลักษณะนี้มีวัตถุประสงค์เพื่อแสดงปริมาณการใช้งานสะสมของเครื่องยนต์ดังสมการต่อไปนี้

$$V_t = M_v \times E_v \quad (5)$$

เมื่อ V_t คือ ภาระการใช้งานสะสม

3.6 ภาระเชิงเวลา: แสดงถึงปริมาณการใช้งานสะสมตามเวลาโดยคำนวณจากอายุรถ สามารถเขียนเป็นสมการได้ดังนี้

$$V_t = M_v \times A_v \quad (6)$$

เมื่อ V_t คือ ภาระเชิงเวลา

4. การแบ่งชุดข้อมูลสำหรับการฝึกสอนและทดสอบแบบจำลอง

งานวิจัยนี้แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลสำหรับการฝึกสอนและการทดสอบ โดยมีสัดส่วนการแบ่งข้อมูลคือ ร้อยละ 80 ต่อ ร้อยละ 20 ชุดข้อมูลฝึกสอนนำไปใช้ในการฝึกแบบจำลองเพื่อให้เรียนรู้ความสัมพันธ์ระหว่างคุณลักษณะและราคาของรถยนต์มือสอง ขณะที่ชุดข้อมูลทดสอบสงวนไว้เพื่อประเมินความสามารถของแบบจำลองในการพยากรณ์ราคาที่ไม่เคยเห็นมาก่อน เพื่อให้ผลการพยากรณ์มีความน่าเชื่อถือและสะท้อนความแม่นยำของแบบจำลองในสถานการณ์จริง

5. การฝึกสอนและทดสอบแบบจำลองการเรียนรู้ของเครื่อง

ผู้วิจัยได้ดำเนินการสร้างแบบจำลองการเรียนรู้ของเครื่องจำนวน 5 แบบจำลอง เพื่อทำการพยากรณ์ราคาของรถยนต์มือสอง โดยแบบจำลองที่เลือกใช้คือ XGBoost, Random Forest, LightGBM, CatBoost และ Gradient Boosting ซึ่งแต่ละแบบจำลองมีคุณลักษณะเชิงโครงสร้างที่แตกต่างกัน และสามารถรองรับข้อมูลที่มีความไม่เป็นเชิงเส้น การเลือกใช้แบบจำลองทั้ง 5 ประเภท มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองในเชิงปฏิบัติ

6. การประเมินประสิทธิภาพของแบบจำลอง

ผลลัพธ์ของแต่ละแบบจำลองจะถูกประเมินโดยใช้ค่าชี้วัด คือค่า RMSE, MAE, MAPE และ R² score เพื่อพิจารณาความแม่นยำในการพยากรณ์และความเหมาะสมของแบบจำลองสำหรับการนำไปประยุกต์ใช้งานจริง

ผลการวิจัยและการอภิปรายผล

งานวิจัยฉบับนี้ได้สรุปผลวิจัยแต่ละขั้นตอนเพื่อให้ได้มาซึ่งข้อมูลที่ต้องการ โดยผลการวิจัยในแต่ละขั้นตอนสรุปได้ดังต่อไปนี้

1. ผลการรวบรวมข้อมูลจากเว็บไซต์

จากการวิจัย ผลการเก็บรวบรวมข้อมูลรถยนต์มือสองมีจำนวนทั้งสิ้น 55,989 รายการจาก 3 แหล่งข้อมูลหลัก ได้แก่ Kaidee.com, One2Car.com และ Chobrod.com โดยเก็บข้อมูลทั้งในด้านยี่ห้อ รุ่น ปีที่ผลิต ระยะทาง ราคา ระบบเกียร์ ระบบขับเคลื่อน และสถานที่ขาย ซึ่งข้อมูลทั้งหมดถูกจัดเก็บในรูปแบบ .csv

2. ผลการทำความสะอาดและเตรียมข้อมูลเบื้องต้น

การตรวจสอบคุณภาพข้อมูลที่ได้จากขั้นตอนที่ 1 พบว่าข้อมูลที่สมบูรณ์และสามารถนำไปใช้งานได้มีจำนวนทั้งสิ้น 42,823 รายการ โดยมีการลบข้อมูลที่ขาดหาย ผิดปกติ หรือซ้ำซ้อน รวมถึงการปรับเปลี่ยนรูปแบบข้อมูลให้สอดคล้องกัน ซึ่งประกอบด้วยคุณลักษณะของข้อมูลดังตารางที่ 2

ตารางที่ 2: คุณลักษณะของข้อมูลที่ใช้ในการพยากรณ์

คุณลักษณะ	คำอธิบาย
brand	ยี่ห้อ
brand_model	ยี่ห้อและรุ่น
cartype	ประเภทรถ
color	สี
fuel	เชื้อเพลิง
engine	ขนาดเครื่องยนต์
gear	ระบบเกียร์
mileage	ระยะทางที่วิ่งทั้งหมด
model	รุ่น
province	จังหวัด
submodel	รุ่นย่อย
WD	ระบบขับเคลื่อน
year	ปีที่ผลิต

3. ผลการแปรรูปและสร้างคุณลักษณะใหม่

ผู้วิจัยได้ทำการสร้างคุณลักษณะใหม่ 6 รายการ ได้แก่ อายุรถ อัตราการใช้งานรถยนต์ต่อปี กำลังสองของระยะทางสะสม กำลังสองของขนาดเครื่องยนต์ ภาระสะสมการใช้งานสะสมของรถยนต์ และภาระเชิงเวลาของรถยนต์ โดยพิจารณาจากนัยสำคัญของคุณลักษณะแต่ละตัว ซึ่งข้อมูลที่ชี้แสดงได้ดังตารางที่ 3

ตารางที่ 3: คุณลักษณะใหม่ของข้อมูลที่ใช้ในการพยากรณ์

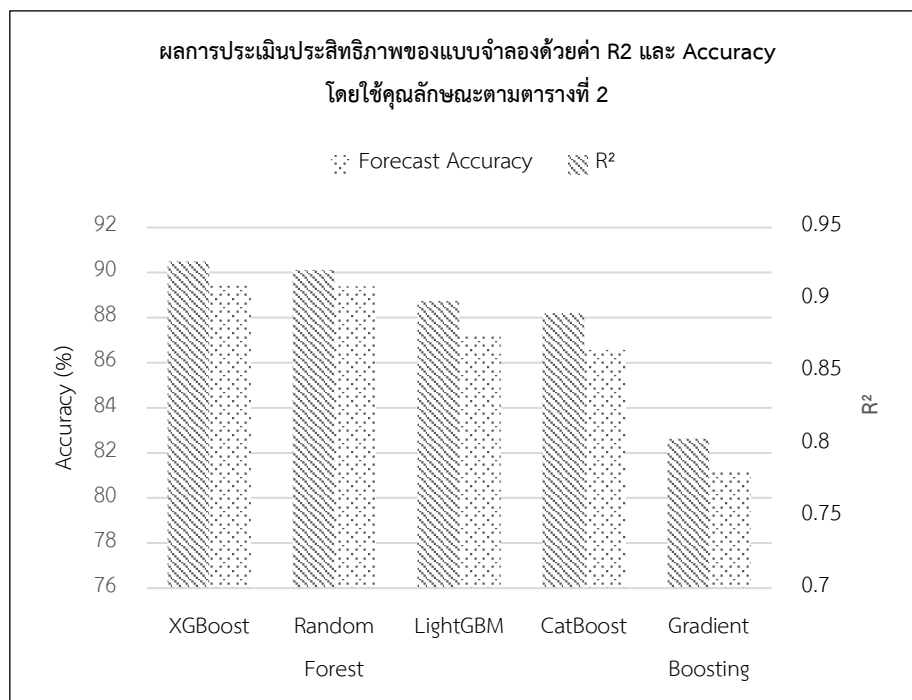
คุณลักษณะ	คำอธิบาย
car_age	อายุรถยนต์
mileage_per_year	อัตราการใช้งานรถยนต์ต่อปี
mileage_squared	กำลังสองของระยะทางสะสม
engine_squared	กำลังสองของขนาดเครื่องยนต์
mileage_engine_interaction	ภาระการใช้งานสะสม
age_mileage_interaction	ภาระเชิงเวลา

4. ผลการประเมินประสิทธิภาพของแบบจำลอง

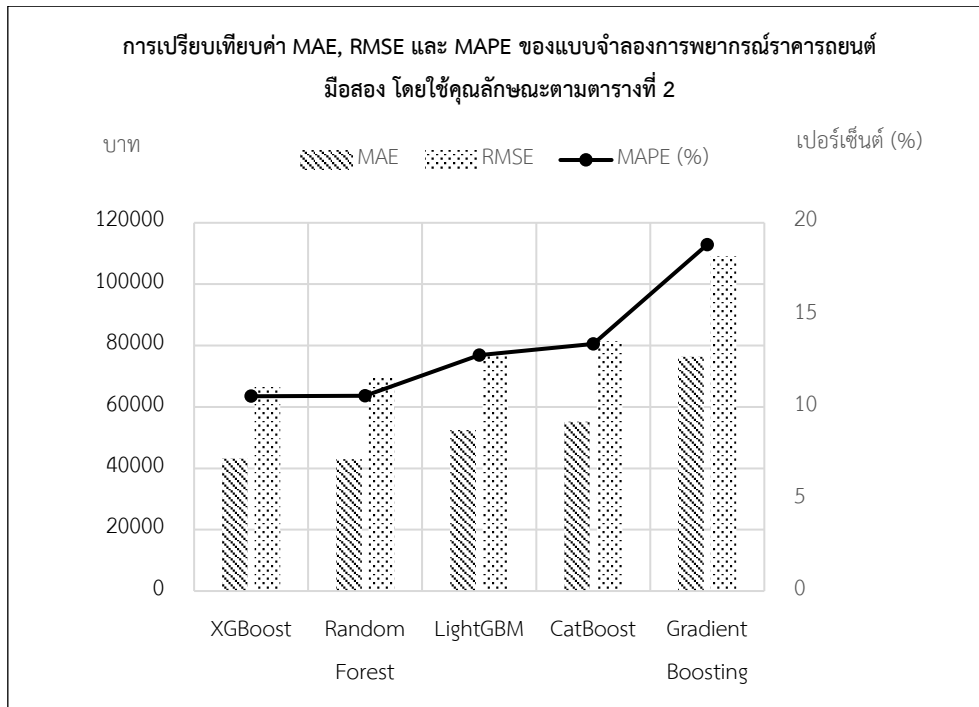
จากข้อมูลทั้งหมด 42,823 รายการ ถูกแบ่งออกเป็นชุดฝึกสอนจำนวน 34,258 รายการและชุดทดสอบจำนวน 8,565 รายการ โดยทดสอบกับแบบจำลองจำนวน 5 แบบ ได้แก่ XGBoost, Random Forest, LightGBM, CatBoost และ Gradient Boosting และตัวชี้วัดได้แก่ ความคลาดเคลื่อนเฉลี่ยแบบสัมบูรณ์ รากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย ค่าสัมประสิทธิ์การตัดสินใจ ความคลาดเคลื่อนเฉลี่ยเชิงเปอร์เซ็นต์ และค่าความแม่นยำของการพยากรณ์ โดยผู้วิจัยได้ทดลอง 2 แบบ คือ 1) ทดสอบโดยใช้คุณลักษณะพื้นฐานในตารางที่ 2 ซึ่งได้ผลการประเมินประสิทธิภาพสรุปได้ดังตารางที่ 4 และรูปที่ 3-4 และ 2) ทดสอบโดยเพิ่มคุณลักษณะในตารางที่ 3 โดยผลการประเมินประสิทธิภาพแสดงได้ดังตารางที่ 5 และรูปที่ 5-6

ตารางที่ 4: ผลการประเมินประสิทธิภาพของแต่ละแบบจำลองโดยใช้คุณลักษณะตามตารางที่ 2

Model	MAE (บาท)	RMSE (บาท)	MAPE (%)	R ²	Accuracy (%)
XGBoost	43,141.89	66,871.3	10.58	0.9262	89.42
Random Forest	43,042.89	69,872.79	10.68	0.9195	89.32
LightGBM	53,140.07	79,131.83	13.09	0.8967	86.91
CatBoost	56,009.61	82,600.02	13.68	0.8875	86.32
Gradient Boosting	77,812.95	111,318.41	19.25	0.7956	80.75



รูปที่ 3: ผลการประเมินเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ด้วยค่า R² และ FA โดยใช้คุณลักษณะตามตารางที่ 2



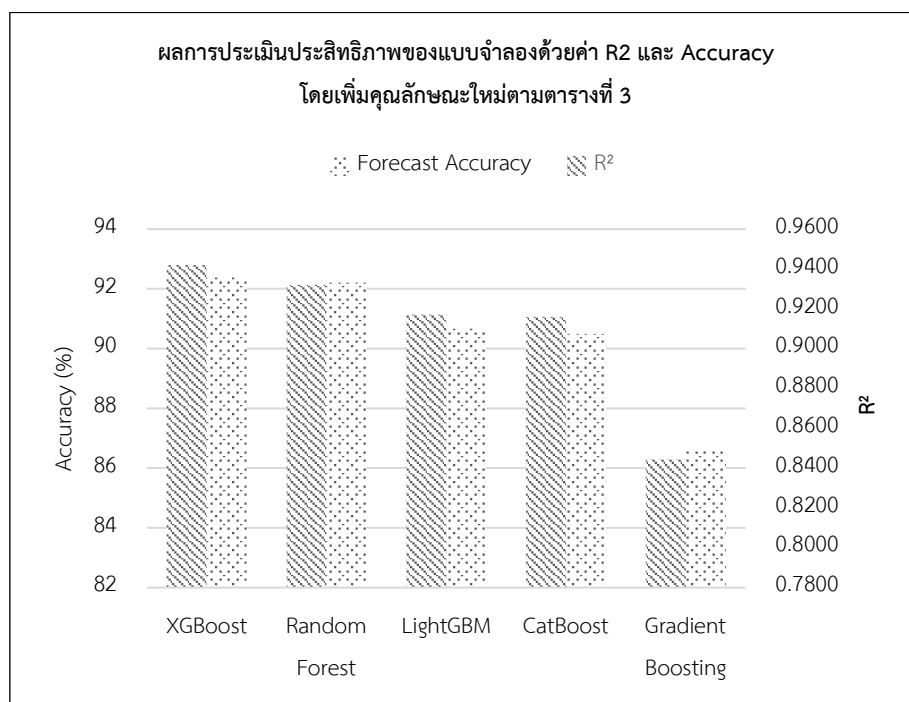
รูปที่ 4: การเปรียบเทียบค่า MAE, RMSE และ MAPE ของแบบจำลองการพยากรณ์ราคารถยนต์มือสอง โดยใช้คุณลักษณะใหม่ตามตารางที่ 2

ผลการประเมินแบบจำลองทั้ง 5 แบบตามตารางที่ 4 พบว่าแบบจำลอง XGBoost มีความแม่นยำมากที่สุด โดยมีค่า R^2 สูงสุดอยู่ที่ 0.9265 และ FA สูงสุดที่ร้อยละ 89.42 ตามที่แสดงในกราฟเปรียบเทียบระหว่าง R^2 และ FA ดังแสดงในรูปที่ 3 ทั้งนี้เมื่อพิจารณาจากกราฟเปรียบเทียบค่า MAE, RMSE และ MAPE ในรูปที่ 4 พบว่าแบบจำลอง XGBoost มีค่าความคลาดเคลื่อนอยู่ในระดับต่ำ โดยเฉพาะค่า RMSE มีค่าเท่ากับ 66,732.61 บาท และ MAPE เท่ากับร้อยละ 10.58 ซึ่งต่ำกว่าแบบจำลองอื่นอย่างชัดเจน ซึ่งแสดงให้เห็นว่าแบบจำลอง XGBoost ไม่เพียงแต่มีความสามารถในการอธิบายข้อมูลได้ดี แต่ยังให้ผลการพยากรณ์ราคารถยนต์มือสองตามชุดข้อมูลที่มีความแม่นยำและความเสถียรสูงกว่าด้วย สำหรับแบบจำลอง Random Forest ที่แม้ค่า MAE ต่ำที่สุดที่ 42,954.29 บาท แต่เมื่อพิจารณาจากกราฟเปรียบเทียบ MAE, RMSE และ MAPE ดังแสดงได้ในรูปที่ 4 จะเห็นว่าค่า RMSE และ MAPE ของแบบจำลอง Random Forest ยังสูงกว่าแบบจำลอง XGBoost เล็กน้อย เป็นผลให้ค่า FA โดยรวมใกล้เคียงกัน โดยแบบจำลอง Random Forest ซึ่งมีค่าเฉลี่ย FA เท่ากับร้อยละ 89.39 จึงกล่าวได้ว่าแบบจำลอง Random Forest มีประสิทธิภาพในการพยากรณ์ราคารถยนต์มือสองสูงรองลงมา ขณะที่แบบจำลอง LightGBM และ CatBoost ให้ค่า R^2 อยู่ในช่วง 0.89–0.90 และ FA อยู่ในช่วงร้อยละ 86–87 แม้ไม่สูงเท่า XGBoost และ Random Forest แต่จากกราฟในรูปที่ 4 จะเห็นว่าแบบจำลองทั้งสองมีแนวโน้มของค่า MAE, RMSE และ MAPE ที่อยู่ในระดับปานกลาง ถือว่าเป็นแบบจำลองที่มีความสมดุลระหว่างความแม่นยำและการคำนวณ ในทางตรงกันข้าม แบบจำลอง Gradient Boosting มีค่า R^2 ต่ำที่สุดที่ 0.8037 และ FA ต่ำสุดเพียงร้อยละ 81.18 อีกทั้งกราฟในรูปที่ 4 ยังแสดงให้เห็นว่ามีค่า MAE, RMSE และ MAPE สูงที่สุดเมื่อเทียบกับแบบจำลองการพยากรณ์แบบอื่น แสดงให้เห็นว่าประสิทธิภาพของแบบจำลองนี้ไม่เหมาะสมกับชุดข้อมูลราคารถยนต์มือสองที่ใช้ในการทดลอง

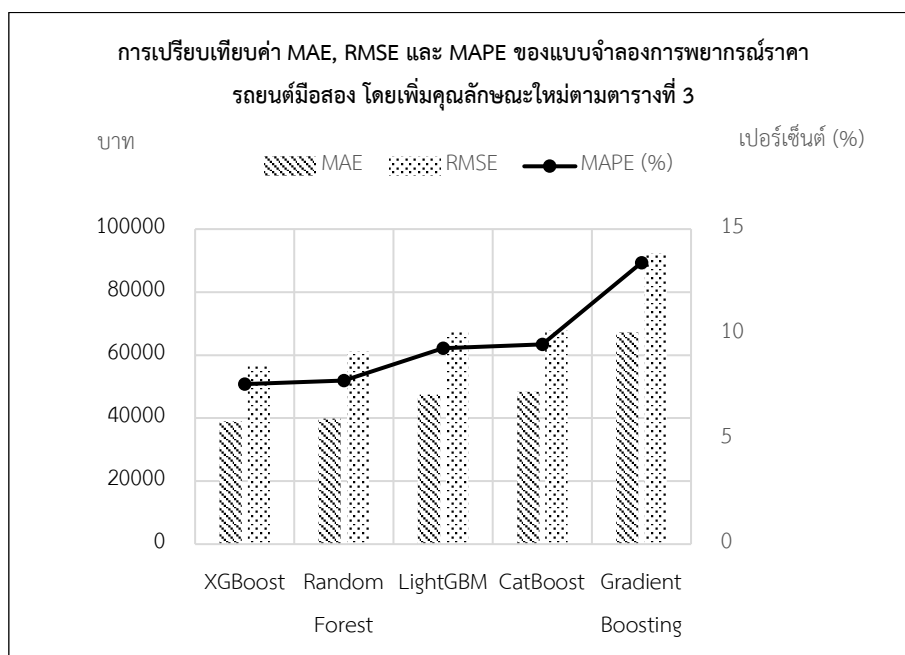
จากผลการทดลองในตารางที่ 4 จะเห็นได้ว่าแบบจำลอง XGBoost และ Random Forest ให้ผลการพยากรณ์ราคาที่แม่นยำกว่าแบบจำลองอื่น โดยมีค่า R^2 และ Accuracy สูงกว่า 0.92 และร้อยละ 89 ตามลำดับ อย่างไรก็ตาม ยังพบว่ามีค่าความคลาดเคลื่อนในตัวชี้วัด RMSE และ MAPE ยังคงอยู่ในระดับค่อนข้างสูง ซึ่งอาจเกิดจากการที่ชุดข้อมูลยังมีคุณลักษณะจำกัด ทำให้แบบจำลองไม่สามารถเรียนรู้ความสัมพันธ์เชิงซับซ้อนของข้อมูลได้อย่างเต็มที่ ดังนั้นเพื่อปรับปรุงประสิทธิภาพของแบบจำลอง ผู้วิจัยจึงได้ทำการเพิ่มจำนวนคุณลักษณะ โดยเฉพาะคุณลักษณะเชิงอนุพันธ์และคุณลักษณะเชิงปฏิสัมพันธ์ที่สร้างขึ้นจากคุณลักษณะพื้นฐานอีก 6 รายการ ซึ่งได้ผลการประเมินประสิทธิภาพดังตารางที่ 5

ตารางที่ 5: ผลการประเมินประสิทธิภาพของแต่ละแบบจำลองโดยเพิ่มคุณลักษณะที่ปรับปรุงใหม่ในตารางที่ 3

Model	MAE (บาท)	RMSE (บาท)	MAPE (%)	R ²	Accuracy (%)
XGBoost	38,900.23	56,447.83	7.62	0.9419	92.38
Random Forest	39,786.85	61,090.38	7.79	0.932	92.21
LightGBM	47,452.14	67,447.82	9.33	0.9171	90.67
CatBoost	48,369.00	67,899.4	9.51	0.916	90.49
Gradient Boosting	67,275.87	92,400.97	13.4	0.8444	86.6



รูปที่ 5: ผลการประเมินเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ด้วยค่า R² และ FA โดยเพิ่มคุณลักษณะใหม่ตามตารางที่ 3

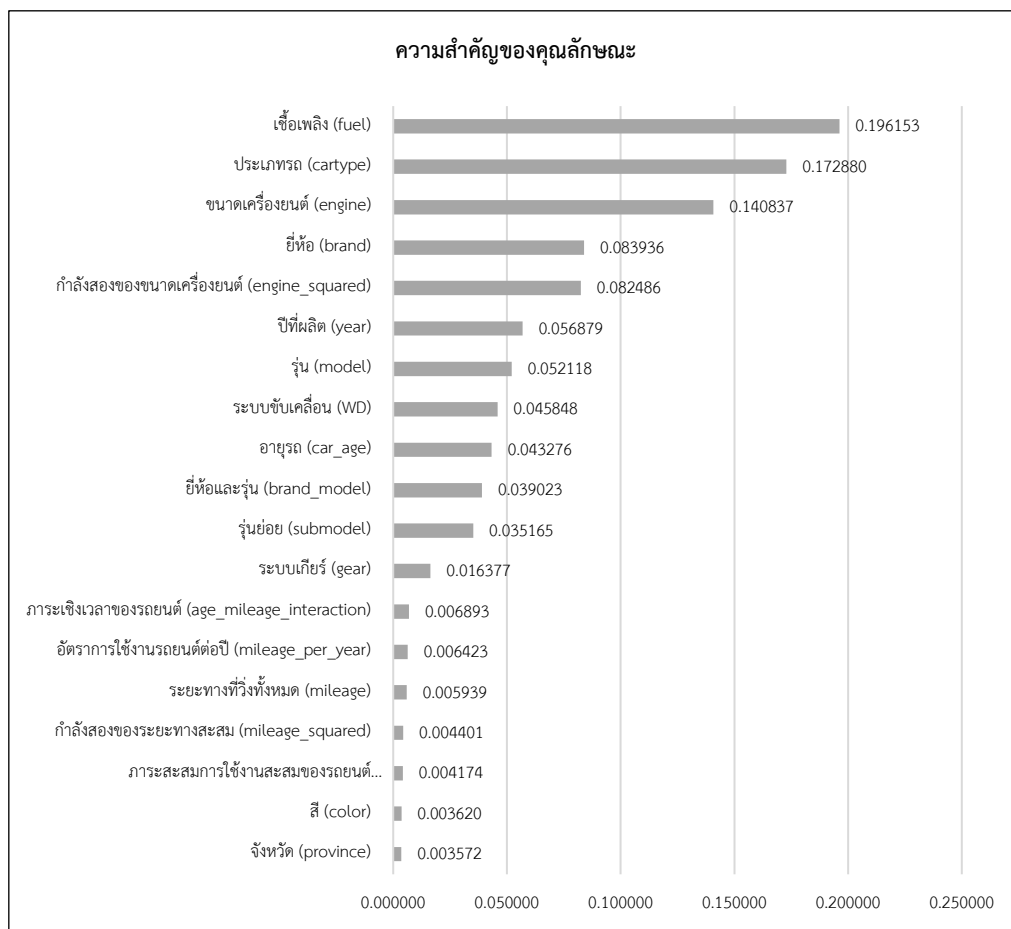


รูปที่ 6: การเปรียบเทียบค่า MAE, RMSE และ MAPE ของแบบจำลองการพยากรณ์ราคารถยนต์มือสองโดยเพิ่มคุณลักษณะใหม่ตามตารางที่ 3

จากตารางที่ 5 แสดงผลการประเมินประสิทธิภาพของแบบจำลองที่ได้รับการปรับปรุงด้วยการเพิ่มคุณลักษณะตามตารางที่ 3 พบว่า ผลลัพธ์โดยรวมของทุกแบบจำลองดีขึ้นอย่างชัดเจนเมื่อเทียบกับตารางที่ 4 ทั้งในด้านค่าความคลาดเคลื่อน และค่าความแม่นยำของการพยากรณ์ ซึ่งแสดงให้เห็นว่าการเพิ่มจำนวนคุณลักษณะสร้างใหม่เข้าไปช่วยให้แบบจำลองสามารถเรียนรู้ความสัมพันธ์เชิงลึกของข้อมูลการรถยนต์มือสองได้ดียิ่งขึ้น โดยในรูปที่ 5 ซึ่งเปรียบเทียบค่า R^2 และ FA ของแบบจำลองทั้ง 5 แบบ พบว่าแบบจำลอง XGBoost ยังคงให้ประสิทธิภาพสูงสุด โดยมีค่า R^2 เท่ากับ 0.9419 และ Accuracy สูงถึงร้อยละ 92.38 ซึ่งเพิ่มขึ้นจากเดิมร้อยละ 89.42 ในตารางที่ 4 ขณะที่แบบจำลอง Random Forest มีประสิทธิภาพรองลงมา ($R^2 = 0.932$, Accuracy = ร้อยละ 92.21) โดยมีความคลาดเคลื่อน (MAE, RMSE, MAPE) ต่ำกว่าแบบจำลองอื่นอย่างต่อเนื่อง และเมื่อพิจารณากราฟในรูปที่ 6 ซึ่งแสดงการเปรียบเทียบค่า MAE, RMSE และ MAPE ของแต่ละแบบจำลอง จะเห็นได้ว่าแบบจำลอง XGBoost และ Random Forest ให้ค่าความคลาดเคลื่อนในระดับต่ำที่สุด โดยเฉพาะ XGBoost ที่มี MAE เพียง 38,900.23 บาท และ MAPE เพียงร้อยละ 7.62 แสดงถึงความแม่นยำในการคาดการณ์ราคาที่มีประสิทธิภาพ ขณะที่ Random Forest แม้มีค่า MAE สูงกว่าเพียงเล็กน้อย แต่ยังคงรักษาความสม่ำเสมอของผลลัพธ์ในเชิงเสถียรภาพได้ดี

ส่วนแบบจำลอง LightGBM และ CatBoost ให้ค่า R^2 อยู่ในช่วง 0.916 – 0.917 และ Accuracy ประมาณร้อยละ 90–91 ถือว่าเป็นแบบจำลองที่มีความสมดุลระหว่างความแม่นยำและประสิทธิภาพในการคำนวณ โดยเฉพาะในกรณีที่ต้องการความเร็วในการประมวลผลมากกว่า ขณะที่ Gradient Boosting ยังคงมีค่าประสิทธิภาพต่ำสุด ($R^2 = 0.8444$, Accuracy = ร้อยละ 86.6) และมีค่า MAE และ RMSE สูงกว่าแบบจำลองอื่น จึงไม่เหมาะสมกับลักษณะของข้อมูลการรถยนต์มือสองที่ใช้ในงานวิจัยนี้

เพื่อทำความเข้าใจปัจจัยที่ส่งผลต่อประสิทธิภาพของแบบจำลอง ผู้วิจัยได้ทำการวิเคราะห์ความสำคัญของคุณลักษณะในแบบจำลอง XGBoost ซึ่งเป็นแบบจำลองที่ให้ผลการพยากรณ์ดีที่สุด ซึ่งผลลัพธ์แสดงได้ดังรูปที่ 7



รูปที่ 7: ความสำคัญของคุณลักษณะในชุดข้อมูล

จากรูปที่ 7 พบว่า คุณลักษณะ 5 อันดับแรกที่มีอิทธิพลต่อการพยากรณ์ราคารถยนต์มือสองมากที่สุด ได้แก่ ประเภทเชื้อเพลิง ประเภทรถ ขนาดเครื่องยนต์ ยี่ห้อ และค่ากำลังสองของขนาดเครื่องยนต์ โดยทั้งหมดมีค่าน้ำหนักรวมมากกว่าร้อยละ 68 ของค่าความสำคัญทั้งหมด สะท้อนว่าปัจจัยด้านเทคนิคของรถยนต์มีอิทธิพลอย่างมากต่อการกำหนดราคาเมื่อสอง ในขณะที่คุณลักษณะรอง เช่น ปีที่ผลิต อายุของรถ และ ระบบขับเคลื่อน มีความสำคัญในระดับปานกลาง ส่วนปัจจัยด้าน สี และ จังหวัด มีผลในระดับต่ำกว่า 0.004 แสดงให้เห็นว่ามีอิทธิพลเพียงเล็กน้อยต่อการประเมินมูลค่ารถยนต์มือสอง

ผลการประเมินประสิทธิภาพของแบบจำลองในตารางที่ 5 และรูปที่ 5-6 ยืนยันได้ว่า คุณลักษณะใหม่ที่เพิ่มเข้าไป ได้แก่ 1) อายุของรถยนต์มือสอง 2) อัตราการใช้งานรถยนต์ต่อปี 3) กำลังสองของระยะทางสะสม 4) กำลังสองของขนาดเครื่องยนต์ 5) ภาระสะสมการใช้งานสะสมของรถยนต์ 6) ภาระเชิงเวลาของรถยนต์ ช่วยเพิ่มประสิทธิภาพการพยากรณ์ราคารถยนต์มือสองตามชุดข้อมูลที่ใช้ในงานวิจัยนี้ โดยเฉพาะในแบบจำลอง XGBoost ที่แสดงศักยภาพสูงสุดในการพยากรณ์ราคารถยนต์มือสองจากข้อมูลจริง

สรุปผล

จากการศึกษาครั้งนี้ ผู้วิจัยได้พัฒนาแบบจำลองการพยากรณ์ราคารถยนต์มือสองโดยใช้เทคนิคการสกัดข้อมูลจากเว็บไซต์ประกาศขายรถยนต์ในประเทศไทย ได้แก่ Kaidee.com, One2Car.com และ Chobrod.com รวมข้อมูลทั้งสิ้น 55,989 รายการ หลังจากผ่านกระบวนการทำความสะอาดและคัดกรองข้อมูลแล้วเหลือข้อมูลที่สามารถนำไปใช้ได้จำนวน 42,823 รายการ ซึ่งถูกแบ่งเป็นชุดฝึกสอนและชุดทดสอบในสัดส่วน 80 ต่อ 20 เพื่อใช้กับแบบจำลองการเรียนรู้ของเครื่อง 5 แบบ ได้แก่ XGBoost, Random Forest, LightGBM, CatBoost และ Gradient Boosting

ในการทดลองแรกที่ใช้เฉพาะคุณลักษณะพื้นฐาน พบว่าแบบจำลอง XGBoost ให้ผลการพยากรณ์แม่นยำที่สุด ($R^2 = 0.9262$, Accuracy = ร้อยละ 89.42) รองลงมาคือ Random Forest ($R^2 = 0.9195$, Accuracy = ร้อยละ 89.32) ส่วน LightGBM และ CatBoost มีประสิทธิภาพระดับปานกลาง และ Gradient Boosting มีค่าความแม่นยำต่ำที่สุด สะท้อนว่าข้อมูลพื้นฐานเพียงอย่างเดียวยังไม่เพียงพอต่อการเรียนรู้ความสัมพันธ์เชิงซับซ้อนของตัวแปรราคา

เมื่อเพิ่มคุณลักษณะใหม่ที่ได้จากกระบวนการแปรรูปและสร้างคุณลักษณะ ประกอบด้วย 1) อายุของรถยนต์ 2) อัตราการใช้งานต่อปี 3) กำลังสองของระยะทางสะสม 4) กำลังสองของขนาดเครื่องยนต์ 5) ภาระสะสมการใช้งาน และ 6) ภาระเชิงเวลา พบว่าแบบจำลองทุกแบบมีประสิทธิภาพสูงขึ้นอย่างมีนัยสำคัญ โดยเฉพาะ XGBoost ซึ่งมีค่า R^2 เพิ่มขึ้นเป็น 0.9419 และ Accuracy เพิ่มขึ้นเป็นร้อยละ 92.38 ขณะที่ค่า MAPE ลดลงเหลือร้อยละ 7.62 แสดงถึงความสามารถของแบบจำลองในการเรียนรู้ความสัมพันธ์เชิงลึกระหว่างคุณลักษณะต่าง ๆ ของรถยนต์มือสอง

ผลการวิเคราะห์ ความสำคัญของคุณลักษณะ ของแบบจำลอง XGBoost ชี้ให้เห็นว่าปัจจัยที่มีอิทธิพลต่อราคาเมื่อสองมากที่สุด 5 อันดับแรก ได้แก่ ประเภทเชื้อเพลิง ประเภทรถ ขนาดเครื่องยนต์ ยี่ห้อ และค่ากำลังสองของเครื่องยนต์ โดยรวมมีค่าน้ำหนักมากกว่าร้อยละ 68 ของความสำคัญทั้งหมด ซึ่งแสดงว่าปัจจัยทางเทคนิคของรถยนต์มีผลต่อมูลค่าตลาดโดยตรง ขณะที่ปีผลิต อายุรถ และระบบขับเคลื่อนมีผลในระดับปานกลาง ส่วนสีและจังหวัดมีอิทธิพลเพียงเล็กน้อย ผลดังกล่าวสอดคล้องกับค่าประสิทธิภาพในตารางที่ 5 ซึ่งแสดงให้เห็นว่าการเพิ่มคุณลักษณะเชิงอนุพันธ์และเชิงปฏิสัมพันธ์ช่วยให้แบบจำลองลดค่าความคลาดเคลื่อน (MAE, RMSE, MAPE) และเพิ่มความแม่นยำของการพยากรณ์ได้อย่างมีนัยสำคัญ

การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องร่วมกับการสร้างคุณลักษณะใหม่ที่เหมาะสมในงานวิจัยนี้สามารถเพิ่มประสิทธิภาพในการพยากรณ์ราคารถยนต์มือสองได้อย่างชัดเจน โดยเฉพาะแบบจำลอง XGBoost ซึ่งให้ผลลัพธ์ที่มีความแม่นยำสูงสุดและความคลาดเคลื่อนต่ำสุด ผลลัพธ์ดังกล่าวสามารถนำไปประยุกต์ใช้เป็นเครื่องมือสนับสนุนการตัดสินใจของผู้ซื้อ ผู้ขาย และผู้ให้บริการสินเชื่อเพื่อประเมินมูลค่ารถยนต์มือสองอย่างเป็นธรรม โปร่งใส และช่วยยกระดับมาตรฐานของตลาดรถยนต์มือสองในประเทศไทย

กิตติกรรมประกาศ

ขอขอบพระคุณ คณะวิทยาศาสตร์และเทคโนโลยี และคณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏภูเก็ต เป็นอย่างยิ่งที่ได้ให้การสนับสนุนทรัพยากรและเอื้อเฟื้อสถานที่ บุคลากร และอำนวยความสะดวกที่ช่วยส่งเสริมบรรยากาศแห่งการเรียนรู้ที่เอื้อต่อการพัฒนาศักยภาพของนักวิจัย และขอขอบคุณ เว็บไซต์ Kaidee.com One2Car.com และ Chobrod.com ที่เป็นแหล่งข้อมูลสำคัญของการสกัดข้อมูลเพื่อใช้ในการวิเคราะห์ที่เป็นประโยชน์อย่างยิ่งต่องานวิจัยในครั้งนี้

REFERENCES

- [1] S. Bergmann and S. Feuerriegel, "Machine learning for predicting used car resale prices using granular vehicle equipment information," *Expert Syst. Appl.*, vol. 263, 2025, Art. no. 125640, doi: 10.1016/j.eswa.2024.125640.
- [2] L. Bukvić, J. P. Škrinjar, T. Fratrović, and B. Abramović, "Price prediction and classification of used-vehicles using supervised machine learning," *Sustainability*, vol. 14, no. 24, 2022, Art. no. 17034, doi: 10.3390/su142417034.
- [3] M. Nandan and D. Ghosh, "Pre-owned car price prediction by employing machine learning techniques," *J. Decis. Anal. Intell. Comput.*, vol. 3, no. 1, pp. 167–184, 2023, doi: 10.31181/jdaic10008102023n.
- [4] A. Theppanya and N. Netpradit, "Forecasting the used car price index using time series forecasting methods," (in Thai), *Maejo Bus. Rev.*, vol. 6, no. 2, pp. 81–95, 2024, doi: 10.14456/mbr.2024.10.
- [5] A. AlShared, "Used cars price prediction and valuation using data mining techniques," M.S. thesis, Rochester Inst. Technol., Rochester, NY, USA, 2021. [Online]. Available: <https://repository.rit.edu/theses/11086>
- [6] V. Singrodia, A. Mitra, and S. Paul, "A review on web scrapping and its applications," in *Proc. Int. Conf. Comput. Commun. Inform. (ICCCI)*, Coimbatore, India, 2019, pp. 1–6, doi: 10.1109/ICCCI.2019.8821809.
- [7] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, 2021, doi: 10.1613/jair.1.12228.
- [8] B. Rolf et al., "A review on unsupervised learning algorithms and applications in supply chain management," *Int. J. Prod. Res.*, vol. 63, no. 5, pp. 1933–1983, 2025, doi: 10.1080/00207543.2024.2390968.
- [9] L.-Z. Guo, L.-H. Jia, J.-J. Shao, and Y.-F. Li, "Robust semi-supervised learning in open environments," *Front. Comput. Sci.*, vol. 19, 2025, Art. no. 198345, doi: 10.1007/s11704-024-40646-w.
- [10] H. Xie et al., "Reinforcement learning for vehicle-to-grid: A review," *Advances Appl. Energy*, vol. 17, 2025, Art. no. 100214, doi: 10.1016/j.adapen.2025.100214.
- [11] J. S. Jhala and D. Anand, "Comparative analysis of supervised learning algorithms for valuating used car prices," in *Proc. Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, Gharuan, India, 2023, pp. 265–270, doi: 10.1109/InCACCT57535.2023.10141827.
- [12] R. Nuzulia, A. Misbullah, L. Farsiah, Rasudin, Husaini, and S. A. Nazhifah, "Comparative analysis of XGBoost and random forest for used car price prediction," in *Proc. Int. Conf. Elect. Eng. Inform. (ICELTICs)*, Banda Aceh, Indonesia, 2024, pp. 125–129, doi: 10.1109/ICELTICs62730.2024.10776051.
- [13] N. O. Idris, A. Achban, S. A. Utiahman, J. Karim, and F. Pontooyo, "Predicting the selling price of cars using business intelligence with the feed-forward backpropagation algorithms," in *Proc. 5th Int. Conf. Inform. Comput. (ICIC)*, Gorontalo, Indonesia, 2020, pp. 1–6, doi: 10.1109/ICIC50835.2020.9288594.
- [14] R. B. A. Supleo, R. G. De Luna, and A. C. Padilla, "Predicting used car prices in Metro Manila using artificial neural networks on web-scraped data," in *Proc. 7th Int. Conf. Inform. Comput. Sci. (ICICoS)*, Semarang, Indonesia, 2024, pp. 30–35, doi: 10.1109/ICICoS62600.2024.10636891.
- [15] F. Wang, X. Zhang, and Q. Wang, "Prediction of used car price based on supervised learning algorithm," in *Proc. Int. Conf. Netw., Commun. Inf. Technol. (NetCIT)*, Manchester, U.K., 2021, pp. 143–147, doi: 10.1109/NetCIT54147.2021.00036.



- [16] J. D. Apeko, I. O. Osunmakinde, M. M. Abdulgader, and K. C. Nwosu, "Predictive analytics on used car prices using business intelligence of Bayesian networks for sales risk reduction," in *Proc. Int. Conf. Elect., Comput. Energy Technol. (ICECET)*, Cape Town, South Africa, 2023, pp. 1–6, doi: 10.1109/ICECET58911.2023.10389200.
- [17] H. Jing, X. Ye, and S. Manoharan, "Residual value of used car analysis and prediction," in *Proc. Int. Conf. Elect. Comput. Energy Technol (ICECET)*, Cape Town, South Africa, 2023, pp. 1–6, doi: 10.1109/ICECET58911.2023.10389355.
- [18] G. Buturac, "Measurement of economic forecast accuracy: a systematic overview of the empirical literature," *J. Risk Financial Manage.*, vol. 15, no. 1, 2022, Art. no. 1, doi: 10.3390/jrfm15010001.
- [19] J. O'Trakoun, "Business forecasting during the pandemic," *Bus. Econ.*, vol. 57, no. 3, pp. 95–110, 2022.