การทำนายราคาอาคารชุดในกรุงเทพมหานคร
ด้วยวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

PRICE PREDICTION OF CONDOMINIUM IN BANGKOK
USING MACHINE LEARNING AND DEEP LEARNING METHODS

สิริวรรณ แสงวงศ์

SIRIWAN SAENGWONG

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์

ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2566

KMITL-2022-SC-M-017-016

PRICE PREDICTION OF CONDOMINIUM IN BANGKOK USING MACHINE LEARNING AND DEEP LEARNING METHODS

SIRIWAN SAENGWONG

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE

IN DATA SCIENCE AND ANALYTICS

KMITL DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2023

KMITL-2022-SC-M-017-016

หัวข้อการค้นคว้าอิสระ การทำนายราคาอาคารชุดในกรุงเทพมหานครด้วยวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

ชื่อนักศึกษา นางสาวสิริวรรณ แสงวงศ์

รหัสประจำตัว 64605113

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์)
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง

พ.ศ. 2566

อาจารย์ที่ปรึกษาการค้นคว้าอิสระ รองศาสตราจารย์สายชล สินสมบูรณ์ทอง

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้ข้อมูลจากแหล่งข้อมูล Baania ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ เลือกใช้คุณสมบัติ 2 วิธี คือ วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด และวิธีการนำตัวแปรเข้าทั้งหมด ในการเปรียบเทียบแบบจำลองการเรียนรู้ทั้งหมด 5 วิธี ซึ่งแบ่งเป็นการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting ส่วนอีก 2 วิธีที่เหลือเป็นการเรียนรู้เชิงลึก คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว เกณฑ์ที่ใช้ในการพิจารณาคือ สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ผลการวิจัยสรุปได้ว่าปัจจัยหลักที่ส่งผลกับราคาอาคารชุดในกรุงเทพมหานครคือ ปัจจัยสถานที่ตั้งของอาคารชุด และการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดได้ผลลัพธ์ดีที่สุดเมื่อเปรียบเทียบกับวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด สำหรับการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้เครื่องพบว่าประสิทธิภาพการทำนายของแบบจำลองวิธี Extreme Gradient Boosting ให้ค่าประสิทธิภาพการทำนายดีที่สุด และการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองของการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกพบว่าวิธีโครงข่ายประสาทแบบคอนโวลูชันให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการทำนายราคาอาคารชุดในกรุงเทพมหานคร

คำสำคัญ : อาคารชุด การเลือกคุณลักษณะ การเรียนรู้ของเครื่อง และการเรียนรู้เชิงลึก

Independent Study Title Price Prediction of Condominium in Bangkok Using Machine Learning and Deep Learning Methods
Student Name Siriwan Saengwong
Student ID 64605113
Degree Master of Science (Data Science and Analytics)
KMITL-Digital Analytics and Intelligence Center
Year 2023
Independent Study Advisor Assoc.Prof. Saichon Sinsomboonthong

Abstract

The objective of this research was to study and compare the effectiveness of predictive models for condominium prices in Bangkok using data from Baania, a real estate database. There were 2 feature selections for this research: best subsets selection and enter regression methods. The research compared 5 learning methods, categorized into 3 machine learning methods: random forest, hedonic price, and extreme gradient boosting methods. The remaining 2 methods are deep learning approaches: convolutional neural network and long short-term memory. The evaluation criteria used for evaluation are coefficient of determination, mean absolute error and root mean square error. The research concluded that the main factors influencing the prices of condominiums in Bangkok are the location factors and the enter algorithm which includes all variables yielded the best results compared to the best subsets selection method. In terms of predictive performance, the Extreme Gradient Boosting outperforms in machine learning method. On the other hand, the convolutional neural network method performs the best predictive performance among all models for price prediction of condominium in Bangkok when using both machine learning and deep learning methods.

Keywords: Condominium, Feature selection, Machine learning, Deep learning

กิตติกรรมประกาศ

งานวิจัยเรื่อง การทำนายราคาอาคารชุดในกรุงเทพมหานครด้วยวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก สามารถดำเนินการจนประสบความสำเร็จลุล่วงไปด้วยดี เนื่องจากได้รับความกรุณาจาก รศ.สายชล สินสมบูรณ์ทอง อาจารย์ที่ปรึกษาการค้นคว้าอิสระที่ให้คำปรึกษา ข้อเสนอแนะ เอื้อเฟื้อเอกสารต่างๆที่ใช้เป็นแนวทางในการวิเคราะห์ข้อมูลและตรวจทานแก้ไขความถูกต้องตลอดจนติดตามงานทุกๆขั้นตอนของการดำเนินงาน จนกระทั่งงานวิจัยครั้งนี้สำเร็จเรียบร้อยด้วยดี ผู้วิจัยขอขอบพระคุณด้วยความเคารพอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณ ผศ.ดร.บุษยมาส พิมพ์พรรณชาติ และรศ.ดร.ฉัฐไชย์ ลีนาวงศ์ คณะกรรมการการค้นคว้าอิสระที่ให้คำปรึกษา และคำแนะนำเพื่อความสมบูรณ์ยิ่งขึ้น

ขอขอบคุณข้อมูลจาก Baania และ Google Maps ที่เผยแพร่ข้อมูลสาธารณะที่ดีและมีประโยชน์จนทำให้งานวิจัยนี้สำเร็จ

ขอขอบคุณคณาจารย์สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้และช่วยเหลือให้คำแนะนำในเรื่องต่างๆมาโดยตลอด

ขอขอบพระคุณบิดามารดา ที่คอยให้กำลังใจและสนับสนุนผู้จัดทำการค้นคว้าอิสระมาโดยตลอด และขอขอบคุณเพื่อนคณะวิทยาศาสตร์สาขาวิทยาการข้อมูลและการวิเคราะห์ทุกท่านที่ให้ความรู้และช่วยเหลือในการวิจัย

สุดท้ายนี้ผู้วิจัยหวังว่างานวิจัยฉบับนี้จะเป็นประโยชน์สำหรับนักลงทุนหรือหน่วยงานที่เกี่ยวข้องและผู้สนใจศึกษาต่อไป

นางสาวสิริวรรณ แสงวงศ์

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการดำรงชีวิตของมนุษย์ ที่อยู่อาศัยหรือบ้านเป็นหนึ่งในปัจจัยสี่ที่จำเป็น นอกเหนือจากอาหาร เครื่องนุ่งห่ม และยารักษาโรค มนุษย์จะแสวงหาถิ่นฐานที่อยู่อาศัยที่มีความมั่นคงปลอดภัย และสามารถอำนวยความสะดวกต่อการดำรงชีวิตมาก โดยการเลือกและตัดสินใจซื้อที่อยู่อาศัยจึงเป็นเรื่องสำคัญ นอกจากจะเป็นปัจจัยพื้นฐานที่จำเป็นของชีวิต ยังเป็นเครื่องชี้วัดคุณภาพชีวิต ฐานะทางสังคมและเศรษฐกิจของประชาชนด้วย รวมถึงเป็นการกระตุ้นการออมและเป็นการลงทุนระยะยาวรูปแบบหนึ่ง ดังนั้นจึงเกิดธุรกิจอสังหาริมทรัพย์ขึ้นเพื่ออำนวยความสะดวกเรื่องที่อยู่อาศัย โดยช่วงแรกเป็นการจัดสรรที่ดินหรือการจำหน่ายที่ดินโดยมีการแบ่งแปลงย่อยๆ ต่อมาเริ่มมีการสร้างบ้านบนที่ดินที่เรียกว่า “บ้านจัดสรร” ได้แก่ บ้านเดี่ยว บ้านแฝด และบ้านแถว เป็นต้น ในส่วนสิ่งก่อสร้างเพื่ออยู่อาศัยสุดท้ายคือ อาคารชุด (Condominium)

ในปัจจุบันสภาพสังคมและเศรษฐกิจของประเทศไทยที่เปลี่ยนแปลงไปตามยุคสมัย ประชากรส่วนมากประกอบอาชีพในเมือง ทำให้ห่างไกลจากที่อยู่อาศัยเดิมมากขึ้น จึงต้องหาที่อยู่อาศัยใหม่เพื่อลดระยะเวลาในการเดินทางสำหรับประกอบอาชีพในแต่ละวัน แต่การก่อสร้างที่อยู่อาศัยด้วยตัวเองหรือซื้อบ้านจัดสรรไม่สามารถตอบสนองความต้องการของกลุ่มประชากรเหล่านี้ได้ เพราะทำให้เกิดค่าใช้จ่ายจำนวนมาก และบ้านจัดสรรเป็นทรัพย์สินที่ราคาสูง จึงต้องคำนึงถึงหลายปัจจัย เช่น ทำเลที่ตั้ง การเลือกโครงการและราคาที่เหมาะสม จากปัจจัยข้างต้นไม่สามารถตอบสนองกลุ่มประชากรที่ต้องการอยู่อาศัยชั่วคราวได้ เพื่อตอบสนองความต้องการของประชากรที่ต้องการที่อยู่อาศัยชั่วคราวสำหรับลดระยะเวลาในการเดินทางเพื่อประกอบภารกิจในแต่ละวัน หรือสำหรับประชากรที่ต้องการลดค่าใช้จ่ายในการเดินทาง ดังนั้นอาคารชุดเป็นตัวเลือกหนึ่งของที่อยู่อาศัยสำหรับกลุ่มประชากรกลุ่มนี้ เพราะทำเลดี ราคาถูกกว่าซื้อบ้านจัดสรร มีระบบขนส่งสาธารณะที่เข้าถึง มีระบบรักษาความปลอดภัยที่เข้มงวด และมีสิ่งอำนวยความสะดวก ส่งผลให้อาคารชุดมีการเติบโตอย่างรวดเร็วและราคาขายเฉลี่ยต่อตารางเมตรของอาคารชุดมีการเพิ่มสูงขึ้น ทำให้ครึ่งปีแรกของปี 2565 สัดส่วนของยอดขายที่อยู่อาศัยใหม่ร้อยละ 60 ของประเภทอสังหาริมทรัพย์เพื่ออยู่อาศัยในกรุงเทพมหานครคือ อาคารชุด (ศูนย์ข้อมูลอสังหาริมทรัพย์, 2565) สาเหตุส่วนหนึ่งเกิดจากความต้องการของประชากร ราคาที่ดินที่ปรับตัวสูงขึ้น ทั้งจากการประเมินราคาที่ดินรายปี เศรษฐกิจ และการพัฒนาระบบขนส่งสาธารณะ ดังนั้นการซื้ออาคารชุดในกรุงเทพมหานครเพื่อการลงทุนเป็นตัวเลือกการลงทุนที่น่าสนใจ ด้วยอัตราการเติบโตของมูลค่าสินทรัพย์และรายได้ที่เกิดจากค่าเช่า เพื่อที่เข้าใจมูลค่าแท้จริงของอาคารชุดในกรุงเทพมหานครหรือสามารถ

ทำนายราคาอาคารชุดได้อย่างมีประสิทธิภาพ และทราบถึงปัจจัยหรือคุณลักษณะที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร ทางผู้วิจัยจึงต้องใช้ขั้นตอนวิธี (Algorithm) การเรียนรู้ของเครื่อง (Machine Learning) มาช่วยในการทำนายราคาอาคารชุดในกรุงเทพมหานครให้มีประสิทธิภาพ

Baldominos et al. (2018) ได้ศึกษาเกี่ยวกับปัจจัยสำคัญในการเติบโตของอสังหาริมทรัพย์โดยใช้การเรียนรู้ของเครื่อง กล่าวถึงข้อมูลของอสังหาริมทรัพย์ในเขตซาลามังกาของเมืองมาดริด ประเทศสเปน โดยเปรียบเทียบด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว (K Nearest Neighbor) วิธีโครงข่ายประสาทเทียม (Artificial Neural Networks) วิธีป่าสุ่ม (Random Forest) และวิธีซัพพอร์ตเวกเตอร์การถดถอย (Support Vector Regression) วัดประสิทธิภาพโดยใช้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error) ผลการศึกษาพบว่าวิธีป่าสุ่มให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำที่สุดคือ 338,715 รองลงมาคือวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีซัพพอร์ตเวกเตอร์การถดถอย และวิธีโครงข่ายประสาทเทียม ตามลำดับ

Piao et al. (2019) ได้ศึกษาเกี่ยวกับการพยากรณ์ราคาที่อยู่อาศัยในต้าเหลียน ประเทศจีน โดยใช้ข้อมูลจากศูนย์ข้อมูลทรัพยากรที่ดินและที่อยู่อาศัยสำหรับเมืองต้าเหลียนที่เป็นฐานข้อมูลภายใต้การจัดการรัฐบาลกลางของสาธารณรัฐประชาชนจีน มีจำนวนโครงการทั้งหมด 171,155 โครงการ เป็นข้อมูลเก็บรวบรวมตั้งแต่ปี ค.ศ. 2013 ถึง 2017 ซึ่งเปรียบเทียบด้วยวิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network) กับวิธี Extreme Gradient Boosting ทำการวัดประสิทธิภาพโดยใช้ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error) และสัมประสิทธิ์การกำหนด (Correlation of Determination) ผลการศึกษาพบว่าวิธีโครงข่ายประสาทแบบคอนโวลูชันให้สัมประสิทธิ์การกำหนดสูงที่สุดคือ 0.9868 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดคือ 0.01057 รองลงมาคือวิธี Extreme Gradient Boosting ให้สัมประสิทธิ์การกำหนดคือ 0.9650 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.1040

ในปีเดียวกัน Jafari and Akhavian (2019) ได้ศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อราคาที่อยู่อาศัยในประเทศสหรัฐอเมริกา โดยใช้ข้อมูลจากสำรวจแบบสอบถามที่อยู่อาศัยจำนวน 13,771 หลัง ซึ่งเปรียบเทียบด้วยวิธีการประเมินราคาแอบแฝง (Hedonic Pricing Method) โดยการเลือกตัวแปรอิสระด้วยวิธีการถดถอยทีละขั้น (Stepwise Regression) และวิธีการประเมินราคาแอบแฝงด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Regression) วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนดผล การศึกษาพบว่าวิธีการประเมินราคาแอบแฝงด้วยการถดถอยชุดข้อมูลย่อยที่ดีที่สุดได้สัมประสิทธิ์การกำหนดคือ 0.4073 ซึ่งดีกว่าวิธีการประเมินราคาแอบแฝงของลักษณะเชิงคุณภาพด้วยการเลือกตัวแปรด้วยวิธีการถดถอยทีละขั้น

ต่อมา พสธร (2563) ได้ศึกษาเกี่ยวกับการประเมินราคาเสนอขายห้องชุดในประเทศไทย ด้วยการเรียนรู้เชิงลึกด้วยวิธีการแบ่งกลุ่มข้อมูลเฉลี่ย k กลุ่ม (K-means Clustering) ซึ่งเปรียบเทียบด้วยวิธีหน่วยความจำระยะสั้น-ยาว (Long Short-Term Memory) โดยใช้ข้อมูลจากวิธีการแบ่งกลุ่มข้อมูล

เฉลี่ย k กลุ่มกับวิธีหน่วยความจำระยะสั้น-ยาว โดยใช้ข้อมูลของอาคารชุดโดยใช้รัศมี 5 กิโลเมตร ทำการวัดประสิทธิภาพโดยใช้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Squared Error) ผลการศึกษาพบว่าวิธีหน่วยความจำระยะสั้น-ยาวโดยใช้ข้อมูลจากวิธีการแบ่งกลุ่มข้อมูลเฉลี่ย k กลุ่มได้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดคือ 9,722 บาท ซึ่งดีกว่าวิธีหน่วยความจำระยะสั้น-ยาวโดยใช้ข้อมูลของอาคารชุดโดยใช้รัศมี 5 กิโลเมตร ให้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยคือ 10,938.34 บาท ถึงแม้ผลการทดลองข้อมูลที่แบ่งกลุ่มเฉลี่ย k กลุ่มดีกว่า แต่ พสธร ได้เสนอแนะว่าเมื่อนำข้อมูลของอาคารชุดโดยใช้รัศมี 5 กิโลเมตร ไปใช้งานจริงสามารถทำได้รวดเร็วกว่า จึงเหมาะสมที่จะประยุกต์นำไปใช้มากกว่า ในปีเดียวกันนั้น Pensri et al. (2020) ได้ศึกษาปัจจัยที่ส่งผลต่อการซื้ออาคารชุดในกรุงเทพมหานคร โดยทำแบบสำรวจและเก็บรวบรวมข้อมูลจำนวน 385 คน วัดประสิทธิภาพด้วยสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient) ผลการศึกษาพบว่า ปัจจัยสถานที่ตั้งของอาคารชุดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันสูงที่สุดคือ 0.898 รองลงมาคือปัจจัยส่วนลดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันคือ 0.721 และปัจจัยราคาของอาคารชุดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันคือ 0.625

ในปีถัดมา Viriya (2021) ได้ศึกษาเกี่ยวกับความสัมพันธ์ระหว่างสิ่งอำนวยความสะดวกรอบอาคารชุดและราคาอาคารชุดด้วยการเรียนรู้ของเครื่อง โดยใช้ชุดข้อมูลราคาอาคารชุดจำนวน 500 โครงการ ในกรุงเทพมหานคร ซึ่งเปรียบเทียบด้วยวิธี Extreme Gradient Boosting และวิธีป่าสุ่ม (Random Forest) วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนด ผลการศึกษาพบว่าวิธี Extreme Gradient Boosting ให้สัมประสิทธิ์การกำหนดสูงที่สุดคือ 0.73 ซึ่งดีกว่าวิธีป่าสุ่มให้สัมประสิทธิ์การกำหนดเพียง 0.66

Guliker et al. (2022) ได้ศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อการประเมินอสังหาริมทรัพย์ในประเทศเนเธอร์แลนด์ ด้วยการเรียนรู้ของเครื่อง โดยใช้ชุดข้อมูลจากการประเมินมูลค่าอสังหาริมทรัพย์ของ 5 เขตใหญ่ ในประเทศเนเธอร์แลนด์ จาก Stater N.V. ที่เป็นบริษัทให้สินเชื่อที่อยู่อาศัยใหญ่ที่สุดในเนเธอร์แลนด์ โดยเปรียบเทียบด้วยวิธีการวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) วิธีการถดถอยแบบถ่วงน้ำหนักทางภูมิศาสตร์ (Geographically Weighted Regression) และวิธี Extreme Gradient Boosting วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนด รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error) ผลการศึกษาพบว่าวิธี Extreme Gradient Boosting ให้สัมประสิทธิ์การกำหนดสูงที่สุดเท่ากับ 0.832 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 65,312 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 43,625 และร้อยละของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 6.35% รองลงมาคือวิธีการวิเคราะห์การถดถอยเชิงเส้น และวิธีการถดถอยแบบถ่วงน้ำหนักทางภูมิศาสตร์ ตามลำดับ

ในปีถัดมา Rupesh and Kumar (2023) ได้ศึกษาเกี่ยวกับการเพิ่มประสิทธิภาพของแบบจำลองสำหรับการทำนายราคาค่าเช่าบ้านหรืออพาร์ตเมนต์ในประเทศอินเดีย โดยใช้ข้อมูลราคาบ้านเช่าที่อยู่อาศัยตั้งแต่ปี ค.ศ. 2016 ถึง 2019 ที่มีข้อมูลทั้งหมด 1,100 โครงการ บ้านหรืออพาร์ตเมนต์จากข้อมูลเปิดใน Kaggle ซึ่งเปรียบเทียบด้วยวิธีโครงข่ายประสาทแบบคอนโวลูชันกับวิธีต้นไม้ตัดสินใจ (Decision Tree) วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนด ผลการศึกษาพบว่าวิธีโครงข่ายประสาทแบบคอนโวลูชันให้สัมประสิทธิ์การกำหนดสูงที่สุดคือ 0.9632 ซึ่งดีกว่าวิธีต้นไม้ตัดสินใจให้สัมประสิทธิ์การกำหนดเพียง 0.9415

ดังนั้นผู้วิจัยมีความสนใจในการทำนายราคาอาคารชุดในกรุงเทพมหานครโดยเปรียบเทียบแบบจำลองทั้งหมด 5 วิธี โดยแบ่งเป็นแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่มจากงานวิจัยของ Baldominos et al.(2018) วิธีการประเมินราคาแอบแฝงจากงานวิจัยของ Jafari and Akhavian (2019) และวิธี Extreme Gradient Boosting จากงานวิจัยของ Viriya (2021) และ Guliker et al. (2022) และแบบจำลองการเรียนรู้เชิงลึก 2 วิธี คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน จากงานวิจัยของ Piao et al. (2019) และ Rupesh and Kumar (2023) และวิธีหน่วยความจำระยะสั้น-ยาวจากงานวิจัยของ พสธร (2563) โดยเลือกคุณลักษณะของแบบจำลองในแต่ละชุดแตกต่างกัน 2 วิธี โดย 1 วิธีจากงานวิจัยของ Jafari and Akhavian (2019) คือ วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด นอกจากนี้ผู้วิจัยสนใจเสนอการเลือกคุณลักษณะของแบบจำลองอีก 1 วิธี คือ วิธีการนำตัวแปรเข้าทั้งหมด (Enter) โดยเปรียบเทียบประสิทธิภาพแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกทั้งหมด 5 วิธี ด้วยสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย สำหรับการศึกษาปัจจัยที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร วัดประสิทธิภาพด้วยสัมประสิทธิ์สหสัมพันธ์เพียร์สัน จากงานวิจัยของ Pensri et al. (2020)

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อศึกษาแบบจำลองที่ช่วยลดความเสี่ยงในการลงทุนกับอาคารชุดในกรุงเทพมหานครโดยเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

2) เพื่อศึกษาปัจจัยที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร

1.3 ขอบเขตของงานวิจัย

ข้อมูลจาก Baania (2565) ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ โดยสนใจเฉพาะอาคารชุดในกรุงเทพมหานคร ที่มี 2,403 โครงการ จากงานวิจัยของ Baldominos et al. (2018), Viriya (2021) และ Guliker et al. (2022) ทำการแบ่งข้อมูลเป็น 2 ชุด โดยมีสัดส่วนของชุดข้อมูลฝึกสอน (Training Data) 80% และชุดข้อมูลทดสอบ (Testing Data) 20% เมื่อได้ชุดข้อมูลฝึกสอนจะใช้วิธีการตรวจสอบไขว้ 5 ชุด (5-Fold Cross Validation) เพื่อแบ่งข้อมูลออกเป็น 5 ชุดเท่าๆกัน แล้วใช้ 4 ชุดสำหรับชุดข้อมูลฝึกสอน และ 1 ชุดสำหรับชุดข้อมูลการตรวจสอบความถูกต้อง (Validation Data) สำหรับการทดสอบแบบจำลองการเรียนรู้ของเครื่องจากงานวิจัยที่ผ่านมาแล้ว 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting เนื่องจากพบว่าประสิทธิภาพของแบบจำลองของการเรียนรู้ของเครื่องไม่ดีเท่าที่ควร จึงเพิ่มการศึกษาแบบจำลองของการเรียนรู้เชิงลึก เพื่อเพิ่มประสิทธิภาพให้มากขึ้น โดยแบบจำลองการเรียนรู้เชิงลึกมี 2 วิธี คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว ทำการแบ่งข้อมูลเป็น 2 ชุด โดยมีสัดส่วนของชุดข้อมูลฝึกสอน 80% และชุดข้อมูลทดสอบ 20% โดยมีการเลือกคุณลักษณะของแบบจำลองทั้งหมด 2 วิธีคือ วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดจากงานวิจัยที่ผ่านมา และวิธีการนำตัวแปรเข้าทั้งหมดเป็นวิธีที่ผู้วิจัยสนใจเสนอ ซึ่งเปรียบเทียบประสิทธิภาพแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกทั้งหมด 5 วิธี ด้วยสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย สำหรับการศึกษาปัจจัยที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร วัดประสิทธิภาพด้วยสัมประสิทธิ์สหสัมพันธ์เพียร์สัน จากงานวิจัยของ Pensri et al. (2020)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) เพื่อได้แบบจำลองที่เหมาะสมในการลงทุนกับอาคารชุดในกรุงเทพมหานคร โดยใช้การเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

2) เพื่อทราบถึงปัจจัยหรือคุณลักษณะที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร

1.5 นิยามคำศัพท์

1) อาคารชุด (Condominium) หมายถึง อาคารที่บุคคลสามารถแยกการถือกรรมสิทธิ์ออกได้เป็นส่วนๆ โดยแต่ละส่วนประกอบด้วยกรรมสิทธิ์ในทรัพย์ส่วนบุคคลและกรรมสิทธิ์ร่วมในทรัพย์ส่วนกลาง

2) การเลือกคุณลักษณะ (Feature Selection) หมายถึง การลดขนาดหรือมิติของข้อมูลและยังคงลักษณะสำคัญของข้อมูล

3) การเรียนรู้ของเครื่อง (Machine Learning) หมายถึง สาขาหนึ่งของปัญญาประดิษฐ์ที่พัฒนามาจากการศึกษาการรู้จำแบบ เกี่ยวข้องกับการศึกษาและการสร้างขั้นตอนวิธีที่สามารถเรียนรู้ข้อมูลและทำนายข้อมูลได้ ขั้นตอนวิธีนั้นจะทำงานโดยอาศัยแบบจำลองที่สร้างมาจากชุดข้อมูลตัวอย่างเพื่อการทำนายหรือตัดสินใจ การเรียนรู้ของเครื่องต้องอาศัยวิธีการทางสถิติศาสตร์เป็นอย่างมาก

4) การเรียนรู้เชิงลึก (Deep Learning) หมายถึง สาขาหนึ่งของการเรียนรู้ของเครื่อง พื้นฐานของการเรียนรู้เชิงลึกคือ ขั้นตอนวิธีที่พยายามจะสร้างแบบจำลองเพื่อแทนความหมายของข้อมูลในระดับสูงโดยการสร้างสถาปัตยกรรมข้อมูลขึ้นมาที่ประกอบไปด้วยโครงสร้างย่อยๆหลายอัน

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เนื้อหาในบทนี้กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการศึกษา ประกอบด้วย 6 ส่วนคือ การจัดเตรียมข้อมูล การเลือกคุณลักษณ การเรียนรู้ของเครื่อง การเรียนรู้เชิงลึก มาตรวัดประสิทธิภาพ และงานวิจัยที่เกี่ยวข้อง

2.1 การจัดเตรียมข้อมูล (Data Preprocessing)

2.1.1 การเก็บรวบรวมข้อมูล (Data Collection)

ข้อมูลจาก Baania (2565) ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ โดยสนใจเฉพาะอาคารชุดในกรุงเทพมหานคร ที่มี 2,403 โครงการ และเก็บข้อมูลตำแหน่งละติจูดและลองจิจูด (Latitude, Longitude) ที่สำคัญของกรุงเทพมหานครจาก Google ได้แก่ สถานีรถไฟฟ้าบนดินและใต้ดินจำนวน 171 ตำแหน่ง โรงพยาบาลจำนวน 51 ตำแหน่ง มหาวิทยาลัยจำนวน 51 ตำแหน่ง และสถานที่ท่องเที่ยวสำคัญจำนวน 52 ตำแหน่ง ด้วยวิธีหาระยะทางยุคลิด (Euclidean Distance) เป็นวิธีหนึ่งในการวัดระยะห่างระหว่างจุดสองจุดในระบบพิกัดสองมิติหรือสูงกว่านั้น เราสามารถคำนวณค่าระยะทางยุคลิดได้โดยใช้สูตรทางคณิตศาสตร์ ซึ่งจะบอกถึงระยะห่างระหว่างจุดสองจุดดังนี้ (Paul, 2005)

สำหรับจุดสองจุด $\mathrm{p}(\mathrm{p}_1, \mathrm{p}_2)$ และ $\mathrm{q}(\mathrm{q}_1, \mathrm{q}_2)$ ในระบบพิกัดสองมิติ ค่าระยะทางยุคลิด คือ

$$\mathrm{d}(\mathrm{p}, \mathrm{q}) = \sqrt{(\mathrm{q}_1 - \mathrm{p}_1)^2 + (\mathrm{q}_2 - \mathrm{p}_2)^2} \qquad (2.1)$$

เมื่อ $\mathrm{d}(\mathrm{p}, \mathrm{q})$ คือระยะห่างระหว่างจุด p และ q ในระบบพิกัดสองมิติ และ √ คือเครื่องหมายรากที่สองในการประยุกต์ใช้งาน ค่าระยะทางยุคลิดสามารถนำมาใช้ในการวัดระยะห่างระหว่างองค์ประกอบต่างๆ เช่น ในการคำนวณความคล้ายคลึงของแบบจำลอง หรือในการหาระยะห่างของข้อมูลจากจุดศูนย์กลาง (Centroid) ในการวิเคราะห์ข้อมูลแบบกลุ่ม (Clustering)

2.1.2 การแทนค่าข้อมูลสูญหาย (Missing Values Replacement)

1) ค่าเฉลี่ยอนุกรม (Series Mean) คือ การแทนที่ค่าสูญหายด้วยค่าเฉลี่ยทั้งชุดข้อมูล (ฐณัฐ, 2559)

$$\hat{X} = \frac{\sum_{i=1}^{k} X_i}{k} \quad (2.2)$$

โดยที่ $\hat{X}$ คือ ค่าประมาณข้อมูลสูญหายที่ถูกประมาณด้วยค่าเฉลี่ยของข้อมูลที่ไม่สูญหาย

X_i คือ ข้อมูลที่ไม่สูญหาย

k คือ จำนวนข้อมูลที่ไม่สูญหาย

2) ค่าเฉลี่ยของค่าใกล้เคียง (Mean of Nearby Points) คือ การแทนที่ค่าสูญหายด้วยค่าเฉลี่ยของค่าที่อยู่ใกล้เคียง ช่วงของจุดใกล้เคียงคือค่าที่อยู่ด้านบนและด้านล่างของค่าที่สูญหายไปจะใช้ในการคำนวณค่าเฉลี่ย

$$\hat{X}_p = \frac{L+U}{2} \quad (2.3)$$

โดยที่ $\hat{X}_p$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p ด้วยค่าเฉลี่ยรอบจุด

L คือ ค่าเฉลี่ยของจุดจำนวน a จุด ที่มีตำแหน่งต่ำกว่าจุด p

U คือ ค่าเฉลี่ยของจุดจำนวน a จุด ที่มีตำแหน่งสูงกว่าจุด p

3) ค่ามัธยฐานของค่าใกล้เคียง (Median of Nearby Points) คือ การแทนที่ค่าสูญหายด้วยค่ามัธยฐานของค่าที่อยู่ใกล้เคียง ช่วงของจุดใกล้เคียงคือค่าที่อยู่ด้านบนและด้านล่างของค่าที่สูญหายไปจะใช้ในการคำนวณค่ามัธยฐาน

$$\hat{X}_p = \frac{T_a+T_{a+1}}{2} \quad (2.4)$$

โดยที่ $\hat{X}_p$ คือ ค่าประมาณของข้อมูลสูญหายที่ตำแหน่ง p ด้วยค่ามัธยฐานรอบจุด

T_a คือ ค่าที่อยู่ก่อนตำแหน่งของมัธยฐาน

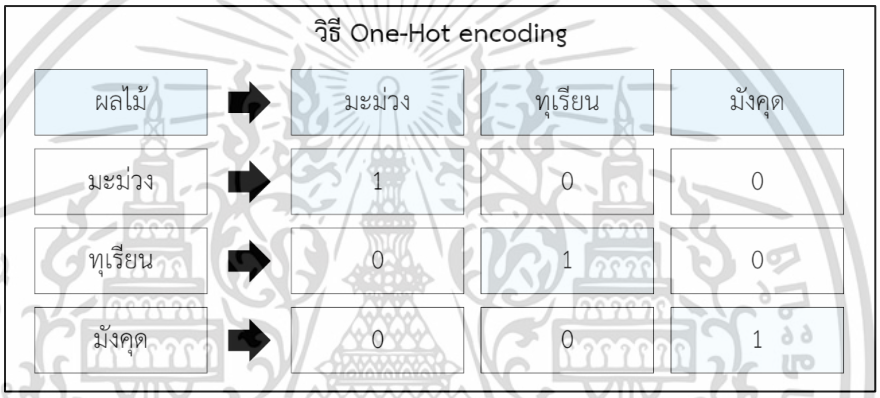
T_{a+1} คือ ค่าที่อยู่หลังตำแหน่งของมัธยฐาน

4) ค่าฐานนิยม (Mode) คือ ค่าของข้อมูลในชุดใดชุดหนึ่ง ซึ่งมีความถี่สูงที่สุดหรือซ้ำกันมากที่สุด ซึ่งการแทนที่ค่าสูญหายด้วยค่าฐานนิยมมักจะใช้กับข้อมูลประเภทเชิงคุณภาพ

2.1.3 การแปลงข้อมูล (Data Transformation)

การทำงานของแบบจำลองส่วนมากทำงานบนตัวเลข ดังนั้นเมื่อมีข้อมูลที่เป็นเชิงคุณภาพทั้งที่มีลำดับหรือไม่มีลำดับ ต้องเปลี่ยนให้อยู่ในรูปแบบของค่าทวิภาค (Binary Values) ที่มีค่า 0 หรือ 1 เท่านั้น การแปลงข้อมูลแบบนี้จะเรียกว่า One-Hot Encoding (Sasiwut, 2020)

วิธี One-Hot Encoding เป็นกระบวนการแปลงข้อมูลจากรูปแบบหนึ่งเป็นรูปแบบอื่น โดยเฉพาะอย่างยิ่งในการแปลงข้อมูลจำแนกประเภท (Categorical Data) เช่น การแปลงค่าของข้อมูลสีจากชื่อสีเป็นตัวเลข โดยที่แต่ละค่าของข้อมูลจำแนกประเภท จะถูกแปลงเป็นเวกเตอร์ (Vector) ของ 0 และ 1



รูปที่ 2.1 วิธี One-Hot Encoding

จากรูปที่ 2.1 ข้อมูลของผลไม้ โดยมีค่าที่เป็นไปได้คือ มะม่วง ทุเรียน และมังคุด เมื่อแปลงเป็นรูปแบบ One-Hot Encoding จะได้ข้อมูลตัวแปรเพิ่มเป็น 3 ตัวแปรคือ มะม่วง ทุเรียน และมังคุด โดยภายใต้ตัวแปรมีค่า 0 หรือ 1 เท่านั้น ซึ่ง 1 จะแทนค่าของค่าที่เป็นไปได้ และ 0 จะแทนค่าของค่าที่ไม่เป็นไปได้ เช่น มะม่วงจะมีค่าเป็น 1 0 0 ทุเรียนจะมีค่าเป็น 0 1 0 และมังคุดจะมีค่าเป็น 0 0 1 เป็นต้น

ดังนั้นเมื่อเราแปลงข้อมูลเชิงคุณภาพด้วย One-Hot Encoding จะช่วยให้แบบจำลองสามารถเรียนรู้และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพมากยิ่งขึ้น โดยไม่เกิดการสับสนกับค่าของข้อมูลเชิงคุณภาพที่มีความหมายต่างกัน และสามารถเรียนรู้ความสัมพันธ์ระหว่างข้อมูลได้อย่างถูกต้องและเหมาะสม

2.1.4 การแบ่งข้อมูล (Splitting Data)

ทำการแบ่งข้อมูลเป็น 2 ชุด โดยมีสัดส่วนของชุดข้อมูลฝึกสอน (Training Data) 80% และชุดข้อมูลทดสอบ (Testing Data) 20% เมื่อได้ชุดข้อมูลฝึกสอนจะใช้วิธีการตรวจสอบไขว้ 5 ชุด (5-Fold Cross Validation) เพื่อแบ่งข้อมูลออกเป็น 5 ชุดเท่าๆกัน แล้วใช้ 4 ชุดสำหรับชุดข้อมูลฝึกสอน และ 1 ชุดสำหรับชุดข้อมูลการตรวจสอบความถูกต้อง (Validation Data)

วิธีการตรวจสอบไขว้ k ชุดจากงานวิจัยของ Baldominos et al. (2018), Viriya (2021) และ Guliker et al. (2022) คือ วิธีการในการคาดการค่าความผิดพลาดของแบบจำลองหรือวิธีการที่นำเสนอโดยพื้นฐานของวิธีการสุ่มตัวอย่าง โดยเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำการตรวจสอบไขว้มักถูกใช้เป็นตัวเลือกในการกำหนดแบบจำลอง เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network Architecture) แบบจำลองในการจำแนก (Classification Model) นั้นจะต้องมีการแบ่งข้อมูลออกเป็นชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบ ทำให้ผลการแบ่งกลุ่มนั้นดีเกินจริง ดังนั้นจะมีการคิดวิธีการตรวจสอบไขว้ k ชุด เพื่อมาแก้ปัญหาการตรวจสอบความถูกต้องของการเรียนรู้ สามารถสังเกตได้จากค่าความแม่นยำหรือค่าความผิดพลาดที่ได้จากการทำการตรวจสอบไขว้ระหว่างการแบ่งชุดทดสอบในการเรียนรู้ ในที่นี้ผู้วิจัยเลือกพิจารณาความแม่นยำซึ่งเป็นสัดส่วนร้อยละของการคัดแยกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยใช้การตรวจสอบไขว้ k ชุด ให้ทำการสุ่มตัวอย่างข้อมูลไปเป็นชุดข้อมูลฝึกสอน และชุดข้องมูลทดสอบแบบสลับกัน โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดข้อมูลทดสอบเพื่อจำลองผลลัพธ์จากการตรวจสอบไขว้ การแบ่งชุดข้อมูลทุกสอบแบบนี้มักถูกใช้เป็นตัวเลือกในการกำหนดแบบจำลอง โดยสังเกตจากผลของทดสอบในแต่ละครั้งของการปรับค่าพารามิเตอร์แต่ละแบบจำลอง (ทรงศักดิ์, 2554)

วิธีการตรวจสอบไขว้ k ชุดเป็นการแบ่งข้อมูลออกเป็น k ชุด เท่าๆกัน และทำการคำนวณค่าความผิดพลาด k รอบ โดยแต่ละรอบการคำนวณ ข้อมูลชุดหนึ่งจากข้อมูล k ชุดจะถูกเลือกออกมาเพื่อเป็นชุดข้อมูลทดสอบ และข้อมูลอีก k-1 ชุดจะถูกใช้เป็นชุดข้อมูลฝึกสอน แล้วนำผลความถูกต้องหรือค่าความผิดพลาดของแต่ละรอบมารวมกันและหาค่าเฉลี่ย เพื่อเป็นค่าสะท้อนประสิทธิภาพของการฝึกสอน โดยผู้วิจัยกำหนดค่า k = 5 หมายถึง การนำข้อมูลทั้งหมดมาแบ่งเป็น 5 ชุด และจะดำเนินการฝึกสอนและทดสอบจำนวน 5 รอบ ดังรูปที่ 2.2

รูปที่ 2.2 วิธีการตรวจสอบไขว้ 5 ชุด

รูปที่ 2.2 แสดงวิธีการตรวจสอบไขว้ 5 ชุด โดยในแต่ละรอบจะเลือกข้อมูลมา 1 ชุดไม่ซ้ำกันมาเป็นชุดข้อมูลทดสอบ ด้วยการเรียนรู้ของชุดข้อมูลฝึกสอนที่เหลือ เมื่อทำครบ 5 รอบ หาผลรวมของค่าความผิดพลาดหรือความแม่นยำจากชุดทดสอบในแต่ละรอบแล้วหารด้วยจำนานรอบ ผลที่ได้เป็นค่าเฉลี่ยของความผิดพลาดในการเรียนรู้ของแบบจำลอง ข้อดีของวิธีการนี้ คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกฝึกสอนทั้งหมด k-1 ครั้ง โดยในขั้นตอนเหล่านี้สามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการคำนวณเป็นจำนวนรอบเท่าใด ซึ่งเหมาะสำหรับการประมวลผลทดสอบกับข้อมูลที่มีจำนวนมาก

2.2 การเลือกคุณลักษณะ (Feature Selection)

2.2.1 วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Selection)

ชุดข้อมูลคือชุดของตัวแปรอิสระ เมื่อมีตัวแปรอิสระจำนวน p ตัว การค้นหาชุดข้อมูลเป็นการค้นหาปริภูมิสถานะ (State space) ซึ่งแต่ละสถานะจะแสดงชุดข้อมูลของตัวแปรอิสระ โดยปริภูมิสถานะของชุดข้อมูลทั้งหมดที่เป็นไปได้คือ 2^p ชุดข้อมูล เช่น เมื่อมีตัวแปรอิสระจำนวน 5 ตัว ชุดข้อมูลทั้งหมดที่เป็นไปได้จะเท่ากับ $2^5 = 32$ ชุดข้อมูล ประกอบด้วย

1. ชุดข้อมูลที่ประกอบด้วยตัวแปรอิสระ 1 ตัวแปร
2. ชุดข้อมูลที่ประกอบด้วยตัวแปรอิสระ 2 ตัวแปร
3. ชุดข้อมูลที่ประกอบด้วยตัวแปรอิสระ 3 ตัวแปร
4. ชุดข้อมูลที่ประกอบด้วยตัวแปรอิสระ 4 ตัวแปร
5. ชุดข้อมูลที่ประกอบด้วยตัวแปรอิสระทั้งหมด

จำนวนชุดข้อมูลทั้งหมดที่เป็นไปได้จะเพิ่มสูงขึ้นตามการเพิ่มขึ้นของจำนวนตัวแปรอิสระ ดังนั้น การค้นหาชุดข้อมูลอาจเป็นไปได้ยาก ชุดข้อมูลที่ดีที่สุดสามารถพิจารณาได้จากเกณฑ์ทางสถิติ เช่น สัมประสิทธิ์การกำหนดปรับแก้ (Adjusted r^2) เกณฑ์สารสนเทศของอะกะอิเกะ (AIC) เกณฑ์สารสนเทศของเบส์ (BIC) และการวิเคราะห์ความแปรปรวน (F-test) เป็นต้น โดยในงานวิจัยนี้จะใช้การวิเคราะห์ความแปรปรวน (F-test) วัดความสำคัญระหว่างตัวแปรอิสระและตัวแปรตาม

$$F = \frac{MSR}{MSE} \tag{2.5}$$

โดยที่ MSR คือ ค่าเฉลี่ยกำลังสองของการถดถอย

MSE คือ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย

สำหรับประเด็นสำคัญที่ควรทราบจากการวิเคราะห์ชุดข้อมูลทั้งหมดที่เป็นไปได้ คือ อาจมีแบบจำลองมากกว่า 1 แบบจำลองที่ขัดแย้งกัน สำหรับทุกขนาดชุดข้อมูลด้วยความแตกต่างเพียงเล็กน้อยของเกณฑ์การเลือกชุดข้อมูลที่ดีที่สุดสำหรับ 2 ชุดข้อมูล หรือ 3 ชุดข้อมูลที่ดีที่สุด จึงมีความเป็นไปได้ที่จะต้องพิจารณาพฤติกรรมของส่วนเหลือ (Residual) หรือความรู้ก่อนหน้าของความสำคัญของตัวแปร (พรชิตา, 2560)

2.2.2 วิธีการนำตัวแปรเข้าทั้งหมด (Enter)

ธิดาเดียว (2559) อธิบายความหมายของวิธีการนำตัวแปรเข้าทั้งหมดเป็นวิธีการเอาตัวแปรอิสระทุกตัวทั้งตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรอิสระอย่างมีนัยสำคัญทางสถิติและไม่มีนัยสำคัญทางสถิติเข้าไปวิเคราะห์ในสมการถดถอย ซึ่งการเลือกตัวแปรอิสระด้วยวิธีการนำตัวแปรเข้าทั้งหมดสรุปเป็นขั้นตอนดังนี้

1. จากตัวแปรอิสระทั้งหมด k ตัว ที่คาดว่ามีความสามารถในการทำนายค่าของตัวแปรตาม นำตัวแปรอิสระทีละ 1 ตัวแปร สร้างแบบจำลองการถดถอยด้วยวิธีกำลังสองต่ำที่สุด แล้วเพิ่มตัวแปรอิสระเข้าไปในแบบจำลองการถดถอยทีละ 1 ตัวแปร ได้แบบจำลองการถดถอยทั้งหมด 2^{k-1} แบบจำลอง

2. คำนวณเกณฑ์ต่างๆ ของแต่ละแบบจำลองการถดถอย ได้แก่ ค่าความคลาดเคลื่อนมาตรฐาน สัมประสิทธิ์การกำหนดพหุคูณ สัมประสิทธิ์การกำหนดพหุคูณปรับแก้ และค่า Mallow's โดยสามารถเขียนตารางหรือกราฟแสดงค่าเกณฑ์ต่างๆ ของแต่ละแบบจำลองการถดถอยเพื่อสะดวกแก่การเปรียบเทียบและการเลือก

3. จากตารางหรือกราฟแสดงค่าเกณฑ์ในขั้นตอนที่ 2 พิจารณาว่ามีแบบจำลองการถดถอยใดบ้างที่อยู่ในข่ายของการถูกเลือก อาจพิจารณาจากเกณฑ์เดียวหรือใช้หลายเกณฑ์ประกอบกัน แบบจำลองที่เลือกอาจมีแบบจำลองเดียวหรือหลายแบบจำลอง ขึ้นอยู่กับความเหมาะสมในการใช้งาน เช่น เมื่อใช้เกณฑ์ของสัมประสิทธิ์การกำหนดพหุคูณจะเลือกแบบจำลองที่ให้สัมประสิทธิ์การกำหนด

พหุคูณสูงสุด หากใช้เกณฑ์ค่าความคลาดเคลื่อนมาตรฐานจะเลือกแบบจำลองที่ให้ค่าความคลาดเคลื่อนมาตรฐานต่ำสุดหรือใกล้เคียงต่ำสุด เป็นต้น จุดบกพร่องของวิธีการนี้คือทำให้ได้แบบจำลองที่จะพิจารณาค่อนข้างหลากหลาย เป็นการเลือกตัวแปรเข้าสู่สมการถดถอยที่ไม่เหมาะสม

2.3 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง คือ การออกแบบโปรแกรมให้สามารถเรียนรู้และพัฒนาตนเองได้จากประสบการณ์ หลักการของการเรียนรู้ของเครื่องคือการนำข้อมูลชุดฝึกสอนและผลลัพธ์มาป้อนเข้าไปให้กับคอมพิวเตอร์เพื่อสอนให้คอมพิวเตอร์เรียนรู้และทำให้เกิดการพัฒนาประสบการณ์ของตัวโปรแกรม เป็นการสร้างแบบจำลองการเรียนรู้ให้คอมพิวเตอร์สามารถทำนายหรือตัดสินใจได้ด้วยตนเองอย่างอัตโนมัติคล้ายมนุษย์ (อรพิน, 2564)

2.3.1 วิธีป่าสุ่ม (Random Forest)

วิธีป่าสุ่มพัฒนาขึ้นมาโดย Tin Kam ในปี ค.ศ. 1995 เป็นการสุ่มเลือกคุณลักษณะต่างๆ ของชุดข้อมูล จากนั้นนำเอาชุดข้อมูลและคุณลักษณะเหล่านี้มาทำการสร้างแบบจำลองการทำนายด้วยเทคนิคต้นไม้ตัดสินใจ (Decision Tree) หลายๆต้น และเลือกใช้แบบจำลองที่มีประสิทธิภาพดีที่สุด โดยรูปแบบของป่าสุ่มประกอบด้วย 3 ปัจจัยหลัก

1. ต้นไม้ในทุกต้นจะถูกฝึกสอน (Train) ด้วยการนำข้อมูลย่อยจากข้อมูลหลัก
2. เมื่อต้นไม้มีขนาดใหญ่ขึ้นก็จะสามารถหาโหนด (Node) ในแต่ละโหนดที่อยู่กิ่งที่ดีที่สุดโดยใช้หลักการสุ่มเลือกคุณลักษณะจาก N คุณลักษณะ
3. ต้นไม้แต่ละต้นจะไม่มีการทิ้ง แต่ทำให้ต้นไม้มีขนาดใหญ่ขึ้นไปเรื่อยๆ จนได้ผลลัพธ์ที่ดีที่สุด หลังจากการสร้างป่า แล้วทำการให้คะแนน (Vote) โดยต้นไม้ภายในป่า หากต้นไม้ต้นใดได้คะแนนสูงสุดก็จะนำเอาต้นไม้นั้นออกมาสร้างเป็นแบบจำลอง (ภูริพัทร์, 2559)

2.3.2 วิธีการประเมินราคาแอบแฝง (Hedonic Price Method)

การประเมินราคาแอบแฝงเป็นวิธีการประเมินราคาแอบแฝง (Implicit Price) ของลักษณะเชิงคุณภาพ คุณลักษณะต่างๆ ที่ประกอบรวมกันเป็นราคาโดยรวมของสินค้าที่มีลักษณะแตกต่างกัน แม้ว่าสินค้าจำนวนหนึ่งหน่วยเท่ากัน แต่ราคาของสินค้ามีความแตกต่างกัน เนื่องจากคุณภาพของสินค้าที่แตกต่างกัน ซึ่งได้รับการพัฒนามาจากการหาผลกระทบของคุณภาพต่อราคาของสินค้าของ Waugh (1928) แนวคิดเกี่ยวกับทฤษฎีผู้บริโภคที่พิจารณาจากคุณลักษณะของสินค้าของ Lancaster (1996) และแนวคิดของ Rosen (1974) เกี่ยวกับการเลือกซื้อสินค้าจากคุณลักษณะที่มีความแตกต่างกัน สุปราณี

(2551) กล่าวว่า การประเมินราคาแอบแฝงเป็นการประยุกต์ทฤษฎีอรรถประโยชน์ (Utility Theory) เป็นการตั้งราคาสินค้าตามความพอใจของผู้บริโภคที่ได้รับจากคุณลักษณะต่างๆ ของสินค้านั้นๆ โดยกำหนดให้อรรถประโยชน์ที่ผู้บริโภคได้รับเกิดจากการบริโภคคุณลักษณะ j จากทุกชนิดสินค้ารวมกันทุกชนิด (X_j) ซึ่งในการตัดสินใจบริโภคนั้น ผู้บริโภคจะตัดสินใจบริโภคภายใต้งบประมาณที่มีอยู่ เพื่อให้ได้อรรถประโยชน์สูงสุด (โชติวุฒิ, 2555)

แบบจำลองวิธีการประเมินราคาแอบแฝงมักถูกนำไปใช้ในการศึกษาตลาดอสังหาริมทรัพย์และตลาดแรงงาน ซึ่งนักเศรษฐศาสตร์ได้นำวิธีนี้มาใช้ในการประเมินมูลค่าคุณภาพสิ่งแวดล้อม โดยนิยมใช้กับอสังหาริมทรัพย์ เช่น บ้าน อาคารชุด เป็นต้น โดยแบบจำลองของวิธีการประเมินราคาแอบแฝงจะใช้ได้ก็ต่อเมื่อลักษณะที่อยู่อาศัยที่มีอยู่จะมีลักษณะเหมือนกันทุกประการ แต่อย่างไรก็ตาม ได้มีข้อโต้แย้งจากนักเศรษฐศาสตร์ที่ทำการวิจัยที่เกี่ยวข้องว่า ลักษณะที่อยู่อาศัยโดยทั่วไปจะมีคุณลักษณะที่แตกต่างกัน ไม่เหมือนกันหมดได้ โดยความแตกต่างนั้นสามารถแบ่งเป็นกลุ่มใหญ่ๆ เพื่อให้เห็นความแตกต่างได้อย่างชัดเจน เช่น ลักษณะทำเลที่ตั้ง ลักษณะโครงสร้างที่อยู่อาศัย และลักษณะสภาพแวดล้อม เป็นต้น ลักษณะของตลาดที่อยู่อาศัยต้องเป็นตลาดแข่งขันสมบูรณ์ จากงานวิจัยที่เกี่ยวข้อง ได้มีการนำทฤษฎีวิธีการประเมินราคาแอบแฝงมาประยุกต์ใช้กับที่อยู่อาศัยต่างๆ ซึ่งปัจจัยหรือคุณลักษณะโดยส่วนใหญ่ของอาคารชุดกับราคาอาคารชุดนั้นจะจำแนกคุณลักษณะที่เกี่ยวข้องกับที่อยู่อาศัยออกเป็น 3 กลุ่ม ได้แก่ ปัจจัยหรือคุณลักษณะด้านทำเล ที่ตั้ง ปัจจัยหรือคุณลักษณะด้านโครงสร้างที่อยู่อาศัย และปัจจัยหรือคุณลักษณะด้านสภาพแวดล้อม ซึ่งเขียนออกมาเป็นความสัมพันธ์ของราคาที่อยู่อาศัยกับปัจจัยหรือคุณลักษณะต่างๆ ได้ดังนี้

$$P = f(L, S, N) \tag{2.6}$$

โดยที่ P คือ ราคาอาคารชุดของแต่ละโครงการ (บาทต่อตารางเมตร)

L คือ ปัจจัยด้านทำเลที่ตั้งของอาคารชุด

S คือ ปัจจัยด้านโครงสร้างของอาคารชุด

N คือ ปัจจัยด้านสภาพแวดล้อมของอาคารชุด

โดยปัจจัยด้านทำเลที่ตั้ง ถือเป็นปัจจัยสำคัญเนื่องจากการเลือกอาคารชุดจะต้องคำนึงถึงระยะการเดินทาง ระบบการขนส่งจากที่พักอาศัยไปยังสถานที่ทำงาน เช่น บริเวณย่านใจกลางเมือง บริเวณใกล้รถไฟฟ้า บริเวณวิวแม่น้ำ เป็นต้น จึงทำให้มีการใช้ที่ดินในบริเวณย่านใจกลางเมือง และในบริเวณใกล้รถไฟฟ้า จึงเกิดการกระจุกตัวของที่พักอาศัยประเภทอาคารชุด ส่วนปัจจัยด้านโครงสร้างของอาคารชุดจะเป็นรูปแบบของตัวอาคารชุด ขนาดพื้นที่ของห้อง จำนวนห้องนอน ห้องน้ำ การตกแต่งเฟอร์นิเจอร์ ความพร้อมในการเข้าอยู่ ชื่อเสียงของผู้ประกอบการ เป็นต้น ส่วนปัจจัยด้านสภาพแวดล้อมจะเป็นพื้นที่ใกล้เคียงบริเวณของอาคารชุดไม่ว่าจะเป็นใกล้โรงพยาบาล สถานศึกษา ห้างสรรพสินค้า เป็นต้น

2.3.3 วิธี Extreme Gradient Boosting

Nonita (2017) อธิบายความหมายของวิธี Extreme Gradient Boosting คือใช้หลักการสร้างต้นไม้แต่ละต้นจะเป็นแบบเรียงลำดับ โดยนำข้อมูลเข้าของต้นไม้แต่ละต้นจะเป็นผลลัพธ์จากต้นไม้ก่อนหน้า โดยหลักการคือ Extreme Gradient Boosting จะทำการสร้างต้นไม้แต่ละต้น เพื่อลดค่าความผิดพลาดที่เกิดจากต้นไม้ก่อนหน้า โดยใช้วิธีการเคลื่อนลงตามความชัน (Gradient Descend) แล้วนำผลลัพธ์ที่ได้มารวมกัน ก็จะทำให้ได้ค่าใกล้เคียงกับค่าทำนาย ซึ่งข้อดีของ Extreme Gradient Boosting คือ ความอคติและความแปรปรวนลดลง เนื่องจากความผิดพลาดก่อนหน้าถูกแก้ไขเพียงแค่ความลึกของชั้นต้นไม้แค่หนึ่งชั้นก็เพียงพอที่จะทำให้ประสิทธิภาพดีขึ้นมากเมื่อเทียบกับต้นไม้ตัดสินใจที่ต้องเพิ่มชั้นความลึกมากขึ้น เพื่อให้ได้ประสิทธิภาพที่ใกล้เคียง

2.4 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึก คือ วิธีการเรียนรู้แบบอัตโนมัติด้วยการเลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ โดยนำระบบโครงข่ายประสาทเทียม (Neural Network) มาซ้อนกันหลายชั้น (Layer) และทำการเรียนรู้ข้อมูลตัวอย่าง ซึ่งข้อมูลดังกล่าวจะถูกนำไปใช้ในการตรวจจับรูปแบบ หรือจำแนกข้อมูล โดยทั่วไปวิธีโครงข่ายประสาทเทียมคือการเรียนรู้เพียงไม่กี่ชั้น เพื่อที่จะทำให้วิธีโครงข่ายประสาทเทียมมีความคิดและประมวลผลซับซ้อนให้เหมือนสมองมนุษย์ ดังนั้นชั้นที่เป็นชั้นซ่อนจะมีหลายๆ ชั้น เพื่อให้มีการส่งข้อมูลประมวลผลเชื่อมต่อกันไป ทำให้มีข้อมูลซ้อนกันหลายชั้น โครงข่ายก็ยิ่งมีความซับซ้อนและลึกขึ้น จึงเป็นที่มาของการเรียนรู้เชิงลึก (ไกรศักดิ์, 2564)

2.4.1 วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

วิธีโครงข่ายประสาทแบบคอนโวลูชัน (CNN) เป็นวิธีการที่ถูกนำมาใช้อย่างกว้างขวางทั้งในคณิตศาสตร์สถิติการประมวลผลสัญญาณ (Signal Processing) เป็นวิธีการศึกษาการเปลี่ยนแปลงของฟังก์ชัน โครงข่ายประสาทแบบคอนโวลูชันถูกนำมาใช้ในรูปแบบจำลองที่สามารถเรียนรู้และเลือกใช้คุณลักษณะของข้อมูลได้ด้วยตัวเอง และจะเรียนรู้การเลือกคุณลักษณะเด่นได้ดีกว่ามนุษย์ ดังนั้นจะทำให้ได้การประมวลผลที่แม่นยำมากกว่าโครงข่ายประสาทเทียมโดยทั่วไป ข้อมูลเข้าของโครงข่ายประสาทเทียมแบบคอนโวลูชันจะเป็นเมทริกซ์ ซึ่งประกอบด้วยชั้นการทำงานจำนวน 2 ชั้น คือ ชั้นคอนโวลูชัน (Convolution Layer) ซึ่งถูกเพิ่มมาเป็นส่วนที่ใช้ในการกรองข้อมูลและแยกองค์ประกอบออกมา และชั้นพูลิ่ง (Pooling Layer) ทำหน้าที่ปรับขนาดข้อมูลให้มีขนาดเล็กลงโดยที่ยังคงรายละเอียดข้อมูลเดิม (ไกรศักดิ์, 2564)

2.4.2 วิธีหน่วยความจำระยะสั้น-ยาว (Long Short-Term Memory)

วิธีหน่วยความจำระยะสั้น-ยาว (LSTM) เป็นวิธีหนึ่งที่ถูกพัฒนามาจากโครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network) โดยโครงข่ายประสาทเทียมแบบวนกลับมีหลักการทำงานคือการนำผลลัพธ์ที่ได้จากการคำนวณจากโหนดก่อนหน้ากลับมาใช้เป็นข้อมูลนำเข้าของโหนดถัดไป ซึ่งแต่ละโหนดของโครงข่ายประสาทเทียมแบบวนกลับจะมีข้อมูลที่เข้ามา 2 ส่วน ได้แก่ ข้อมูลจากการนำเข้าของโหนดนั้นๆ กับข้อมูลจากผลลัพธ์ โครงข่ายประสาทเทียมแบบวนกลับนั้นเหมาะนำมาใช้งานกับข้อมูลที่มีลักษณะเป็นลำดับ หรือข้อมูลที่มีความต่อเนื่อง ในปี 1997 นักวิทยาศาสตร์ Hochreiter และ Schmidhuber ได้คิดค้นวิธีหน่วยความจำระยะสั้น-ยาว เพื่อพัฒนาโครงข่ายประสาทเทียมแบบวนกลับให้มีความเสถียรและประสิทธิภาพมากขึ้น โดยมีหลักการทำงานคือ สามารถเก็บสถานะหรือข้อมูลของแต่ละโหนดเอาไว้เพื่อที่เวลาย้อนกลับไปดูจะได้ทราบถึงที่มาของข้อมูลดังกล่าว ซึ่งจะมีประตู (Gate) ที่คอยควบคุมข้อมูลที่จะเข้ามาในแต่ละโหนด ประกอบไปด้วยประตูลืม(Forget Gate Layer) ประตูทางเข้า (Input Gate Layer) และประตูทางออก (Output Gate Layer) (ไกรศักดิ์, 2564)

2.5 มาตรวัดประสิทธิภาพ (Efficiency Measure)

2.5.1 สัมประสิทธิ์การกำหนด (Coefficient of Determination)

สัมประสิทธิ์การกำหนด (r^2) เป็นมาตรวัดความเหมาะสมของการถดถอยในการประมาณค่าของความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระและตัวแปรตาม (สายชล, 2559)

$$r^2 = \frac{SSR}{SST} \tag{2.7}$$

โดยที่ SSR คือ ผลบวกกำลังสองของการถดถอย

SST คือ ผลบวกกำลังสองทั้งหมด

เนื่องจากสัมประสิทธิ์การกำหนดเป็นอัตราส่วนระหว่าง SSR และ SST ดังนั้นสัมประสิทธิ์การกำหนดเป็นสัดส่วนของความแปรผันในตัวแปร Y ที่อธิบายได้โดยการถดถอย นั่นคือโดยความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระและตัวแปรตาม

สัมประสิทธิ์การกำหนดจะมีค่าอยู่ระหว่าง 0 และ 1 สัมประสิทธิ์การกำหนดที่ใกล้เคียง 1 แสดงว่าการถดถอยมีความเหมาะสมกับข้อมูลดีมาก ส่วนสัมประสิทธิ์การกำหนดที่ใกล้เคียง 0 แสดงว่าการถดถอยไม่มีความเหมาะสมกับข้อมูล

2.5.2 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error)

ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) หรือส่วนเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute Deviation) เป็นวิธีที่การวัดค่าความคลาดเคลื่อนที่นิยมอีกวิธีหนึ่ง ซึ่งวิธีนี้จะช่วยบอกถึงขนาดของความคลาดเคลื่อนรวมได้ โดยมีสมการดังสมการที่ 2.7 ในการวัดค่าความแม่นยำจากวิธีการนี้ ยิ่งค่าที่ได้มีค่าน้อยแสดงว่าแบบจำลองที่ได้จะมีความแม่นยำมาก (สายชล, 2559)

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t| \tag{2.8}$$

โดยที่ n คือ จำนวนข้อมูลที่ใช้

Y_t คือ ค่าจริงที่เวลา t ใดๆ

$\hat{Y}_t$ คือ ค่าที่ได้จากการทำนายที่เวลา t ใดๆ

2.5.3 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error)

รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) เป็นวิธีการวัดค่าความคลาดเคลื่อนแบบมาตรฐานที่นิยมใช้กันอย่างแพร่หลาย โดยค่าที่ได้ยิ่งน้อยจะยิ่งแสดงว่าแบบจำลองที่ได้มีความแม่นยำมาก (สายชล, 2559)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \tag{2.9}$$

โดยที่ n คือ จำนวนข้อมูลที่ใช้

Y_t คือ ค่าจริงที่เวลา t ใดๆ

$\hat{Y}_t$ คือ ค่าที่ได้จากการทำนายที่เวลา t ใดๆ

2.5.4 สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient)

สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (ρ) เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างตัวแปร 2 ตัว แปรว่ามีความสัมพันธ์กันหรือไม่ ซึ่งในทางปฏิบัติเก็บรวบรวมข้อมูลจากตัวอย่าง ดังนั้นจะประมาณสัมประสิทธิ์สหสัมพันธ์เพียร์สันด้วย r โดยสัมประสิทธิ์สหสัมพันธ์จะมีค่าอยู่ระหว่าง -1 ถึง 1 (ยุทธ, 2549)

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{[\sum_{i=1}^{n}(X_i-\bar{X})^2][\sum_{i=1}^{n}(Y_i-\bar{Y})^2]} \tag{2.10}$$

โดยที่ n คือ จำนวนข้อมูลที่ใช้

X_i คือ ค่าตัวแปร X ณ ชุดข้อมูลที่ i

$\bar{X}$ คือ ค่าเฉลี่ยของตัวแปร X

Y_i คือ ค่าตัวแปร Y ณ ชุดข้อมูลที่ i

$\overline{Y}$ คือ ค่าเฉลี่ยของตัวแปร Y

ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 1 หมายถึงมีความสัมพันธ์เชิงเส้นบวกแบบเสริมกันที่แน่นอนระหว่างตัวแปรสองตัว ค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ -1 หมายถึงมีความสัมพันธ์เชิงเส้นลบแบบเสริมกันที่แน่นอนระหว่างตัวแปรสองตัว ค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0 หมายถึงไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรสองตัว

เพื่อทดสอบความสัมพันธ์ให้แม่นยำมากขึ้นจึงทำการทดสอบสมมติฐาน (hypothesis) โดยเริ่มจากการตั้งสมมติฐานว่างคือตัวแปรอิสระไม่ส่งผลต่อตัวแปรตาม และสมมติฐานทางเลือกคือตัวแปรอิสระส่งผลต่อตัวแปรตาม ด้วยสถิติทดสอบที (t-test) ดังนี้

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{2.11}$$

หลังจากนั้นนำสถิติทดสอบทีไปหาค่าพี (p-value) และนำค่าพีไปเปรียบเทียบกับระดับนัยสำคัญที่กำหนดไว้ ถ้าค่าพีน้อยกว่าระดับนัยสำคัญ แสดงว่าปฏิเสธสมมติฐานว่าง คือตัวแปรอิสระส่งผลต่อตัวแปรตาม ถ้าค่าพีมากกว่าระดับนัยสำคัญ แสดงว่ายอมรับสมมติฐานว่าง คือตัวแปรอิสระไม่ส่งผลต่อตัวแปรตาม

2.6 งานวิจัยที่เกี่ยวข้อง (Literature review)

Baldominos et al. (2018) ได้ศึกษาเกี่ยวกับปัจจัยสำคัญในการเติบโตของอสังหาริมทรัพย์โดยใช้การเรียนรู้ของเครื่อง กล่าวถึงข้อมูลของอสังหาริมทรัพย์ในเขตซาลามังกาของเมืองมาดริด ประเทศสเปน ที่ลงประกาศขายและเช่าทางออนไลน์ และนำข้อมูลที่ได้สร้างแบบจำลองโดยเปรียบเทียบด้วยวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีป่าสุ่ม และวิธีซัพพอร์ตเวกเตอร์การถดถอย วัดประสิทธิภาพโดยใช้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ผลการศึกษาพบว่าวิธีป่าสุ่มให้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำที่สุดคือ 338,715 รองลงมาคือวิธีเพื่อนบ้านใกล้สุด k ตัว วิธีซัพพอร์ตเวกเตอร์การถดถอย และวิธีโครงข่ายประสาทเทียม ตามลำดับ

Piao et al. (2019) ได้ศึกษาเกี่ยวกับการพยากรณ์ราคาที่อยู่อาศัยในต้าเหลียน ประเทศจีน โดยใช้ข้อมูลจากศูนย์ข้อมูลทรัพยากรที่ดินและที่อยู่อาศัยสำหรับเมืองต้าเหลียนที่เป็นฐานข้อมูลภายใต้การจัดการรัฐบาลกลางของสาธารณรัฐประชาชนจีน มีจำนวนโครงการทั้งหมด 171,155 โครงการ และจำนวนตัวแปรอิสระ 16 ตัวแปร เป็นข้อมูลเก็บรวบรวมตั้งแต่ปี ค.ศ. 2013 ถึง 2017 และมีข้อมูลเพิ่มเติมคือ ข้อมูลเศรษฐกิจของเมืองต้าเหลียน ประเทศจีน จากสำนักงานสถิติจำนวน 6 ตัวแปร ซึ่งเปรียบเทียบด้วยวิธีโครงข่ายประสาทแบบคอนโวลูชันกับวิธี Extreme Gradient Boosting ทำการวัด

ประสิทธิภาพโดยใช้ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และสัมประสิทธิ์การกำหนด ผลการศึกษาพบว่าวิธีโครงข่ายประสาทแบบคอนโวลูชันให้สัมประสิทธิ์การกำหนดสูงที่สุดคือ 0.9868 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดคือ 0.01057 รองลงมาคือวิธี Extreme Gradient Boosting ให้สัมประสิทธิ์การกำหนดคือ 0.9650 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.1040 สำหรับปัจจัยที่ส่งผลกับราคาที่อยู่อาศัยในต้าเหลียนสูงสุด 3 อันดับแรกคือ พื้นที่บ้าน ราคาทำธุรกรรมที่อยู่อาศัย และผลิตภัณฑ์มวลรวมของประเทศ

Jafari and Akhavian (2019) ได้ศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อราคาที่อยู่อาศัยในประเทศสหรัฐอเมริกา โดยใช้ข้อมูลจากสำรวจแบบสอบถามที่อยู่อาศัยจำนวน 13,771 หลัง ของปี 2013 ซึ่งมี 22 หลังที่อาศัยอยู่ในเมืองซานฟรานซิสโกรวมอยู่ในชุดข้อมูล และสร้างแบบจำลองด้วยวิธีการประเมินราคาแอบแฝง โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยทีละขั้นและวิธีการประเมินราคาแอบแฝงด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนดผลการศึกษาพบว่าวิธีการประเมินราคาแอบแฝงด้วยการถดถอยชุดข้อมูลย่อยที่ดีที่สุดได้สัมประสิทธิ์การกำหนดคือ 0.4073 ซึ่งดีกว่าวิธีการประเมินราคาแอบแฝงของลักษณะเชิงคุณภาพด้วยการเลือกตัวแปรด้วยวิธีการถดถอยทีละขั้น ในงานวิจัยกล่าวเพิ่มเติมว่า เนื่องจากมีเวลาจำกัดในการทำงานวิจัย ส่งผลให้ไม่ได้พิจารณาบางข้อมูล เช่น อัตราเงินเฟ้อ อัตราดอกเบี้ย อุปสงค์อุปทานของบ้าน และอัตราการตกงาน เป็นต้น ซึ่งข้อมูลที่กล่าวมาอาจส่งผลกับการเพิ่มประสิทธิภาพของแบบจำลอง

พสธร (2563) ได้ศึกษาเกี่ยวกับการประเมินราคาเสนอขายห้องชุดในประเทศไทย จาก ZmyHome ซึ่งประกอบไปด้วยข้อมูลจำนวน 11,062 ชุด ปัจจัยที่นำมาศึกษาแบ่งเป็นปัจจัยด้านตัวโครงการ ปัจจัยด้านทำเล และปัจจัยด้านสภาพแวดล้อม ซึ่งเปรียบเทียบด้วยวิธีหน่วยความจำระยะสั้น-ยาว โดยใช้ข้อมูลจากวิธีการแบ่งกลุ่มข้อมูลเฉลี่ย k กลุ่มกับวิธีหน่วยความจำระยะสั้น-ยาว โดยใช้ข้อมูลของอาคารชุดโดยใช้รัศมี 5 กิโลเมตร ทำการวัดประสิทธิภาพโดยใช้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ผลการศึกษาพบว่าวิธีหน่วยความจำระยะสั้น-ยาวโดยใช้ข้อมูลจากวิธีการแบ่งกลุ่มข้อมูลเฉลี่ย k กลุ่มได้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดคือ 9,722 บาท ซึ่งดีกว่าวิธีหน่วยความจำระยะสั้น-ยาวโดยใช้ข้อมูลของอาคารชุดโดยใช้รัศมี 5 กิโลเมตร ให้รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยคือ 10,938.34 บาท ถึงแม้ผลการทดลองข้อมูลที่แบ่งกลุ่มเฉลี่ย k กลุ่มดีกว่า แต่ผู้ทำวิจัยได้เสนอแนะว่าเมื่อนำข้อมูลของอาคารชุดโดยใช้รัศมี 5 กิโลเมตร ไปใช้งานจริงสามารถทำได้รวดเร็วกว่า จึงเหมาะสมที่จะประยุกต์นำไปใช้มากกว่า ส่วนการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์พบว่า ปัจจัยด้านสภาพแวดล้อม ปัจจัยด้านตัวโครงการ และปัจจัยสถานที่ตั้ง มีความสัมพันธ์ต่อราคาเสนอขายมากไปน้อยที่สุด ตามลำดับ

Pensri et al. (2020) ได้ศึกษาปัจจัยที่ส่งผลต่อการซื้ออาคารชุดในกรุงเทพมหานคร โดยทำแบบสำรวจและเก็บรวบรวมข้อมูลจำนวน 385 คน วัดประสิทธิภาพด้วยสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ผลการศึกษาพบว่า ปัจจัยสถานที่ตั้งของอาคารชุดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันสูงที่สุดคือ 0.898

รองลงมาคือปัจจัยส่วนลดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันคือ 0.721 และปัจจัยราคาของอาคารชุดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันคือ 0.625 นอกจากนี้ปัจจัยสถานที่ตั้งของอาคารชุดที่ส่งผลให้ราคาอาคารชุดมีราคาที่สูงคือ สถานที่ตั้งใกล้กับระบบขนส่งมวลชน เช่น รถไฟฟ้าบนดินและใต้ดิน เป็นต้น สถานที่ตั้งใกล้กับทำเลทอง สถานที่ตั้งใกล้กับศูนย์กลางธุรกิจ และสถานที่ตั้งใกล้กับแม่น้ำหรือสิ่งแวดล้อมที่ดี ตามลำดับ

Viriya (2021) ได้ศึกษาเกี่ยวกับความสัมพันธ์ระหว่างสิ่งอำนวยความสะดวกรอบอาคารชุดและราคาอาคารชุดด้วยการเรียนรู้ของเครื่อง โดยใช้ชุดข้อมูลจาก ddproperty จำนวน 1,911โครงการ และเลือกข้อมูลอาคารชุดจำนวน 500 โครงการอันดับแรกที่ราคาใกล้เคียงกับราคาขายจริงมากที่สุด ในกรุงเทพมหานคร และข้อมูลสถานที่สำคัญรอบๆอาคารชุดในรัศมี 400 เมตร และเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่อง 2 วิธีคือ Extreme Gradient Boosting และวิธีป่าสุ่ม โดยวัดประสิทธิภาพด้วยสัมประสิทธิ์การกำหนด ผลการศึกษาพบว่าวิธี Extreme Gradient Boosting ให้สัมประสิทธิ์การกำหนดสูงที่สุดคือ 0.73 ซึ่งดีกว่าวิธีป่าสุ่มให้สัมประสิทธิ์การกำหนดเพียง 0.66 นอกจากนี้พบว่ามีตัวแปรอิสระที่สำคัญ 36 ตัวแปร จากแบบจำลองวิธี Extreme Gradient Boosting

Guliker et al. (2022) ได้ศึกษาเกี่ยวกับปัจจัยที่ส่งผลต่อการประเมินอสังหาริมทรัพย์ในประเทศเนเธอร์แลนด์ ด้วยการเรียนรู้ของเครื่อง โดยใช้ชุดข้อมูลจากการประเมินมูลค่าอสังหาริมทรัพย์ของ 5 เขตใหญ่ ในประเทศเนเธอร์แลนด์ จาก Stater N.V. ที่เป็นบริษัทให้สินเชื่อที่อยู่อาศัยใหญ่ที่สุดในเนเธอร์แลนด์ โดยเปรียบเทียบด้วยวิธีการวิเคราะห์การถดถอยเชิงเส้น วิธีการถดถอยแบบถ่วงน้ำหนักทางภูมิศาสตร์ และวิธี Extreme Gradient Boosting วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนด รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และร้อยละของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย ผลการศึกษาพบว่าวิธี Extreme Gradient Boosting ให้สัมประสิทธิ์การกำหนดสูงที่สุดเท่ากับ 0.832 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 65,312 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 43,625 และร้อยละของค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 6.35% รองลงมาคือวิธีการวิเคราะห์การถดถอยเชิงเส้น และวิธีการถดถอยแบบถ่วงน้ำหนักทางภูมิศาสตร์ ตามลำดับ ส่วนปัจจัยสำคัญของแบบจำลอง Extreme Gradient Boosting คือ ปัจจัยสถานที่ตั้ง และค่าภาษี ซึ่งข้อมูลทั้งสองเป็นข้อมูลสาธารณะ

Rupesh and Kumar (2023) ได้ศึกษาเกี่ยวกับการเพิ่มประสิทธิภาพของแบบจำลองสำหรับการทำนายราคาค่าเช่าบ้านหรืออพาร์ตเมนต์ในประเทศอินเดีย โดยใช้ข้อมูลราคาบ้านเช่าที่อยู่อาศัยตั้งแต่ปี ค.ศ. 2016 ถึง 2019 ที่มีข้อมูลทั้งหมด 1,100 โครงการ บ้านหรืออพาร์ตเมนต์จากข้อมูลเปิดใน Kaggle และมีตัวแปรทั้งหมด 60 ตัวแปร ประกอบด้วยข้อมูลเชิงปริมาณ เช่น ราคา จำนวนห้องน้ำ จำนวนห้องนั่งเล่น จำนวนห้องนอน และรวมถึงจำนวนห้องอื่นๆ ส่วนข้อมูลเชิงคุณภาพคือ ข้อมูลพื้นที่

และข้อมูลจำแนกภาค ซึ่งเปรียบเทียบด้วยวิธีโครงข่ายประสาทแบบคอนโวลูชันกับวิธีต้นไม้ตัดสินใจ วัดประสิทธิภาพโดยใช้สัมประสิทธิ์การกำหนด ผลการศึกษาพบว่าวิธีโครงข่ายประสาทแบบคอนโวลูชันให้สัมประสิทธิ์การกำหนดสูงที่สุดคือ 0.9632 ซึ่งดีกว่าวิธีต้นไม้ตัดสินใจให้สัมประสิทธิ์การกำหนดเพียง 0.9415

บทที่ 3

วิธีการดำเนินงานวิจัย

งานวิจัยนี้ศึกษาการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้การเลือกคุณลักษณะ 2 วิธี คือ วิธีการนำตัวแปรเข้าทั้งหมด (Enter) และวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Selection) ซึ่งทำการเปรียบเทียบแบบจำลองการเรียนรู้ทั้งหมด 5 วิธี แบ่งเป็นการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม (Random Forest) วิธีการประเมินราคาแอบแฝง (Hedonic Price Method) และวิธี Extreme Gradient Boosting ส่วนอีก 2 วิธีที่เหลือเป็นการเรียนรู้เชิงลึก คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network) และวิธีหน่วยความจำระยะสั้น-ยาว (Long Short-Term Memory) ซึ่งใช้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง โดยรายละเอียดจะกล่าวถึงในหัวข้อย่อยๆตามลำดับต่อไป

รูปที่ 3.1 กระบวนการทำงานการเปรียบเทียบประสิทธิภาพวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกของการทำนายราคาอาคารชุดในกรุงเทพมหานคร

จากรูปที่ 3.1 แสดงให้เห็นถึงขั้นตอนการดำเนินงานโดยเริ่มจากนำข้อมูลมาจาก Baania (2566) ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ จากนั้นทำเก็บรวบรวมข้อมูลพิกัดละติจูดและลองจิจูดเพิ่มเติมจาก Google Maps เพื่อเพิ่มข้อมูลจำนวนสถานที่สำคัญในบริเวณโครงการ ขั้นตอนต่อไปทำการแทนค่าข้อมูลสูญหายโดยใช้ค่าเฉลี่ยในการแทนค่าข้อมูลสูญหายกับข้อมูลเชิงปริมาณ และใช้ค่าฐานนิยมในการแทนค่าข้อมูลสูญหายกับข้อมูลเชิงคุณภาพ นำชุดข้อมูลที่ไม่มีข้อมูลสูญหายมากำจัดค่านอกเกณฑ์ (Outlier) โดยใช้แผนภาพกล่อง (Boxplot) และขั้นตอนสุดท้ายของการเตรียมข้อมูลคือ การแปลงข้อมูล

ซึ่งจะทำการแปลงข้อมูลที่เป็นข้อมูลเชิงคุณภาพให้อยู่ในรูปแบบของค่าทวิภาค ที่มีค่า 0 หรือ 1 เท่านั้น ด้วยวิธี One-Hot Encoding ขั้นตอนต่อไปทำการออกแบบจำลองด้วยการเลือกคุณลักษณะ 2 วิธี คือ วิธีการนำตัวแปรเข้าทั้งหมด (Enter) และวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Selection) เมื่อทำการเลือกคุณลักษณะ จะได้ชุดข้อมูลทั้งหมด 2 ชุด นำชุดข้อมูลที่ได้มานั้นสร้างแบบจำลองทั้งหมด 5 วิธี โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม (Random Forest) วิธีการประเมินราคาแอบแฝง (Hedonic Price Method) และวิธี Extreme Gradient Boosting ซึ่งจะใช้วิธีการแบ่งชุดข้อมูลออกเป็น 3 ชุด ชุดข้อมูลฝึกสอน ชุดข้อมูลทดสอบ และชุดข้อมูลตรวจสอบ โดยผู้ทำวิจัยกำหนดการแบ่งสัดส่วนของชุดข้อมูลฝึกสอน (Training Data) 80% และชุดข้อมูลทดสอบ (Testing Data) 20% หลังจากได้ชุดข้อมูลฝึกสอนจะใช้วิธีการตรวจสอบไขว้ 5 ชุด (5-Fold Cross Validation) เพื่อแบ่งข้อมูลออกเป็น 5 ชุดเท่าๆกัน แล้วใช้ 4 ชุดสำหรับชุดข้อมูลฝึกสอน และ 1 ชุดสำหรับชุดข้อมูลการตรวจสอบ (Validation Data) และอีก 2 วิธีด้วยแบบจำลองการเรียนรู้เชิงลึก คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network) และวิธีหน่วยความจำระยะสั้น-ยาว (Long Short-Term Memory) ซึ่งทำการแบ่งชุดข้อมูลออกเป็น 2 ชุด ชุดข้อมูลฝึกสอน 80% และชุดข้อมูลทดสอบ 20% ขั้นตอนสุดท้าย คือ การเปรียบเทียบประสิทธิภาพแบบจำลอง ผู้ทำวิจัยกำหนดใช้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ส่วนการศึกษาปัจจัยที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร ผู้วิจัยกำหนดใช้สัมประสิทธิ์สหสัมพันธ์เพียร์สันในการหาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม

3.1 การจัดเตรียมข้อมูล (Data Preprocessing)

3.1.1 การเก็บรวบรวมข้อมูล (Data Collection)

ผู้วิจัยได้ทำการเก็บรวบรวมชุดข้อมูลมาจาก Baania (2566) ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ โดยสนใจเฉพาะอาคารชุดในกรุงเทพมหานคร ที่มี 2,403 โครงการ โดยชุดข้อมูลดังกล่าวเปิดให้เข้าใช้งานในรูปแบบของสาธารณะ ภายในชุดข้อมูลประกอบด้วยตัวแปรทั้งหมด 29 ตัว คือ ตัวแปรตาม 1 ตัว และตัวแปรอิสระ 28 ตัว ซึ่งตัวแปรตามเป็นข้อมูลประเภทเชิงปริมาณ ส่วนตัวแปรอิสระเป็นข้อมูลเชิงคุณภาพ 11 ตัว และข้อมูลเชิงปริมาณ 17 ตัว

การทำให้แบบจำลองมีประสิทธิภาพมากขึ้น ผู้วิจัยจึงทำการเก็บรวบรวมข้อมูลพิกัดละติจูดและลองจิจูด เพิ่มเติมจาก Google Maps โดยมี 4 ประเภทสถานที่ ดังนี้ ข้อมูลพิกัดของสถานีรถไฟฟ้าบนดินและใต้ดิน รวมถึงรถไฟฟ้าเชื่อมท่าอากาศยานสุวรรณภูมิ จำนวน 171 สถานี ข้อมูลพิกัดของโรงพยาบาล 51 สถานที่ ข้อมูลพิกัดของมหาวิทยาลัย 50 สถานที่ และข้อมูลพิกัดของสถานที่ท่องเที่ยวสำคัญ 52 สถานที่ หลังจากนั้นนำข้อมูลพิกัดของสถานที่สำคัญมาใช้วิเคราะห์หาระยะทางเทียบกับแต่ละ

โครงการด้วยวิธีระยะทางยุคลิด เพื่อหาจำนวนสถานที่สำคัญในบริเวณโครงการ โดยกำหนดให้จำนวนสถานีรถไฟฟ้าบนดินและใต้ดิน รวมถึงรถไฟฟ้าเชื่อมท่าอากาศยานสุวรรณภูมิในบริเวณโครงการรัศมี 0.4 กิโลเมตร ส่วนจำนวนโรงพยาบาล มหาวิทยาลัย และสถานที่ท่องเที่ยวสำคัญในบริเวณโครงการรัศมี 3 กิโลเมตร ตามงานวิจัยของ Viriya (2021) การเลือกรัศมี 0.4 กิโลเมตรสำหรับสถานีรถไฟฟ้าเนื่องจากเป็นระยะทางที่ใช้เวลาเดินทางด้วยเท้าประมาณ 5 นาที ซึ่งเป็นระยะเวลาที่ยอมรับได้ในการเดินทางด้วยเท้าเพื่อขึ้นสถานีรถไฟฟ้า ส่วนสำหรับสถานที่ท่องเที่ยวสำคัญเลือกสถานที่สำคัญที่อยู่ใกล้โครงการรัศมี 3 กิโลเมตร เนื่องจากเป็นระยะทางที่ใช้เวลาเดินทางด้วยรถจักรยานยนต์หรือรถยนต์ประมาณ 5 นาที หลังจากได้รวมข้อมูลจำนวนสถานที่สำคัญในบริเวณโครงการกับข้อมูลอาคารชุด จะได้ตัวแปร ชื่อหัวข้อ คำอธิบาย หน่วย และประเภท ดังตารางที่ 3.1

ตารางที่ 3.1 ตัวแปร ชื่อหัวข้อ คำอธิบาย หน่วย และประเภท ของข้อมูล Baania รวมกับข้อมูลที่เพิ่มเติมเกี่ยวกับจำนวนสถานที่สำคัญรอบโครงการ

ตัวแปร	ชื่อหัวข้อ	คำอธิบาย	หน่วย	ประเภท
Y	Price per sqm	ราคาต่อตารางเมตร	บาท	ปริมาณ
X_1	facility_clubhouse	ห้องสนทนา	0(ไม่มี),1(มี)	คุณภาพ
X_2	facility_fitness	ห้องออกกำลังกาย	0(ไม่มี),1(มี)	คุณภาพ
X_3	facility_meeting	ห้องประชุม	0(ไม่มี),1(มี)	คุณภาพ
X_4	facility_park	สวน	0(ไม่มี),1(มี)	คุณภาพ
X_5	facility_playground	สนามเด็กเล่น	0(ไม่มี),1(มี)	คุณภาพ
X_6	facility_pool	สระว่ายน้ำ	0(ไม่มี),1(มี)	คุณภาพ
X_7	facility_security	หน่วยรักษาความปลอดภัย	0(ไม่มี),1(มี)	คุณภาพ
X_8	zipcode	รหัสไปรษณีย์	รหัสไปรษณีย์	คุณภาพ
X_9	area_usable_min	พื้นที่ใช้สอย ต่ำสุด	ตารางเมตร	ปริมาณ
X_{10}	count_elevator	จำนวนลิฟต์ในบ้าน	ลิฟต์	ปริมาณ
X_{11}	count_floor	จำนวนชั้นของตัวโครงการ	ชั้น	ปริมาณ
X_{12}	count_parking	จำนวนที่จอดรถของโครงการ	ที่จอดรถ	ปริมาณ
X_{13}	count_room_bath	จำนวนห้องน้ำ	ห้อง	ปริมาณ
X_{14}	count_room_bed	จำนวนห้องนอน	ห้อง	ปริมาณ
X_{15}	count_room_dinning	จำนวนห้องทานข้าว	ห้อง	ปริมาณ

ตารางที่ 3.1 ตัวแปร ชื่อหัวข้อ คำอธิบาย หน่วย และประเภท ของข้อมูล Baania รวมกับข้อมูลที่เพิ่มเติมเกี่ยวกับจำนวนสถานที่สำคัญรอบโครงการ (ต่อ)

ตัวแปร	ชื่อหัวข้อ	คำอธิบาย	หน่วย	ประเภท
X_{16}	count_room_guest	จำนวนห้องพักแขก	ห้อง	ปริมาณ
X_{17}	count_room_kitchen	จำนวนห้องครัว	ห้อง	ปริมาณ
X_{18}	count_room_living	จำนวนห้องนั่งเล่น	ห้อง	ปริมาณ
X_{19}	count_room_maid	จำนวนห้องแม่บ้าน	ห้อง	ปริมาณ
X_{20}	count_room_storage	จำนวนห้องเก็บของ	ห้อง	ปริมาณ
X_{21}	count_room_utility	จำนวนห้องซักล้าง	ห้อง	ปริมาณ
X_{22}	count_unit	จำนวนห้องในโครงการ	ห้อง	ปริมาณ
X_{23}	count_unittype	จำนวนประเภทอาคารชุดในโครงการ	ประเภท	ปริมาณ
X_{24}	date_updated	วันที่ข้อมูลล่าสุด	วันที่	คุณภาพ
X_{25}	year finish	จำนวนปีตั้งแต่โครงการสร้างเสร็จ	ปี	คุณภาพ
X_{26}	year update	จำนวนปีข้อมูลล่าสุด	ปี	คุณภาพ
X_{27}	latitude	พิกัดโครงการ latitude	องศา	คุณภาพ
X_{28}	longitude	พิกัดโครงการ longitude	องศา	คุณภาพ
X_{29}	Train	จำนวนสถานีรถไฟฟ้าบนดินและใต้ดินภายในรัศมี 0.4 กิโลเมตร	สถานี	ปริมาณ
X_{30}	Hospital	จำนวนโรงพยาบาลภายในรัศมี 3 กิโลเมตร	โรงพยาบาล	ปริมาณ
X_{31}	University	จำนวนมหาวิทยาลัยภายในรัศมี 3 กิโลเมตร	มหาวิทยาลัย	ปริมาณ
X_{32}	Attraction	จำนวนสถานที่ท่องเที่ยวภายในรัศมี 3 กิโลเมตร	สถานที่ท่องเที่ยว	ปริมาณ

จากตารางที่ 3.1 แสดงตัวแปรทั้งหมด 33 ตัว โดยมีตัวแปรตามเป็นข้อมูลประเภทเชิงปริมาณ และตัวแปรอิสระ 32 ตัว แบ่งเป็นตัวแปรอิสระ 28 ตัวจากการรวมข้อมูลของอาคารชุดในกรุงเทพมหานคร และตัวแปรอิสระที่เพิ่มเติมอีก 4 ตัว จากข้อมูลจำนวนสถานที่สำคัญรอบโครงการ ซึ่งตัวแปรอิสระที่เพิ่มเติมคือ จำนวนสถานีรถไฟฟ้าบนดินและใต้ดินภายในรัศมี 0.4 กิโลเมตร (Train, X_{29})

จำนวนโรงพยาบาลภายในรัศมี 3 กิโลเมตร (Hospital, X_{30}) จำนวนมหาวิทยาลัยภายในรัศมี 3 กิโลเมตร (University, X_{31}) และจำนวนสถานที่ท่องเที่ยวภายในรัศมี 3 กิโลเมตร (Attraction, X_{32})

3.1.2 การแทนค่าข้อมูลสูญหาย (Missing Values Replacement)

การแทนค่าข้อมูลสูญหายเป็นการแก้ไขค่าสูญหายโดยนำหลักการทางคณิตศาสตร์มาประมาณค่าและแทนที่ค่าข้อมูลสูญหาย ทำให้ผลลัพธ์สุดท้ายคล้ายเดิมหรือทำให้ไม่มีการเปลี่ยนแปลงในผลการวิเคราะห์ จากข้อมูลชุดที่ได้รวบรวม มีตัวแปรทั้งหมด 33 ตัว โดยมีตัวแปรอิสระ 1 ตัวเป็นข้อมูลเชิงปริมาณ และตัวแปรอิสระ 32 ตัว เป็นข้อมูลเชิงคุณภาพ 11 ตัว และข้อมูลเชิงปริมาณ 21 ตัว ซึ่งจากข้อมูลทั้งหมดพบว่าจำนวนข้อมูลสูญหายทั้งหมด 6,840 จำนวน ในตัวแปรอิสระ โดยแบ่งเป็นข้อมูลเชิงคุณภาพ 5,279 จำนวน และข้อมูลเชิงปริมาณ 1,561 จำนวน ดังตารางที่ 3.2

ตารางที่ 3.2 ตัวแปร ชื่อหัวข้อ ประเภทข้อมูล จำนวนข้อมูลสูญหาย วิธีการแทนค่า และค่าที่แทน ของตัวแปรที่มีข้อมูลสูญหาย

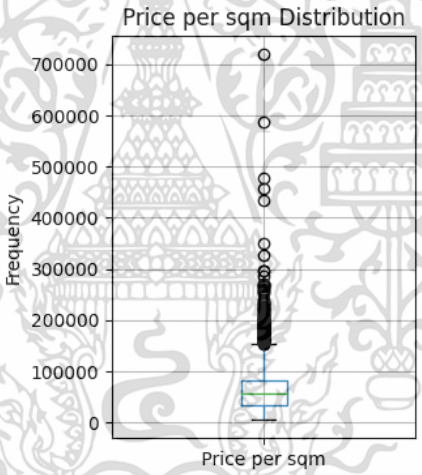
ตัวแปร	ชื่อหัวข้อ	ประเภทข้อมูล	จำนวนข้อมูลสูญหาย	วิธีการแทนค่า	ค่าที่แทน
X_1	facility_clubhouse	คุณภาพ	1,679	ค่าฐานนิยม	0 (ไม่มี)
X_2	facility_fitness	คุณภาพ	435	ค่าฐานนิยม	1 (มี)
X_3	facility_meeting	คุณภาพ	1,781	ค่าฐานนิยม	0 (ไม่มี)
X_4	facility_park	คุณภาพ	805	ค่าฐานนิยม	1 (มี)
X_6	facility_pool	คุณภาพ	429	ค่าฐานนิยม	1 (มี)
X_7	facility_security	คุณภาพ	150	ค่าฐานนิยม	1 (มี)
X_{10}	count_elevator	ปริมาณ	1,219	ค่าเฉลี่ย	3.3959
X_{22}	count_unit	ปริมาณ	47	ค่าเฉลี่ย	404.44
X_{23}	count_unittype	ปริมาณ	295	ค่าเฉลี่ย	2.5054

ตารางที่ 3.2 แสดงตัวแปร ชื่อหัวข้อ ประเภทข้อมูล จำนวนข้อมูลสูญหาย วิธีการแทนค่า และค่าที่แทน ของตัวแปรที่มีข้อมูลสูญหาย โดยทางผู้วิจัยใช้ค่าฐานนิยมในการแทนค่าข้อมูลสูญหายกับข้อมูลเชิงคุณภาพ ดังนั้นค่าฐานนิยมของห้องสนทนา (facility_clubhouse, X_1) คือ 0 หรือไม่มี ค่าฐานนิยมของห้องออกกำลังกาย (facility_fitness, X_2) คือ 1 หรือมี ค่าฐานนิยมของห้องประชุม (facility_meeting, X_3) คือ 0 หรือไม่มี ค่าฐานนิยมของสวน (facility_park, X_4) คือ 1 หรือมี ค่าฐาน

นิยมของสระว่ายน้ำ (facility_pool, X_6) คือ 1 หรือมี และค่าฐานนิยมของหน่วยรักษาความปลอดภัย (facility_security, X_7) คือ 1 หรือมี และใช้ค่าเฉลี่ยในการแทนค่าข้อมูลสูญหายกับข้อมูลเชิงปริมาณ ดังนั้นค่าเฉลี่ยของจำนวนลิฟต์ในบ้าน (count_elevator, X_{10}) คือ 3.3959 ค่าเฉลี่ยของจำนวนห้องในโครงการ (count_unit, X_{22}) คือ 404.44 และค่าเฉลี่ยของจำนวนประเภทอาคารชุดในโครงการ (count_unittype, X_{23}) คือ 2.5054 หลังจากเราทำการแทนค่าข้อมูลสูญหายแล้วจะพบว่าทุกตัวแปรจะไม่มีข้อมูลสูญหายปรากฏขึ้น

3.1.3 การกำจัดค่านอกเกณฑ์ (Outlier Deletion)

หลังจากทำการแทนค่าข้อมูลสูญหาย ขั้นตอนถัดไปคือ การกำจัดค่านอกเกณฑ์ ในที่นี้ผู้วิจัยเลือกการกำจัดค่านอกเกณฑ์ด้วยวิธีแผนภาพกล่อง โดยทำการตรวจสอบข้อมูลของราคาอาคารชุดต่อตารางเมตร ดังรูปที่ 3.2



รูปที่ 3.2 แผนภาพกล่องของราคาอาคารชุดต่อตารางเมตร

จากรูปที่ 3.2 พบว่าค่านอกเกณฑ์ Lower Anomaly คือ -37,447.6925 และ Upper Anomaly คือ 153,770.8475 ด้วยวิธีแผนภาพกล่อง ซึ่งมีค่านอกเกณฑ์ของราคาอาคารชุดต่อตารางเมตรทั้งหมด 117 โครงการ ทำการตัดค่านอกเกณฑ์ออกแล้ว ส่งผลให้ข้อมูลอาคารชุดในกรุงเทพมหานครเหลือ 2,403 - 117 = 2,286 โครงการ

3.1.4 การแปลงข้อมูลเชิงคุณภาพโดยวิธี One-Hot Encoding

One-Hot Encoding เป็นกระบวนการที่ใช้ในการแปลงข้อมูลที่มีลักษณะเป็นตัวแปรแบบแบ่งกลุ่มหรือแบบสัญลักษณ์ เป็นรูปแบบที่เหมาะสมสำหรับการประมวลผลโดยอัตโนมัติ โดยพิจารณาตัวแปรที่มีการแบ่งกลุ่มหรือสัญลักษณ์เป็นประเภทข้อมูลที่สามารถนับได้จำกัด จากตารางที่ 3.1 ตัวแปร ชื่อหัวข้อ คำอธิบาย หน่วย และประเภทของข้อมูล Baania รวมกับข้อมูลที่เพิ่มเติมเกี่ยวกับจำนวนสถานที่สำคัญรอบโครงการ พบว่ามีข้อมูลเชิงคุณภาพคือ ตัวแปรรหัสไปรษณีย์ ดังรูปที่ 3.3

Price per sqm	zipcode	c_room_bath	c_room_bed	c_floor	c_parking
25806.45	10260	1	1	8	90.000000
7791.67	10220	1	0	5	47.923413
30833.33	10230	2	2	8	140.000000
143817.53	10110	2	2	8	70.000000
101666.67	10900	1	0	32	237.000000
14615.38	10260	1	2	8	245.076692
62686.57	10210	1	1	8	208.679163
71428.57	10120	3	2	24	247.075518
19000.00	10230	2	3	8	59.000000
98039.22	10110	1	2	51	426.000000
54264.71	10330	2	2	33	313.018745
83076.92	10500	2	2	40	465.000000

รูปที่ 3.3 ข้อมูลตัวอย่างของชุดข้อมูล

รูปที่ 3.3 แสดงข้อมูลตัวอย่างของชุดข้อมูล ซึ่งจะเห็นว่าตัวแปรรหัสไปรษณีย์เป็นข้อมูลเชิงคุณภาพ ดังนั้นจะทำการแปลงข้อมูลรหัสไปรษณีย์โดยใช้วิธี One-Hot Encoding หรือการทำข้อมูลที่ถูกเก็บในลักษณะเชิงคุณภาพ ทั้งในลักษณะที่มีลำดับและไม่มีลำดับเปลี่ยนให้อยู่ในรูปแบบของค่าทวิภาคที่มีค่า 0 หรือ 1 เท่านั้น ดังรูปที่ 3.4

zipcode_10400	zipcode_10500	zipcode_10510	zipcode_10520	zipcode_10530	zipcode_10600	zipcode_10700	zipcode_10800	zipcode_10900	zipcode_11000
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0

รูปที่ 3.4 ข้อมูลของรหัสไปรษณีย์หลังแปลงข้อมูลด้วยวิธี One-Hot Encoding

รูปที่ 3.4 แสดงข้อมูลของรหัสไปรษณีย์หลังแปลงข้อมูลด้วยวิธี One-Hot Encoding กล่าวคือ จากข้อมูลของตัวแปรรหัสไปรษณีย์ ที่มีรหัสไปรณีย์ 31 รหัส เมื่อแปลงข้อมูลด้วยวิธี One-Hot Encoding ส่งผลให้มีจำนวนหลักหรือจำนวนตัวแปรอิสระเพิ่มขึ้นมา 31 ตัว หรือตัวแปรอิสระทั้งหมดมี 62 ตัว ในตัวแปรของรหัสไปรณีย์ที่เพิ่มมานั้น จะมีค่า 0 กับ 1

3.1.5 การแบ่งข้อมูล (Data Splitting)

การแบ่งข้อมูลด้วยวิธีตรวจสอบไขว้ คือกระบวนการในการประเมินแบบแบบจำลองโดยใช้ข้อมูลที่มีอยู่ให้ได้มากที่สุด โดยแบ่งข้อมูลออกเป็นส่วนย่อยๆ ซึ่งทางผู้วิจัยกำหนดค่า K คือ 5 และแต่ละชุดข้อมูลจะถูกเลือกขึ้นโดยสุ่มและไม่ซ้ำกัน ทำให้การประเมินประสิทธิภาพของแบบจำลองได้มากขึ้น โดยทำการแบ่งข้อมูลเป็น 2 ชุด โดยมีสัดส่วนของชุดข้อมูลฝึกสอน 80% และชุดข้อมูลทดสอบ 20% เมื่อได้ชุดข้อมูลฝึกสอนจะใช้วิธีการตรวจสอบไขว้ 5 ชุด เพื่อแบ่งข้อมูลออกเป็น 5 ชุดเท่าๆกัน แล้วแต่ละชุดจะได้ 4 ชุดข้อมูลฝึกสอน และ 1 ชุดข้อมูลตรวจสอบ ที่ไม่ซ้ำกันกับชุดถัดไป ดังรูปที่ 3.5

```
Fold 1:
Training set size: 1462
Validation set size: 366
==================================================
Fold 2:
Training set size: 1462
Validation set size: 366
==================================================
Fold 3:
Training set size: 1462
Validation set size: 366
==================================================
Fold 4:
Training set size: 1463
Validation set size: 365
==================================================
Fold 5:
Training set size: 1463
Validation set size: 365
==================================================
Test set size: 458
```

รูปที่ 3.5 จำนวนชุดข้อมูลทดสอบ ข้อมูลฝึกสอน และข้อมูลตรวจสอบ

รูปที่ 3.5 แสดงชุดข้อมูลทดสอบ 458 จำนวน ที่คิดเป็น 20% ของชุดข้อมูลทั้งหมด และชุดข้อมูลฝึกสอน 1,828 จำนวน ซึ่งจะถูกแบ่งข้อมูลด้วยการตรวจสอบไขว้ 5 ชุด พบว่าการตรวจสอบไขว้รอบที่ 1 ถึง 3 ได้ชุดข้อมูลฝึกสอนเท่ากันคือ 1,462 จำนวน และชุดข้อมูลตรวจสอบเท่ากันคือ 366 จำนวน ส่วนรอบที่ 4 ถึง 5 ได้ชุดข้อมูลฝึกสอน 1,463 จำนวน และชุดข้อมูลตรวจสอบ 365 จำนวน โดยชุดข้อมูลที่ถูกเลือกด้วยวิธีการตรวจสอบแบบไขว้จะไม่ซ้ำกันในแต่ละรอบ ทำให้แบบจำลองมีประสิทธิภาพมากขึ้น

3.2 การออกแบบแบบจำลอง

การออกแบบแบบจำลองประกอบด้วยการเลือกคุณลักษณะและการสร้างแบบจำลอง

3.2.1 การเลือกคุณลักษณะ

ผู้วิจัยได้ทำการเลือกคุณลักษณะเป็น 2 วิธี ซึ่งวิธีที่ 1 คือวิธีการเลือกคุณลักษณะจากงานวิจัยที่ผ่านมาได้แก่ วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด และวิธีที่ 2 คือวิธีเลือกคุณลักษณะที่ผู้วิจัยเสนอได้แก่ วิธีการนำตัวแปรเข้าทั้งหมด ดังนี้

1. วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Selection)

วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดมีตัวแปรอิสระที่ถูกเลือกจากวัดความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรอิสระตามด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดมี 31 ตัวแปร จาก 62 ตัวแปรทั้งหมด ดังตารางที่ 3.3

ตารางที่ 3.3 ชื่อตัวแปร และค่าเอฟระหว่างตัวแปรอิสระกับตัวแปรตามด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

ชื่อตัวแปร	ค่าเอฟระหว่างตัวแปรอิสระและตัวแปรตาม
Hospital	361.86
Count_floor	214.59
Attraction	211.07
Train	173.57
Year finish	165.09
University	101.53
count_unittype	77.30
facility_meeting	74.94
zipcode_10330	65.21
year update	56.53
count_elevator	52.19
count_room_kitchen	50.32
count_room_dinning	46.26
zipcode_10500	29.25
count_room_living	27.90
count_unit	26.79

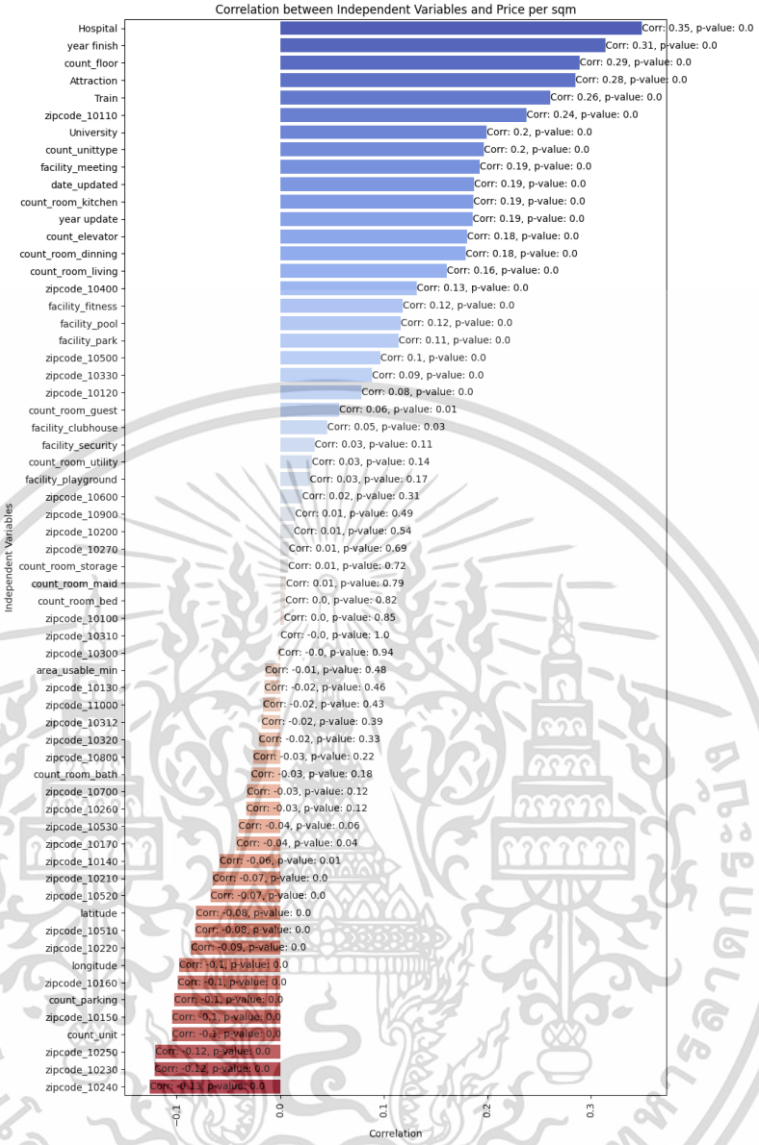
ตารางที่ 3.3 ชื่อตัวแปร และค่าเอฟระหว่างตัวแปรอิสระกับตัวแปรตามด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด (ต่อ)

ชื่อตัวแปร	ค่าเอฟระหว่างตัวแปรอิสระและตัวแปรตาม
zipcode_10400	23.28
facility_pool	19.27
zipcode_10250	18.20
longitude	16.39
count_parking	15.90
zipcode_10220	15.88
latitude	15.61
count_room_guest	12.86
zipcode_10800	4.52
count_room_maid	3.11
Count_room_bath	3.08
date_updated	2.76
area_usable_min	1.54
count_room_bed	1.42
zipcode_10312	1.06

ตารางที่ 3.3 แสดงชื่อตัวแปร และการเรียงลำดับค่าเอฟระหว่างตัวแปรอิสระกับตัวแปรตามด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด 31 ตัวแปร โดยตัวแปรอิสระที่ส่งผลต่อตัวแปรตามมากเป็น 3 อันดับแรก คือ จำนวนโรงพยาบาลภายในรัศมี 3 กิโลเมตร (Hospital) ด้วยค่าเอฟ 361.86 รองลงมาคือจำนวนชั้นของโครงการ (Count_floor) ด้วยค่าเอฟ 214.59 และจำนวนสถานที่ท่องเที่ยวภายในรัศมี 3 กิโลเมตร (Attraction) ด้วยค่าเอฟ 211.07

2. วิธีการนำตัวแปรเข้าทั้งหมด (Enter)

วิธีการนำตัวแปรเข้าทั้งหมดเป็นวิธีการนำตัวแปรอิสระทุกตัวทั้งที่มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติและไม่มีนัยสำคัญทางสถิติเข้าไปวิเคราะห์ในสมการถดถอย ตัวแปรอิสระมีทั้งหมด 62 ตัว เพื่อศึกษาปัจจัยที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร ในขั้นตอนนี้ได้ทำการหาสัมประสิทธิ์สหสัมพันธ์เพิ่มเติม เพื่อวัดความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม ดังรูปที่ 3.6



รูปที่ 3.6 สัมประสิทธิ์สหสัมพันธ์และค่าพีระหว่างตัวแปรอิสระและตัวแปรตาม

รูปที่ 3.6 แสดงการเรียงลำดับสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยตัวแปรอิสระที่ส่งผลต่อตัวแปรตามมีจำนวน 24 ตัวแปร เช่น จำนวนโรงพยาบาลภายในรัศมี 3 กิโลเมตร (Hospital) มีค่าสัมประสิทธิ์สหสัมพันธ์มากที่สุดคือ 0.35 รองลงมาคือ จำนวนปีตั้งแต่โครงการสร้างเสร็จ (Year finish) มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.31 จำนวนชั้นของโครงการ (Count_floor) มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.29 จำนวนสถานที่ท่องเที่ยวภายในรัศมี 3 กิโลเมตร (Attraction) มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.28 จำนวนสถานีรถไฟฟ้าบนดินและใต้ดินภายในรัศมี 0.4 กิโลเมตร (Train) มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.26 โดยตัวแปรอิสระทุกตัวที่กล่าวข้างต้นมีค่าพีเท่ากับ 0.00 ซึ่งหมายถึงค่าพีที่น้อยกว่าระดับนัยสำคัญ 0.05 ดังนั้นจึงสามารถปฏิเสธสมมติฐานว่างและสรุปได้ว่ามีความสัมพันธ์

ระหว่างตัวแปรอิสระที่กล่าวข้างต้นทั้ง 24 ตัวแปรกับราคาอาคารชุดในกรุงเทพมหานครที่ระดับนัยสำคัญ 0.05

จากการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามพบว่า ตัวแปรอิสระที่เพิ่มเติมจาก Google Maps อีก 4 ตัว จากข้อมูลจำนวนสถานที่สำคัญรอบโครงการ ได้แก่ จำนวนสถานีรถไฟฟ้าบนดินและใต้ดินภายในรัศมี 0.4 กิโลเมตร (Train, X_{29}) จำนวนโรงพยาบาลภายในรัศมี 3 กิโลเมตร (Hospital, X_{30}) จำนวนมหาวิทยาลัยภายในรัศมี 3 กิโลเมตร (University, X_{31}) และจำนวนสถานที่ท่องเที่ยวภายในรัศมี 3 กิโลเมตร (Attraction, X_{32}) ส่งผลต่อตัวแปรตามสูง ซึ่งข้อมูลดังกล่าวเป็นข้อมูลเกี่ยวกับสถานที่ตั้ง (Location) สรุปได้ว่าข้อมูลสถานที่ตั้งเป็นปัจจัยหลักที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร

3.2.2 การสร้างแบบจำลอง

การแบ่งการทดลองออกเป็น 5 รูปแบบดังนี้ ซึ่งวิธีที่ 1-3 คือวิธีการทดสอบโดยการเรียนรู้ของเครื่อง สำหรับวิธีที่ 4-5 คือวิธีการทดสอบโดยการเรียนรู้เชิงลึก

1. วิธีป่าสุ่ม (Random Forest)
2. วิธีการประเมินราคาแอบแฝง (Hedonic Price Method)
3. วิธี Extreme Gradient Boosting
4. วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)
5. วิธีหน่วยความจำระยะสั้น-ยาว (Long Short-Term Memory)

3.3 การเปรียบเทียบประสิทธิภาพแบบจำลอง

การเปรียบเทียบประสิทธิภาพของแบบจำลองทั้งหมด 5 วิธี ด้วยมาตรวัดประสิทธิภาพจำนวน 3 ค่า คือ สัมประสิทธิ์การกำหนด โดยพิจารณาจากสัมประสิทธิ์การกำหนดที่มีค่าสูงสุดหรือเข้าใกล้ 1 ส่วนค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยและรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย โดยพิจารณาจากค่าที่ต่ำสุด จะมีประสิทธิภาพของแบบจำลองดีที่สุด

บทที่ 4
ผลการวิจัยและการอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยวิธีการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก นำข้อมูลมาจาก Baania (2566) ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ จากนั้นทำเก็บรวบรวมข้อมูลพิกัดละติจูดและลองจิจูดเพิ่มเติมจาก Google Maps เพื่อเพิ่มข้อมูลจำนวนสถานที่สำคัญในบริเวณโครงการ และทำการเตรียมข้อมูล เพื่อที่จะไปสู่ขั้นตอนการออกแบบจำลองด้วยการเลือกคุณลักษณะ 2 วิธี คือ วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด และวิธีการนำตัวแปรเข้าทั้งหมด เมื่อทำการเลือกคุณลักษณะ จะได้ชุดข้อมูลทั้งหมด 2 ชุด นำชุดข้อมูลที่ได้มานั้นสร้างแบบจำลองทั้งหมด 5 วิธี โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting และอีก 2 วิธีด้วยแบบจำลองการเรียนรู้เชิงลึก คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว ขั้นตอนสุดท้ายทำการเปรียบเทียบประสิทธิภาพแบบจำลอง ผู้ทำวิจัยกำหนดใช้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย

4.1 ผลการวิเคราะห์ของการเรียนรู้ของเครื่อง

4.1.1 วิธีป่าสุ่ม (Random Forest)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีป่าสุ่มในการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด ได้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังตารางที่ 4.1

ตารางที่ 4.1 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีป่าสุ่มโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด

การเลือกคุณลักษณะ	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด	0.5967	**15,579.75**	20,219.40
วิธีการนำตัวแปรเข้าทั้งหมด	**0.5996**	15,617.50	**20,147.78**

จากตารางที่ 4.1 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีป่าสุ่มโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีป่าสุ่มโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดมีสัมประสิทธิ์การกำหนดสูงที่สุดเท่ากับ 0.5997 และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดเท่ากับ 20,147.78 รองลงมาคือการทำนายของแบบจำลองวิธีป่าสุ่มโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำที่สุดเท่ากับ 15,579.75 จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีป่าสุ่มโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.1.2 วิธีการประเมินราคาแอบแฝง (Hedonic Price Method)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีการประเมินราคาแอบแฝงในการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด ได้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยและรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังตารางที่ 4.2

ตารางที่ 4.2 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีการประเมินราคาแอบแฝงโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด

การเลือกคุณลักษณะ	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด	0.4583	18,234.58	23,433.85
วิธีการนำตัวแปรเข้าทั้งหมด	**0.4618**	**18,227.33**	**23,358.77**

จากตารางที่ 4.2 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีการประเมินราคาแอบแฝงโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีการประเมินราคาแอบแฝงโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดเท่ากับ 0.4618, 18,227.33 และ 23,358.77 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีการประเมินราคาแอบแฝงโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.1.3 วิธี Extreme Gradient Boosting

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธี Extreme Gradient Boosting ในการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด ได้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังตารางที่ 4.3

ตารางที่ 4.3 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด

การเลือกคุณลักษณะ	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด	0.6076	**15,210.41**	19,944.64
วิธีการนำตัวแปรเข้าทั้งหมด	**0.6079**	15,243.88	**19,936.09**

จากตารางที่ 4.3 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนดสูงที่สุดและรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดเท่ากับ 0.6079 และ 19,936.09 ตามลำดับ รองลงมาคือการทำนายของแบบจำลองวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด มีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยต่ำที่สุดเท่ากับ 15,210.41 จึงสรุปได้ว่าการทำนายของแบบจำลองวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.2 ผลการวิเคราะห์ของการเรียนรู้เชิงลึก

4.2.1 วิธีโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีโครงข่ายประสาทแบบคอนโวลูชันในการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด ได้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังตารางที่ 4.4

ตารางที่ 4.4 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด

การเลือกคุณลักษณะ	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด	0.5916	15,439.85	20,347.63
วิธีการนำตัวแปรเข้าทั้งหมด	**0.6160**	**15,042.54**	**19,728.69**

จากตารางที่ 4.4 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดเท่ากับ 0.6160, 15,042.54 และ 19,728.69 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.2.2 วิธีหน่วยความจำระยะสั้น-ยาว (Long Short-Term Memory)

จากการสร้างแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีหน่วยความจำระยะสั้น-ยาว ในการทำนายราคาอาคารชุดในกรุงเทพมหานคร โดยใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด ได้สัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ดังตารางที่ 4.5

ตารางที่ 4.5 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด

การเลือกคุณลักษณะ	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด	0.5922	16,128.00	20,332.98
วิธีการนำตัวแปรเข้าทั้งหมด	**0.6158**	**15,439.82**	**19,736.09**

จากตารางที่ 4.5 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดเท่ากับ 0.6158, 15,439.82 และ 19,736.09 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีหน่วยความจำระยะสั้น-ยาวโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุด

4.3 การเปรียบเทียบแบบจำลอง

4.3.1 การเปรียบเทียบแบบจำลองของการเรียนรู้ของเครื่อง

จากตาราง 4.1, 4.2 และ 4.3 แสดงผลการเปรียบเทียบประสิทธิภาพของการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดกับการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดในการทำนายของแบบจำลองวิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting ในการทำนายราคาอาคารชุดในกรุงเทพมหานคร พบว่า การใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดดีกว่าการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด ดังนั้นทำการเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่องด้วยวิธีการนำตัวแปรเข้าทั้งหมด ดังตาราง 4.6

ตารางที่ 4.6 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boostingโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด

การเรียนรู้ของเครื่อง	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีป่าสุ่ม	0.5996	15,617.50	20,147.78
วิธีการประเมินราคาแอบแฝง	0.4583	18,234.58	23,433.85
วิธี Extreme Gradient Boosting	**0.6079**	**15,243.88**	**19,936.09**

จากตารางที่ 4.6 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมดมีประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่อง มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดเท่ากับ 0.6079, 15,243.88 และ 19,936.09 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่อง

4.3.2 การเปรียบเทียบแบบจำลองของการเรียนรู้เชิงลึก

จากผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว ในการทำนายราคาอาคารชุดในกรุงเทพมหานคร พบว่าการใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ประสิทธิภาพการทำนายของแบบจำลองของการเรียนรู้เชิงลึกดีที่สุด ดังตารางที่ 4.7

ตารางที่ 4.7 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด

การเรียนรู้ของเครื่อง	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีโครงข่ายประสาทแบบคอนโวลูชัน	**0.6160**	**15,042.54**	**19,728.69**
วิธีหน่วยความจำระยะสั้น-ยาว	0.6158	15,439.82	19,736.09

จากตารางที่ 4.7 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดเท่ากับ 0.6160, 15,042.54 และ 19,728.69 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้เชิงลึก

4.3.3 การเปรียบเทียบแบบจำลองของการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

จากผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองทั้งหมด 5 วิธี โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting และแบบจำลองการเรียนรู้เชิงลึก 2 วิธี คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว ในการทำนายราคาอาคารชุดในกรุงเทพมหานคร พบว่าการใช้ข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด ให้ประสิทธิภาพการทำนายของแบบจำลองของการเรียนรู้เชิงลึกดีที่สุด ดังตารางที่ 4.8

ตารางที่ 4.8 ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองทั้งหมด 5 วิธี โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting และแบบจำลองการเรียนรู้เชิงลึก 2 วิธี คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด

การเรียนรู้ของเครื่อง	ผลการทำนายเทียบกับข้อมูลทดสอบ		
	สัมประสิทธิ์การกำหนด	ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย	รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย
วิธีป่าสุ่ม	0.5996	15,617.50	20,147.78
วิธีการประเมินราคาแอบแฝง	0.4618	18,227.33	23,358.77
วิธี Extreme Gradient Boosting	0.6079	15,243.88	19,936.09
วิธีโครงข่ายประสาทแบบคอนโวลูชัน	**0.6160**	**15,042.54**	**19,728.69**
วิธีหน่วยความจำระยะสั้น-ยาว	0.6158	15,439.82	19,736.09

จากตารางที่ 4.8 แสดงผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองทั้งหมด 5 วิธี โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting และแบบจำลองการเรียนรู้เชิงลึก 2 วิธี คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุดเท่ากับ 0.6160, 15,042.54 และ 19,728.69 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือก

คุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

4.4 การอภิปรายผล

จากการศึกษาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามพบว่า ข้อมูลสถานที่ตั้งเป็นปัจจัยหลักที่ส่งผลกับราคาอาคารชุดในกรุงเทพมหานคร ซึ่งสอดคล้องกับงานวิจัยที่กล่าวมาของ Pensri et al. (2020) พบว่าปัจจัยที่ส่งผลต่อการซื้ออาคารชุดในกรุงเทพมหานครคือ ปัจจัยสถานที่ตั้งของอาคารชุดให้สัมประสิทธิ์สหสัมพันธ์เพียร์สันสูงที่สุด และ Guliker et al. (2022) พบว่าปัจจัยที่ส่งผลต่อการประเมินอสังหาริมทรัพย์ในประเทศเนอเธอร์แลนด์ คือ ปัจจัยสถานที่ตั้ง สำหรับงานวิจัยของพสธร (2563) พบว่าปัจจัยด้านสภาพแวดล้อม ปัจจัยด้านตัวโครงการ และปัจจัยสถานที่ตั้ง มีความสัมพันธ์ต่อราคาเสนอขายมากไปน้อยที่สุด ตามลำดับ ซึ่งไม่สอดคล้องกับงานวิจัย และงานวิจัยของ Piao et al. (2019) ได้ศึกษาเกี่ยวกับการพยากรณ์ราคาที่อยู่อาศัยในต้าเหลียน ประเทศจีน พบว่าปัจจัยที่ส่งผลกับราคาที่อยู่อาศัยในต้าเหลียนสูงสุด 3 อันดับแรกคือ พื้นที่บ้าน ราคาทำธุรกรรมที่อยู่อาศัย และผลิตภัณฑ์มวลรวมของประเทศ ซึ่งไม่สอดคล้องกับงานวิจัยของผู้วิจัย เนื่องจากไม่ได้ศึกษาปัจจัยสถานที่ตั้ง

การเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองกล่าวได้ว่าการใช้ชุดข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดได้ค่าประสิทธิภาพที่ดีกว่าชุดข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด ซึ่งไม่สอดคล้องกับงานวิจัยที่กล่าวมาของ Jafari and Akhavian (2019) เนื่องจากงานวิจัยนี้ไม่ได้ศึกษาการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด

จากการศึกษาแบบจำลองของการเรียนรู้ของเครื่อง 3 วิธี คือวิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมดมีประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่อง มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด เท่ากับ 0.6079, 15,243.88 และ 19,936.09 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่อง ซึ่งสอดคล้องกับงานวิจัยของ Viriya (2021) และ Guliker et al. (2022) พบว่าแบบจำลองวิธี Extreme Gradient Boosting ดีสุด ส่วนงานวิจัยของ Baldominos et al. (2018) พบว่าวิธีป่าสุ่มดีที่สุด และงานวิจัยของ Jafari and Akhavian (2019) พบว่าแบบจำลองวิธีการประเมินราคาแอบแฝงดีที่สุด ซึ่งได้ผลที่ไม่สอดคล้องกัน เนื่องจากไม่ได้ทำนายด้วยวิธี Extreme Gradient Boosting

นอกจากนี้ยังศึกษาแบบจำลองของการเรียนรู้เชิงลึก 2 วิธี คือวิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด เท่ากับ 0.6160, 15,042.54 และ 19,728.69 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้เชิงลึก ซึ่งสอดคล้องกับงานวิจัยของ Piao et al. (2019) และ Rupesh and Kumar (2023) ซึ่งพบว่าแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันมีประสิทธิภาพดีสุด ส่วนงานวิจัยของ พสธร (2563) ไม่ได้ทำวิธีโครงข่ายประสาทแบบคอนโวลูชัน จึงได้ผลที่ไม่สอดคล้องกัน

เมื่อนำแบบจำลองของการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกมารวมกันทั้งหมด 5 วิธี โดยแบ่งเป็นแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting และแบบจำลองการเรียนรู้เชิงลึก 2 วิธี คือ วิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยดีที่สุด เท่ากับ 0.6160, 15,042.54 และ 19,728.69 ตามลำดับ จึงสรุปได้ว่าการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ซึ่งประสิทธิภาพการทำนายของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมดสอดคล้องกับงานวิจัยของ Piao et al. (2019) และ Rupesh and Kumar (2023) พบว่าแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันดีที่สุด

จากการเปรียบเทียบแบบจำลองทั้งหมด สรุปได้คือแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันดีที่สุด ให้ผลการทำนายราคาอาคารชุดในกรุงเทพมหานคร ดังรูปที่ 4.1

Pricepersqm	Predicted_Price	count_room_bath	count_room_bed	count_floor	count_parking	count_unit	count_unittype	facility_clubhouse	facility_fitness	...	zipcode_10400	zipcode_10500
102325.58	93848.984375	3.0	3.0	33.0	495.000000	231.0	2.0	1.0	1.0	...	0.0	0.0
78681.32	98218.906250	1.0	1.0	30.0	518.058156	854.0	2.0	1.0	1.0	...	0.0	0.0
17241.38	22600.244141	1.0	2.0	6.0	349.416274	576.0	1.0	0.0	0.0	...	0.0	0.0
30769.23	33304.339844	2.0	2.0	8.0	60.000000	79.0	3.0	1.0	1.0	...	0.0	0.0
52500.00	45294.890625	1.0	1.0	8.0	389.000000	1114.0	1.0	1.0	1.0	...	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
37894.74	45464.523438	1.0	2.0	8.0	59.000000	158.0	2.0	1.0	1.0	...	0.0	0.0
40625.00	42537.136719	1.0	0.0	8.0	202.006283	333.0	1.0	0.0	1.0	...	0.0	0.0
68627.45	51336.519531	2.0	3.0	8.0	44.890285	74.0	3.0	1.0	1.0	...	0.0	0.0
88556.20	79248.015625	1.0	2.0	8.0	115.000000	147.0	2.0	1.0	1.0	...	0.0	0.0
33823.53	43590.531250	1.0	0.0	9.0	278.441093	459.0	2.0	0.0	0.0	...	1.0	0.0

รูปที่ 4.1 ผลการทำนายราคาอาคารชุดในกรุงเทพมหานครของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมด

จากรูปที่ 4.1 แสดงผลการทำนายราคาอาคารชุดในกรุงเทพมหานครของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมด เช่น จำนวนห้องน้ำ (count_room_bath) เท่ากับ 3 ห้อง จำนวนห้องนอน (count_room_bed) เท่ากับ 3 ห้อง จำนวนชั้นของโครงการ (count_floor) เท่ากับ 33 ชั้น จนถึง รหัสไปรษณีย์ที่ 10500 (zipcode_10500) เท่ากับ 0 พบว่าค่าทำนายราคาอาคารชุด (Predicted_Price) เท่ากับ 93,848.98 บาทต่อตารางเมตร ส่วนค่าทำนายราคาอาคารชุดอื่นๆ ก็สามารถหาในทำนองเดียวกัน

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาการทำนายราคาอาคารชุดในกรุงเทพมหานครโดยใช้การเลือกคุณลักษณะ 2 วิธี คือ วิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด และวิธีการนำตัวแปรเข้าทั้งหมด ทำให้มีชุดข้อมูล 2 ชุดในการสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง ซึ่งมีวัตถุประสงค์ดังนี้ เพื่อศึกษาแบบจำลองที่ช่วยลดความเสี่ยงในการลงทุนกับอาคารชุดในกรุงเทพมหานคร โดยเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก และเพื่อศึกษาปัจจัยที่ส่งผลต่อราคาอาคารชุดในกรุงเทพมหานคร

การศึกษาครั้งนี้ได้นำข้อมูลมาจาก Baania (2566) ที่เป็นแหล่งรวมโครงการอสังหาริมทรัพย์ทั้งหมด 2,403 โครงการ และเก็บรวบรวมข้อมูลเพิ่มเติมจาก Google Maps เพื่อเพิ่มข้อมูลจำนวนสถานที่สำคัญในบริเวณโครงการ และทำการเตรียมข้อมูล เพื่อที่จะไปสู่ขั้นตอนการสร้างแบบจำลองด้วยการเลือกคุณลักษณะ จากนั้นนำเข้าแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก และทำการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกในการทำนายราคาอาคารชุดในกรุงเทพมหานคร ซึ่งสามารถแสดงการสรุปผลการวิจัยและข้อเสนอแนะได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 คุณลักษณะที่สำคัญ

จากสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม พบว่าข้อมูลสถานที่ตั้งมีสัมประสิทธิ์สหสัมพันธ์สูงสุดกับตัวแปรตาม ดังนั้นปัจจัยหลักที่ส่งผลกับราคาอาคารชุดในกรุงเทพมหานครคือ ปัจจัยสถานที่ตั้งของอาคารชุด

5.1.2 การเลือกคุณลักษณะ

ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองทั้งหมด 5 วิธี โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง 3 วิธี คือ วิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting และแบบจำลองการเรียนรู้เชิงลึก 2 วิธีคือ วิธีโครงข่ายประสาทแบบคอนโวลูชันและวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดและวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองทั้งหมด 5 วิธี โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด มีสัมประสิทธิ์การกำหนด และรากของค่าคลาดเคลื่อน

กำลังสองเฉลี่ยดีที่สุด ถึงแม้ว่าการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดได้สัมประสิทธิ์การกำหนดและรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่า แต่เทียบกับจำนวนตัวแปรอิสระของวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดน้อยกว่าวิธีการนำตัวแปรเข้าทั้งหมด 31 ตัวแปร ส่งผลให้ระยะเวลาในการสร้างแบบจำลองน้อยลง เมื่อใช้งานจริงเห็นสมควรนำชุดข้อมูลจากการเลือกคุณลักษณะด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุดจะเหมาะสมมากกว่า

5.1.3 การเรียนรู้ของเครื่อง

ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีป่าสุ่ม วิธีการประเมินราคาแอบแฝง และวิธี Extreme Gradient Boosting โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมดมีประสิทธิภาพการทำนายดีที่สุดสำหรับการเรียนรู้ของเครื่อง

5.1.4 การเรียนรู้เชิงลึก

ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชัน และวิธีหน่วยความจำระยะสั้น-ยาว โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดดีที่สุดสำหรับการเรียนรู้เชิงลึก

5.1.5 การเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก

ผลการเปรียบเทียบประสิทธิภาพการทำนายของแบบจำลองของการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึก โดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมด พบว่าประสิทธิภาพการทำนายของแบบจำลองวิธีโครงข่ายประสาทแบบคอนโวลูชันโดยการเลือกคุณลักษณะด้วยวิธีการนำตัวแปรเข้าทั้งหมดให้ค่าประสิทธิภาพการทำนายดีที่สุดสำหรับการทำนายราคาอาคารชุดในกรุงเทพมหานคร

5.2 ข้อจำกัดและข้อเสนอแนะ

5.2.1 ข้อจำกัด

1. เนื่องจากเวลาในการทำวิจัยมีจำกัด จึงไม่สามารถคัดคุณสมบัติหรือเลือกตัวแปรอิสระเพื่อสร้างแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกได้ดี

2. การสร้างแบบจำลองบางวิธีจะต้องใช้การทำนายที่นาน อุปกรณ์ที่ใช้ในการวิเคราะห์ควรมีความพร้อมและประสิทธิภาพสูงในการทดสอบ

5.2.2 ข้อเสนอแนะ

1. เนื่องจากข้อมูลมีจำนวน 2,403 โครงการ ซึ่งข้อมูลมีจำนวนน้อย ต้องหาข้อมูลโครงการอาคารชุดเพิ่มเพื่อทำให้แบบจำลองมีประสิทธิภาพในการทำนายมากขึ้น และควรเพิ่มข้อมูลเกี่ยวกับเหตุการณ์ไม่ปกติเกิดขึ้น ซึ่งอาจส่งผลกับราคาอาคารชุดในกรุงเทพมหานครได้ เช่น การเกิดโรคระบาดและสงครามต่างประเทศ เป็นต้น เพื่อให้แบบจำลองสามารถปรับปรุงและทำนายราคาอาคารชุดในสถานการณ์ที่เปลี่ยนแปลงได้มากยิ่งขึ้น

2. เนื่องจากข้อมูลมีตัวแปรจำนวนมากถึง 62 ตัวแปร จึงมีความจำเป็นที่จะต้องมีการเลือกคุณลักษณะ เช่น การวิเคราะห์ส่วนประกอบหลัก (Principal Component Analysis) หรือการวิเคราะห์ปัจจัย (Factor Analysis) เป็นต้น เพื่อลดจำนวนตัวแปรให้เหลือเพียงแค่ตัวแปรที่มีความสำคัญ และมีผลต่อผลลัพธ์ที่ต้องการในการทำนายราคาอาคารชุดในกรุงเทพมหานคร

3. ควรมีการทดสอบแบบจำลองที่มากกว่านี้เพื่อพิจารณาเปรียบเทียบว่าแบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกวิธีไหนดีกว่ากันในการทำนายราคาอาคารชุดในกรุงเทพมหานคร เช่น วิธีต้นไม้ตัดสินใจ (Decision Tree) และวิธีหน่วยเวียนกลับมีประตู (Gated Recurrent Unit) เป็นต้น

เอกสารอ้างอิง

ศูนย์ข้อมูลอสังหาริมทรัพย์. 2565. **จำนวนหน่วยของยอดขายที่อยู่อาศัยใหม่ (กรุงเทพฯ - ปริมณฑล).** [Online]. Available : https://www.reic.or.th/Product/AllChart.

Baldominos, A. Blanco, I. Moreno, A.J. Iturrarte, R. Bernárdez, Ó. and Afonso, C. 2018. "Identifying Real Estate Opportunities Using Machine Learning." *International Journal of Data Science and Analytics*. 8(11) : 2321. https://doi.org/10.1007/s41060-018-00170-0.

Piao, Y. Chen, A. and Shang, Z. 2019. "Housing Price Prediction Based on CNN." 2019 9th International Conference on Information Science and Technology (ICIST). 9 : 491-495. https://doi.org/10.1109/ICIST.2019.8836731.

Jafari, A. and Akhavian, R. 2019. "Driving Forces for the US Residential Housing Price: A Predictive Analysis." *Built Environment Project and Asset Management*. 9(4) : 515-529. https://doi.org/10.1108/BEPAM-07-2018-0100.

พสธร วิทยาปรีชาพล. 2563. "การประเมินราคาเสนอขายห้องชุดด้วย Deep Neural Network และ K-Means Clustering Algorithm" วิทยานิพนธ์มหาบัณฑิต สาขาวิชาธุรกิจอสังหาริมทรัพย์ คณะพาณิชยศาสตร์และการบัญชี, มหาวิทยาลัยธรรมศาสตร์.

Pensri, B. Thanakorn, T. Yannakorn, T. and Suriya, K. 2020. "Marketing Factors That Affecting The Purchase Of Condominium In Bangkok Thailand" PSYCHOLOGY AND EDUCATION. 58(1) : 4434-4438.

Viriya, T. 2021. "Google Maps Amenities and Condominium Prices : Investigating The Effects and Relationships Using Machine Learning." *Habitat International*. 2021(118) : 1-12. https://doi.org/10.1016/j.habitatint.2021.102463.

Guliker, E. Folmer, E. and Sinderen, M.V. 2022. "Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach." *International Journal of Geo Information*. 11(2) : 125. https://doi.org/10.3390/ijgi11020125.

เอกสารอ้างอิง (ต่อ)

Rupesh, S. and Kumar, S. 2023. “CNN predicts house or apartment rent more accurately than Naive Bayes does in major cities.” Journal of Survey in Fisheries Sciences. 10(1S) : 2663-2674. https://doi.org/10.17762/sfs.v10i1S.496.

Baania. 2565. **DATA API.** [Online]. Available : https://baaniathailand.com/data.

ฐณัฐ วงศ์สายเชื้อ. 2559. **Replace Missing Value – การแทนค่าสูญหายในโปรแกรม SPSS.** [Online]. Available : https://www.youtube.com/watch ?v=WzaeJ_HAqtk.

Paul Barrett. 2005. **Euclidean Distance (raw, normalized, and double-scaled coefficients).** [Online]. Available : https://www.pbarrett.net/techpapers/euclid.

Sasiwut Chaiyadecha. 2020. **One-Hot Encoding สร้างตัวแปร Dummies สำหรับ Classification model.** [Online]. Available : https://lengyi.medium.com/one-hot-encoding-737c66e5b1bd.

ทรงศักดิ์ ภูสีอ่อน. 2554. **การประยุกต์ใช้ SPSS วิเคราะห์ข้อมูลงานวิจัย**. มหาสารคาม : สำนักพิมพ์มหาวิทยาลัยมหาสารคาม.

พรชิตา ธนากร. 2560. “โครงข่ายประสามเทียมเชิงพยากรณ์แบบปรับปรุงโดยใช้การเลือกสับเซตที่ดีที่สุด”. วิทยานิพนธ์มหาบัณฑิต สาขาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี, มหาวิทยาลัยธรรมศาสตร์.

ธิดาเดียว มยุรีสวรรค์. 2559. **การวิเคราะห์การถดถอย (Regression Analysis).** พิมพ์ครั้งที่ 1. ขอนแก่น : บริษัท เพ็ญพรินติ้ง จำกัด.

อรพิน ประวัติบริสุทธิ์. 2564. **Python สำหรับงาน Data Science Data Visualization และ Machine Learning**. กรุงเทพฯ : โปรวิชั่น.

ภูริพัทร์ ทองคำ. 2559. “อัลกอริทึมแบบรวมสำหรับการเลือกคุณลักษณะของข้อมูล”. วิทยานิพนธ์มหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี, มหาวิทยาลัยธรรมศาสตร์.

โชติวุฒิ เหล่าไพโรจน์. 2555. “ปัจจัยกำหนดราคาคอนโดมิเนียมในเขตกรุงเทพมหานคร”. วิทยานิพนธ์มหาบัณฑิต สาขาเศรษฐศาสตร์ธุรกิจ คณะเศรษฐศาสตร์, มหาวิทยาลัยธรรมศาสตร์.

เอกสารอ้างอิง (ต่อ)

Nonita, S. 2017. **Extreme Gradient Boosting for Data Mining Applications.** European Union : LAP LAMBERT Academic Publishing.

ไกรศักดิ์ เกษร. 2564. **วิทยาศาสตร์ข้อมูล (Data Science).** พิษณุโลก : โรงพิมพ์มหาวิทยาลัยนเรศวร.

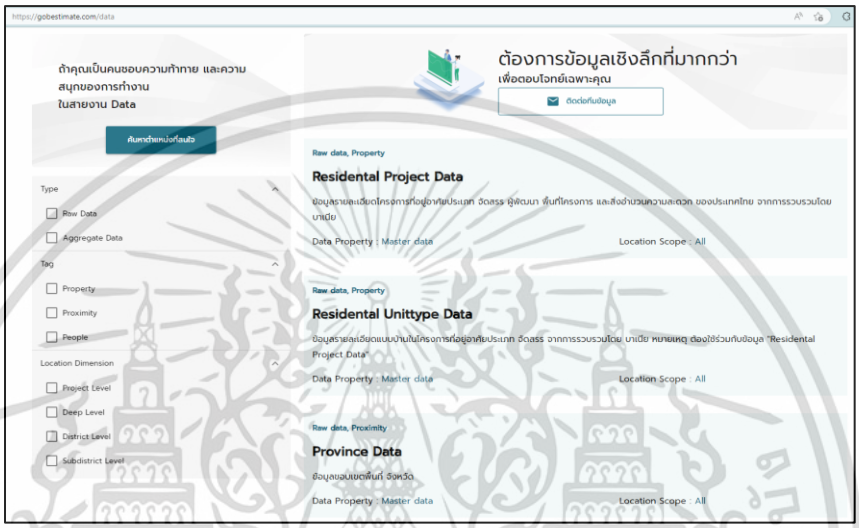
สายชล สินสมบูรณ์ทอง. 2559. **การทำเหมืองข้อมูล เล่ม 2 วิธีการและตัวแบบ Data Mining 2 : Methods and Models**. 1. กรุงเทพฯ : ศูนย์หนังสือจุฬาลงกรณ์มหาวิทยาลัย.

ยุทธ ไกยวรรณ์. 2549. **สถิติเพื่อการวิจัย.** พิมพ์ครั้งที่ 2. กรุงเทพฯ : ศูนย์สื่อเสริมกรุงเทพ.

ภาคผนวก

ภาคผนวก ก

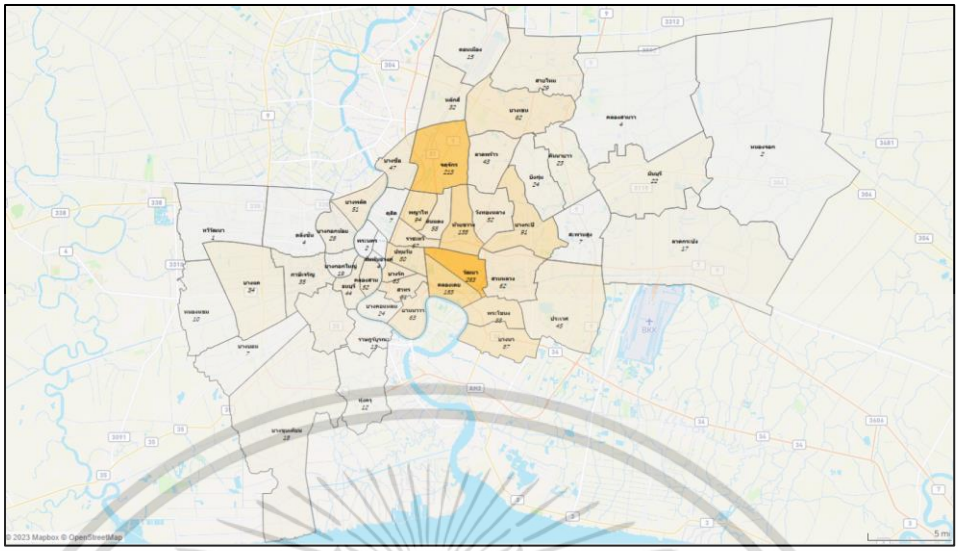
ตัวอย่างขั้นตอนการเตรียมข้อมูล



รูปที่ ก.1 แหล่งที่มาข้อมูลจาก Baania จากเว็บไซต์

Price per sqm	zipcode	c_room_bath	c_room_bed	c_floor	c_parking	c_unit	c_unittype	f_clubhouse	f_fitness	...	count_elevator	count_room_dinning	count_room_guest	count_room_kitchen	count_room_living	count_room_maid
25806.45	10260	1	1	8	90.000000	260.000000	3	1	1	...	1	0	0	0	0	0
7791.67	10220	2	0	5	47.923413	79.000000	1	1	1	...	0	0	0	0	0	0
30833.33	10230	2	2	8	140.000000	140.000000	2	1	1	...	1	0	0	0	0	0
143817.53	10110	2	2	8	70.000000	138.000000	3	1	1	...	2	0	0	0	0	0
101666.67	10900	1	0	32	237.000000	322.000000	3	0	1	...	3	0	0	0	0	0
14615.38	10260	1	2	8	245.076692	404.000000	3	1	1	...	0	0	0	0	0	0
62686.57	10210	1	1	8	208.679163	344.000000	2	1	1	...	4	0	0	0	0	0
71428.57	10120	3	2	24	247.075518	407.294991	1	0	1	...	0	1	0	1	1	0
19000.00	10230	2	3	8	59.000000	127.000000	3	1	1	...	0	1	0	1	1	0
98039.22	10110	1	2	51	426.000000	833.000000	3	1	1	...	2	1	1	0	0	0
54264.71	10330	2	2	33	313.018745	516.000000	3	1	1	...	0	1	0	1	1	0
83076.92	10500	2	2	40	465.000000	560.000000	3	1	1	...	6	0	0	0	0	0
10000.00	10510	1	1	8	970.600760	1600.000000	1	1	1	...	0	0	0	0	0	0
22307.69	10310	2	2	35	231.730931	382.000000	3	1	1	...	2	1	0	1	1	0
34444.44	10160	2	2	24	306.000000	811.000000	3	1	1	...	0	1	0	1	1	0
37894.74	10260	1	2	8	59.000000	158.000000	2	1	1	...	0	0	0	0	0	0
40625.00	10900	1	0	8	202.006283	333.000000	1	0	1	...	0	0	0	0	0	0
68627.45	10110	2	3	8	44.890285	74.000000	3	1	1	...	0	1	0	1	1	0
88556.20	10900	1	2	8	115.000000	147.000000	2	1	1	...	2	1	1	1	0	0

รูปที่ ก.2 ข้อมูลจาก Baania จากเว็บไซต์



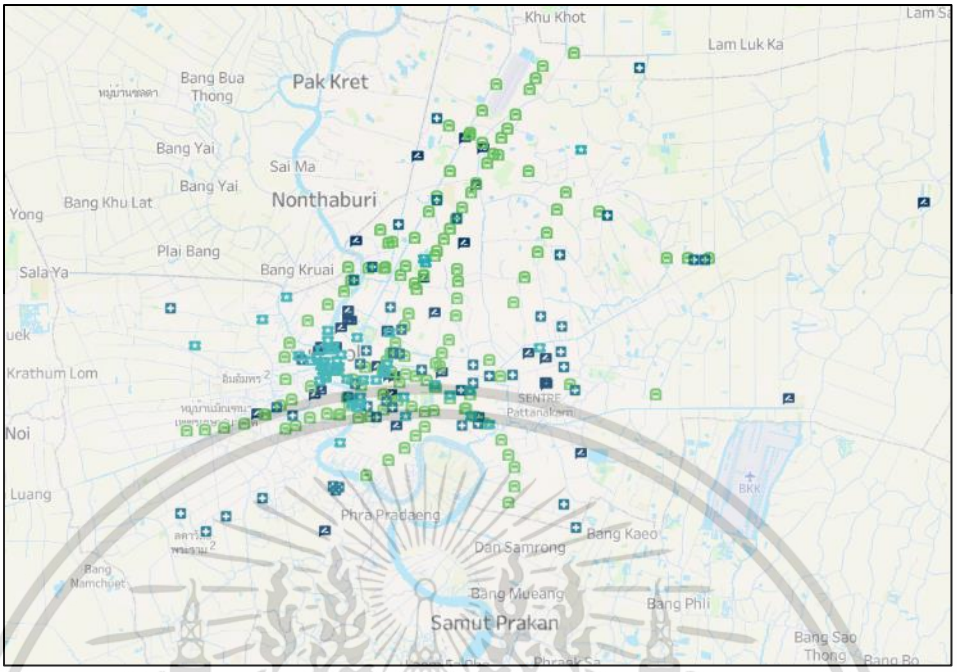
รูปที่ ก.3 จำนวนข้อมูลโครงการเมื่อแบ่งเป็นเขต

```python
from geopy.geocoders import Nominatim

geolocator = Nominatim(user_agent="attraction_geocoder")

for location in locations:
    try:

        geocode = geolocator.geocode(location)
        lat = geocode.latitude
        lon = geocode.longitude
        print(f"Location: {location}")
        print(f"Latitude: {lat}")
        print(f"Longitude: {lon}")
        print("----------------------")

    except Exception as e:
        print(f"Error occurred for location: {location}")
        print(str(e))
        print("----------------------")
```

✓ 18.2s

```
Location: มหาวิทยาลัยรามคำแหง
Latitude: 13.7562028
Longitude: 100.61744936124273
----------------------
Location: สถาบันเทคโนโลยีปทุมวัน
Latitude: 13.7482977
Longitude: 100.5258472675923
----------------------
Location: มหาวิทยาลัยราชภัฏจันทรเกษม
Latitude: 13.8204665
Longitude: 100.5781065
----------------------
Location: มหาวิทยาลัยราชภัฏธนบุรี
Latitude: 13.734146899999999
Longitude: 100.49169359843563
----------------------
Location: มหาวิทยาลัยราชภัฏบ้านสมเด็จเจ้าพระยา
Latitude: 13.73217475
Longitude: 100.48803726602786
----------------------
```

รูปที่ ก.4 โปรแกรมเก็บรวบรวมพิกัดสถานที่สำคัญ



รูปที่ ก.5 ข้อมูลพิกัดละติจูดและลองจิจูดที่เพิ่มเติมจาก Google Maps

SUM: =(SQRT((($M6-AW$2)^2)+(($N6-AW$3)^2))*60*1852)/1000

	J	M	N	AW	AX	AY	AZ	BA	BB	BC	BD
1											
2				13.73829	13.87366	13.80792	13.8601	13.74307	13.74849	13.69692	13.71926
3				100.6422	100.5957	100.5431	100.5155	100.549	100.5185	100.6053	100.5215
4				Airport-Re	BTS	SRT	MRT	BTS	BTS	BTS	BTS
5	Price per sq	latitud	longitu	หัวหมา	วัดพระค	บางซื่อ	ศูนย์รา	เพลินจิ	ยศเส	บางจาก	สุรศักดิ์
6	102,325.58	13.73599	100.5431	1000	16.37649	7.992195	14.1261	1.024906	3.066302	8.162259	3.033092
7	78,681.32	13.74421	100.603	4.40866	14.40725	9.718714	16.13433	6.003381	9.405958	5.261806	9.471322
8	17,241.38	13.67057	100.6747	8.345384	24.21605	21.13876	27.50269	16.12489	19.39964	8.251714	17.86184
9	30,769.23	13.78801	100.5832	8.57726	9.61808	4.976265	10.9867	6.275856	8.427226	10.41547	10.26395
10	52,500.00	13.74972	100.6076	4.055574	13.83576	9.653255	15.97069	6.551524	9.902554	5.873156	10.14521
11	116,438.36	13.73109	100.576	7.400665	15.9927	9.289369	15.83389	3.285728	6.681853	4.997625	6.199245
12	79,555.56	13.66215	100.6174	8.901083	23.62701	18.18054	24.73708	11.77219	14.58876	4.091014	12.39981
13	35,087.72	13.8424	100.5392	16.27823	7.1761	3.855604	3.282517	11.09126	10.68555	17.7559	13.82264
14	105,000.00	13.76432	100.4841	17.80934	17.36026	8.150851	11.20169	7.588438	4.205237	15.40783	6.507508
15	114,838.71	13.73921	100.589	5.916712	14.95924	9.183408	15.72006	4.466084	7.904957	5.034994	7.820357
16	52,452.83	13.89629	100.5656	19.51074	4.178187	10.13492	6.869797	17.12639	17.24006	22.58754	20.27351
17	137,857.14	13.75395	100.5645	8.806953	13.74492	6.453797	12.99212	2.109149	5.155718	7.787521	6.142297
18	26,607.14	13.67432	100.6459	7.120422	22.84289	18.73188	25.21884	13.2009	16.38204	5.164135	14.69489

รูปที่ ก.6 การหาระยะทางระหว่างโครงการกับสถานที่สำคัญจากการคำนวณระยะทางยุคลิด

#	Column
0	Price per sqm
1	zipcode
2	count_room_bath
3	count_room_bed
4	count_floor
5	count_parking
6	count_unit
7	count_unittype
8	facility_clubhouse
9	facility_fitness
10	facility_meeting
11	facility_park
12	facility_playground
13	facility_pool
14	facility_security
15	date_updated
16	Train
17	Hospital
18	University
19	Attraction
20	area_usable_min
21	year finish
22	year update
23	count_elevator
24	count_room_dinning
25	count_room_guest
26	count_room_kitchen
27	count_room_living
28	count_room_maid
29	count_room_storage
30	count_room_utility
31	latitude
32	longitude

รูปที่ ก.7 ข้อมูลตัวอย่างหลังจากเพิ่มข้อมูลจำนวนสถานที่สำคัญในบริเวณโครงการ

```python
import pandas as pd

# Replace missing values with the mode
condo_df['facility_clubhouse'] = condo_df['facility_clubhouse'].fillna(condo_df['facility_clubhouse'].mode()[0])
condo_df['facility_fitness'] = condo_df['facility_fitness'].fillna(condo_df['facility_fitness'].mode()[0])
condo_df['facility_meeting'] = condo_df['facility_meeting'].fillna(condo_df['facility_meeting'].mode()[0])
condo_df['facility_park'] = condo_df['facility_park'].fillna(condo_df['facility_park'].mode()[0])
condo_df['facility_pool'] = condo_df['facility_pool'].fillna(condo_df['facility_pool'].mode()[0])
condo_df['facility_security'] = condo_df['facility_security'].fillna(condo_df['facility_security'].mode()[0])

# Replace missing values with the mean
condo_df['count_elevator'] = condo_df['count_elevator'].fillna(condo_df['count_elevator'].mean())
condo_df['count_unit'] = condo_df['count_unit'].fillna(condo_df['count_unit'].mean())
condo_df['count_unittype'] = condo_df['count_unittype'].fillna(condo_df['count_unittype'].mean())
```

รูปที่ ก.8 การแทนค่าข้อมูลสูญหาย

```
q1 = condo_df[' Price per sqm '].quantile(0.25)
q3 = condo_df[' Price per sqm '].quantile(0.75)
iqr = q3 - q1
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr
num_outliers = len(condo_df[(condo_df[' Price per sqm '] < lower_bound) | (condo_df[' Price per sqm '] > upper_bound)])
print(f"Number of outliers: {num_outliers}")
```

```
Number of outliers: 117
```

```
lower_bound
```

```
-37447.6925
```

```
upper_bound
```

```
153770.8475
```

```
condo_df = condo_df[(condo_df[' Price per sqm '] < upper_bound)]
```

รูปที่ ก.9 กำจัดค่านอกเกณฑ์โดยใช้วิธีแผนภาพกล่อง (Boxplot)

```
condo_df = pd.get_dummies(condo, columns = ['zipcode'])
print(condo_df)
```

```
      Price per sqm   c_room_bath  c_room_bed  c_floor   c_parking  c_unit  \
0         102325.58             3           3       33  495.000000   231.0
1          78681.32             1           1       30  518.058156   854.0
2          17241.38             1           2        6  349.416274   576.0
3          30769.23             2           2        8   60.000000    79.0
4          52500.00             1           1        8  389.000000  1114.0
...             ...           ...         ...      ...         ...     ...
2398       37894.74             1           2        8   59.000000   158.0
2399       40625.00             1           0        8  202.006283   333.0
2400       68627.45             2           3        8   44.890285    74.0
2401       88556.20             1           2        8  115.000000   147.0
2402       33823.53             1           0        9  278.441093   459.0

      c_unittype  f_clubhouse  f_fitness  f_meeting  ...  zipcode_10400
0              2            1          1          1  ...          False  \
1              2            1          1          0  ...          False
2              1            0          0          0  ...          False
3              3            1          1          0  ...          False
4              1            1          1          0  ...          False
...          ...          ...        ...        ...  ...            ...
2398           2            1          1          0  ...          False
2399           1            0          1          0  ...          False
2400           3            1          1          0  ...          False
2401           2            1          1          1  ...          False
2402           2            0          0          0  ...           True
...
2401         False
2402         False
```

รูปที่ ก.10 การแปลงข้อมูลด้วยวิธี One-Hot Encoding

```
print(condo_df.columns)
```

```
Index([' Price per sqm ', 'c_room_bath', 'c_room_bed', 'c_floor', 'c_parking',
       'c_unit', 'c_unittype', 'f_clubhouse', 'f_fitness', 'f_meeting',
       'f_park', 'f_playground', 'f_pool', 'f_security', 'date_updated',
       'Train <= 1 km', 'Hospital <= 3 km', 'University <= 3 km',
       'Attraction<= 3 km', 'area_usable_min', 'year finish', 'year update',
       'count_elevator', 'count_room_dinning', 'count_room_guest',
       'count_room_kitchen', 'count_room_living', 'count_room_maid',
       'count_room_storage', 'count_room_utility', 'zipcode_10100',
       'zipcode_10110', 'zipcode_10120', 'zipcode_10130', 'zipcode_10140',
       'zipcode_10150', 'zipcode_10160', 'zipcode_10170', 'zipcode_10200',
       'zipcode_10210', 'zipcode_10220', 'zipcode_10230', 'zipcode_10240',
       'zipcode_10250', 'zipcode_10260', 'zipcode_10270', 'zipcode_10300',
       'zipcode_10310', 'zipcode_10312', 'zipcode_10320', 'zipcode_10330',
       'zipcode_10400', 'zipcode_10500', 'zipcode_10510', 'zipcode_10520',
       'zipcode_10530', 'zipcode_10600', 'zipcode_10700', 'zipcode_10800',
       'zipcode_10900', 'zipcode_11000'],
      dtype='object')
```

รูปที่ ก.11 แสดงจำนวนตัวแปรอิสระและตัวแปรอิสระของชุดข้อมูล

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import Pipeline
from sklearn.feature_selection import SelectKBest, f_regression

# Define the target variable and the independent variables
y = condo_df['Pricepersqm'].values
X = condo_df.drop('Pricepersqm', axis=1)

# split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=123)

# define the linear regression model
lr = LinearRegression()

# define the parameter grid for best subset selection
param_grid = {
    'selector__k': range(1, len(X_train.columns) + 1),
    'selector__score_func': [f_regression],
}

# define the pipeline for best subset selection
pipe = Pipeline([
    ('selector', SelectKBest()),
    ('lr', lr),
])

# perform grid search cross-validation for best subset selection
grid = GridSearchCV(pipe, param_grid=param_grid, scoring='neg_mean_squared_error', cv=None)
grid.fit(X_train, y_train)

# get the best estimator from the grid search
best_estimator = grid.best_estimator_

# get the selected features and their corresponding scores from the best estimator
selected_features = X_train.columns[best_estimator.named_steps['selector'].get_support()]
scores = best_estimator.named_steps['selector'].scores_[best_estimator.named_steps['selector'].get_support()]

# print the selected features and their corresponding scores
for feature, score in zip(selected_features, scores):
    print(f"{feature}: {score}")
```

รูปที่ ก.12 โปรแกรมการเลือกคุณลักษณะของวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import pearsonr

# กำหนดตัวแปรอิสระและตัวแปรตาม
y = condo_df[' Price per sqm ']
X = condo_df.drop(' Price per sqm ', axis=1)

# คำนวณ correlation และ p-value ระหว่างตัวแปรอิสระและตัวแปรตาม
correlation_matrix = []
p_values = []
for column in X.columns:
    correlation, p_value = pearsonr(X[column], y)
    correlation_matrix.append(correlation)
    p_values.append(p_value)

# แปลงเป็น DataFrame สำหรับการแสดงผล
correlation_df = pd.DataFrame({'Correlation': correlation_matrix, 'p-value': p_values}, index=X.columns)

# เรียงลำดับตามค่า correlation มากที่สุด
correlation_df = correlation_df.sort_values(by='Correlation', ascending=False)

# แสดงกราฟ barplot
plt.figure(figsize=(10, 20))
sns.barplot(x='Correlation', y=correlation_df.index, data=correlation_df, palette='coolwarm')
plt.xticks(rotation=90)
plt.xlabel('Correlation')
plt.ylabel('Independent Variables')
plt.title('Correlation between Independent Variables and Price per sqm')

# เพิ่มค่า correlation และ p-value ลงในกราฟ
for i, (correlation, p_value) in enumerate(zip(correlation_df['Correlation'], correlation_df['p-value'])):
    plt.text(correlation, i, f'Corr: {round(correlation, 2)}, p-value: {round(p_value, 2)}', va='center')

plt.show()
```

รูปที่ ก.13 โปรแกรมการหาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม

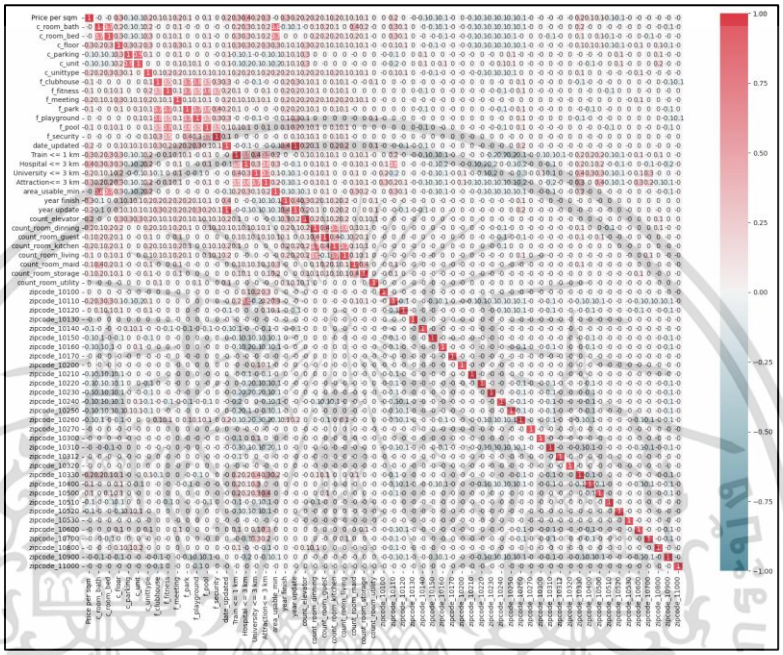
```python
from sklearn.model_selection import train_test_split, KFold

# split the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(condo_df[X], condo_df['Pricepersqm'], test_size=0.2, random_state=42)

# use 5-fold cross-validation to further split the training set into train and validation sets
kf = KFold(n_splits=5, shuffle=True, random_state=42)
for fold, (train_idx, val_idx) in enumerate(kf.split(X_train)):
    X_train_fold, y_train_fold = X_train.iloc[train_idx], y_train.iloc[train_idx]
    X_val_fold, y_val_fold = X_train.iloc[val_idx], y_train.iloc[val_idx]

    # print the sizes of the training and validation sets for each fold
    print(f"Fold {fold + 1}:")
    print(f"Training set size: {X_train_fold.shape[0]}")
    print(f"Validation set size: {X_val_fold.shape[0]}")
    print("="*50)

# print the size of the test set
print(f"Test set size: {X_test.shape[0]}")
```

```
Fold 1:
Training set size: 1462
Validation set size: 366
==================================================
Fold 2:
Training set size: 1462
Validation set size: 366
==================================================
Fold 3:
Training set size: 1462
Validation set size: 366
==================================================
Fold 4:
Training set size: 1463
Validation set size: 365
==================================================
Fold 5:
Training set size: 1463
Validation set size: 365
==================================================
Test set size: 458
```

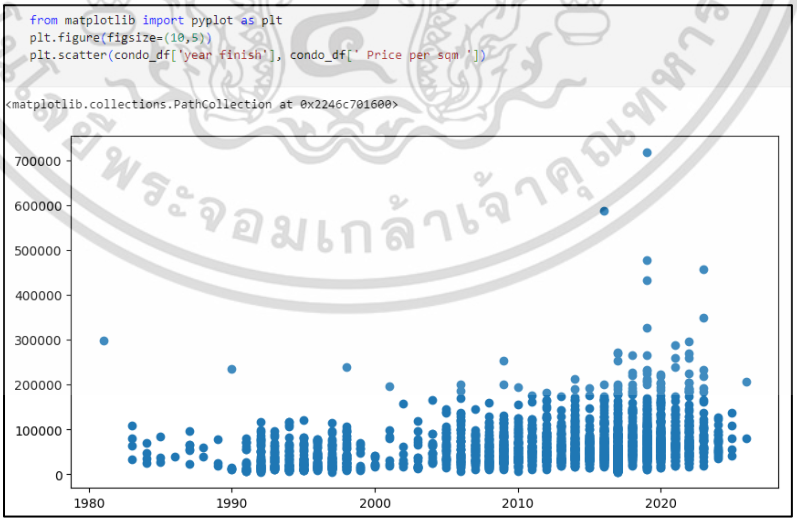
รูปที่ ก.14 โปรแกรมการแบ่งข้อมูลด้วยวิธีการตรวจสอบไขว้ 5 ชุด

ภาคผนวก ข

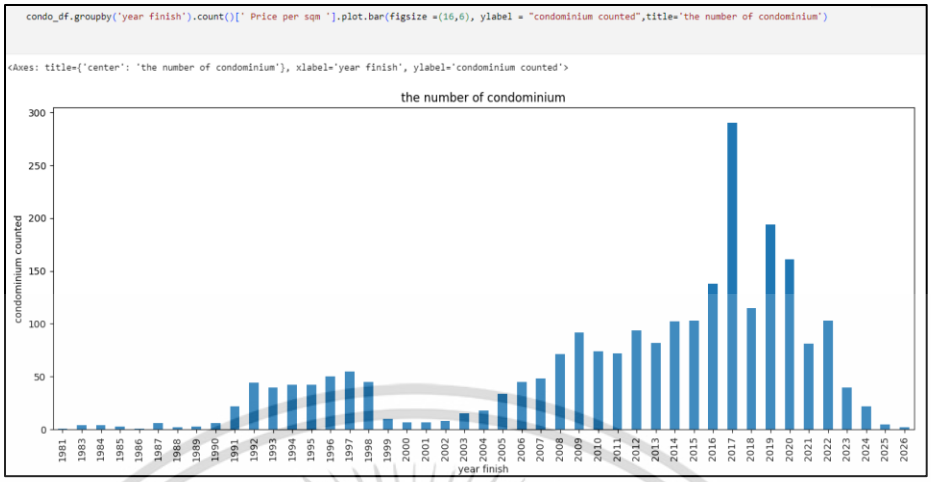
ตัวอย่างผลการวิเคราะห์ของชุดข้อมูล



รูปที่ ข.1 แผนภูมิความร้อนของสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทั้งหมด



รูปที่ ข.2 แสดงข้อมูลระหว่างราคาอาคารชุดต่อตารางเมตรกับเดือนปีที่สร้างเสร็จ



รูปที่ ข.3 แสดงข้อมูลระหว่างจำนวนอาคารชุดกับปีที่สร้างเสร็จ

ภาคผนวก ค

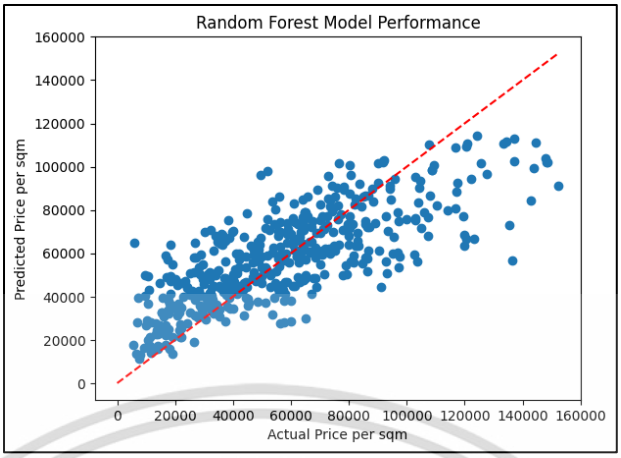
ตัวอย่างโปรแกรมสร้างแบบจำลองและผลลัพธ์

```python
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error,mean_absolute_percentage_error
import numpy as np
from sklearn.model_selection import KFold

rf_model = RandomForestRegressor(random_state=42)
kf = KFold(n_splits=5, shuffle=True, random_state=42)

for fold, (train_idx, val_idx) in enumerate(kf.split(X_train, y_train)):
    print(f"Fold {fold+1}:")
    X_train_fold, y_train_fold = X_train.iloc[train_idx], y_train.iloc[train_idx]
    X_val_fold, y_val_fold = X_train.iloc[val_idx], y_train.iloc[val_idx]

    rf_model.fit(X_train_fold, y_train_fold)
    y_pred_fold = rf_model.predict(X_val_fold)
    r2_fold = r2_score(y_val_fold, y_pred_fold)
    mse_fold = mean_squared_error(y_val_fold, y_pred_fold)
    rmse_fold = np.sqrt(mse_fold)
    mae_fold = mean_absolute_error(y_val_fold, y_pred_fold)
    rmae_fold = np.sqrt(mae_fold)
    r2_list.append(r2_fold)
    mse_list.append(mse_fold)
    rmse_list.append(rmse_fold)
    mae_list.append(mae_fold)
    rmae_list.append(rmae_fold)

print(f"Average R2 score: {np.mean(r2_list)}")
print(f"Average MSE score: {np.mean(mse_list)}")
print(f"Average RMSE score: {np.mean(rmse_list)}")
print(f"Average MAE score: {np.mean(mae_list)}")
print(f"Average RMAE score: {np.mean(rmae_list)}")
y_pred_test = rf_model.predict(X_test)
r2_test = r2_score(y_test, y_pred_test)
mse_test = mean_squared_error(y_test, y_pred_test)
rmse_test = np.sqrt(mse_test)
mae_test = mean_absolute_error(y_test, y_pred_test)
rmae_test = np.sqrt(mae_test)
mape1_score = mean_absolute_percentage_error(y_test, y_pred_test)
print("================================================")
print("Overall Evaluation Metrics:")
print(f"R2 Score: {r2_test:.4f}")
print(f"MSE: {mse_test:.4f}")
print(f"RMSE: {rmse_test:.4f}")
print(f"MAE: {mae_test:.4f}")
print(f"RMAE: {rmae_test:.4f}")
print(f"MAPE: {mape1_score:.4f}")
```

รูปที่ ค.1 โปรแกรมการสร้างแบบจำลองด้วยวิธีป่าสุ่ม

```
Average R2 score: 0.5585604750724528
Average MSE score: 443284181.5960995
Average RMSE score: 21038.864900875313
Average MAE score: 16098.620169726777
Average RMAE score: 126.85842735625086
================================================
Overall Evaluation Metrics:
R2 Score: 0.5967
MSE: 408823998.9300
RMSE: 20219.3966
MAE: 15579.7458
RMAE: 124.8189
MAPE: 0.4230
```

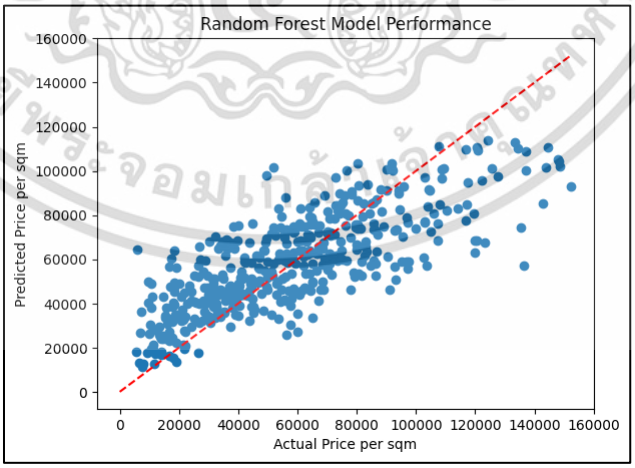
รูปที่ ค.2 ผลประสิทธิภาพของวิธีป่าสุ่มด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด



รูปที่ ค.3 แผนภาพแสดงความสัมพันธ์ของค่าทำนายกับค่าจริงของวิธีป่าสุ่มด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

```
Average R2 score: 0.55819720901272
Average MSE score: 443478544.78438014
Average RMSE score: 21047.02558272495
Average MAE score: 16090.104606592411
Average RMAE score: 126.82395814760471
==========================================
Overall Evaluation Metrics:
R2 Score: 0.5996
MSE: 405932867.2496
RMSE: 20147.7757
MAE: 15617.4951
RMAE: 124.9700
MAPE: 0.4246
```

รูปที่ ค.4 ผลประสิทธิภาพของวิธีป่าสุ่มด้วยวิธีการนำตัวแปรเข้าทั้งหมด



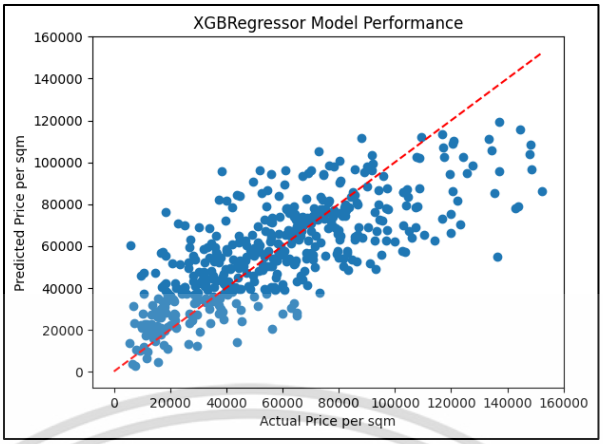
รูปที่ ค.5 แผนภาพแสดงความสัมพันธ์ของค่าทำนายกับค่าจริงของวิธีป่าสุ่มด้วยวิธีการนำตัวแปรเข้าทั้งหมด

```python
from xgboost import XGBRegressor
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error,mean_absolute_percentage_error
from sklearn.model_selection import KFold

xgb = XGBRegressor(n_estimators=100, learning_rate=0.1, random_state=42, max_depth=3, min_child_weight=5, eval_metric='rmse', feature_names=feature_names)
kf = KFold(n_splits=5, shuffle=True, random_state=42)

for fold, (train_idx, val_idx) in enumerate(kf.split(X_train, y_train)):
    print(f"Fold {fold+1}:")
    X_train_fold, y_train_fold = X_train.iloc[train_idx], y_train.iloc[train_idx]
    X_val_fold, y_val_fold = X_train.iloc[val_idx], y_train.iloc[val_idx]
    xgb.fit(X_train_fold, y_train_fold)
    y_pred_fold = xgb.predict(X_val_fold)
    r2_fold = r2_score(y_val_fold, y_pred_fold)
    mse_fold = mean_squared_error(y_val_fold, y_pred_fold)
    rmse_fold = np.sqrt(mse_fold)
    mae_fold = mean_absolute_error(y_val_fold, y_pred_fold)
    rmae_fold = np.sqrt(mae_fold)

    r2_list.append(r2_fold)
    mse_list.append(mse_fold)
    rmse_list.append(rmse_fold)
    mae_list.append(mae_fold)
    rmae_list.append(rmae_fold)

print(f"Average R2 score: {np.mean(r2_list)}")
print(f"Average MSE score: {np.mean(mse_list)}")
print(f"Average RMSE score: {np.mean(rmse_list)}")
print(f"Average MAE score: {np.mean(mae_list)}")
print(f"Average RMAE score: {np.mean(rmae_list)}")
y_pred_test = xgb.predict(X_test)
r2_test = r2_score(y_test, y_pred_test)
mse_test = mean_squared_error(y_test, y_pred_test)
rmse_test = np.sqrt(mse_test)
mae_test = mean_absolute_error(y_test, y_pred_test)
rmae_test = np.sqrt(mae_test)
mape1_score = mean_absolute_percentage_error(y_test, y_pred_test)
print("============================================")
print("Overall Evaluation Metrics:")
print(f"R2 Score: {r2_test:.4f}")
print(f"MSE: {mse_test:.4f}")
print(f"RMSE: {rmse_test:.4f}")
print(f"MAE: {mae_test:.4f}")
print(f"RMAE: {rmae_test:.4f}")
print(f"MAPE: {mape1_score:.4f}")
```

รูปที่ ค.6 โปรแกรมการสร้างแบบจำลองด้วยวิธี Extreme Gradient Boosting

```
Average R2 score: 0.5743463037900571
Average MSE score: 427528792.9879551
Average RMSE score: 20665.627873513975
Average MAE score: 15648.872183628051
Average RMAE score: 125.08660265531314
============================================
Overall Evaluation Metrics:
R2 Score: 0.6076
MSE: 397788712.0948
RMSE: 19944.6412
MAE: 15210.4124
RMAE: 123.3305
MAPE: 0.3971
```

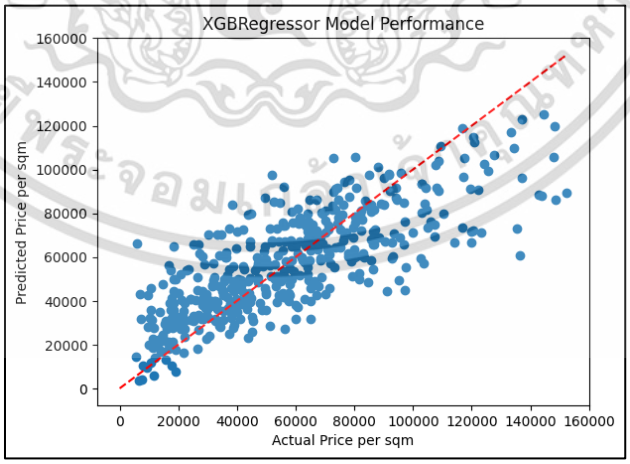
รูปที่ ค.7 ผลประสิทธิภาพของวิธี Extreme Gradient Boosting ด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด



รูปที่ ค.8 แผนภาพแสดงความสัมพันธ์ของค่าทำนายกับค่าจริงของวิธี Extreme Gradient Boosting ด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

```
Average R2 score: 0.5689238169283648
Average MSE score: 432653039.8704008
Average RMSE score: 20792.21475910104
Average MAE score: 15885.456757634074
Average RMAE score: 126.02249982210739
==========================================
Overall Evaluation Metrics:
R2 Score: 0.6079
MSE: 397447609.0858
RMSE: 19936.0881
MAE: 15243.8801
RMAE: 123.4661
MAPE: 0.4025
```

รูปที่ ค.9 ผลประสิทธิภาพของวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมด



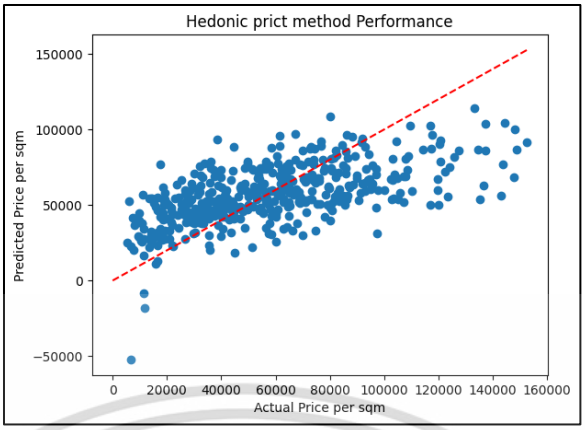
รูปที่ ค.10 แผนภาพแสดงความสัมพันธ์ของค่าทำนายกับค่าจริงของวิธี Extreme Gradient Boosting ด้วยวิธีการนำตัวแปรเข้าทั้งหมด

```python
import pandas as pd
from sklearn.model_selection import KFold
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error, mean_absolute_percentage_error

model = LinearRegression()
kf = KFold(n_splits=5, shuffle=True, random_state=42)

for fold, (train_idx, val_idx) in enumerate(kf.split(X_train, y_train)):
    X_train_fold, y_train_fold = X_train.iloc[train_idx], y_train.iloc[train_idx]
    X_val_fold, y_val_fold = X_train.iloc[val_idx], y_train.iloc[val_idx]
    model.fit(X_train, y_train)
    y_pred_fold = model.predict(X_val_fold)
    r2_fold = r2_score(y_val_fold, y_pred_fold)
    mse_fold = mean_squared_error(y_val_fold, y_pred_fold)
    rmse_fold = np.sqrt(mse_fold)
    mae_fold = mean_absolute_error(y_val_fold, y_pred_fold)
    rmae_fold = np.sqrt(mae_fold)
    r2_list.append(r2_fold)
    mse_list.append(mse_fold)
    rmse_list.append(rmse_fold)
    mae_list.append(mae_fold)
    rmae_list.append(rmae_fold)
print(f"Average R2 score: {np.mean(r2_list)}")
print(f"Average MSE score: {np.mean(mse_list)}")
print(f"Average RMSE score: {np.mean(rmse_list)}")
print(f"Average MAE score: {np.mean(mae_list)}")
print(f"Average RMAE score: {np.mean(rmae_list)}")
y_pred_test = model.predict(X_test)
r2_test = r2_score(y_test, y_pred_test)
mse_test = mean_squared_error(y_test, y_pred_test)
rmse_test = np.sqrt(mse_test)
mae_test = mean_absolute_error(y_test, y_pred_test)
rmae_test = np.sqrt(mae_test)
mape1_score = mean_absolute_percentage_error(y_test, y_pred_test)
print("==========================================")
print("Overall Evaluation Metrics:")
print(f"R2 Score: {r2_test:.4f}")
print(f"MSE: {mse_test:.4f}")
print(f"RMSE: {rmse_test:.4f}")
print(f"MAE: {mae_test:.4f}")
print(f"RMAE: {rmae_test:.4f}")
print(f"MAPE: {mape1_score:.4f}")
```

รูปที่ ค.11 โปรแกรมการสร้างแบบจำลองด้วยวิธีการประเมินราคาแอบแฝง

```
Average R2 score: 0.44574272484821
Average MSE score: 555947065.7999837
Average RMSE score: 23572.50876029674
Average MAE score: 18344.039997077667
Average RMAE score: 135.4282240976382
==========================================
Overall Evaluation Metrics:
R2 Score: 0.4583
MSE: 549145340.3694
RMSE: 23433.8503
MAE: 18234.5845
RMAE: 135.0355
MAPE: 0.4838
```

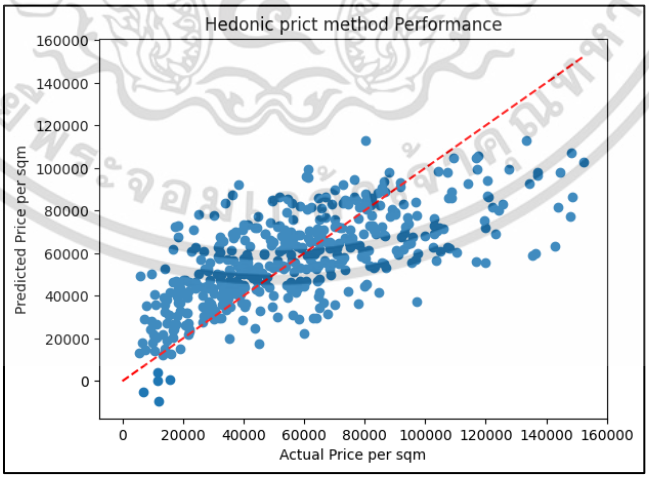
รูปที่ ค.12 ผลประสิทธิภาพของวิธีการประเมินราคาแอบแฝงด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด



รูปที่ ค.13 แผนภาพแสดงความสัมพันธ์ของค่าทำนายกับค่าจริงของวิธีการประเมินราคาแอบแฝงด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

```
Average R2 score: 0.4622961774176349
Average MSE score: 539316795.6076043
Average RMSE score: 23210.765141676682
Average MAE score: 17900.58983101017
Average RMAE score: 133.77933278329465
================================================
Overall Evaluation Metrics:
R2 Score: 0.4618
MSE: 545631926.7444
RMSE: 23358.7655
MAE: 18227.3334
RMAE: 135.0086
MAPE: 0.4739
```

รูปที่ ค.14 ผลประสิทธิภาพของวิธีการประเมินราคาแอบแฝงด้วยวิธีการนำตัวแปรเข้าทั้งหมด



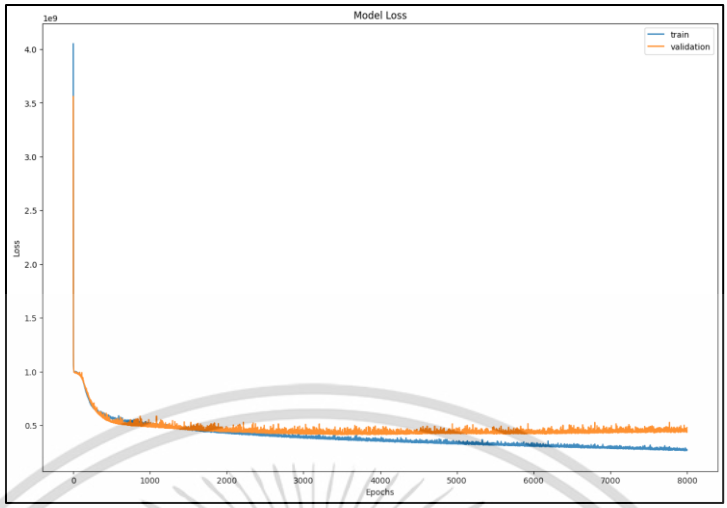
รูปที่ ค.15 แผนภาพแสดงความสัมพันธ์ของค่าทำนายกับค่าจริงของวิธีการประเมินราคาแอบแฝงด้วยวิธีการนำตัวแปรเข้าทั้งหมด

```python
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score, mean_absolute_percentage_error
from sklearn.model_selection import train_test_split
import pandas as pd
import numpy as np

class MetricsCallback(keras.callbacks.Callback):
    def __init__(self, validation_data):
        super().__init__()
        self.validation_data = validation_data

    def on_epoch_end(self, epoch, logs=None):
        X_val, y_val = self.validation_data
        y_pred = self.model.predict(X_val)
        r2_metric = r2_score(y_val, y_pred)
        print(f"R2: {r2_metric:.4f}")

X_train, X_test, y_train, y_test = train_test_split(condo_df[X], condo_df['Pricepersqm'], test_size=0.2, random_state=42)
X_train = np.array(X_train)
X_test = np.array(X_test)
y_train = np.array(y_train)
y_test = np.array(y_test)
X_train = X_train.reshape(X_train.shape[0], X_train.shape[1], 1)
X_test = X_test.reshape(X_test.shape[0], X_test.shape[1], 1)

model = keras.Sequential([
    layers.Conv1D(filters=64, kernel_size=7, activation='relu', input_shape=(X_train.shape[1], 1)),
    layers.MaxPooling1D(pool_size=2),
    layers.Flatten(),
    layers.Dense(64, activation='relu'),
    layers.Dense(1)
])

print(model.summary())

model.compile(loss='mse', optimizer='adam', metrics=['mse', 'mae', 'mape'])

checkpoint_callback = keras.callbacks.ModelCheckpoint(
    'best_model_CNN_New_MD.h5',
    monitor='val_loss',
    save_best_only=True,
    save_weights_only=False,
    mode='min',
    verbose=1
)
# Create an instance of MetricsCallback with validation_data
metrics_callback = MetricsCallback(validation_data=(X_test, y_test))

# Train the model with the callback
hist = model.fit(X_train, y_train, epochs=15000, batch_size=64, validation_data=(X_test, y_test), shuffle=True,
callbacks=[checkpoint_callback,metrics_callback])

best_epoch = np.argmin(hist.history['val_mse'])
best_mse = hist.history['val_mse'][best_epoch]
best_rmse = np.sqrt(best_mse)
best_mae = hist.history['val_mae'][best_epoch]
best_rmae = np.sqrt(best_mae)
best_mape = hist.history['val_mape'][best_epoch]
print("Best Evaluation Metrics:")
print(f"Epoch: {best_epoch+1}")
print(f"MSE: {best_mse:.4f}")
print(f"RMSE: {best_rmse:.4f}")
print(f"MAE: {best_mae:.4f}")
print(f"RMAE: {best_rmae:.4f}")
print(f"MAPE: {best_mape:.4f}")
```

รูปที่ ค.16 โปรแกรมการสร้างแบบจำลองด้วยวิธีโครงข่ายประสาทแบบคอนโวลูชัน

Best	Evaluation
Epoch	3,417
R2:	0.5916
RMSE:	20,348
MAE:	15,440
RMAE:	124
MAPE:	41

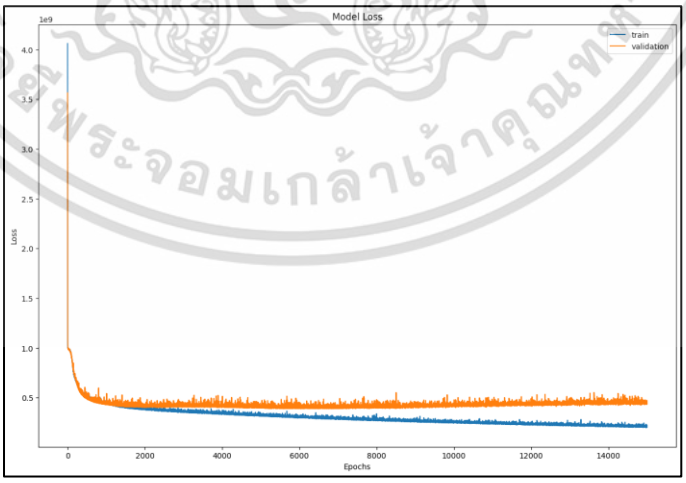
รูปที่ ค.17 ผลประสิทธิภาพของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด



รูปที่ ค.18 แผนภาพแสดงค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าทำนายกับค่าจริงของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

Best	Evaluation
Epoch	5,598
R2:	0.6160
RMSE:	19,729
MAE:	15,043
RMAE:	123
MAPE:	38

รูปที่ ค.19 ผลประสิทธิภาพของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมด



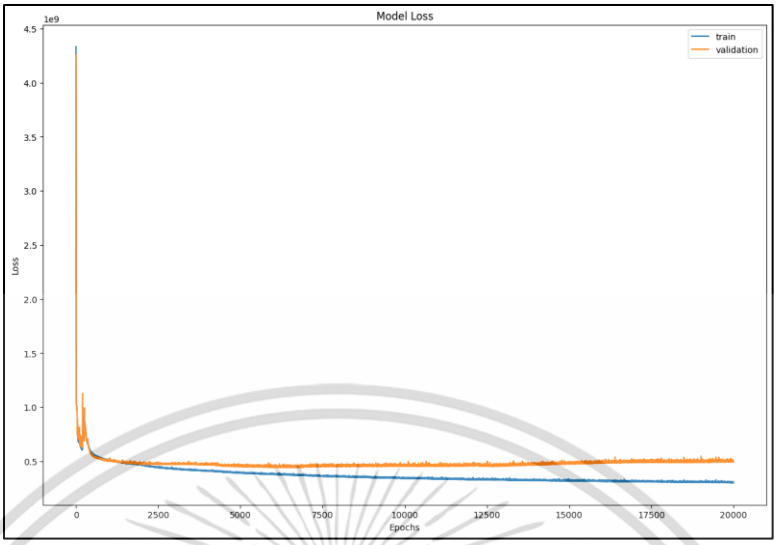
รูปที่ ค.20 แผนภาพแสดงค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าทำนายกับค่าจริงของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมด

```python
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score, mean_absolute_percentage_error
from sklearn.model_selection import train_test_split
import pandas as pd
import numpy as np

class MetricsCallback(keras.callbacks.Callback):
    def __init__(self, validation_data):
        super().__init__()
        self.validation_data = validation_data

    def on_epoch_end(self, epoch, logs=None):
        X_val, y_val = self.validation_data
        y_pred = self.model.predict(X_val)
        r2_metric = r2_score(y_val, y_pred)
        print(f"R2: {r2_metric:.4f}")

X_train, X_test, y_train, y_test = train_test_split(condo_df[X], condo_df['Pricepersqm'], test_size=0.2, random_state=42)
X_train = X_train.values.reshape((X_train.shape[0], 1, X_train.shape[1]))
X_test = X_test.values.reshape((X_test.shape[0], 1, X_test.shape[1]))

model = keras.Sequential([
    keras.layers.LSTM(64, input_shape=(1, X_train.shape[2]), activation='relu', return_sequences=True),
    keras.layers.LSTM(32, activation='relu'),
    keras.layers.Dense(1)
])

print(model.summary())
model.compile(loss='mse', optimizer='adam', metrics=['mse', 'mae', 'mape'])
checkpoint_callback = keras.callbacks.ModelCheckpoint(
    'best_model_LSTM_New_MD.h5',
    monitor='val_loss',
    save_best_only=True,
    save_weights_only=False,
    mode='min',
    verbose=1
)
metrics_callback = MetricsCallback(validation_data=(X_test, y_test))

# Train model
hist = model.fit(X_train, y_train, epochs=20000, batch_size=64, validation_data=(X_test, y_test), shuffle=True,
callbacks=[checkpoint_callback,metrics_callback])
y_pred = model.predict(X_test)
best_epoch = np.argmin(hist.history['val_mse'])
best_mse = hist.history['val_mse'][best_epoch]
best_rmse = np.sqrt(best_mse)
best_mae = hist.history['val_mae'][best_epoch]
best_rmae = np.sqrt(best_mae)
best_mape = hist.history['val_mape'][best_epoch]
print("Best Evaluation Metrics:")
print(f"Epoch: {best_epoch+1}")
print(f"MSE: {best_mse:.4f}")
print(f"RMSE: {best_rmse:.4f}")
print(f"MAE: {best_mae:.4f}")
print(f"RMAE: {best_rmae:.4f}")
print(f"MAPE: {best_mape:.4f}")
```

รูปที่ ค.21 โปรแกรมการสร้างแบบจำลองด้วยวิธีหน่วยความจำระยะสั้น-ยาว

Best	Evaluation
Epoch	6,756
R2:	0.5922
RMSE:	20,333
MAE:	16,128
RMAE:	127
MAPE:	43

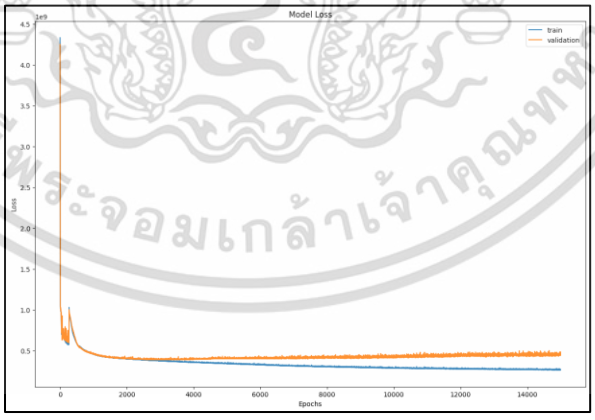
รูปที่ ค.22 ผลประสิทธิภาพของวิธีหน่วยความจำระยะสั้น-ยาวด้วยวิธีด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด



รูปที่ ค.23 แผนภาพแสดงค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าทำนายกับค่าจริงของวิธีหน่วยความจำระยะสั้น-ยาวด้วยวิธีการถดถอยชุดข้อมูลย่อยที่ดีที่สุด

Best	Evaluation
Epoch	3,472
R2:	0.6158
RMSE:	19,736
MAE:	15,440
RMAE:	124
MAPE:	41

รูปที่ ค.24 ผลประสิทธิภาพของวิธีหน่วยความจำระยะสั้น-ยาวด้วยวิธีการนำตัวแปรเข้าทั้งหมด



รูปที่ ค.25 แผนภาพแสดงค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าทำนายกับค่าจริงของวิธีหน่วยความจำระยะสั้น-ยาวด้วยวิธีการนำตัวแปรเข้าทั้งหมด

```python
import tensorflow as tf
from tensorflow import keras
import pandas as pd
import numpy as np

# Load the saved model
saved_model = keras.models.load_model('Enter_model_CNN_New.h5')

# Prepare the data for prediction
X_pred = condo_df[X].values
X_pred = X_pred.reshape(X_pred.shape[0], X_pred.shape[1], 1)

# Make predictions
predictions = saved_model.predict(X_pred)

# Add predicted values as a new column to condo_df
condo_df['Predicted_Price'] = predictions.flatten()

# Save the DataFrame as a CSV file
condo_df.to_csv('condo_df_predicted.csv', index=False)

condo_df
```

รูปที่ ค.26 โปรแกรมการทำนายราคาอาคารชุดในกรุงเทพมหานครของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมด

Pricepersqm	Predicted_Price	count_room_bath	count_room_bed	count_floor	count_parking	count_unit	count_unittype	facility_clubhouse	facility_fitness	...	zipcode_10400	zipcode_10500
102325.58	93848.984375	3.0	3.0	33.0	495.000000	231.0	2.0	1.0	1.0	...	0.0	0.0
78681.32	98218.906250	1.0	1.0	30.0	518.058156	854.0	2.0	1.0	1.0	...	0.0	0.0
17241.38	22600.244141	1.0	2.0	6.0	349.416274	576.0	1.0	0.0	0.0	...	0.0	0.0
30769.23	33304.339844	2.0	2.0	8.0	60.000000	79.0	3.0	1.0	1.0	...	0.0	0.0
52500.00	45294.890625	1.0	1.0	8.0	389.000000	1114.0	1.0	1.0	1.0	...	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
37894.74	45464.523438	1.0	2.0	8.0	59.000000	158.0	2.0	1.0	1.0	...	0.0	0.0
40625.00	42537.136719	1.0	0.0	8.0	202.006283	333.0	1.0	0.0	1.0	...	0.0	0.0
68627.45	51336.519531	2.0	3.0	8.0	44.890285	74.0	3.0	1.0	1.0	...	0.0	0.0
88556.20	79248.015625	1.0	2.0	8.0	115.000000	147.0	2.0	1.0	1.0	...	0.0	0.0
33823.53	43590.531250	1.0	0.0	9.0	278.441093	459.0	2.0	0.0	0.0	...	1.0	0.0

รูปที่ ค.27 ผลการทำนายราคาอาคารชุดในกรุงเทพมหานครของวิธีโครงข่ายประสาทแบบคอนโวลูชันด้วยวิธีการนำตัวแปรเข้าทั้งหมด

ประวัติผู้เขียน

ชื่อ	นางสาวสิริวรรณ แสงวงศ์
วัน เดือน ปีเกิด	5 มิถุนายน 2536
ที่อยู่ปัจจุบัน	457/299 ซอยเจริญกรุง 107 แยก 15/1 ถนนเจริญกรุง แขวงบางโคล่ เขตบางคอแหลม กรุงเทพฯ
ประวัติการศึกษา	(2558) วิทยาศาสตรบัณฑิต สาขาสถิติประยุกต์ เกรดเฉลี่ย 3.36 มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี (2566) วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์ เกรดเฉลี่ย 3.90 สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง