การพัฒนาแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อพยากรณ์ความนิยมอาคารชุดของผู้บริโภคออนไลน์ด้วยภาพถ่าย

วรวุฒิ สว่างอัม

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

THE ONLINE CONSUMER HITS FOR CONDOMINIUM PHOTO USING MACHINE LEARNING TECHNIQUE

WORRAWOOT SAWANGUM

A Thematic Paper Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Big Data Engineering,
College of Innovative Technology and Engineering,
Dhurakij Pundit University
Academic Year 2022

ใบรับรองสารนิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

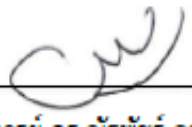
หัวข้อสารนิพนธ์ การพัฒนาแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อพยากรณ์ความนิยมอาคารชุดของผู้บริโภคออนไลน์ด้วยภาพถ่าย

เสนอโดย วรวุฒิ สว่างอัม

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาสารนิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบสารนิพนธ์แล้ว

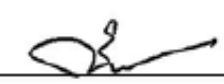


(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพัชร์ อารีรัชกุลกานต์) ประธานกรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา) กรรมการที่ปรึกษาสารนิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว



(ดร.ชัยพร เขมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

หัวข้อสารนิพนธ์	การพัฒนาแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องเพื่อพยากรณ์ความสนใจอาคารชุดของผู้บริโภคออนไลน์ด้วยภาพถ่าย
ชื่อผู้เขียน	วรวุฒิ สว่างอัม
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
หลักสูตร	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2565

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยจากภาพถ่ายสำหรับการพยากรณ์ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง ในการพัฒนาแบบจำลองเพื่อจัดทำเครื่องมือที่แสดงข้อมูลการพยากรณ์ความสนใจอาคารชุดของผู้บริโภคออนไลน์ด้วยภาพถ่าย ซึ่งผู้บริโภคในยุคใหม่นี้ได้มีการเปลี่ยนแปลงพฤติกรรมในการเลือกซื้อหรือเช่าอาคารชุดเพิ่มจำนวนมากขึ้น ฉะนั้นภาพถ่ายอาคารชุดจึงเป็นอีกมิติหนึ่งที่ส่งผลต่อผู้บริโภคออนไลน์ในการทำความเข้าใจเกี่ยวกับคุณลักษณะและองค์ประกอบของอาคารชุด

ผลของการศึกษาด้วยการได้มาซึ่งข้อมูลผู้วิจัยได้ทำการเก็บรวบรวมข้อมูลจากเว็บอสังหาริมทรัพย์ที่เกี่ยวกับอาคารชุด แล้วมาแบ่งกลุ่มความสนใจเป็น 2 กลุ่มเพื่อพยากรณ์ความสนใจของผู้บริโภคออนไลน์ ผลการวิจัยพบว่า การวัดประสิทธิภาพได้ค่า Accuracy ที่ 82.56 เปอร์เซ็นต์ ได้ค่า Class Recall 78.84 เปอร์เซ็นต์ และได้ค่า Class Precision 79.17 เปอร์เซ็นต์ โดยพบว่าฟีเจอร์ที่มีความสำคัญ 5 ลำดับแรก โดยประกอบไปด้วย entropy photo, lightness photo, contrast photo, price และ area จากการวัดประสิทธิภาพอยู่ในเกณฑ์ที่น่าพอใจและนำแบบจำลองที่ได้ไปประยุกต์ใช้กับข้อมูลเพื่อพัฒนาแสดงผลในรูปแบบ Web Application

คำสำคัญ : คอนโดมิเนียม, สื่อออนไลน์, การดึงข้อมูลจากเว็บไซต์, เทคนิคการเรียนรู้ของเครื่อง

.....

อาจารย์ที่ปรึกษา

Thematic Paper Title	THE ONLINE CONSUMER HITS FOR CONDOMINIUM PHOTO USING MACHINE LEARNING TECHNIQUE
Author	WORRAWOOT SAWANGUM
Thematic Paper Advisor	Dr. Eakasit Pacharawongsakda
Program	Big Data Engineering
Academic Year	2022

ABSTRACT

This study was to find the variables or the factors from the pictures for predicting of applying Machine Learning technique in developing the model for making the tool that display the predicted information about the online customers' interest by the pictures. The customers' behavior about purchasing or renting condominium in present time had been changed. Therefore, the pictures of the condominium were one of the aspect that affected the online customers. To understand the properties and the components of the condominium, the data was collected from real estate website about condominium. The data was divided into two groups to be predicted the interest of the online customers.

The study found that 1) the accuracy was at 82.56 percent 2) class recall was at 78.84 percent 3) class precision 79.17 percent. The first five important features consisted of entropy photo, lightness photo, contrast photo, price, and area. From measuring the efficiency, the result was at the satisfied rate and the model was used with data to develop the display in Web Application

Keywords: Condominium, Social media, Web scraping, Machine learning

Advisor

กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยกรุณาของ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา อาจารย์ที่ปรึกษาสารนิพนธ์ ที่ได้กรุณาให้คำปรึกษาแนะนำ ตรวจสอบ การปรับปรุงแก้ไขสารนิพนธ์ฉบับนี้เป็นอย่างดีมาโดยตลอด

ผู้เขียนขอกราบขอบพระคุณ ผศ.ดร.ณัฐพัชร์ อารีรัชกุลกานต์ ที่กรุณาให้เกียรติเป็นประธานโดยมี ผศ.ดร. ดวงใจ จิตคงชื่น เป็นกรรมการในการสอบสารนิพนธ์ ซึ่งได้ให้คำแนะนำแนวทางที่เป็นประโยชน์ต่องานสารนิพนธ์ และตรวจแก้ไขสารนิพนธ์ฉบับนี้ให้ถูกต้องสมบูรณ์ยิ่งขึ้น ตลอดจน นางสาวกุลธิดา รอดบุญ รวมถึงเจ้าหน้าที่บัณฑิตมหาวิทยาลัยธุรกิจบัณฑิตย์ ทุกท่านที่ช่วยอำนวยการ และประสานงานในการทำสารนิพนธ์ให้ผู้เขียนตลอดมา ส่งผลให้การจัดทำสารนิพนธ์ของผู้เขียนครั้งนี้สำเร็จลุล่วงไปด้วยดี

สุดท้ายนี้ผู้วิจัยต้องขอขอบคุณ บิดา มารดา ครอบครัวและเพื่อนๆ ที่คอยช่วยส่งเสริม สนับสนุนและให้กำลังใจ ทำให้การศึกษาสารนิพนธ์ในครั้งนี้สำเร็จลุล่วงไปด้วยดี ทั้งนี้ทางผู้วิจัยต้องกราบขออภัยเป็นอย่างสูงมา ณ โอกาสนี้หากมีสิ่งใดที่ผู้วิจัยได้ทำผิดพลาด หรือบกพร่องประการใด และผู้วิจัยหวังเป็นอย่างยิ่งว่าสารนิพนธ์ฉบับนี้จะเป็นพื้นฐานในการต่อยอดองค์ความรู้ของผู้ที่สนใจศึกษาในงานด้านนี้ต่อไป

วรวุฒิ สว่างอัม

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันคนไทยกว่า 51 ล้านคนเป็นผู้ใช้งาน Social Network และมีสถิติการใช้งาน Smart Phone สูงที่สุดในโลก โดยใช้เวลาเฉลี่ยในการเสพสื่อออนไลน์ หรือท่องโลกอินเตอร์เน็ต เฉลี่ยประมาณวันละ 5 ชั่วโมง ส่งผลให้โลกออนไลน์กลายเป็นช่องทางหลักที่เข้าถึงผู้คนได้มากที่สุด หลายแวดวงจึงมีการปรับตัวให้เข้ากับพฤติกรรมของผู้บริโภค ไม่เว้นแม้แต่วงการอสังหาริมทรัพย์ที่มีการปรับตัวให้ตรงกับความต้องการของผู้บริโภคในยุคใหม่ ไม่ว่าจะเป็นการออกแบบ Product ให้ตอบโจทย์ความต้องการของผู้บริโภคในยุค Digital โดยการแสวงหานวัตกรรมใหม่ๆ เข้ามาใช้งานในโครงการ เช่น ระบบลงประกาศขายหรือเช่าอสังหาริมทรัพย์หรือระบบการจองอสังหาริมทรัพย์ออนไลน์

นอกจากนี้ผู้บริโภคในยุคใหม่นี้ยังมีการเปลี่ยนแปลงพฤติกรรมในการหาข้อมูลเลือกซื้อและเปรียบเทียบอาคารชุด ซึ่งตัวกำหนดการตัดสินใจซื้อของผู้บริโภคเป็นสิ่งที่น่าสนใจ และมีการระบุปัจจัยหลายประการ เช่น แบรนด์ผู้พัฒนาอสังหาริมทรัพย์ คุณภาพโครงการ ขนาดห้อง ฟังก์ชันการทำงาน และราคา นอกจากนี้ผลกระทบของสื่อออนไลน์ได้รับความสนใจในตลาดมากขึ้น ด้วยสภาพของสื่อออนไลน์นำมาซึ่งความท้าทายของตลาด ปัจจุบันกลุ่มผู้บริโภคออนไลน์นิยมหาข้อมูลของโครงการที่สนใจก่อนตัดสินใจซื้อหรือเช่า โดยให้ความสนใจในหลายประเด็น เช่น ตำแหน่งที่ตั้งของโครงการ ราคา ขนาดของห้อง คุณภาพของวัสดุและสิ่งอำนวยความสะดวกภายในโครงการรวมถึงกระแสการตอบรับในตลาด เพื่อนำมาเปรียบเทียบกับโครงการอื่นๆ ซึ่งการเลือกสรรโครงการต่างๆต้องตอบโจทย์การใช้ชีวิตของผู้บริโภค การรับรู้ที่มีอิทธิพลต่อคุณภาพของผลิตภัณฑ์ และความตั้งใจในการซื้อหรือเช่า ฉะนั้นภาพถ่ายอาคารชุดจึงเป็นอีกมิติหนึ่งที่ส่งผลต่อผู้บริโภคออนไลน์ในการทำความเข้าใจเกี่ยวกับคุณลักษณะและองค์ประกอบของอาคารชุด

ดังนั้น ผู้วิจัยจึงมีความสนใจที่จะศึกษาเกี่ยวกับภาพถ่ายส่วนใหญ่ที่ใช้วิธีการทดลองเป็นหลัก เพื่อให้ผลที่ชัดเจนต่อการทดสอบทฤษฎีแต่ครอบคลุมลักษณะเฉพาะของภาพถ่ายเพียงเล็กน้อย และพัฒนาคุณสมบัติมิติสูงเพื่อทำความเข้าใจความหมายของภาพที่จะนำไปสู่การพัฒนาแบบจำลองข้อมูลการพยากรณ์ความนิยมของผู้บริโภคออนไลน์ด้วยภาพถ่ายอาคารชุด

1.2 วัตถุประสงค์ของการศึกษาหรือวิจัย

1. เพื่อหาตัวแปรหรือปัจจัยสำหรับการพยากรณ์
2. เพื่อประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องในการพัฒนาแบบจำลอง
3. เพื่อจัดทำเครื่องมือที่แสดงข้อมูลการพยากรณ์ความสนใจของอาคารชุด

1.3 ขอบเขตของงานวิจัย

1. ข้อมูลในภาพถ่าย เช่น TV , Refrigerator, chair, couch, bed, table, entropy photo, Lightness photo, contrast photo
2. ข้อมูลพื้นฐาน คือ ราคา, เนื้อที่, ห้องนอน, ห้องน้ำ
3. ข้อมูลจำนวนผู้เข้าชม

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้แบบจำลองที่สามารถพยากรณ์ข้อมูลความสนใจของอาคารชุด
2. ได้เครื่องมือที่แสดงข้อมูลการพยากรณ์ความสนใจของอาคารชุด

1.5 นิยามศัพท์

1. คอนโดมิเนียม หรืออาคารชุด หมายถึง เป็นที่อยู่อาศัยรูปแบบหนึ่งที่บุคคลสามารถแยกการถือกรรมสิทธิ์ออกได้เป็นส่วนๆ โดยแต่ละส่วนประกอบด้วยกรรมสิทธิ์ในทรัพย์ส่วนบุคคลและกรรมสิทธิ์ร่วมในทรัพย์ส่วนกลาง เช่น ทางเดิน บริเวณห้องโถง บันได ลิฟต์ โรงจอดรถ สระว่ายน้ำ

2. สื่อออนไลน์ หมายถึง สิ่งที่ผู้ส่งสารแบ่งปันสาร ซึ่งอยู่ในรูปแบบต่างๆ เช่น ข้อความ รูปภาพ เสียง วีดีโอ เป็นต้น ไปยังผู้รับสารผ่านเครือข่ายออนไลน์โดยสามารถโต้ตอบกันระหว่างผู้ส่งสารและผู้รับสาร หรือผู้รับสารด้วยกันเอง ซึ่งมี ผู้ประกอบการเข้ามาสร้างสารในรูปแบบต่างๆ และสามารถสร้างรายได้จากธุรกิจบนสื่อสังคมออนไลน์

3. การดึงข้อมูลจากเว็บไซต์ หมายถึง เทคนิคการเก็บรวบรวมข้อมูลจากเว็บไซต์มาใช้ประโยชน์ในการวิเคราะห์ข้อมูลที่ได้มาเพื่อนำไปต่อยอด เช่น นำข้อมูลไปพัฒนาเครื่องมือเพื่อพยากรณ์ในรูปแบบต่างๆ รวมไปถึงการดึงข้อมูลเพื่อจะหาเหตุผล หรือ Insight บางอย่างที่จะได้จากข้อมูลบนเว็บไซต์

4. เทคนิคการเรียนรู้ของเครื่อง หมายถึง ระบบที่สามารถเรียนรู้จากข้อมูลที่มีอยู่ได้ด้วยตนเอง ประกอบด้วยข้อมูลและเครื่องมือทางสถิติเพื่อช่วยพัฒนากระบวนการแก้ปัญหาและทำนายผลลัพธ์ออกมาอย่างเหมาะสม

บทที่ 2

ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมโดยใช้เทคนิคการเรียนรู้ของเครื่องในการพัฒนาแบบจำลองที่แสดงข้อมูลการพยากรณ์ความสนใจของอาคารชุด โดยจำเป็นต้องศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

2.1 ความหมายและลักษณะของอาคารชุด

2.2 การหาค่าข้อมูลผิดปกติด้วยวิธีการ Interquartile range (IQR)

2.3 เทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

2.4 แรนดอมฟอเรส (Random Forest)

2.5 ไลท์กราเดี้ยนบูทติ้งแมชชิน (Light Gradient Boosting Machine)

2.6 การจัดการกับข้อมูลที่ไม่สมดุล

2.7 ตัววัดประสิทธิภาพของโมเดล (Confusion Matrix)

2.8 งานวิจัยที่เกี่ยวข้อง

2.1 ความหมายและลักษณะของอาคารชุด

อาคารชุดตามพระราชบัญญัติอาคารชุด พ.ศ. 2522 มาตรา 4 ได้ให้คำจำกัดความไว้ว่า “อาคารชุด” หมายความว่า “อาคารที่บุคคลสามารถแยกการถือกรรมสิทธิ์ได้เป็นส่วน ๆโดยแต่ละส่วนประกอบด้วยกรรมสิทธิ์ในทรัพย์ส่วนบุคคลและกรรมสิทธิ์ร่วมในทรัพย์ส่วนกลาง” จากคำจำกัดความดังกล่าวข้างต้น จะเห็นได้ว่าลักษณะสำคัญของระบบกรรมสิทธิ์ในอาคารชุดนั้นจะต้องประกอบด้วยกรรมสิทธิ์ 2 ประเภทคือ กรรมสิทธิ์ส่วนบุคคลในทรัพย์สินส่วนที่จัดไว้เพื่อประโยชน์หรือให้ไว้เป็นของส่วนบุคคลหนึ่ง ๆ โดยเฉพาะและกรรมสิทธิ์ในทรัพย์สินส่วนกลางซึ่งมีไว้เพื่อประโยชน์หรือเพื่อใช้ร่วมกัน ซึ่งยอมให้ผู้เป็นเจ้าของทรัพย์ส่วนบุคคลได้ภาระจำยอมหรือสิทธิในการใช้สอยเหนือทรัพย์ส่วนกลางเหล่านั้นก็ตามหรือแม้แต่กรรมสิทธิ์ในทรัพย์ส่วนกลางจะเป็นของสมาคม สหกรณ์บริษัทหรือนิติบุคคลที่เรียกชื่ออย่างอื่นซึ่งมีผู้เป็นเจ้าของทรัพย์ส่วนบุคคลเป็นสมาชิกหรือเป็นผู้ถือหุ้นอยู่เท่านั้นก็ตามลักษณะ ดังกล่าวนี้ก็ยังไม่ถือว่าเป็นอาคารชุดหรือคอนโดมิเนียม อาคารชุดหนึ่งอาจจะประกอบด้วยตัวอาคารหลังเดียวหรือหลายหลังก็ได้และจะเป็นอาคารชั้นเดียวอาคารหลายชั้นก็ได้เช่นเดียวกันกฎหมายไม่ได้จำกัดได้แต่ตัวอาคารนั้นจะต้องมีการแบ่งออกเป็นส่วนๆ ให้บุคคลแยกถือกรรมสิทธิ์ตามสิทธิได้เฉพาะบุคคลและต้องมีส่วนที่เป็นส่วนกลางซึ่งเป็นกรรมสิทธิ์ร่วมระหว่างผู้เป็นเจ้าของกรรมสิทธิ์ส่วนเฉพาะบุคคลเท่านั้น ส่วนเฉพาะบุคคลอาจจะแบ่งเป็นห้องๆ หรือแบ่งเป็นชั้น ๆ หรือแบ่งอย่างไร

ก็ได้โดยไม่จำกัดเนื้อที่และไม่จำเป็นต้องอยู่ติดกันทั้งหมด แต่อย่างน้อยจะต้องมี ส่วนเฉพาะบุคคลตั้งแต่สองส่วนขึ้นไปจึงจะเป็นอาคารชุดได้อย่างไรก็ตามเงื่อนไขสำคัญที่กฎหมาย กำหนดไว้ก็คือ อาคารจะต้องจดทะเบียนดังนั้น ตราบใดที่ยังมิได้จดทะเบียนเป็นอาคารชุด แม้ว่า 13 ลักษณะของอาคารจะเป็นอาคารชุดดังได้กล่าวมาแล้วข้างต้นก็ตาม

อาคารดังกล่าวนั้นย่อมไม่เป็นอาคารชุดและไม่อยู่ในบังคับของกฎหมายอาคารชุด 10 อาคารชุดในปัจจุบันแบ่งได้ตามลักษณะการใช้งาน ดังนี้

1) ใช้เป็นที่อยู่อาศัย (Residential) เป็นอาคารชุดที่สร้างขึ้นมาทั้งสถาปัตยกรรม และการออกแบบเพื่อการอยู่อาศัยทั้งรองรับกลุ่มลูกค้าที่ไม่สามารถซื้อบ้านเดี่ยวในราคาแพง และต้องการเลี่ยงปัญหารถติดขณะเดินทาง

2) ใช้เป็นสำนักงาน (Commercial) เป็นอาคารชุดที่ใช้เป็นที่ทำการเพื่อดำเนินการทางธุรกิจ ห้ามพักอาศัยลูกค้ากลุ่มนี้มองเห็นว่าซื้อดีกว่าเช่าและได้กรรมสิทธิ์ในทรัพย์สินด้วย

3) ใช้เป็นที่พักผ่อน (Resort) เป็นอาคารชุดที่รองรับลูกค้าที่อยากจะมีที่พักผ่อน เป็นของตนเองและอาจจะนำไปให้เช่าต่อได้เพราะมักจะมีบรรยากาศและทำเลดี

4) ประเภทคอมเพล็กซ์ (Complex) อาคารชุดประเภทนี้ใช้เป็นที่อยู่อาศัยและที่ทำการค้าได้ด้วยเพราะสะดวกไม่ต้องเสียเวลาในการเดินทางและได้รับความนิยมจากผู้ซื้อพอสมควร

5) ประเภทอุตสาหกรรม (Industrial) อาคารชุดประเภทนี้มีวัตถุประสงค์เพื่อให้ อุตสาหกรรมขนาดย่อมมาอยู่ร่วมกัน และเฉลี่ยกันออกค่าสาธารณูปโภคต่าง ๆ เช่น ค่าน้ำประปา ค่ากระแสไฟฟ้า เป็นการลดต้นทุน ซึ่งในประเทศไทยก็มีอยู่บ้างในย่านของด้านเขตอุตสาหกรรม

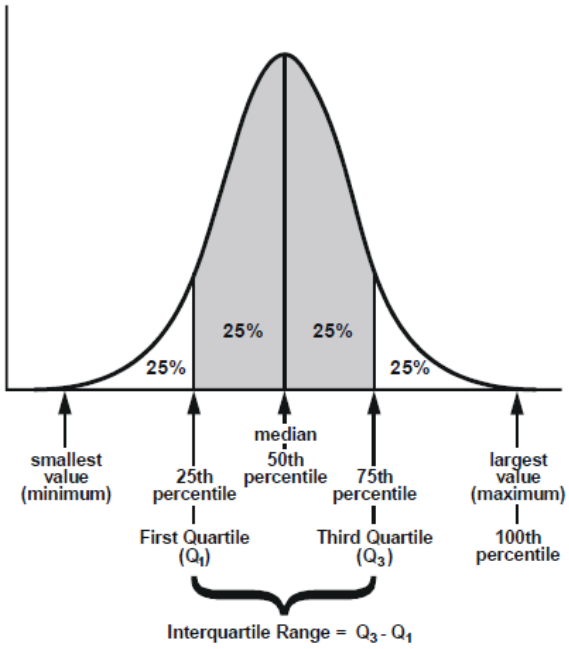
ภาพที่ 2.1 ตัวอย่างรูปอาคารชุด

ที่มา: https://www.smartservice.co.th/th/เลือกบริษัทบริหารนิติบ/

2.2 การหาค่าข้อมูลผิดปกติด้วยวิธี Interquartile Range (IQR)

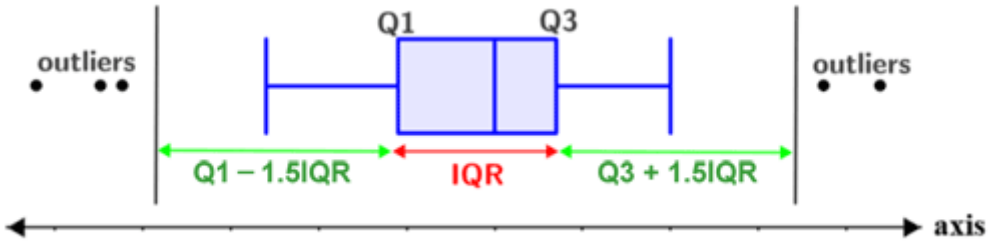
Interquartile range (IQR) คือความแตกต่างระหว่าง Quartile ที่ 1 และ 3 หรือ Q3 - Q1 ขั้นตอนการหาทำได้โดยการนาข้อมูลทั้งหมดมาเรียงกันจากน้อยไปมาก และแล้วแบ่งข้อมูล ออกเป็น 4 ช่วง ตามภาพที่ 2.1 จากนั้นนำ Q3 - Q1 เพื่อจะได้ความแตกต่างระหว่าง Quartile ที่ 1 และ 3

การคำนวณ Outlier จะได้จากสมการ Q1 - (1.5 × IQR) และ Q3 + (1.5 × IQR) โดยค่าที่น้อยกว่า Q1 - (1.5×IQR) หรือมากกว่า Q3 + (1.5 × IQR) จะถือว่าเป็นข้อมูลที่ค่าแตกต่างไปจากปกติ หรือ Outlier ซึ่ง IQR นิยมเสนอในรูปแบบของ Box Plot ตัวอย่างดังภาพที่ 2.2



ภาพที่ 2.2 แสดงการหา IQR

ที่มา: https://dcmlearning.ie/lean-course-content/lean-six-sigma-measure-interquartiles.html



ภาพที่ 2.3 ตัวอย่างการหา Outlier ด้วยวิธี IQR

ที่มา: https://www.math.net/interquartile-range

2.3 เทคนิคการเรียนรู้ของเครื่อง

คือระบบที่สามารถเรียนรู้จากขอ้มูลที่มีอยู่ได้ด้วยตนเอง ประกอบด้วยข้อมูลและเครื่องมือทางสถิติ เพื่อช่วยพัฒนากระบวนการแก้ปัญหาและทำนายผลลัพธ์ออกมาอย่างเหมาะสม

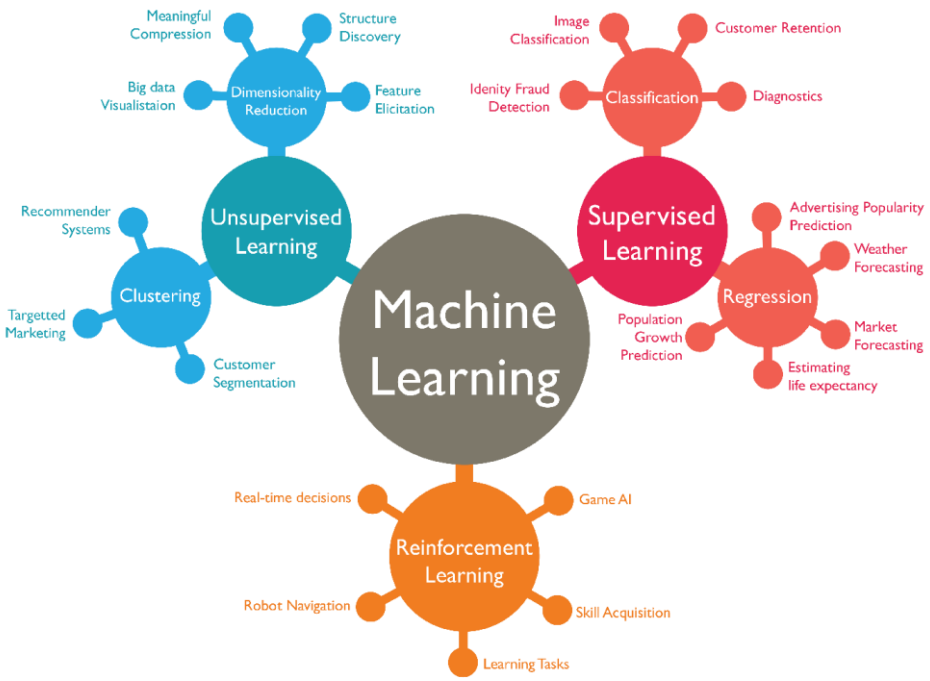
Machine learning จะทำงานคล้ายกับการเรียนรู้ของมนุษย์โดยการป้อนชุดข้อมูลและชุดคำสั่งให้คอมพิวเตอร์ “เรียนรู้” เพื่อจำแนกแยกแยะวัตถุต่างๆ รวมถึงบุคคล สิ่งของ ฯลฯ

Machine Learning algorithms สามารถแบ่งได้ 3 ประเภท

1. Supervised Learning เครื่องเรียนรู้ด้วยข้อมูล คือ ใส่ข้อมูล (input) เข้าไปแล้วมีผลลัพธ์ (output) ออกมา

2. Unsupervised Learning เครื่องเรียนรู้โดยไม่มีข้อมูล โดยที่เครื่องจะเรียนรู้และค้นพบรูปแบบด้วยตัวเอง

3. Reinforcement learning เครื่องเรียนรู้ด้วยการกระทำ เสมือนเด็กเพิ่งเกิดใหม่ ค่อยๆเรียนรู้ตามการกระทำหรือสภาพแวดล้อมที่เจอ โดยจะมีการเรียนรู้เพื่อปรับปรุงและพัฒนาอย่างต่อเนื่อง

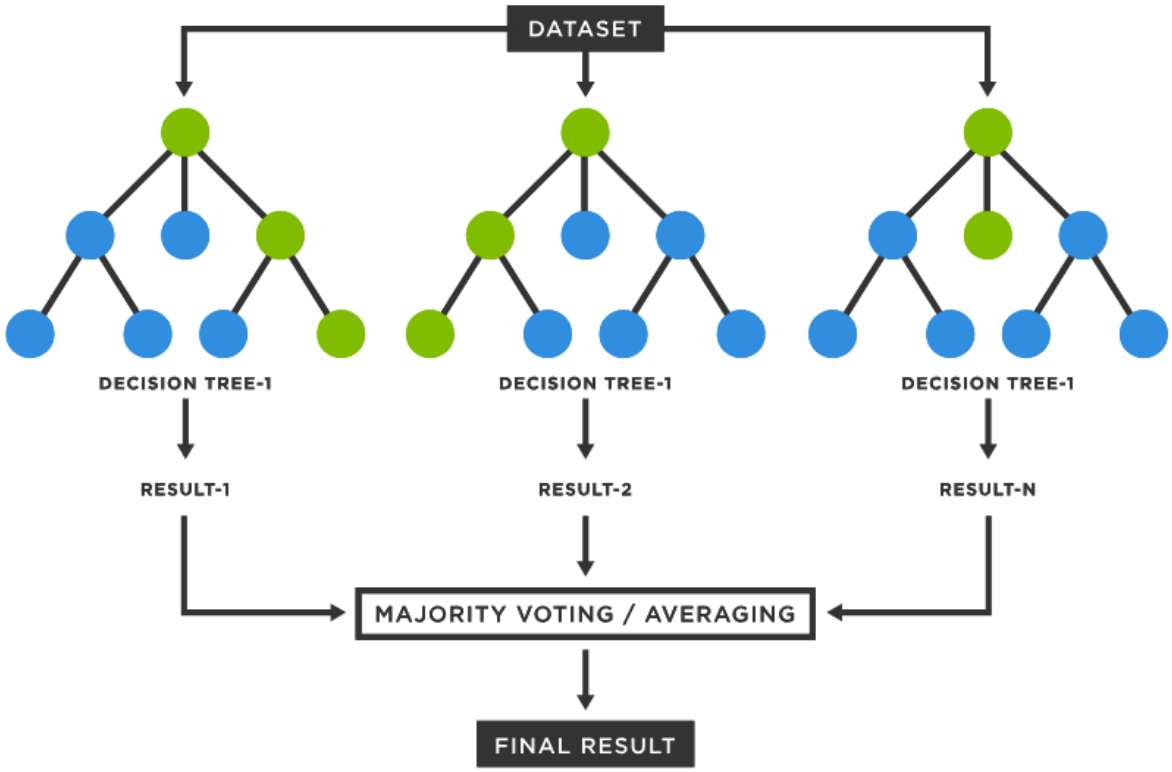


ภาพที่ 2.4 ตัวอย่างหลักการเรียนรู้ข้อมูลของ Machine Learning

ที่มา: https://medium.com/@pradyasin/machine-learning

2.4 แรนดอมฟอเรส (Random Forest)

แรนดอมฟอเรส (Random Forest) เป็นหนึ่งในแบบจำลองที่อยู่ในกลุ่มของ Ensemble Learning ซึ่งมีพื้นฐานมาจากต้นไม้ตัดสินใจ (Decision Tree) ซึ่งจะทาการสร้างแบบจาลอง Decision Tree หลายๆต้นโดยสร้างจากการสุ่มข้อมูลตัวอย่างจากชุดข้อมูลฝึกฝน (Training data) แบบเลือกแล้วใส่กลับเพื่อให้มีโอกาสมีโอกาสถูกเลือกอีกครั้ง ซึ่งจะสุ่มเลือกข้อมูลให้ได้จำนวน N ตัวอย่าง และ สุ่มเลือกแอตทริบิวต์เป็นจำนวนที่น้อยกว่าจำนวนแอตทริบิวต์ ทั้งหมดโดยแบบจำลองจะมีการทำนายผลออกมาซึ่งจะนาผลการทานายที่ได้มาโหวตหาผลการทานายที่ได้รับการโหวตมากที่สุด (Majority Voting)

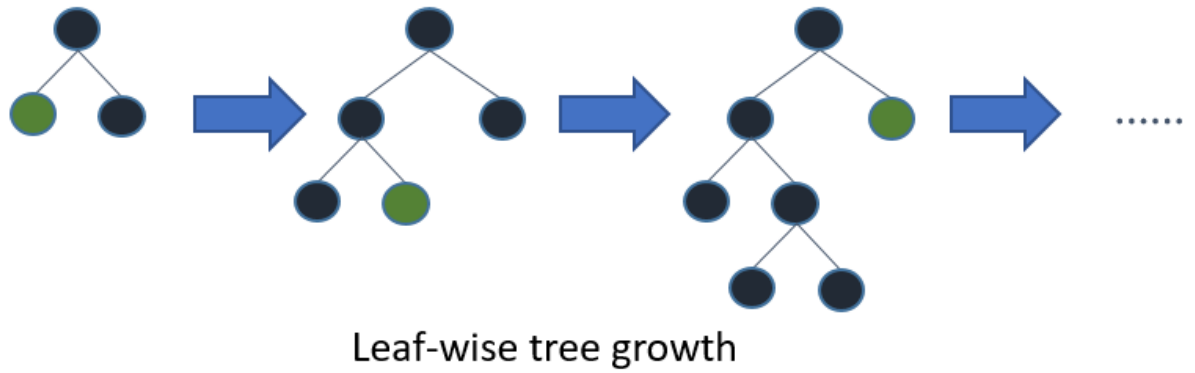


ภาพที่ 2.5 โครงสร้างของ Random Forest

ที่มา: https://www.tibco.com/reference-center/what-is-a-random-forest

2.5 ไลท์กราเดี้ยนบูทติ้งแมชชิน (Light Gradient Boosting Machine)

ไลท์กราเดี้ยนบูทติ้งแมชชิน (Light Gradient Boosting Machine) เป็นโมเดลทางคณิตศาสตร์ที่มีโครงสร้างเป็นแบบต้นไม้ตัดสินใจสำหรับทำ Classification หรือ Regression หลาย ๆ ต้น (trees) โดยต้นไม้เหล่านี้จะถูกสร้างขึ้นจากข้อมูลที่ใช้สอนโมเดล Light GBM เป็น GBM เฟรมเวิร์กที่มีประสิทธิภาพสูง โดยที่ Light GBM ใช้อัลกอริทึม Leaf-wise ซึ่งสามารถลดการสูญเสียได้ มากกว่าอัลกอริทึม Level-wise ดังนั้นจึงให้ผลลัพธ์ที่แม่นยำและมีความเร็วมากกว่า



ภาพที่ 2.6 โครงสร้างของ Light Gradient Boosting Machine

ที่มา: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

2.6 การจัดการกับข้อมูลที่ไม่สมดุล

ในการสร้างโมเดลจำเป็นต้องมีชุดข้อมูลเรียนรู้ (Training data) เพื่อเรียนรู้ แอตทริบิวต์ ทั่วไปคือแอตทริบิวซ์หรือตัวแปรที่ใช้ในการสร้างแบบจำลองแอตทริบิวต์ประเภทคำตอบ (Label) คือแอตทริบิวต์ที่เป็นคำตอบที่เราสนใจในการสร้างแบบจำลอง ซึ่งชุดข้อมูลเรียนรู้ควรจะมีข้อมูล แต่ละคลาสคำตอบเท่ากันหรือใกล้เคียงกัน (Balance data) เพื่อให้แบบจำลองสามารถเรียนรู้ได้จากทุกคลาสคำตอบ แต่โดยส่วนใหญ่แล้วข้อมูลจะเป็นลักษณะของ (Imbalanced data) หมายถึง ข้อมูลคำตอบของแต่ละคลาสมีจานวนไม่เท่ากัน จะเรียกข้อมูลที่มีจำนวนมากกว่าว่า Majority class และเรียกข้อมูลที่มีจำนวนน้อยกว่าว่า Minority class เทคนิคในการจัดการกับข้อมูลที่ไม่สมดุลแบ่งเป็น 3 วิธีหลักๆ

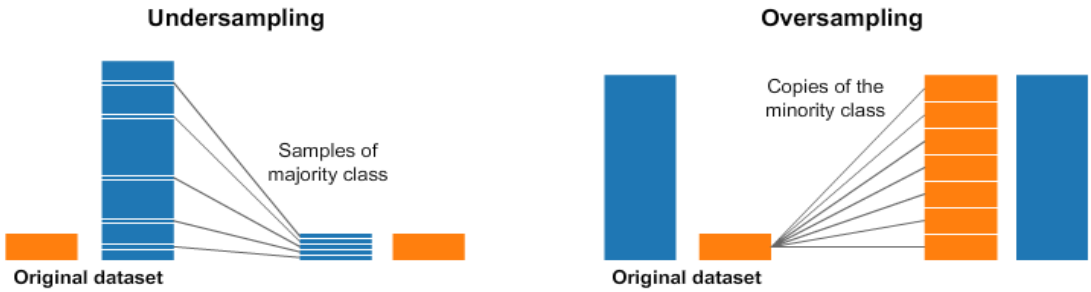
1. Sampling methods เป็นวิธีการสุ่มตัวอย่างซึ่งเป็นวิธีทางสถิติ มีจุดประสงค์เพื่อทำให้ข้อมูลแต่ละกลุ่มมีประมาณที่สมดุลกัน โดยแบ่งเป็น 2 วิธีหลักๆ คือ

1.1 Under-sampling สุ่มลดจำนวนข้อมูลกลุ่มหลัก (Major class) ให้พอๆ กับข้อมูลกลุ่มน้อย (Minor class)

1.2 Over-sampling สุ่มเพิ่มจำนวนข้อมูลกลุ่มน้อย (Minor class) ขึ้นให้พอๆ กับข้อมูลกลุ่มหลัก (Major class)

2. Cost-sensitive methods หลักการคือพิจารณาจากค่าความผิดพลาดจากการแบ่งกลุ่ม (Misclassifiying examles)

3. Kernel-based methods หลักการนี้คือการย้ายตำแหน่งข้อมูลที่ไม่สามารถแบ่งกลุ่มได้ในระนาบปกติ โดยการเพิ่มมิติให้สูงขึ้นจนสามารถแบ่งกลุ่มข้อมูลออกจากกันได้



ภาพที่ 2.7 ลักษณะของข้อมูลที่ไม่สมดุล

ที่มา: https://www.linkedin.com/pulse/some-tricks-handling-imbalanced-dataset-image-m-farhan-tandia/

2.7 ตัววัดประสิทธิภาพของโมเดล (Confusion Matrix)

การจะนำแบบจำลองไปใช้จริงจำเป็นต้องมีการวัดประสิทธิภาพของแบบจำลองก่อนว่าแบบจำลองที่สร้างขึ้นมานั้นมีประสิทธิภาพเพียงพอที่จะนำมาพัฒนาต่อหรือใช้งานจริงหรือไม่ซึ่งส่วนใหญ่นั้นจะใช้ Confusion Matrix ในการวัดประสิทธิภาพ

Confusion Matrix คือ ตารางที่มีแนวคิดมาจากสิ่งที่แบบจำลองทำนายได้กับสิ่งที่เกิดขึ้นจริงนั้นมีสัดส่วนเป็นอย่างไร

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

ภาพที่ 2.8 แสดงตัวอย่างตาราง Confusion Matrix

ที่มา: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

True Positive (TP) คือ สิ่งที่แบบจา ลองพยากรณ์ว่า “จริง” และสิ่งนั้น เกิดขึ้น “จริง”
True Negative (TN) คือ สิ่งที่แบบจา ลองพยากรณ์ว่า “ไม่จริง” และสิ่งนั้น เกิดขึ้น “ไม่จริง”
False Positive (FP) คือ สิ่งที่แบบจา ลองพยากรณ์ว่า “จริง” แต่สิ่งนั้น เกิดขึ้น “ไม่จริง”
False Negative (FN) คือ สิ่งที่แบบจา ลองพยากรณ์ว่า “ไม่จริง” แต่สิ่งนั้น เกิดขึ้น “จริง”
ทั่วไปแล้วตัววัดที่นิยมใช้ในการวิจัยและงานต่างๆ มีหลักๆอยู่ 3 ตัวได้แก่

1) Precision
2) Recall
3) Accuracy

โดยแต่ละค่ามักจะนำไปใช้ต่างกันตามความเหมาะสม

2.8 การจัดการกับข้อมูลที่ไม่สมดุล (Imbalance Data)

Xin Li, Mengyue Wang, Yubo Chen (2014) นำเสนอปัจจัยในการตัดสินใจซื้อของผู้บริโภคทางออนไลน์เป็นสิ่งที่นักวิจัยและ ผู้ปฏิบัติงานให้ความสนใจในระยะยาว เนื่องจากภาพถ่ายของผลิตภัณฑ์ช่วยให้ผู้บริโภคเข้าใจผลิตภัณฑ์ได้โดยตรง ผู้ค้าปลีกมักจะใช้ความพยายามอย่างมากในการขัดเกลาให้สวยงาม อย่างไรก็ตาม ยังมีงานวิจัยจำกัดเกี่ยวกับผลกระทบของภาพถ่ายผลิตภัณฑ์ต่อการตัดสินใจซื้อ การศึกษาก่อนหน้านี้ส่วนใหญ่ใช้วิธีการทดลอง ซึ่งนำเสนอทฤษฎีที่เคร่งครัดเกี่ยวกับภาพถ่ายผลิตภัณฑ์บางแง่มุม งานวิจัยนี้ใช้ประโยชน์จากเทคนิคการประมวลผลภาพเพื่อศึกษาผลกระทบของภาพถ่ายผลิตภัณฑ์ เทคนิค เหล่านี้ช่วยให้

เราสามารถตรวจสอบคุณลักษณะของภาพถ่ายชุดใหญ่ได้พร้อมๆ กันในการศึกษา เชิงประจักษ์ เพื่อขจัดปัจจัยที่อาจก่อให้เกิดความสับสน เรารวบรวมชุดข้อมูลจากเว็บไซต์โซเชียลช้อปปิ้งซึ่งมีอินเทอร์เฟซที่เรียบง่ายทำให้ผู้ใช้สามารถตัดสินผลิตภัณฑ์จากภาพถ่ายเป็นหลัก เราตรวจสอบลักษณะเฉพาะของภาพถ่ายผลิตภัณฑ์จากแง่มุมของข้อมูล อารมณ์ สุนทรียภาพ และการแสดงตน ทางสังคม พบว่าผู้บริโภคชอบภาพถ่ายผลิตภัณฑ์ที่มีวัตถุหลักขนาดใหญ่กว่า เอนโทรปีของวัตถุหลักที่ต่ำกว่า สีที่อุ่นกว่า คอนทราสต์สูงกว่า ความชัดลึกที่สูงกว่า และการแสดงตนทางสังคมมากกว่า งานวิจัยนี้แนะนำวิธีการที่ใช้ข้อมูลขนาดใหญ่เพื่อศึกษาผลกระทบของคุณสมบัติภาพของระบบอีคอมเมิร์ซที่มี ต่อผู้บริโภค

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dolla. (2015). นำเสนอชุดข้อมูลใหม่ซึ่งมีเป้าหมายในการเสนอการจดจำวัตถุในบริบทของปัญหาความเข้าใจในสถานการณ์ที่กว้างขึ้น สิ่งนี้ทำได้โดยการรวบรวมภาพที่ซับซ้อนฉากในชีวิตประจำวันที่มีวัตถุทั่วไปในสภาพแวดล้อมทางธรรมชาติ ใช้แต่ละอินสแตนซ์แบ่งส่วนเพื่อทำเครื่องหมายวัตถุเพื่อช่วยการวางตำแหน่งของวัตถุที่แม่นยำ ชุดข้อมูลของเรามีภาพถ่ายของวัตถุ 91 ชนิดที่เด็กอายุ 4 ขวบสามารถจดจำได้ง่าย การใช้ด้วยอินสแตนซ์แท็กทั้งหมด 2.5 ล้านภาพจาก 328k การสร้างชุดข้อมูลของเราใช้ประโยชน์จากคนช่วยจำนวนมากผ่านส่วนต่อประสานผู้ใช้ที่แปลกใหม่สำหรับการตรวจจับหมวดหมู่การค้นพบอินสแตนซ์และการแยกอินสแตนซ์ เราสนใจชุดข้อมูลเปรียบเทียบกับ PASCAL, ImageNet และ SUN

Moaiad Ahmad Khder. (2021). นำเสนอการรวบรวมข้อมูลเครือข่ายหรือการรวบรวมข้อมูลเว็บโดยใช้ซอฟต์แวร์เพื่อดึงข้อมูลจากเว็บไซต์โดยอัตโนมัติ ซึ่งช่วยทำให้เราสามารถดึงข้อมูลที่มีโครงสร้างจากข้อความเช่น html และจะเป็นประโยชน์มากหากมีข้อมูลในรูปแบบที่อ่านได้โดยเครื่องเช่น JSON หรือXML การรวบรวมข้อมูล เราสามารถรวบรวมราคาได้เกือบตามเวลาจริงจากเว็บไซต์ออนไลน์ และให้รายละเอียดเพิ่มเติม นอกจากนี้การใช้โปรแกรมรวบรวมข้อมูลเว็บจะสร้างข้อมูลเพิ่มเติมอย่างละเอียดถูกต้องและสอดคล้องกันมากกว่าการป้อนด้วยมือ กล่าวคือการดึงข้อมูลเว็บคืออะไรและเป็นอย่างไร ผลงาน, ขั้นตอนการดึงข้อมูลเว็บ, เทคโนโลยี, มันเกี่ยวข้องกับธุรกิจอย่างไร ปัญญาประดิษฐ์วิทยาศาสตร์ข้อมูลข้อมูลขนาดใหญ่เครือข่ายการรักษาความปลอดภัยและวิธีการทำในภาษา Python

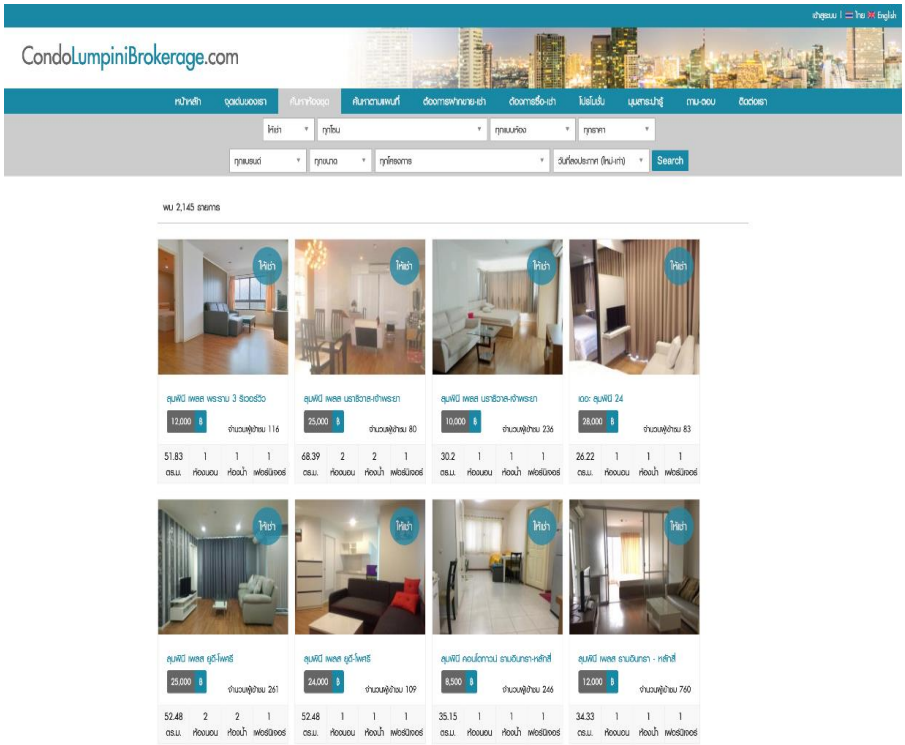
บทที่ 3

วิธีวิจัย

งานวิจัยนี้ เป็นการศึกษาเพื่อหาตัวแปรหรือปัจจัยที่มีความสำคัญและเหมาะสมโดยใช้เทคนิคการเรียนรู้ของเครื่องในการพัฒนาแบบจำลองที่แสดงข้อมูลการพยากรณ์ความสนใจของอาคารชุด โดยมีขั้นตอนการดำเนินการดังต่อไปนี้

3.1 การเก็บรวบรวมข้อมูล

3.1.1 การเก็บข้อมูลจากแหล่งออนไลน์ ด้วยเทคนิค Web scraping



ภาพที่ 3.1 แสดงตัวอย่างหน้าเว็บไซต์ www.condolumpinibrokerage.com

มีลักษณะข้อมูลดังนี้

1. เป็นข้อมูลที่ค้นหา ด้วยเงื่อนไข ประเภท ให้เช่า และ เรียงลำดับจากวันที่ลงข้อมูล วันที่ลงประกาศ (ใหม่-เก่า)

2. เป็นข้อมูลห้องชุดจำนวนข้อมูล 8,824 ตัวอย่าง โดยเก็บข้อมูล ตั้งแต่ 27 สิงหาคม 2564 ถึง 07 กันยายน 2565

3. คุณลักษณะที่สนใจได้แก่

- ไอดี (listing_id)
- ประเภท (contract_type)
- ชื่อห้องชุด (title)
- ที่อยู่เว็บ (url)
- ภาพถ่าย (image)
- ราคา (price)
- ขนาดห้องชุด (area)
- จำนวนที่นอน (beds)
- จำนวนห้องน้ำ (baths)
- จำนวนผู้ชมครั้งแรก (hits_start)
- จำนวนผู้ชมล่าสุด (hits_update)
- วันที่เริ่มที่เก็บข้อมูล (date_time_start)
- วันที่ล่าสุดที่เก็บข้อมูล (date_time_update)
- โซน (zone)

	listing_id	contract_type	title	url	image	price	area	beds	baths	hits_start	hits_update	date_time_start	date_time_update	zone
0	62207	ให้เช่า	ลุมพินี พาร์ค ริเวอร์ไซด์ พระราม 3	http://condolumpinibrokerage.com/index.php/pro...	http://condolumpinibrokerage.com/assets/upload...	14000	33.01	1	1	95	155	2020-08-27 00:00:00	2022-09-07 00:00:00	สาธร, พระราม 4, บางรัก, พระราม 3, สาทุประดิษฐ์
1	63620	ให้เช่า	ลุมพินี พาร์ค ริเวอร์ไซด์ พระราม 3	http://condolumpinibrokerage.com/index.php/pro...	http://condolumpinibrokerage.com/assets/upload...	9000	26.48	1	1	213	314	2020-08-27 00:00:00	2022-09-07 00:00:00	สาธร, พระราม 4, บางรัก, พระราม 3, สาทุประดิษฐ์
2	63832	ให้เช่า	ลุมพินี เพลส พระราม 4 กล้วยน้ำไท	http://condolumpinibrokerage.com/index.php/pro...	http://condolumpinibrokerage.com/assets/upload...	12000	29.08	1	1	175	187	2020-08-27 00:00:00	2022-02-18 00:00:00	สาธร, พระราม 4, บางรัก, พระราม 3, สาทุประดิษฐ์
3	63988	ให้เช่า	ลุมพินี เพลส พระราม 4 - สาทร	http://condolumpinibrokerage.com/index.php/pro...	http://condolumpinibrokerage.com/assets/upload...	20000	61.91	1	1	102	208	2020-08-27 00:00:00	2021-07-12 00:00:00	สาธร, พระราม 4, บางรัก, พระราม 3, สาทุประดิษฐ์
4	64117	ให้เช่า	ลุมพินี พาร์ค ริเวอร์ไซด์ พระราม 3	http://condolumpinibrokerage.com/index.php/pro...	http://condolumpinibrokerage.com/assets/upload...	12000	32.43	1	1	5	11	2020-08-27 00:00:00	2021-01-04 00:00:00	สาธร, พระราม 4, บางรัก, พระราม 3, สาทุประดิษฐ์

ภาพที่ 3.2 แสดงตัวอย่างข้อมูลห้องชุด ด้วย Pandas

ภาพที่ 3.3 แสดงตัวอย่างภาพถ่าย Thumbnail ที่เก็บข้อมูลจากแหล่งออนไลน์

3.2 การเตรียมข้อมูล (Data Preprocessing)

3.2.1 ข้อมูลทั่วไป

3.2.1.1 เลือกข้อมูลทั่วไปที่จะทำการเตรียมข้อมูล ได้แก่

- ราคา (price)
- ขนาดห้องชุด (area)
- จำนวนที่นอน (beds)
- จำนวนห้องน้ำ (baths)
- จำนวนผู้ชมครั้งแรก (hits_start)
- จำนวนผู้ชมล่าสุด (hits_update)
- วันที่เริ่มที่เก็บข้อมูล (date_time_start)
- วันที่ล่าสุดที่เก็บข้อมูล (date_time_update)

	listing_id	price	area	beds	baths	date_time_start	date_time_update	hits_start	hits_update
0	62207	14000	33.01	1	1	2020-08-27	2022-09-07	95	155
1	63620	9000	26.48	1	1	2020-08-27	2022-09-07	213	314
2	63832	12000	29.08	1	1	2020-08-27	2022-02-18	175	187
3	63988	20000	61.91	1	1	2020-08-27	2021-07-12	102	208
4	64117	12000	32.43	1	1	2020-08-27	2021-01-04	5	11

ภาพที่ 3.4 แสดงตัวอย่างข้อมูลห้องชุดที่เลือกมาเพื่อเตรียมข้อมูล ด้วย Pandas

3.2.1.2 สร้าง คอลัมน์ จำนวนวัน (date_count) โดย คำนวณมาจาก วันที่ล่าสุดที่เก็บข้อมูล (date_time_update) ลบด้วย วันที่เริ่มที่เก็บข้อมูล (date_time_start) แล้วบวกด้วย 1 วัน

	listing_id	price	area	beds	baths	date_time_start	date_time_update	hits_start	hits_update	date_count
0	62207	14000	33.01	1	1	2020-08-27	2022-09-07	95	155	742
1	63620	9000	26.48	1	1	2020-08-27	2022-09-07	213	314	742
2	63832	12000	29.08	1	1	2020-08-27	2022-02-18	175	187	541
3	63988	20000	61.91	1	1	2020-08-27	2021-07-12	102	208	320
4	64117	12000	32.43	1	1	2020-08-27	2021-01-04	5	11	131

ภาพที่ 3.5 แสดงตัวอย่างข้อมูลห้องชุดที่เพิ่มคอลัมน์ จำนวนวัน (count_date) ด้วย Pandas

3.2.1.3 สร้าง คอลัมน์ ค่าเฉลี่ยจำนวนผู้ชม (avg_hits) โดย คำนวณมาจาก จำนวนผู้ชมล่าสุด (hits_update) ลบด้วย จำนวนผู้ชมครั้งแรก (hits_start) แล้วหารด้วย จำนวนวัน (date_count)

	listing_id	price	area	beds	baths	date_time_start	date_time_update	hits_start	hits_update	date_count	avg_hits
0	62207	14000	33.01	1	1	2020-08-27	2022-09-07	95	155	742	0.08
1	63620	9000	26.48	1	1	2020-08-27	2022-09-07	213	314	742	0.14
2	63832	12000	29.08	1	1	2020-08-27	2022-02-18	175	187	541	0.02
3	63988	20000	61.91	1	1	2020-08-27	2021-07-12	102	208	320	0.33
4	64117	12000	32.43	1	1	2020-08-27	2021-01-04	5	11	131	0.05

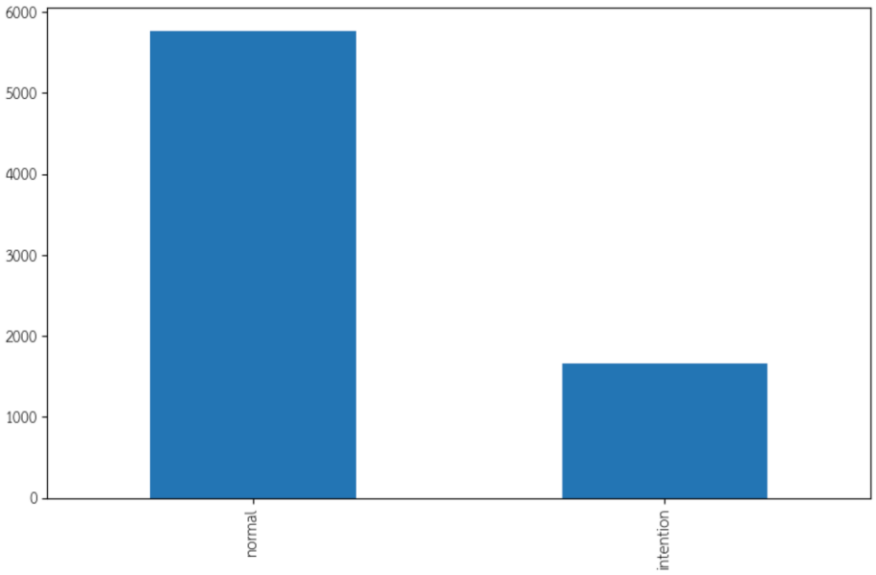
ภาพที่ 3.6 แสดงตัวอย่างข้อมูลห้องชุดที่เพิ่มคอลัมน์ ค่าเฉลี่ยจำนวนผู้ขม (avg_hits) ด้วย Pandas

3.2.1.4 ตัดข้อมูลผิดปกติ (Outlier) เนื่องจากการที่มีข้อมูลผิดปกติจะมีผลต่อการ วิเคราะห์ข้อมูล ซึ่งงานวิจัย นี้ได้ทำการตรวจสอบข้อมูลผิดปกติด้วยวิธีการหา IQR (Inter Quartile Range)

3.2.1.5 แบ่งกลุ่มจำนวนผู้ชม (hits_label) จากค่าเฉลี่ยจำนวนผู้ชม เป็น 2 กลุ่ม ได้แก่ กลุ่มผู้ชม (normal) และกลุ่มผู้ชม (intention) จากข้อมูลที่ได้ 8,824 ตัวอย่าง หลังจากทำการเตรียมข้อมูลไปข้างต้น จะได้เป็น กลุ่มผู้ชมปกติ 5,771 ตัวอย่าง และกลุ่มผู้ชมสนใจ 1,654 ตัวอย่าง

	listing_id	price	area	beds	baths	hits_label
0	62207	14000	33.01	1	1	normal
1	63620	9000	26.48	1	1	normal
2	63832	12000	29.08	1	1	normal
3	63988	20000	61.91	1	1	normal
4	64117	12000	32.43	1	1	normal

ภาพที่ 3.7 แสดงตัวอย่างข้อมูลห้องชุดที่แบ่งกลุ่ม (hits_label) ด้วย Pandas



ภาพที่ 3.8 แสดงตัวอย่างกราฟที่ได้หลังจาก แบ่งกลุ่มจำนวนผู้ชมด้วย Matplotlib

3.2.2 ข้อมูลภาพถ่าย

3.2.2.1 นำภาพถ่ายเข้าไปประมวลผลกับโมเดล YOLOv8 เพื่อได้มาซึ่งข้อมูลคุณลักษณะ ที่สนใจ คุณลักษณะที่สนใจได้แก่

- เก้าอี้ (chair)
- โซฟา (couch)
- เตียงนอน (bed)
- โต๊ะทานข้าว (dining table)
- โต๊ะ (table)
- ทีวี (tv)
- ไมโครเวฟ (microwave)
- เตาอบ (oven)
- ตู้เย็น (refrigerator)

3.2.2.2 นำภาพถ่าย เข้าไปคำนวณ หาค่า entropy, lightness และcontrast

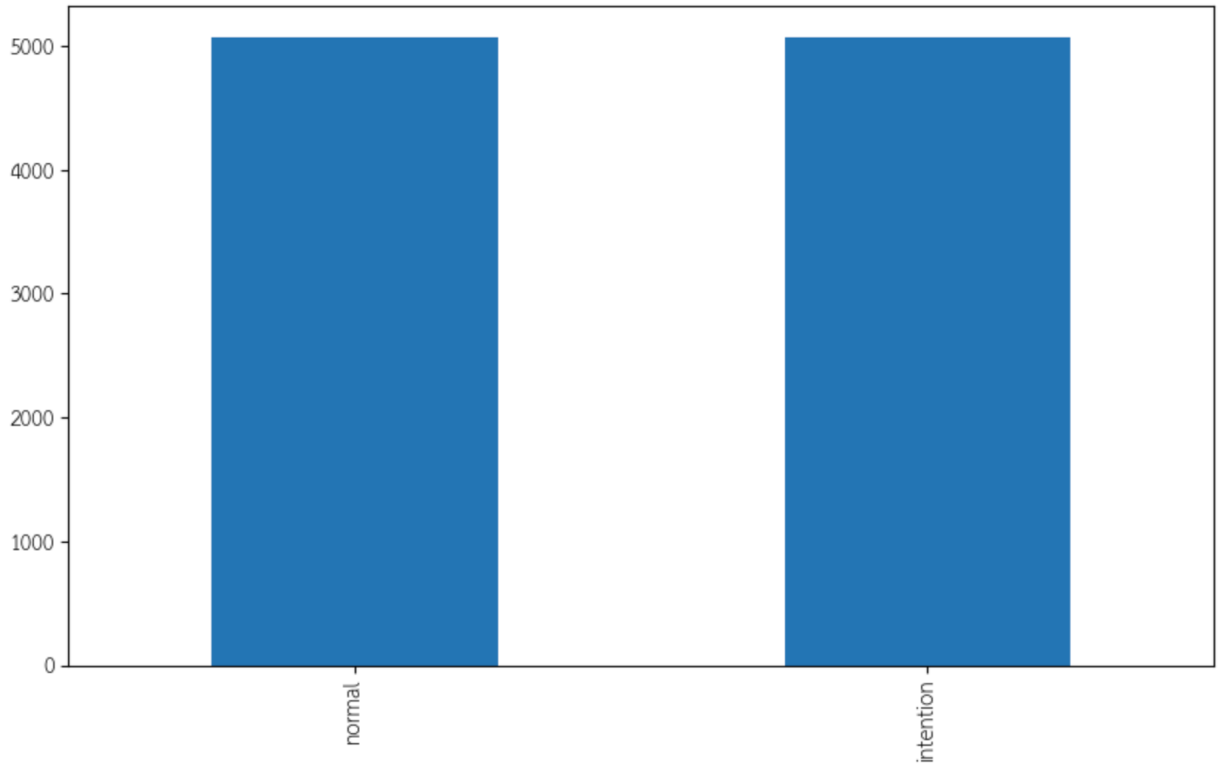
ภาพที่ 3.9 แสดงตัวอย่างภาพถ่ายที่แสดงถึงการได้ข้อมูลคุณลักษณะที่สนใจของภาพถ่าย ด้วย YOLOv8

3.2.3 นำข้อมูลทั่วไป และข้อมูลภาพถ่าย มารวมกัน โดยอ้างอิง ไอดี (listing_id)

	hits_label	price	area	beds	baths	chair	couch	bed	dining table	table	tv	microwave	oven	refrigerator	entropy photo	lightness photo	contrast photo
0	normal	14000	33.01	1	1	1	1	0	1	0	1	0	0	0	6.17	169.49	11.89
1	normal	9000	26.48	1	1	0	1	0	0	0	0	0	0	0	6.85	128.21	19.20
2	normal	12000	29.08	1	1	1	0	0	0	0	0	0	0	0	5.55	213.62	6.38
3	normal	20000	61.91	1	1	1	1	0	1	0	0	0	0	0	6.46	129.44	23.67
4	normal	12000	32.43	1	1	1	1	0	0	0	1	0	0	0	6.72	128.87	19.99

ภาพที่ 3.10 แสดงตัวอย่างข้อมูลที่ได้จากการนำข้อมูลทั่วไป และข้อมูลภาพถ่าย มารวมกัน ด้วย Pandas

3.2.4 ข้อมูลที่ได้มีลักษณะเป็น Imbalance data ไม่เหมาะต่อการนำมาสร้างแบบจำลองจึงต้องมีการทำ over sampling จึงมีการปรับความสมดุลโดยใช้วิธี SMOTE ให้ Label normal และ Label intention มีจำนวนเท่ากัน



ภาพที่ 3.11 แสดงตัวอย่างกราฟที่ได้หลังจาก Imbalance data ด้วย Matplotlib

3.3 การสร้างแบบจำลอง

ทำการสร้างแบบจำลองแบบ Supervised Learning Model (Binary Classification) ในการทำนายความสนใจของผู้บริโภค โดยการนำข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลทำการเปลี่ยน Data type ของ คุณลักษณะ chair, couch, bed, dining table, table, TV, microwave, oven, refrigerator จาก Numeric Data ให้เป็น Categorical Data แล้วนำเข้าเครื่องมือ Pycaret ด้วยการกำหนด Output Label เป็น hits_label ซึ่งประกอบด้วย ปกติ (normal) และสนใจ (intention) และทำการแบ่ง training 70% และ validation 30% โดยเปรียบเทียบ 2 โมเดล ดังนี้ Random Forest Classifier, Light Gradient Boosting Machine

3.4 การวัดประสิทธิภาพแบบจำลอง

วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพการทำงานของแบบจำลอง ที่สร้างจากแต่ละอัลกอริทึมสามารถพิจารณาได้จากค่า (Accuracy, Precision, Recall) ด้วยการใช้ Confusion Matrix โดยที่ความถูกต้องของค่าที่ทำนายได้จะมีค่ามากหรือน้อยขึ้นอยู่กับ ค่าความอ่อนไหว และ ค่าความจำเพาะ ของการทำนาย

3.5 เครื่องมือที่ใช้ในงานวิจัย

3.5.1 Visual Studio Code เป็นโปรแกรม Code Editor ที่ใช้ในการแก้ไขและปรับแต่งโค้ด จากค่ายไมโครซอฟท์ มีการพัฒนาออกมาในรูปแบบของ Open Source จึงสามารถนำมาใช้งานได้แบบฟรี ๆ ที่ต้องการความเป็นมืออาชีพ

3.5.2 Python เป็นภาษาโปรแกรมคอมพิวเตอร์ระดับสูงโดยถูกออกแบบมาให้เป็นภาษาสคริปต์ ที่อ่านง่ายโดยตัดความซับซ้อนของโครงสร้างและไวยกรณ์ของภาษา ออกไป ในส่วนของการแปลงชุดคำสั่ง ที่เราเขียนให้เป็นภาษาเครื่อง

3.5.3 Request เป็นโมดูลเสริมของ Python ใช้สำหรับการอ่านหน้าเว็บทำ

3.5.4 BeautifulSoup เป็นโมดูลเสริมของ Python ใช้สำหรับการแยกวิเคราะห์เอกสาร HTML และ XML

3.5.5 Miniconda เป็นตัวจัดการแพคเกจของ Python ใช้สำหรับติดตั้งแพ็คเกจ และ จัดสภาพแวดล้อม

3.5.6 Pandas เป็นโมดูลเสริมของ Python ใช้สำหรับการจัดการและวิเคราะห์ข้อมูลประสิทธิภาพสูง

3.5.7 Numpy เป็นโมดูลเสริมของ Python ใช้สำหรับการคำนวนทางคณิตศาสตร์

3.5.8 Cv2 เป็นโมดูลเสริมของ Python ใช้สำหรับการจัดการเกี่ยวกับข้อมูลภาพถ่าย

3.5.9 Scipy เป็นโมดูลเสริมของ Python ส่วนขยายต่อจาก NumPy ใช้สำหรับการคำนวณ ทางคณิตศาสตร์

3.5.10 Matplotlib เป็นโมดูลเสริมของ Python ใช้สำหรับการแสดงผลในรูปแบบ visualization

3.5.11 Ultralytics เป็นโมดูลเสริมของ Python ใช้สำหรับการทำ Image detection

3.5.12 PIL เป็นโมดูลเสริมของ Python ใช้สำหรับการจัดการเกี่ยวกับข้อมูลภาพถ่าย

3.5.13 Imblearn เป็นโมดูลเสริมของ Python ใช้สำหรับการเพื่อจัดการกับ Dataset ที่ไม่มี ความสมดุลกัน

3.5.14 Pycaret เป็นโมดูลเสริมของ Python ใช้สำหรับการช่วยให้การเปรียบเทียบผลลัพธ์ของ Machine Learning Model แต่ละโมเดลนั้นรวดเร็วยิ่งขึ้น โดยรวมเอา Machine Learning Library อื่น ๆ เข้ามาอยู่ด้วยกัน ตัวอย่าง Library ที่นำมารวมด้วย เช่น scikit-learn, XGBoost, LightGMB

3.5.15 Flask เป็นโมดูลเสริมของ Python ใช้สำหรับการสร้าง Web Application

3.5.16 HTML เป็นภาษาคอมพิวเตอร์ที่ใช้ในการแสดงผลของเอกสารบนเว็บไซต์

3.5.17 CSS เป็นภาษาคอมพิวเตอร์ที่ใช้สำหรับตกแต่งเอกสาร HTML

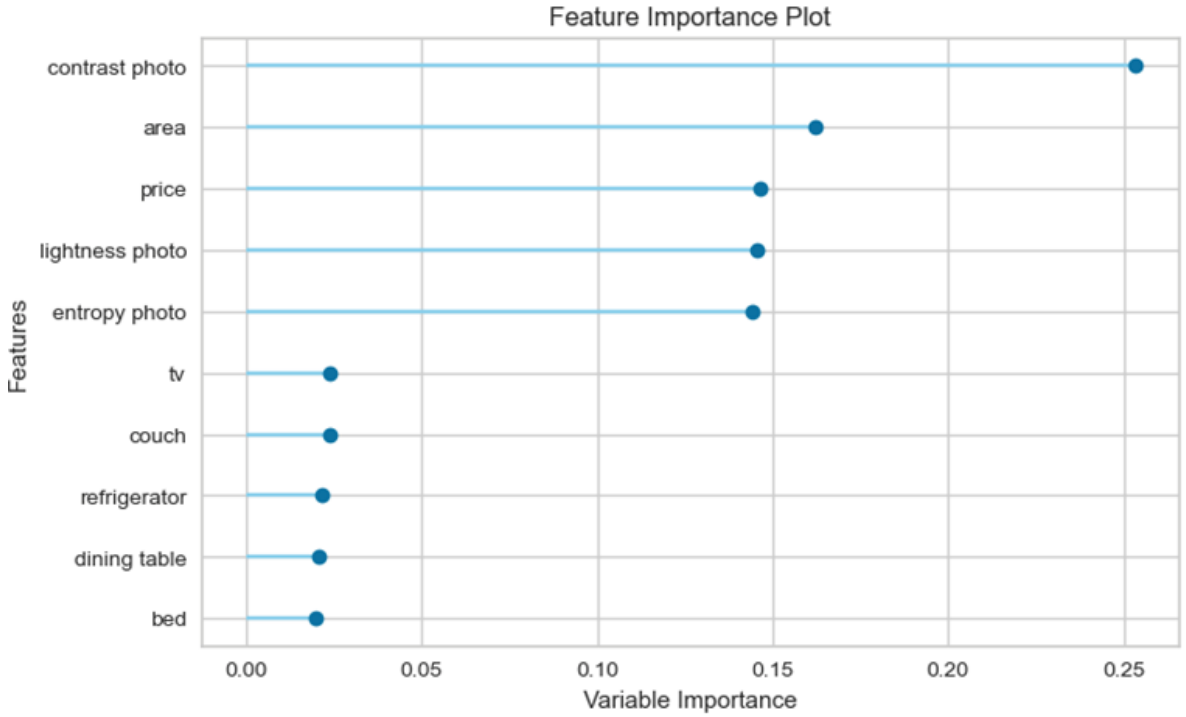
3.5.18 Javascript เป็นภาษาคอมพิวเตอร์ร์ที่ใช้สำหรับจัดการเอฟเฟต์หรือพฤติกรรม ของหน้าเว็บไซต์

บทที่ 4

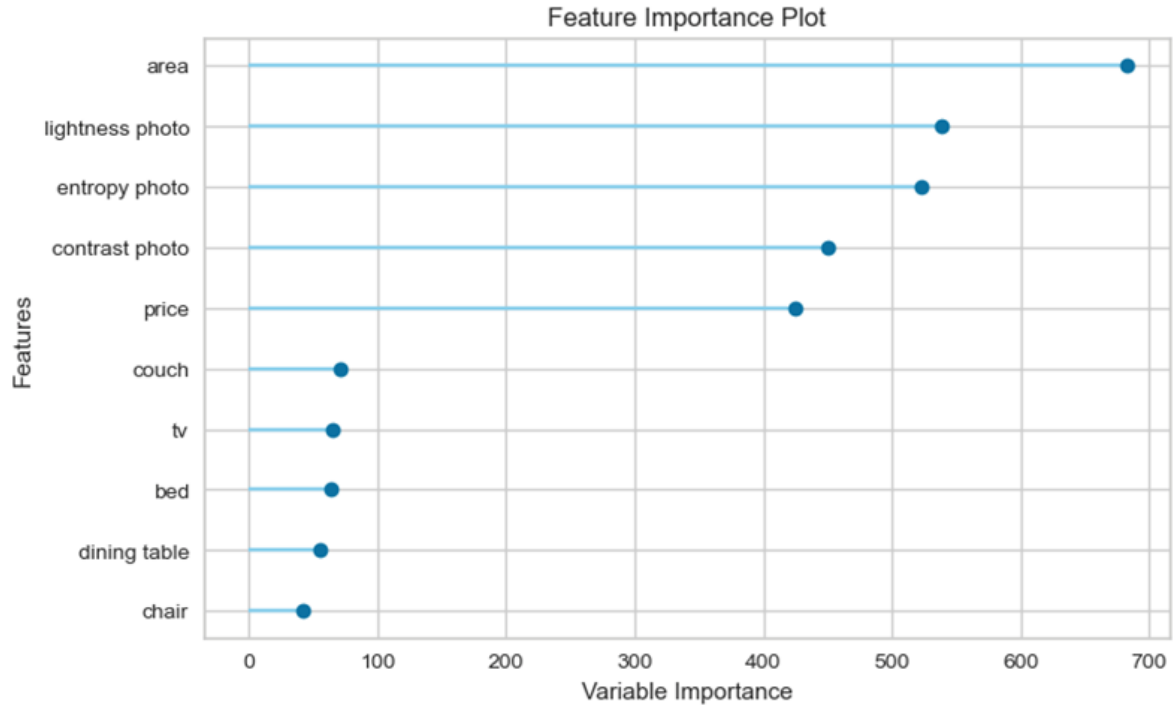
ผลการศึกษา

งานวิจัยนี้มีวัตถุประสงค์หาตัวแปรหรือปัจจัยสำหรับการพยากรณ์โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการพัฒนาแบบจำลองที่แสดงข้อมูลการพยากรณ์ความสนใจของอาคารชุดด้วยภาพถ่าย โดยมีรายละเอียดของผลการศึกษาดังต่อไปนี้

4.1 ตัวแปรหรือปัจจัยที่สำคัญสำหรับการพยากรณ์



ภาพที่ 4.1 แสดงตัวอย่างกราฟที่ที่แสดงฟีเจอร์ที่สำคัญ Random Forest Classifier ด้วย Pycaret



ภาพที่ 4.2 แสดงตัวอย่างกราฟที่ที่แสดงฟีเจอร์ที่สำคัญ Light Gradient Boosting Machine ด้วย Pycaret

จากรูปที่ 4.1 และ 4.2 แบบจำลอง Random Forest Classifier และ แบบจำลอง Light Gradient Boosting Machine มีฟีเจอร์ที่มีความสำคัญ 5 ลำดับแรก โดยประกอบไปด้วย entropy photo, lightness photo, contrast photo, price และ area ซึ่งจะเห็นว่า มีข้อมูลที่เกี่ยวกับคุณลักษณะภาพถ่าย 3 ฟีเจอร์ คือ entropy photo, lightness photo, contrast photo

4.2 ผลจากการวัดผลแบบจำลอง

ตารางที่ 4.1 ตารางการวัดผลแบบจำลองด้วย Predict Model ด้วย Pycaret

Model	Accuracy	AUC	Recall	Prec.	F1
Random Forest Classifier	0.8256	0.9031	0.7884	0.8517	0.8188
Light Gradient Boosting Machine	0.8180	0.8968	0.7812	0.8433	0.8111

ตารางที่ 4.2 ตาราง Confusion Matrix จากการวัดผล Random Forest Classifier ด้วย Pycaret

	True normal	True intention	Class Precision
Pred. normal	1313	209	**80.3%**
Pred. intention	322	1200	**85.2%**
Class Recall	**86.3%**	**78.8%**	

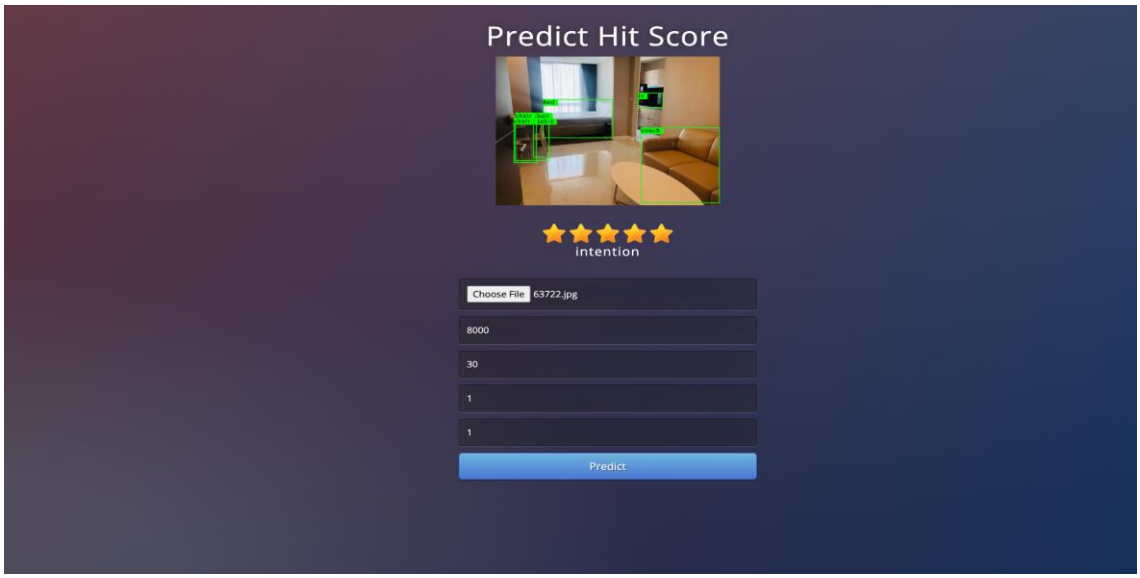
ตารางที่ 4.3 ตาราง Confusion Matrix จากการวัดผล Light Gradient Boosting Machine ด้วย Pycaret

	True normal	True intention	Class Precision
Pred. normal	1301	221	**79.6%**
Pred. intention	333	1189	**84.3%**
Class Recall	**85.5%**	**78.1%**	

จากตารางที่ 4.1 การวัดผล แบบจำลอง Random Forest Classifier และ แบบจำลอง Light Gradient Boosting Machine ได้ Accuracy 82.56 เปอร์เซ็นต์ และ 81.80 เปอร์เซ็นต์ ตามลำดับ ดังนั้น งานวิจัยนี้จะเลือก โมเดล แบบจำลอง Random Forest Classifier ไปใช้สำหรับ Web Application ในการแสดงผล

4.3 การแสดงผล

นำแบบจำลองที่ได้ไปประยุกต์ใช้กับข้อมูลเพื่อพัฒนา และแสดงผล ในรูปแบบ Web Application



ภาพที่ 4.3 Web Application

จาก Web Application ที่ได้พยากรณ์ จะได้ว่าบริโภคนิยมภาพถ่ายที่มีค่า entropy photo ที่ต่ำกว่า, lightness สูงกว่า photo และcontrast photo ที่สูงกว่า ตัวอย่างดังนี้

ภาพที่ 4.4 แสดงตัวอย่างรูปห้องชุด พยากรณ์สนใจ (intention)

ภาพที่ 4.5 แสดงตัวอย่างรูปห้องชุด พยากรณ์ปกติ (normal)

แต่ถ้ามีปัจจัยเท่านี้อาจทำให้การพยากรณ์คาดเคลื่อนได้ จึงต้องมีปัจจัยพื้นฐาน **price** และ **area** เป็นปัจจัยที่มาช่วยทำให้ การพยากรณ์ แม่นยำมากขึ้น

บทที่ 5

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้ได้วัตถุประสงค์หาตัวแปรหรือปัจจัยสำหรับการพยากรณ์โดยการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการพัฒนาแบบจำลองที่แสดงข้อมูลการพยากรณ์ความสนใจของอาคารชุดด้วยภาพถ่าย โดยสามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการศึกษา

5.1.1 ได้พัฒนา แบบจำลอง Random Forest Classifier และ แบบจำลอง Light Gradient Boosting Machine ได้ Accuracy 82.56 เปอร์เซ็นต์ และ 81.80 เปอร์เซ็นต์ ตามลำดับ โดยจะใช้ โมเดล แบบจำลอง Random Forest Classifier ในการพยากรณ์

5.1.2 ตัวแปรหรือปัจจัยที่สำคัญสำหรับการพยากรณ์ 5 ลำดับแรก โดยประกอบไปด้วย contrast photo, area, price, lightness photo และ entropy photo

5.1.3 ได้พัฒนา Web Application สำหรับพยากรณ์ความสนใจของอาคารชุด

5.2 ข้อเสนอแนะ

5.2.1 พิจารณาปัจจัย อื่นเพิ่มเติมในภาพถ่าย เช่น saturation, color

5.2.2 เพิ่มจำนวน class ให้มีมากกว่า 2 class เพื่อให้ได้ผลลัพธ์ที่ละเอียดมากขึ้น

5.2.3 พัฒนาประสิทธิภาพของแบบจำลอง หรือทดลองทำแบบจำลองอื่นเพิ่มเติม เพื่อให้ได้ผล การพยากรณ์ที่แม่นยำมากขึ้น

5.2.4 พัฒนา Web Application ให้สามารถ upload ได้ที่ละหลายภาพถ่าย แล้วแสดงผลออกมาว่ารูปไหน ปกติ (normal) หรือ สนใน (intention)

บรรณานุกรม

บรรณานุกรม

[1] Xin Li, Mengyue Wang and Yubo Chen, *The Impact Of Product Photo On Online Consumer Purchase Intention: An Image-processing Enabled Empirical Study*, CityUniversity of Hong Kong, Hong Kong, 2014.

[2] Fournier. S, "Consumers and their brands: Developing relationship theory in consumer research", *Journal of Consumer Research*, vol. 24 no. 4, pp.343-353, 1998.

[3] Hassanein K. and Head. M, "Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping", *International Journal of Human Computer Studies*, vol. 65 no. 8, pp. 689-708, 2007.

[4] Lin, TY. *et al,* "Microsoft COCO: Common Objects in Context", Computer Vision – ECCV 2014. ECCV 201, *Lecture Notes in Computer Science*, vol. 8693, Springer, Cham, doi: org/10.1007/978-3-319-10602-1_48

[5] M. A. Khder, "*Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application*" , Department of Computer Science, College of Arts and Science, Applied Science University, Bahrain, 2021.

[6] K.W. Bowyer *et al*. , "SMOTE: Synthetic Minority Over-sampling Technique" , 2011.

[7] กิตติพงศ์ สินจิตต์, "*ปัญหาข้อกฎหมายเกี่ยวกับอาคารชุดที่มีผลกระทบต่อผู้บริโภค,*" วิทยานิพนธ์ น.ม., มหาวิทยาลัยรามคำแหง, กรุงเทพฯ, 2541.

ภาคผนวก

ภาคผนวก ก

โค้ดการเก็บรวบรวมข้อมูล

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

from bs4 import BeautifulSoup
import requests
from datetime import timedelta, date, datetime
from time import sleep
import random
import subprocess
import re
import os
import json
import csv
import pandas as pd
import numpy as np
import sys, getopt
import urllib.request

def getData(url, contract_type, sort_by, zone_id, page, today):
    ### Get Api ###
    param = {}
    param['contract_type'] = contract_type
    param['zone_id'] = zone_id
    param['sort_by'] = sort_by
    param['page'] = page

    i = 0

    while True:
        if i > 2:
            break
        html = {}
        html['data'] = {}
        try:
            r = requests.post(url, param)
            soup = BeautifulSoup(r.text,"html.parser")
```

```
        # lastPage
        try:
            lastpage = [li for li in soup.find('ul',{'pagination'}).findAll('li')]
            lastpage = lastpage[len(lastpage)-1].find('a', href=True)['href']
            html['lastpage'] = int(re.findall("(\d+)",lastpage)[0])
        except:
            pass

        # data
        for li in soup.find('ul').findAll('li'):
            no = re.findall("\d+", li.find('a', href=True)['href'])[0]
            html['data'][no] = {}
            html['data'][no]['id'] = int(no)

            html['data'][no]['contract_type'] = li.find('a', {'property-featured-image'}).find('span',
{'badges'}).text

            html['data'][no]['title'] = li.find('div', {'property-info'}).find('a').text
            html['data'][no]['url'] = li.find('a', href=True)['href']
            html['data'][no]['image'] = li.find('a').find('img')['src']

            html['data'][no]['price'] = int(li.find('div', {'price'}).find('span').text.replace(',', ''))
            html['data'][no]['area'] = float(li.find('div', {'property-amenities'}).find('span',
{'area'}).find('strong').text)
            html['data'][no]['beds'] = int(li.find('div', {'property-amenities'}).find('span',
{'beds'}).find('strong').text)
            html['data'][no]['baths'] = int(li.find('div', {'property-amenities'}).find('span',
{'baths'}).find('strong').text)

            html['data'][no]['hits_start'] = int(re.findall("\d+", li.find('div', {'hits'}).find('span').text)[0])
            html['data'][no]['hits_update'] = int(re.findall("\d+", li.find('div', {'hits'}).find('span').text)[0])

            html['data'][no]['date_time_start'] = today
            html['data'][no]['date_time_update'] = today
```

```
            break
        except:
            i += 1
            rand = random.randint(1, 5)
            print("Sleep Time : " + str(rand) + " sec. :: Error Request " )
            sleep(rand)
            pass

    return html

def job():

    backdir = ''
    config_file = open(backdir + 'config/config.json')
    conf = json.load(config_file)

    pd_data = pd.read_csv(backdir + 'tmp/listing.tsv', sep='\t')
    tsv_transection = open(backdir + 'tmp/listing_transection.tsv', 'a')
    tsv_transection_writer = csv.writer(tsv_transection, delimiter='\t', lineterminator='\n', quotechar = '"')
    tsv_new = open(backdir + 'tmp/new_listing.tsv', 'a')
    tsv_new_writer = csv.writer(tsv_new, delimiter='\t', lineterminator='\n', quotechar = '"')
    tsv_count = open(backdir + 'tmp/count_listing.tsv', 'a')
    tsv_count_writer = csv.writer(tsv_count, delimiter='\t', lineterminator='\n', quotechar = '"')

    for zone in conf['zone']:

        print(zone['title'])

        today = str(date.today()) + ' 00:00:00'
        page = 1
        already_listing = 0
        new_listing = 0
        html = getData(conf['url'], conf['contract_type'], conf['sort_by'], zone['id'], page, today)

        for page in range(1, html['lastpage'] + 1):
            page_count = page
```

```
        html = getData(conf['url'], conf['contract_type'], conf['sort_by'], zone['id'], page, today)

        for listing_id, v in html['data'].items():
            lid = pd_data['listing_id'].tolist()
            price_status = 0

            try:
                # Already Listing
                lid.index(int(listing_id))
                pd_data.loc[pd_data['listing_id'] == int(listing_id), 'hits_update'] = v['hits_update']
                pd_data.loc[pd_data['listing_id'] == int(listing_id), 'date_time_update'] = today
                already_listing += 1
            except:
                # New Listing
                tmp = []
                for key, value in v.items():
                    tmp.append(value)
                tmp.append(zone['title'])

                try:
                    urllib.request.urlretrieve(v['image'], backdir + 'image/thumbnail/' + str(listing_id) + '.' +
v['image'].split('.')[2])
                    urllib.request.urlretrieve(v['image'].replace('thumb_', ''), backdir + 'image/original/' +
str(listing_id) + '.' + v['image'].split('.')[2])
                    if len(pd_data.index.values) == 0:
                        pd_data.loc[1] = tmp
                    else:
                        pd_data.loc[pd_data.index.values[len(pd_data.index.values)-1]+1] = tmp

                    # Write New Listing to new_listing.tsv
                    tsv_new_writer.writerow([tmp[0], tmp[10], today, zone['title']])
                    tsv_new.flush()

                    new_listing += 1
                except:
                    price_status = 1
```

```
                    pass

            if price_status == 0:
                # Write Listing transection to listing_transection.tsv
                tsv_transection_writer.writerow([int(listing_id), v['hits_update'], today])
                tsv_transection.flush()

        print("date:", today, ", page count:", page_count, ", already:", already_listing, ", new:", new_listing)

        # Write Listing Count (already, new) to count_listing.tsv
        tsv_count_writer.writerow([today, zone['title'], already_listing, new_listing])
        tsv_count.flush()

    # Write Listing to listing.tsv
    with open(backdir + 'tmp/listing.tsv' ,'w') as write_tsv:
        write_tsv.write(pd_data.to_csv(sep='\t', index=False))

    tsv_transection.close()
    tsv_new.close()
    tsv_count.close()

    config_file.close()

try:
    opts, args = getopt.getopt(sys.argv[1:] ,"h",["help"])
except getopt.GetoptError:
    print('')
    print('==== Program Scrap CondoLumpiniBrokerage ====')
    print('')
    print('Error! Please Check Command.')
    print('')
    print('Command : ')
    print('"python condolumpinibrokerage.py -h" or "python condolumpinibrokerage.py --help"')
    print('')
    sys.exit(2)
```

```
for opt, arg in opts:
    if opt in ("-h", "--help"):
        print('')
        print('==== Program Scrap CondoLumpiniBrokerage ====')
        print('')
        print('Command : ')
        print('python condolumpinibrokerage.py')
        print('')
        print('Example : ')
        print('python condolumpinibrokerage.py')
        print('')
        sys.exit()
    # elif opt in ("-t", "--task"):
    #     _task = arg
    # elif opt in ("-u", "--user"):
    #     _user = arg
    # elif opt in ("-p", "--type"):
    #     _type = arg
    # elif opt in ("-g", "--group"):
    #     _group = arg
    # elif opt in ("-y", "--year"):
    #     _year = arg

job()
```

2. โค้ดการดึงข้อมูลคุณลักษณะที่สนใจของภาพถ่าย (object detection)

```
#!/usr/bin/env python
import os
from ultralytics import YOLO
from PIL import Image
import pandas as pd
import numpy as np
import json
import csv

out_obj = {}
item = ['chair','couch','bed','dining table','table','tv','microwave','oven','refrigerator']

def detect():
    tsv_obj_detect = open('listing_obj_detect.tsv', 'w')
    tsv_obj_detect_writer = csv.writer(tsv_obj_detect, delimiter='\t', lineterminator='\n', quotechar = """)

    tsv_obj_detect_writer.writerow(['id'] + item)
    tsv_obj_detect.flush()

    path = "../../scrap_data/image/thumbnail"
    dir = os.listdir(path)

    for file in dir:
        # print(file)
        result = detect_objects_on_image(Image.open(path + '/' + file))
        tsv_obj_detect_writer.writerow([file.split('.')[0]] + result)
        tsv_obj_detect.flush()

    tsv_obj_detect.close()

def detect_objects_on_image(buf):
    model = YOLO("yolov8m.pt")
    results = model.predict(buf)
    result = results[0]
```

```
    output = []
    value = [0,0,0,0,0,0,0,0,0,0]
    for box in result.boxes:
        # x1, y1, x2, y2 = [
        #   round(x) for x in box.xyxy[0].tolist()
        # ]
        # class_id = box.cls[0].item()
        # prob = round(box.conf[0].item(), 2)
        # output.append([
        #   x1, y1, x2, y2, result.names[class_id], prob
        # ])
        # print(result.names[class_id])
        class_id = box.cls[0].item()
        try:
            if result.names[class_id] == 'chairs':
                obj_name = 'chair'
            else:
                obj_name = result.names[class_id]
            index = item.index(obj_name)
            value[index] = 1
        except:
            if obj_name in out_obj.keys():
                out_obj[obj_name] += 1
            else:
                out_obj[obj_name] = 1
            print(out_obj)
    return value


detect()
print(out_obj)
```

3. โค้ดการดึงข้อมูลคุณลักษณะที่สนใจของภาพถ่าย (entropy photo, lightness photo, contrast photo)

```
#!/usr/bin/env python
import os
import cv2
from scipy.stats import entropy
from numpy.linalg import norm
import math
import pandas as pd
import numpy as np
import json
import csv

out_obj = {}
item = ['entropy photo','lightness photo','contrast photo']

def asthetics():
    tsv_asthetics = open('listing_asthetics.tsv', 'w')
    tsv_asthetics_writer = csv.writer(tsv_asthetics, delimiter='\t', lineterminator='\n', quotechar = "'")

    tsv_asthetics_writer.writerow(['id'] + item)
    tsv_asthetics.flush()

    path = "../../scrap_data/image/thumbnail"
    dir = os.listdir(path)

    for file in dir:
        # print(file)
        result = asthetic_on_image(cv2.imread(path + '/' + file))
        tsv_asthetics_writer.writerow([file.split('.')[0]] + result)
        tsv_asthetics.flush()

    tsv_asthetics.close()

def asthetic_on_image(buf):
    value = [
```

```
        round(image_entropy(buf), 2),
        round(lightness(buf), 2),
        round(contrast(buf), 2)
    ]
    return value


def image_entropy(buf):
    gray_image = cv2.cvtColor(buf, cv2.COLOR_BGR2GRAY)
    _bins = 128
    hist, _ = np.histogram(gray_image.ravel(), bins=_bins, range=(0, _bins))
    prob_dist = hist / hist.sum()
    image_entropy = entropy(prob_dist, base=2)

    return image_entropy

def lightness(buf):
    img_hsv = cv2.cvtColor(buf, cv2.COLOR_BGR2LAB)
    lightness = img_hsv[:, :, 0].mean()
    return lightness

def contrast(buf):
    lab = cv2.cvtColor(buf,cv2.COLOR_BGR2LAB)

    # separate channels
    L,A,B=cv2.split(lab)

    # compute minimum and maximum in 5x5 region using erode and dilate
    kernel = np.ones((5,5),np.uint8)
    min = cv2.erode(L,kernel,iterations = 1)
    max = cv2.dilate(L,kernel,iterations = 1)

    # convert min and max to floats
    min = min.astype(np.float64)
    max = max.astype(np.float64)
```

```
    # compute local contrast
    contrast = (max-min)/(max+min)

    # get average across whole image
    average_contrast = 100*np.mean(contrast)

    return average_contrast

asthetics()
```

4. โค้ดเตรียมข้อมูลทั่วไปที่ตัดข้อมูลที่ผิดปกติ และแบ่งกลุ่มจำนวนผู้ชม แล้วนำข้อมูลข้อมูลทั่วไปที่ได้เตรียมมารวมกับข้อมูลคุณลักษะภาพถ่าย จากนั้นทำการจัดการปรับความสมดุลโดยเทคนิค Over-sampling

```
import pandas as pd
import datetime
import matplotlib.pyplot as plt
import matplotlib as mpl
import numpy as np
import datetime
from imblearn.over_sampling import SMOTE
from collections import Counter
from matplotlib import pyplot
from sklearn import preprocessing
from numpy import where

# %%
df = pd.read_csv('../scrap_data/tmp/listing.tsv', sep='\t')

# %%
df_obj = pd.read_csv('../extract_features/object_detection/listing_obj_detect.tsv', sep='\t')

# %%
df_asthetics = pd.read_csv('../extract_features/extract_features/listing_asthetics.tsv', sep='\t')

# %%
# เปลี่ยน date_time_start, date_time_update เป็น datatype datetime
```

```
df['date_time_start'] = df['date_time_start'].astype('datetime64[ns]')
df['date_time_update'] = df['date_time_update'].astype('datetime64[ns]')
df.dtypes

# %%
# เลือกข้อมูลที่จะมาทำการวิเคราะห์
# ราคา (price)
# ขนาดห้องชุด (area)
# จำนวนที่นอน (beds)
# จำนวนห้องน้ำ (baths)
# จำนวนผู้ชมครั้งแรก (hits_start)
# จำนวนผู้ชมล่าสุด (hits_update)
# วันที่เริ่มที่เก็บข้อมูล (date_time_start)
# วันที่ล่าสุดที่เก็บข้อมูล (date_time_update)

df = df[['listing_id', 'price', 'area', 'beds', 'baths', 'date_time_start', 'date_time_update', 'hits_start',
'hits_update']]

# %%
# เพิ่ม column date_count คือ การคำนวณจำนวนวัน ระหว่าง date_time_start ถึง date_time_update
df['date_count'] = df.apply(lambda x: abs((x['date_time_update'] - x['date_time_start']).days)+1, axis=1)

# %%
df['avg_hits'] = df.apply(lambda x: "{:.2f}".format((x['hits_update']-x['hits_start'])/x['date_count']) if
x['date_count'] > 0 else 0, axis=1)
df['avg_hits'] = df['avg_hits'].astype('float64')

# %%
np.mean(df['avg_hits'])

# %%
df['avg_hits_cut2'] = pd.cut(df['avg_hits'], [0,1,3])
df.to_csv('data_prepare_class.tsv', sep="\t")

# %%
df = pd.read_csv('data_prepare_class.tsv', sep='\t')
```

```
df.rename(columns = {'avg_hits_cut2':'hits_label'}, inplace = True)
df = df.replace(['(0, 1]'], 'normal')
df = df.replace(['(1, 3]'], 'intention')

# %%
df.sort_values(by=['hits_label']).head(5)

# %%
# เลือกข้อมูลที่จะมาทำการวิเคราะห์
df = df[['listing_id', 'price', 'area', 'beds', 'baths', 'hits_label', 'date_time_start']]

# %%
df_join = df.join(df_obj.set_index('id'), on='listing_id')
df_join = df_join.join(df_asthetics.set_index('id'), on='listing_id')

# %%
df_final = df_join[['hits_label','price', 'area', 'beds', 'baths', 'chair', 'couch', 'bed', 'dining table', 'table', 'tv',
'microwave', 'oven', 'refrigerator', 'entropy photo', 'lightness photo', 'contrast photo', 'date_time_start']]

df_final = df_final.dropna()

# %%
df_final_feature = df_final[['hits_label','price', 'area', 'beds', 'baths', 'chair', 'couch', 'bed', 'dining table',
'table', 'tv', 'microwave', 'oven', 'refrigerator', 'entropy photo', 'contrast photo', 'date_time_start']]

# %%
# train data
# df2 = df.loc[(df['date_time_start'] >= "2022-01-01") & (df['date_time_start'] <= "2022-07-31")]
df_final_train = df_final.loc[df_final['date_time_start'] <= "2022-07-31"]
df_final_train = df_final_train[['hits_label', 'price', 'area', 'beds', 'baths', 'chair', 'couch', 'bed', 'dining table',
'table', 'tv', 'microwave', 'oven', 'refrigerator', 'entropy photo', 'lightness photo', 'contrast photo']]

# %%
# summarize class distribution
counter = Counter(df_final_train['hits_label'])
```

```
# %%
oversample = BorderlineSMOTE(random_state=123)
X_smote, y_smote = oversample.fit_resample(df_final_train[['price', 'area', 'beds', 'baths', 'chair', 'couch',
'bed', 'dining table', 'table', 'tv', 'microwave', 'oven', 'refrigerator', 'entropy photo', 'lightness photo', 'contrast
photo']], df_final_train[['hits_label']])
smote_array = np.concatenate([X_smote, y_smote], axis=1)
df2 = pd.DataFrame(smote_array, columns=['price', 'area', 'beds', 'baths', 'chair', 'couch', 'bed', 'dining table',
'table', 'tv', 'microwave', 'oven', 'refrigerator', 'entropy photo', 'lightness photo', 'contrast photo', 'hits_label'])

# %%
df2.to_csv('data_train.tsv', sep="\t")

# %%
# train data
df_final_test = df_final.loc[(df_final['date_time_start'] >= "2022-08-01")]
df_final_test = df_final_test[['hits_label', 'price', 'area', 'beds', 'baths', 'chair', 'couch', 'bed', 'dining table',
'table', 'tv', 'microwave', 'oven', 'refrigerator', 'entropy photo', 'lightness photo', 'contrast photo']]

df_final_test.to_csv('data_test.tsv', sep="\t")
```

5. โค้ดสร้างแบบจำลองแบบ Supervised Learning Model (Binary Classification)

```
# %%
# object
import csv
import pandas as pd
import pandas_profiling
import cv2
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as mpl
from pycaret.classification import *
from pycaret.utils.generic import check_metric
from sklearn.metrics import classification_report
```

```
# %%
df_train = pd.read_csv('data_train.tsv', sep='\t')
df_train = df_train.dropna()
cols = ['hits_label', 'price', 'area', 'beds', 'baths', 'entropy photo', 'lightness photo', 'contrast photo', 'chair',
'couch', 'bed', 'dining table', 'table', 'tv', 'microwave', 'oven', 'refrigerator']

df_train = df_train[cols]
df_train['chair'] = df_train['chair'].astype(int).astype(str)
df_train['couch'] = df_train['couch'].astype(int).astype(str)
df_train['bed'] = df_train['bed'].astype(int).astype(str)
df_train['dining table'] = df_train['dining table'].astype(int).astype(str)
df_train['table'] = df_train['table'].astype(int).astype(str)
df_train['tv'] = df_train['tv'].astype(int).astype(str)
df_train['microwave'] = df_train['microwave'].astype(int).astype(str)
df_train['oven'] = df_train['oven'].astype(str)
df_train['refrigerator'] = df_train['refrigerator'].astype(int).astype(str)
df_train.head(5)

# %%
report = pandas_profiling.ProfileReport(df_train)
report

# %%
df_test = pd.read_csv('data_test.tsv', sep='\t')
df_test = df_test[cols]
df_test.rename(columns = {'hits_label':'hits_label_2'}, inplace = True)

# %%
exp_mclf101 = setup(data = df_train, target = 'hits_label', train_size = 0.7, session_id=123)

# %%
compare_models()

# %%
rf = create_model('rf')
```

```
# %%
plot_model(rf, plot = 'confusion_matrix')

# %%
plot_model(rf, plot = 'class_report')

# %%
plot_model(rf, plot = 'feature')

# %%
plot_model(rf, plot = 'pr')

# %%
plot_model(rf, plot = 'auc')

# %%
evaluate_model(rf)

# %%
predict_model(rf)

# %%
final_rf = finalize_model(rf)

# %%
lightgbm = create_model('lightgbm')

# %%
plot_model(lightgbm, plot = 'confusion_matrix')

# %%
plot_model(lightgbm, plot = 'class_report')

# %%
plot_model(lightgbm, plot = 'feature')
```

```
# %%
evaluate_model(lightgbm)


# %%
predict_model(lightgbm)


# %%
final_lightgbm = finalize_model(lightgbm)


# %%
save_model(final_rf,'Final rf Model Object')
```

ประวัติผู้เขียน

ชื่อ – นามสกุล วรวุฒิ สว่างอัม

ประวัติการศึกษา

พ.ศ. 2551 - ปริญญาตรี สาขาสาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยโยนก

ประสบการณ์ทำงาน

พ.ศ. 2566 - Senior Programmer,
บริษัท อาร์ ที เอส (2003) จำกัด