การทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง

ชัรวรรค์ ปูเตะ[1*], จันตรี ผลประเสริฐ[2]

บทคัดย่อ

งานวิจัยมีนี้วัตถุประสงค์เพื่อสร้างแบบจำลองที่สามารถทำนายราคารถยนต์มือสองได้อย่างแม่นยำโดยใช้เทคนิคการเรียนรู้ของเครื่องและศึกษาปัจจัยที่มีผลต่อราคารถยนต์มือสองจากชุดข้อมูลรายการรถยนต์มือสองของตลาดรถยนต์ในสหราชอาณาจักร งานวิจัยนี้สามารถเป็นประโยชน์ต่อผู้ขาย ผู้ซื้อ และผู้ผลิตรถยนต์ในตลาดรถมือสอง ในการคาดการณ์ราคาที่เหมาะสมอย่างแม่นยำ งานวิจัยนี้จึงพัฒนาแบบจำลองการทำนายราคารถยนต์มือสองของหลากหลายยี่ห้อและหลากหลายรุ่น ด้วยแบบจำลองการเรียนรู้ด้วยเครื่อง โดยพิจารณาจากคุณลักษณะต่าง ๆ ของรถยนต์ เช่น รุ่นรถ ปีที่ผลิต ขนาดเครื่องยนต์ ประเภทของเชื้อเพลิงที่ใช้ ภาษี-ถนน ประเภทของเกียร์ และเลขไมล์รถยนต์ เพื่อนำมาใช้วิเคราะห์ข้อมูลและเปรียบเทียบหลากหลายแบบจำลองและหาแบบจำลองที่แม่นยำที่สุด เช่น การถดถอยต้นไม้การตัดสินใจ Decision Tree Regression, การถดถอยแบบเชิงเส้น Linear Regression,การถดถอยแบบ Ridge, ลาสโซ่ LassoและRandom forest เป็นต้น จากการทดลองครั้งนี้ Random forest เป็นแบบจำลองที่มีประสิทธิภาพที่ดีที่สุดโดยมีผลลัพธ์จาการวัดประสิทธิภาพดังนี้ MAE 1139.238, MAPE 0.073,MSE 3496491.842 และR-Squared score 0.96 รองลงมาคือแบบจำลอง Linear Regression MAE 0.13,MAPE 0.95, MSE 0.030 และR-Squared score 0.96 ถัดมาคือแบบจำลอง DecisionTreeRegression MAE 1417.302, MAPE 0.90, MSE 6752453.883 และR-square 0.96 ถัดมาคือแบบจำลอง Ridge MAE 2254.940, MAPE 0.180, MSE 13459786.977และ R-square 0.86 และแบบจำลองสุดท้ายได้ผลลัพธ์การวัดประสิทธิภาพที่น้อยที่สุดคือ Lasso MAE 2309.359, MAPE 0.186, MSE 13761927.624และ R-square 0.85

คำสำคัญ : การเรียนรู้ของเครื่อง, ปัจจัยที่มีผลต่อราคารถยนต์มือสอง, MAE,MAPE,R-Squared

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 099-2341870 E-mail address: sarwan.puteh@g.swu.ac.th

Prediction of used car price using machine learning

Sarwan Puteh[1*], chantri polprasert[2]

Abstract

The objective of this research is to develop a predictive model that can accurately predict the prices of used cars using machine learning techniques and studying the factors affecting the prices of used cars from a dataset of used cars in the UK car market. This research can be useful for car sellers, buyers, and manufacturers in the used car market to predict appropriate prices with accuracy. The research developed a predictive model for various makes and models of cars using a machine learning model considering various features of the car such as car model, year of manufacture, engine size, fuel type, tax, transmission type, and mileage to analyze the data and compare various models to find the most accurate model. The models tested include Decision Tree Regression, Linear Regression, Ridge Regression, Lasso, and Random Forest. Random Forest was found to be the most accurate model with the following performance metrics: MAE 1139.238, MAPE 0.073, MSE 3496491.842, and R-Squared score 0.96. The next best model was Linear Regression with the following performance metrics: MAE 0.13, MAPE 0.95, MSE 0.030, and R-Squared score 0.96. The other models tested had lower performance metrics, with Decision Tree Regression having a MAE of 1417.302, MAPE 0.90, MSE 6752453.883, and R-Squared score 0.96; Ridge Regression having a MAE of 2254.940, MAPE 0.180, MSE 13459786.977, and R-Squared score 0.86; and Lasso having the lowest performance metrics with a MAE of 2309.359, MAPE 0.186, MSE 13761927.624, and R-Squared score 0.85.

Keywords : machine learning, factors that affect the prcie ,MAE,MAPE,R-Squared

---

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

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

* Corresponding author:Tel.: 099-2341870 E-mail address: sarwan.puteh@g.swu.ac.th

บทนำ

รถยนต์มือสอง คือ รถยนต์ที่ผ่านการใช้งานมาแล้วจากบุคคลใดบุคคลหนึ่ง อาจมีการใช้งานมาก น้อยหรือหนัก เบา แตกต่างกันออกไป รถยนต์มีส่วนสำคัญในการดำรงชีวิตของมนุษย์เป็นอย่างมาก ตอบสนองความสะดวกสบายในการเดินทางในชีวิตประจำวันไม่ว่าจะเป็นการเดินทางไปทำงาน ประกอบธุรกิจ หรือท่องเที่ยวเป็นต้น เนื่องด้วยเศรษฐกิจในปัจจุบันรถยนต์มือสองหรือรถยนต์ใช้แล้ว เข้ามามีบทบาททดแทนรถยนต์มือหนึ่ง จุดเด่นสำคัญอยู่ที่เรื่องของราคาที่ถูกกว่ารถมือหนึ่งซึ่งนับเป็นอีกทางเลือกหนึ่งที่จะหาซื้อรถยนต์ไว้ในครอบครองได้ในราคาที่เพียงพอกับรายได้ ปัจจุบันรถมือสองมีจำหน่ายอย่างแพร่หลายทั้งจากบริษัท เต็นท์รถ หรือบุคคลทั่วไปที่ต้องการขายรถ ทำให้ผู้ที่กำลังจะซื้อรถมือสองมีตัวเลือกในการเปรียบเทียบเรื่องของราคา รวมไปถึงคุณภาพของรุ่นรถที่สนใจได้ แม้จะเป็นรถมือสอง แต่สามารถนำไปขายต่อได้หากต้องการรถยนต์รุ่นที่ใหม่กว่า อีกทั้งรถมือสองโอกาสผ่านการอนุมัติค่อนข้างง่าย เนื่องจากวงเงินในการกู้ซื้อรถไม่สูงมากเท่ารถยนต์มือหนึ่ง งานวิจัยของ Ken Research บริษัทวิจัยด้านการตลาดในอินเดียที่เปิดเผยล่าสุดเมื่อเดือนสิงหาคมในปี 2564 เกี่ยวกับธุรกิจรถยนต์มือสองที่กำลังได้รับความนิยมไปทั่วโลก ไม่ใช่เพียงในอินเดีย หรือเอเชียเท่านั้น แม้แต่ประเทศในยุโรป หรือสหรัฐอเมริกาเองได้ให้ความสนใจตั้งแต่โลกเข้าสู่การแพร่ระบาดโควิด โดยข้อมูลในรายงานของ Ken Research[3] ได้ให้ข้อมูลว่าประเทศไทย (Thailand Used Car Market Outlook To 2025) เป็นหนึ่งในประเทศที่มีอัตราการซื้อรถยนต์มือสองสูงในช่วงสถานการณ์โควิด (เทียบกับตลาด SEA) ถึงแม้ว่าธุรกิจรถยนต์มือสองในประเทศไทยจะได้รับความนิยมอยู่แล้วตั้งแต่ก่อนที่จะมีโรคระบาดก็ตาม ภาพรวมตลาดรถยนต์มือสองในปีที่ผ่านมา แม้ได้รับผลกระทบจากสถานการณ์โควิดแต่กลับทำให้เกิดความต้องการรถยนต์ส่วนตัวเพิ่มมากขึ้นเพื่อหลีกเลี่ยงการใช้บริการระบบขนส่งสาธารณะ สมาคมผู้ประกอบการรถยนต์ใช้แล้วระบุว่าตลาดธุรกิจรถยนต์ใช้แล้ว มีมูลค่าประมาณ 8 หมื่นล้านบาท ในปี 2564 เติบโตประมาณ 10 - 15% สำหรับรถยนต์นั่ง และขยายตัว 1.5 - 3.5% สำหรับรถบรรทุก อ้างอิงข้อมูลจากกรุงเทพกิจโดยในปี 2565 ราคารถมือสองของไทยยังมีโอกาสปรับขึ้นต่อเนื่อง ทิศทางตลาดรถยนต์มือสองในปี 2565 จะยังมีแนวโน้มที่ดี ผู้บริโภคยังคงมีความต้องการซื้ออย่างต่อเนื่อง เพราะได้เปรียบเรื่องราคาเมื่อเทียบกับรถใหม่ อีกทั้งรถยนต์มือสองในปัจจุบันมีอายุการใช้งานไม่มาก ยังอยู่ในเงื่อนไขการรับประกัน ทำให้ลูกค้ามีทางเลือกและมั่นใจในการซื้อรถมือสองมากขึ้นและยังช่วยในการประหยัดค่าใช้จ่าย ช่วงที่ผ่านมาแก้ปัญหาของการซื้อรถยนต์มือสองที่มีราคาไม่สมเหตุสมผล โดยใช้คนเข้าไปดูสภาพรถตามร้านขายรถมือสองหรือเต็นขายรถยนต์มือสองและประเมินราคา ซึ่งรถยนต์บางคันก็จะไม่สามารถประเมินราคาตามลักษณะภายนอกได้ เนื่องจากรถยนต์คันนั้นๆอาจจะถูกดัดแปลงสภาพให้ดูเหมือนใหม่และกลับเลขไมล์ให้แสดงเลขไมล์น้อย มีผลเสียคือนอกจากจะไม่สามารถการันตีได้ว่าจะได้รถยนต์มีค่าใช้จ่ายและเสียเวลา จึงมีแนวทางแก้ไขให้ผู้ซื้อรถยนต์ใช้แล้วได้รับสินค้าที่มีคุณภาพและเพิ่มความเชื่อมั่นให้กับผู้ซื้อรถยนต์ โดยใช้การเรียนรู้ด้วยเครื่องในการแก้ปัญหา และการนำผลลัพธ์ไปใช้กับเว็บ E-commerce ด้านรถยนต์เพื่อให้ข้อมูลที่เป็นประโยชน์ในการตัดสินใจซื้อ-ขายรถยนต์มือสอง

จากพฤติกรรมการซื้อรถยนต์มือสองของผู้บริโภคสิ่งสำคัญที่สุดในการเลือกซื้อรถยนต์มือสองคือการมองหาราคาที่สมเหตุสมผลและสภาพรถยนต์ ปัญหาของผู้ต้องการจะขายต่อรถยนต์ที่ใช้แล้วแต่ไม่รู้จะเสนอขายในราคาเท่าใด หากนำรถยนต์ที่ใช้แล้วไปขายที่เต็นรถยนต์ก็อาจจะถูกกดราคา ซึ่งก่อนหน้านี้ผู้ที่ต้องการจะขายต่อรถยนต์ใช้วิธีการเทียบราคารถยนต์รุ่นเดียวกันกับรถยนต์ของตนทางเว็บไซต์ E-commerce ด้านรถยนต์ ซึ่งจะไม่สามารถเปรียบเทียบได้อย่างถูกต้องและแม่นยำเนื่องจากราคารถยนต์มาจากหลายๆปัจจัยเช่นระยะทางที่ใช้งาน ,ประวัติการชนและสภาพรถยนต์เพราะว่าการใช้งานของแต่ละบุคคลแตกต่างกัน งานวิจัยนี้จึงช่วยแก้ปัญหาทั้งฝ่ายผู้ซื้อและผู้ขายรถยนต์มือสอง โดยใช้โมเดลการเรียนรู้ของเครื่องทำนายราคารถยนต์มือสอง การหาข้อมูลรถยนต์จากแพลต์ฟอร์ม E-commerce ทางด้านรถยนต์ และการเปรียบเทียบราคารถยนต์มือสองที่เหมาะสมกับคุณภาพของรถยนต์

เป็นเรื่องที่ยากและใช้เวลา งานวิจัยนี้ยังช่วยให้ผู้ประกอบการรถยนต์มือสองในด้านการให้ราคาประมูลรถยนต์ที่ไม่สูงเกินกว่าราคาตลาดและคุณภาพตามการใช้งาน งานวิจัยนี้จึงพัฒนาแบบจำลองการทำนายราคารถยนต์มือสองของหลากหลายยี่ห้อและหลากหลายประเภท [1]ด้วยแบบจำลองการเรียนรู้ด้วยเครื่อง โดยพิจารณาจากคุณลักษณะต่าง ๆ ของรถยนต์ เช่น รุ่นรถ จำนวนปีที่ใช้ ประเภทของเชื้อเพลิงที่ใช้ ประเภทของผู้ขาย ประเภทของเกียร์ และจำนวนกิโลเมตรที่ขับรถมาจนถึงปัจจุบัน เพื่อนำมาใช้วิเคราะห์ข้อมูลและเปรียบเทียบหลากหลายแบบจำลอง เช่น การถดถอยต้นไม้การตัดสินใจ Decision Tree Regression, การถดถอยแบบเชิงเส้น Linear Regression, การถดถอยแบบ Ridge, ลาสโซ่ Lasso, XGBoost, และ XGBoost_ HYPEROPT เป็นต้น แล้วนำแบบจำลองที่แม่นยำที่สุด นำข้อมูลมาแสดงผลลัพธ์เป็น Dashboard ให้ลูกค้านำข้อมูลไปตัดสินใจซื้อ – ขายเพื่อลดความเสี่ยงในการลงทุน

วิธีดำเนินการ

ขั้นตอนที่ 1 : แนะนำชุดข้อมูลที่ใช้ในการศึกษานี้

ผู้วิจัยนำข้อมูลเกี่ยวกับรายการรถยนต์มือสองในตลาดรถยนต์มือสองมาวิเคราะห์ในการทำวิจัยครั้งนี้ โดยใช้ชุดข้อมูลการรถยนต์มือสองในสหราชอาณาจักร จำนวนข้อมูลทั้งหมด 100,000 แถว จากฐานข้อมูล 100,000 UK Used Car Data set [2] ประกอบด้วย 10 คอลัมน์ ดังตารางที่ 1-2

ตาราง 1 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description)

No	Name	Data type	Description
1	Model	Data type is string	รุ่นรถยนต์เช่น Ford Focus
2	Year	Data type is intiger	รุ่นปี
3	Price	Data type is Float	ราคารถยนต์อยู่ในช่วง450 ถึง 159,999 ดอลลาร์สหรัฐ
4	Transmission	Data type is string	เกียร์รถประกอบด้วย Manual,Automatic Semi-Auto
5	Mileage	Data type is intiger	ไมล์สะสม
6	fuelType	Data type is string	ชนิดเชื้อเพลิง
7	engineSize	Data type is Float	ขนาดเครื่องยนต์

ตาราง 2 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description)

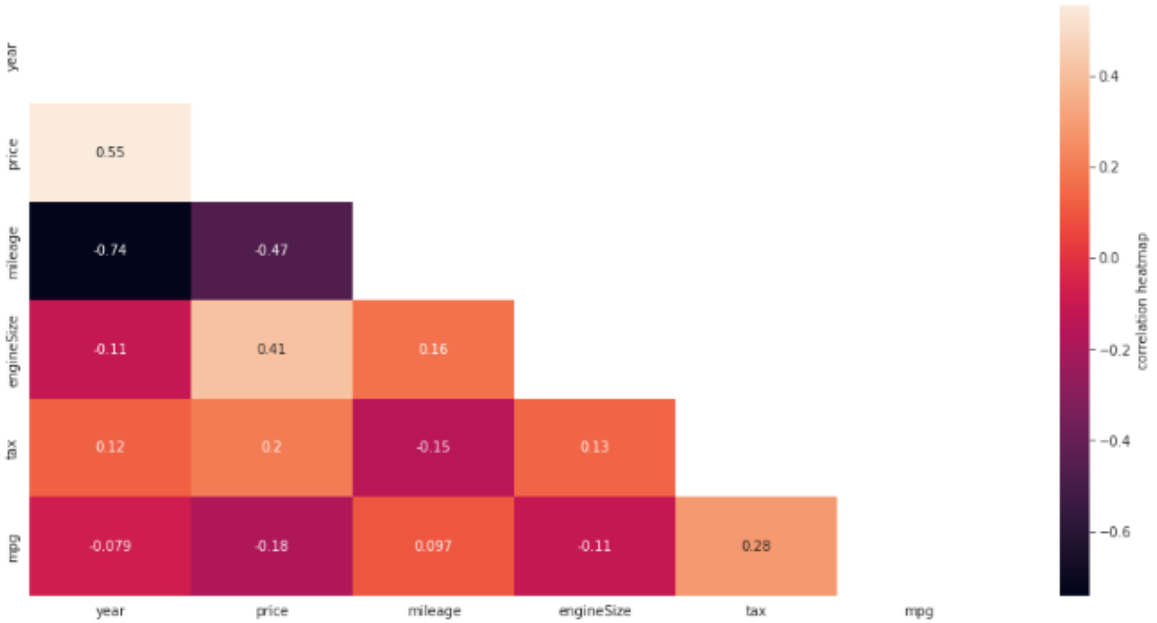
No	Name	Data type	Description
8	Company	Data type is string	บริษัทผู้ผลิตเช่น Ford,Toyota,bmw, audi,merczedes benz
9	Tax	Data type is intiger	ภาษีรถยนต์
10	Mpg	Data type is Float	เป็นตัววัดความประหยัดน้ำมันเชื้อเพลิงของรถยนต์

ขั้นตอนที่ 2 : การนำเข้าข้อมูล ตรวจสอบข้อมูล และพิจารณาข้อมูล

นำเข้าข้อมูลรายการรถยนต์ ทำการตรวจสอบข้อมูลโดยใช้ชุดคำสั่งภาษาไพธอน จากนั้นทำการรวบรวมข้อมูลทั้ง 9 ที่สมบูรณ์มารวมไว้ในไฟล์เดียว ทำความสะอาดข้อมูล จัดการ nan ลบข้อมูลที่ไม่ถูกต้อง และจัดFormat การจัดฟิวเจอร์ให้ง่ายต่อการใช้งานเช่นการหาอายุของรถจากปีผลิตเพื่อนำข้อมูลไปใช้ในขั้นตอนต่อไป

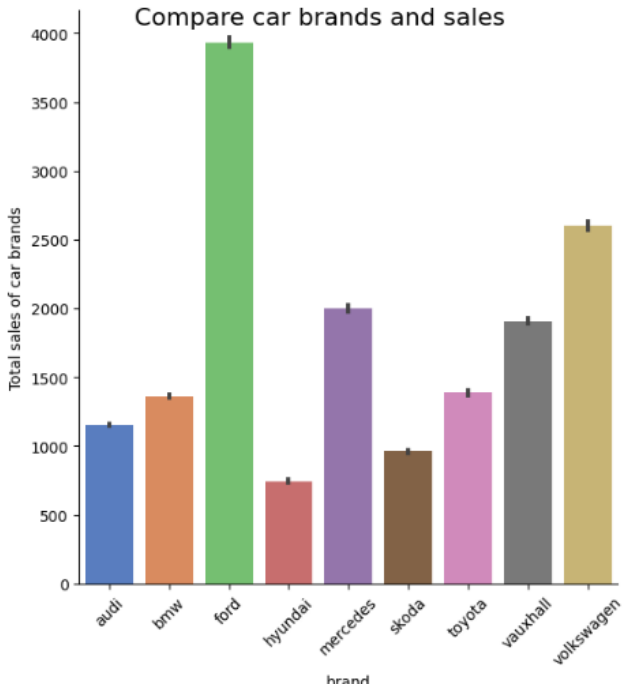
ขั้นตอนที่ 3 : การสำรวจข้อมูลและการวิเคราะห์ข้อมูล

จากการสำรวจและวิเคราะห์ข้อมูลคอลัมน์ที่ส่งผลต่อราคามากที่สุด พบว่าคอลัมน์ ปีรถ (year)หรืออายุรถ,ขนาดเครื่องยนต์ (engine Size), เลขไมล์สะสม (mileage) ส่งผลต่อราคารถยนต์มือสองมากที่สุดตามลำดับ Correlation หรือ ค่าสหสัมพันธ์ เป็นการดูทิศทางความสัมพันธ์ระหว่างตัวแปร 2 ตัว โดยมี Correlation Coefficient (r) หรือ ค่าสัมประสิทธิ์สหสัมพันธ์เป็นตัวบ่งชี้ถึงความสัมพันธ์นี้ ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์นี้จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 ซึ่งหากมีค่าใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากมีค่าใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันโดยตรงอย่างมาก และหากมีค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกันดังภาพประกอบที่ 1 และผู้วิจัยได้ทำการวิเคราะห์ข้อมูลเพื่อเปรียบเทียบระหว่างคุณลักษณะของข้อมูลต่างๆในชุดข้อมูลเพื่อศึกษาปัจจัยที่เกี่ยวข้องกับราคารถยนต์



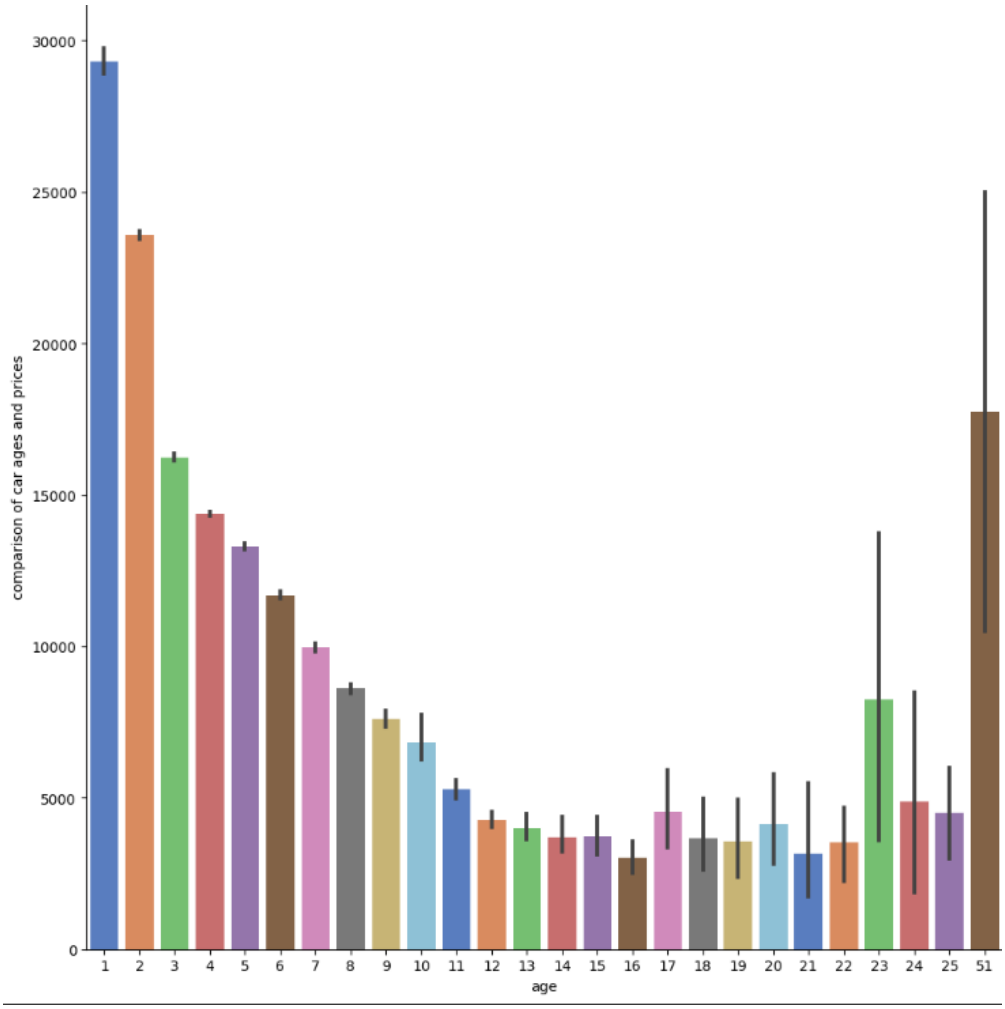
ภาพประกอบที่ 1 ความสัมพันธ์ Correlation heatmap แสดงปัจจัยที่มีผลต่อราคารถยนต์มือสอง

การสำรวจข้อมูลยอดขายรถยนต์แบรนด์ยอดนิยมจากชุดข้อมูล โดยนับตามจำนวนคันเพื่อหาว่ารถยนต์แบรนด์ยอดนิยมในชุดข้อมูลนี้สำหรับการนำไปใช้วิเคราห์แต่ละแบรนด์ในขั้นตอนต่อไปดังรูปภาพที่ 2



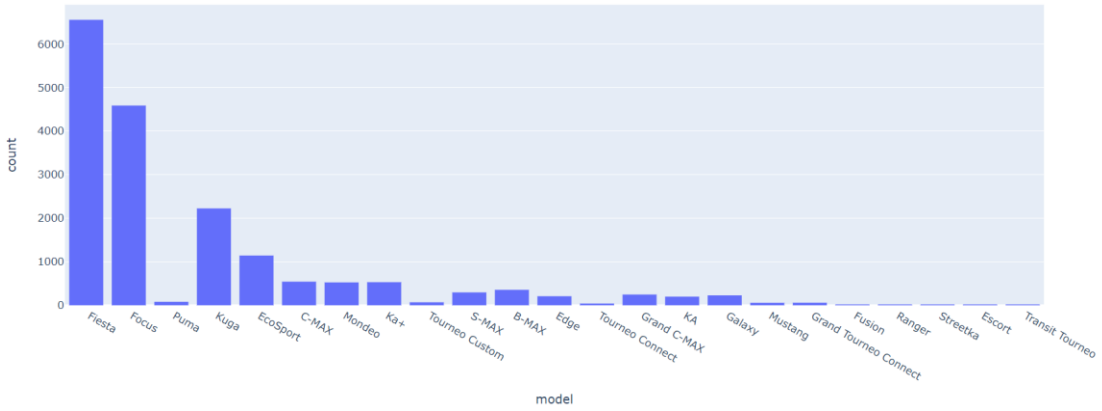
ภาพประกอบที่ 2 ภาพเป็นรูปภาพการเปรียบเทียบจำนวนการใช้งานกับแบรนด์ผู้ผลิต

การเปรียบเทียบอายุรถยนต์กับราคารถยนต์มือสอง โดยสกุลเงินของประเทศอังกฤษ (British pound) ซึ่งเป็นสกุลเงินที่ใช้ในการค้าขายและการเงินในประเทศอังกฤษ และบางประเทศในยุโรป โดยใช้ seaborn พลอตกราฟให้แกน xคืออายุรถยนต์และแกน Y คือราคารถยนต์มีองดังรูปภาพที่ 3

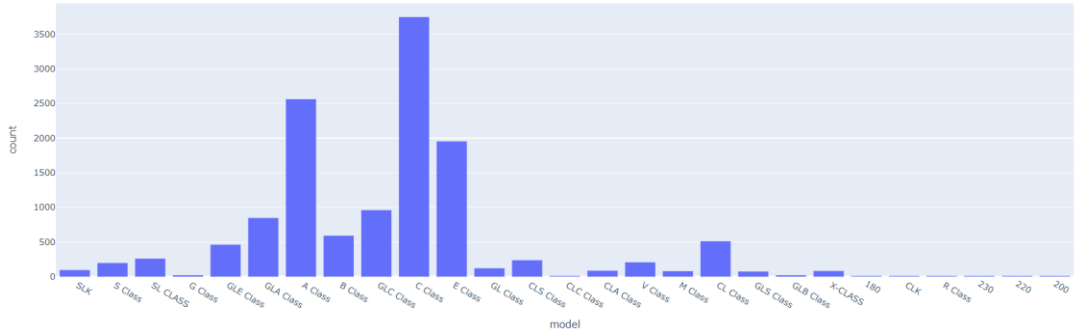


ภาพประกอบที่ 3 การเปรียบเทียบอายุรถยนต์กับราคารถยนต์มือสอง

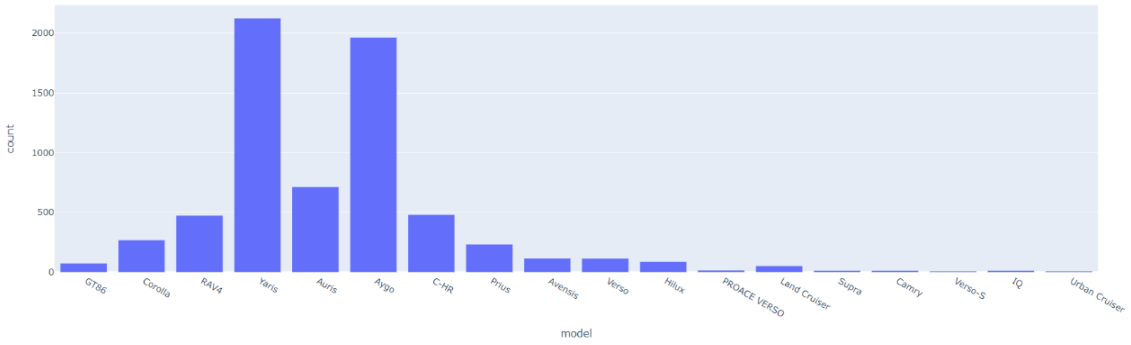
การสำรวจข้อมูลยอดขายรถยนต์รุ่นต่างๆของแบรนด์ยอดนิยมจากชุดข้อมูล โดยนับตามจำนวนคันในชุดข้อมูลโดยได้ยกตัวอย่าง 3 แบรนด์ยอดนิยมเพื่อจะได้เห็นคุณลักษณะต่างๆที่เกี่ยวข้องกับราคารถยนต์มือสองได้ชัดขึ้น จากผลลัพธ์รูปภาพที่4-6 สามารถอธิบายได้ว่า แบรนด์Ford รุ่นยอดนิยมคือ Focus ,Benz รุ่นยอดนิยมคือ C-class และToyota รุ่นยอดนิยมคือ Yaris



รูปภาพประกอบที่ 4 ผลลัพธ์การวิเคราะห์ข้อมูลความนิยมการใช้งานรถยนต์รุ่นต่างๆของแบรนด์Ford

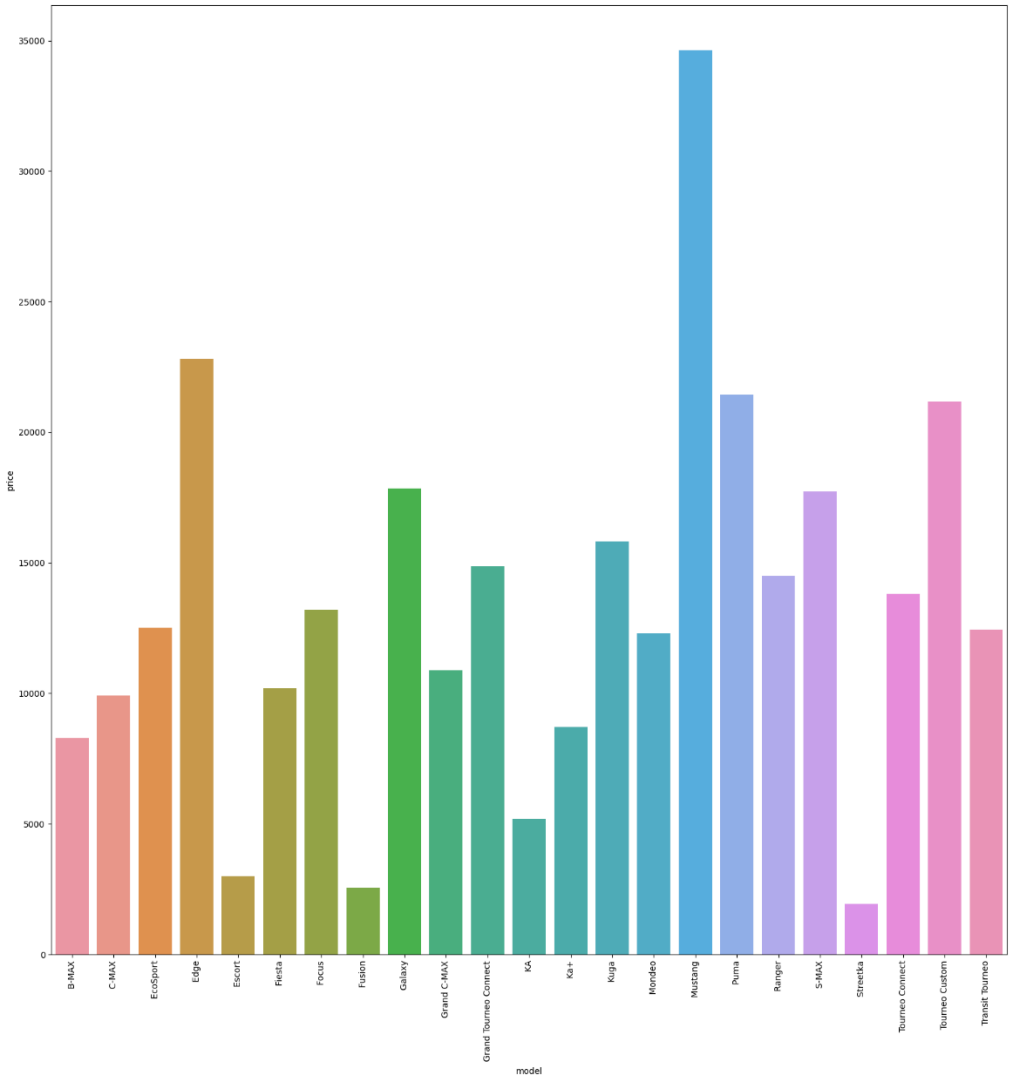


รูปภาพประกอบที่ 5 ผลลัพธ์การวิเคราะห์ข้อมูลความนิยมการใช้งานรถยนต์รุ่นต่างๆของแบรนด์Mercedez-benz

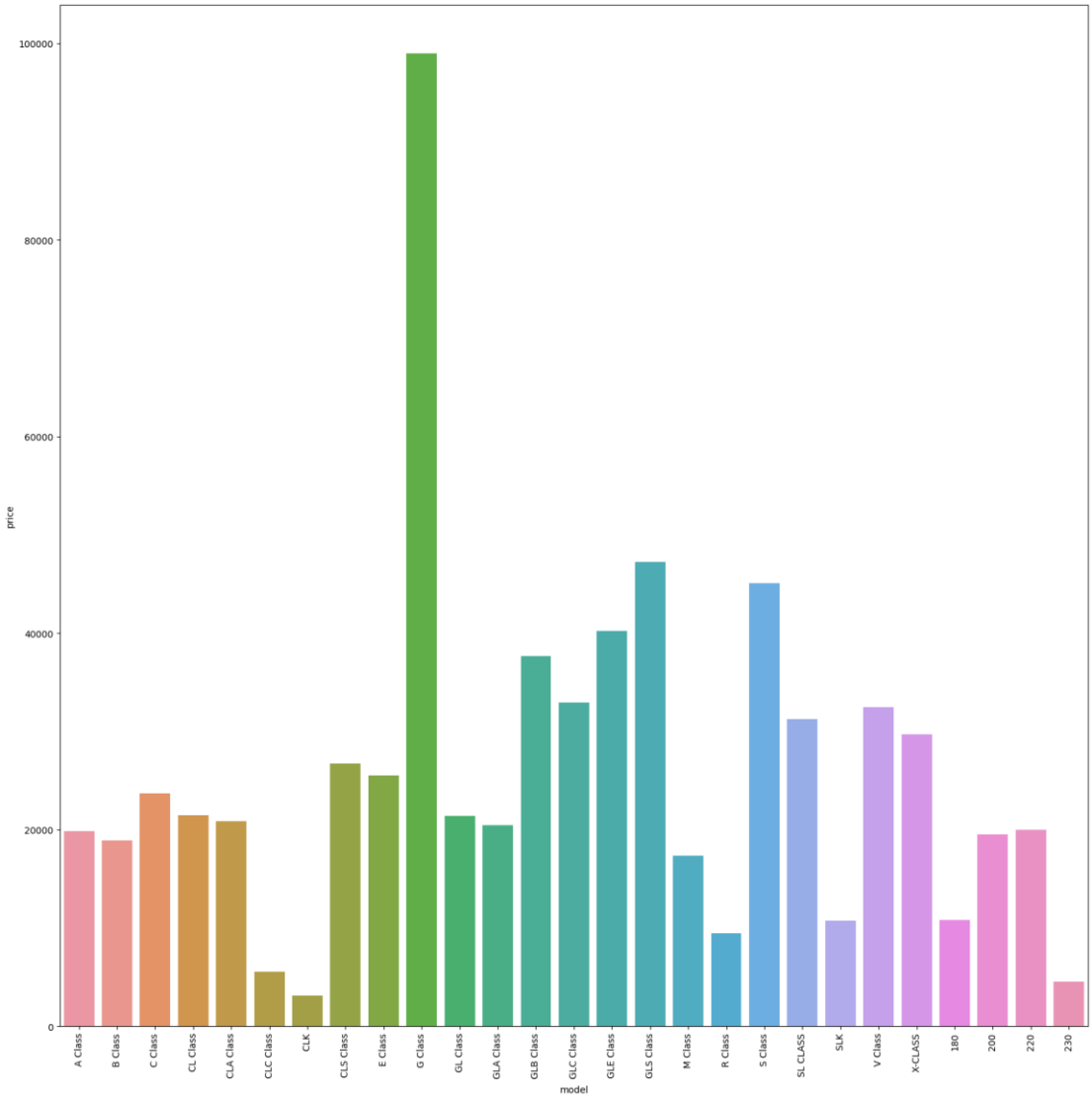


รูปภาพประกอบที่ 6 ผลลัพธ์การวิเคราะห์ข้อมูลความนิยมการใช้งานรถยนต์รุ่นต่างๆของแบรนด์ Toyota

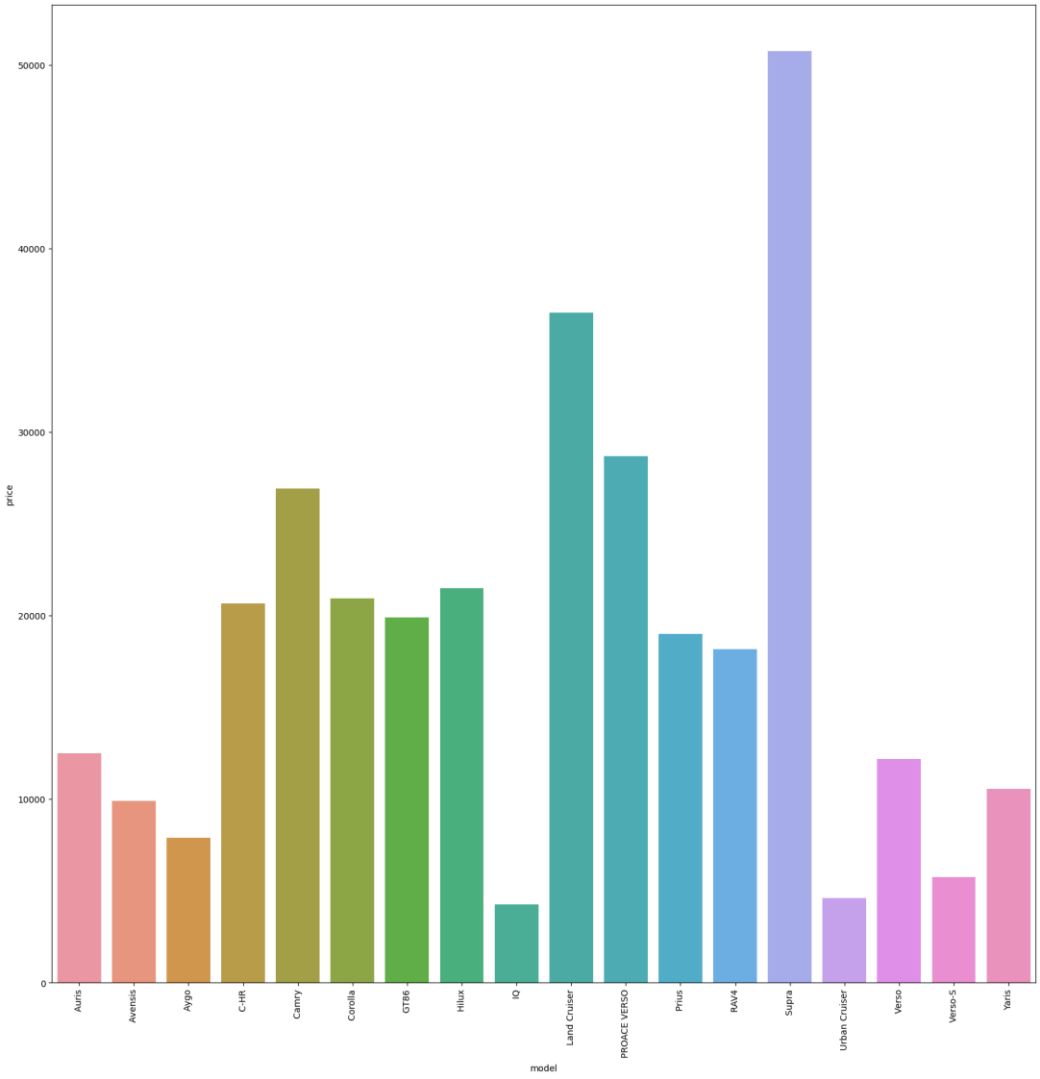
จากรูปภาพที่ 4-6 การเปรียบเทียบรุ่นรถยอดนิยมของ แบรนด์ Ford,BenzและToyota จะทำการวิเคราะห์ต่อเพื่อให้ทราบว่ารุ่นยอดนิยมโดยแยกตามแบรนด์ต่างๆนั้น มีความเกี่ยวข้องกับคุณลักษณะราคามากน้อยเพียงใด จากผลลัพธ์ดังรูปที่ 7-9 สามารถอธิบายได้ว่ารถรุ่นยอดนิยมของแบรนด์ Fordคือ Focus และ Fiesta นั้นจะมีราคาไม่สูงเช่นเดียวกับรุ่น C-Class ของแบรนด์ Benz และรุ่น Yaris ของแบรนด์ Toyota ซึ่งเป็นรถ Eco car ใช้งานทั่วไป ราคาไม่แพงเมื่อเทียบกับรุ่นอื่นๆและประหยัดน้ำมัน



รูปภาพประกอบ 7 ผลลัพธ์การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคาของแบรนด์ Ford

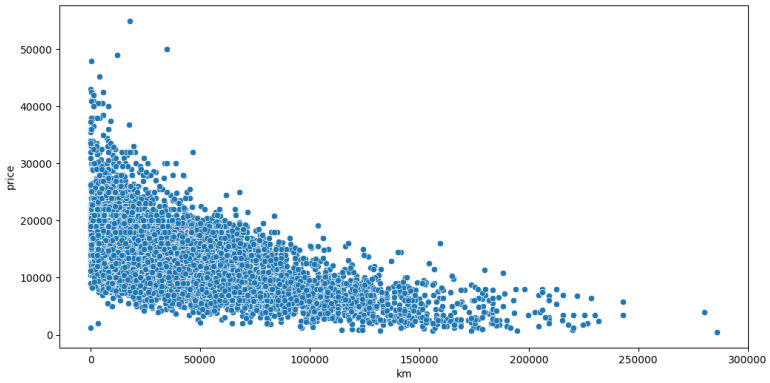


รูปภาพประกอบ 8 ผลลัพธ์การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคาของแบรนด์ Benz

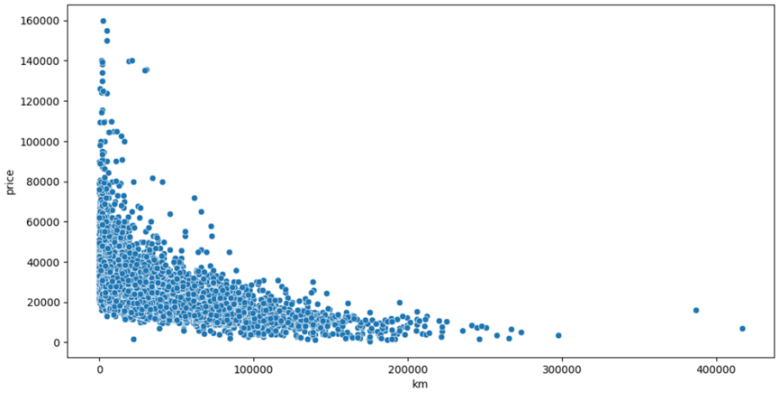


รูปภาพประกอบ 9 ผลลัพธ์การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคาของแบรนด์ Toyota

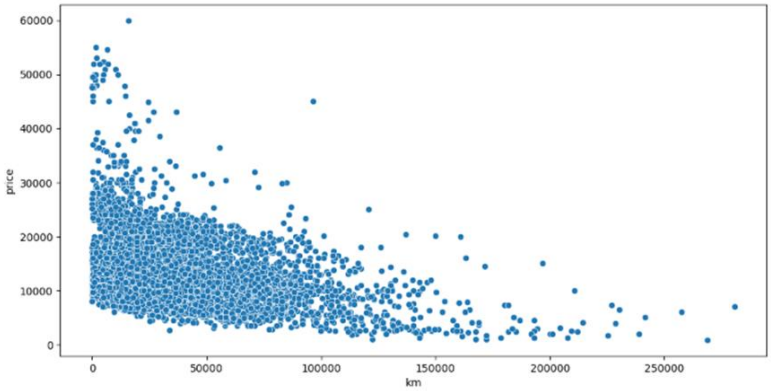
การวิเคราะห์ความสัมพันธ์ระหว่างคุณลักษณะราคากับจำนวนระยะทางที่ใช้จะทำให้ทราบถึงว่าแบรนด์ใดบ้างเมื่อขับแล้วได้เลขไมล์สะสมระยะหนึ่ง หากขายต่อแล้วจะได้ราคาดีโดยทำการเปรียบเทียบแบรนด์ยอดนิยมคือ Ford, Mercedez-benzและToyota ดังรูปภาพที่ 10-12



รูปภาพประกอบที่10ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้แบรนด์ Ford



รูปภาพประกอบที่11ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้แบรนด์ Ford



รูปภาพประกอบที่12ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้แบรนด์ Ford

จากนั้นได้ทำการพล็อตการกระจายของแต่ละฟีเจอร์ต่อเนื่อง และสามารถสรุปได้ว่ารถยนต์ที่ใช้เชื้อเพลิงเบนซินมีราคาสูงกว่าและมีค่า outlier มากขึ้น โดยเฉพาะอย่างยิ่งรถยนต์เบนซินที่มีระบบเกียร์อัตโนมัติและเกียร์ Semi-Auto จะมีราคาแพงที่สุด ในขณะที่รถยนต์เบนซินที่มีระบบเกียร์มือจะถูกกว่าดังรูปภาพที่ 13

รูปภาพ 13 การพลอตกราฟจากฟิวเจอร์ต่างๆเพื่อสรุปภาพรวมของการวิเคราะห์ข้อมูล

ขั้นตอนที่ 4 : การสร้างแบบจำลอง

ขั้นตอนนี้เป็นขั้นตอนของการนำข้อมูลที่จัดเตรียมไว้มาวิเคราะห์ด้วยเทคนิคต่างๆ ตามความเหมาะสม โดยมากจะเป็นการวิเคราะห์ด้วยเทคนิคมากกว่าหนึ่งแบบเพื่อประเมินหาเทคนิคที่ให้ค่าการวิเคราะห์ที่เป็น ไปตามวัตถุประสงค์และมีประสิทธิภาพมากที่สุด จากการผ่านขั้นตอนการเตรียมข้อมูลจะได้ชุดข้อมูลจำนวน 100,000 แถวและ 10 คอลัมน์ จากนั้นทำการแบ่งข้อมูล(split data)ออกเป็น 2 ส่วนคือ Training Set คือชุดข้อมูลสำหรับสอนแบบจำลอง และ Test setคือชุดข้อมูลสำหรับทดสอบ โดย Train set 70: Test set 30 โดย Algorithm ที่เลือกใช้คือ Linear Regression, Random Forest, Decision Tree, LASSOและ RIDGE

ขั้นตอนที่ 5 : การประเมินแบบจำลอง

สำหรับการประเมินประสิทธิภาพแบบจำลองผลตัวแปรที่เป็นข้อมูลเชิงปริมาณ ใช้สูตรในการวัดคือMAE, MAPE,MSEและ R-Square

MAE (Mean Absolute Error) เป็นค่าเฉลี่ยของความคลาดเคลื่อนในการทำนายของโมเดล คือค่าสัมบูรณ์ของความแตกต่างระหว่างค่าที่ทำนายได้กับค่าจริง โดยจะมีหน่วยเดียวกับข้อมูลที่ใช้ในการทำนาย เช่น หน่วยเงินบาท หรือ หน่วยกิโลกรัม เป็นต้น สูตรคำนวณ MAE คือ **MAE = (1/n) * Σ|i=1 to n|(|yi - xi|)**

โดยที่ yi คือค่าจริง (actual values) และ xi คือการทำนาย (predicted values)

MAPE (Mean Absolute Percentage Error) เป็นการคำนวณความคลาดเคลื่อนของโมเดลในรูปแบบเปอร์เซ็นต์ คือค่าเฉลี่ยของอัตราส่วนระหว่างค่าที่ทำนายได้กับค่าจริง โดยนำค่าความคลาดเคลื่อนมาหารด้วยค่าจริงแล้วคูณ 100%

สูตรคำนวณ MAPE คือ MAPE = (1/n) * Σ|i=1 to n|(|(yi - xi)/yi| * 100%)[4]

โดยที่ yi คือค่าจริง (actual values) และ xi คือการทำนาย (predicted values)

MSE (Mean Squared Error) เป็นค่าเฉลี่ยของความคลาดเคลื่อนในการทำนายของโมเดลโดยยกกำลังสองของค่าผลต่างระหว่างค่าที่ทำนายได้กับค่าจริง โดยจะมีหน่วยเป็นหน่วยของข้อมูลที่ใช้ในการทำนาย ยกเว้นถ้าค่าที่ใช้ในการทำนายมีหน่วยไม่ใช่ตัวเลข เช่น สีของสินค้า หรือ ประเภทของสินค้า เป็นต้น หากค่า MSE น้อยลงแสดงว่าโมเดลทำนายได้ดีขึ้น

สูตรของ MSE คือ MSE = (1/n) * $\Sigma(y_i - \hat{y}_i)^2$ [5]

เมื่อ

n = จำนวนข้อมูลในชุดข้อมูล

y_i = ค่าเป้าหมายของข้อมูลคือค่าที่จริงหรือค่าที่ต้องการทำนาย

$\hat{y}_i$ = ค่าที่โมเดลทำนายขึ้นมา

R-Square (Coefficient of Determination) เป็นค่าวัดประสิทธิภาพของโมเดล โดยค่า R-Square จะบอกว่า โมเดลสามารถอธิบายความแปรปรวนของข้อมูลได้เท่าใด โดยค่า R-Square จะอยู่ในช่วง 0 ถึง 1 โดยค่าที่ใกล้เคียงกับ 1 จะแสดงว่าโมเดลสามารถอธิบายความแปรปรวนของข้อมูลได้ดีมาก สูตรของ R-Square คือ

R-Squared Score = 1 - (SSR/SST)

เมื่อSSR = ผลรวมของค่า residuals ยกกำลังสอง หรือ $(y_i - \hat{y}_i)^2$ หรือ ผลรวมของความคลาดเคลื่อนระหว่างการทำนายและค่าเฉลี่ยของค่า

SST = ผลรวมของค่าตัวแปรตาม $(y_i - \bar{y})^2$ หรือผลรวมของค่าความแปรปรวนทั้งหมด (Total Sum of Squares)

y_i = ค่าเป้าหมายของข้อมูลคือค่าที่จริงหรือค่าที่ต้องการทำนาย

$\hat{y}_i$ = ค่าที่โมเดลทำนายขึ้นมา

$\bar{y}$ = ค่าเฉลี่ยของค่าตัวแปรตาม

ผลการวิจัยและอภิปรายผลการวิจัย

ตารางที่ 1 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์จากรายการราคารถยตร์มือสองทั้งหมด 9 แบรนด์

Model	MAE	MAPE	MSE	R-square
LinearRegressionModel	0.13	0.95	0.03	0.96
DecisionTreeRegression	1417.302	0.090	6752453.883	0.96
Random forest	1138.238	0.073	3496491.842	0.96
RIDGE	2254.940	0.180	13459786.977	0.86
LASSO	2309.359	0.186	13761927.624	0.85

จากผลการทดสอบโมเดลที่แตกต่างกันตามตารางที่ 1 พบว่าโมเดล Random Forest ได้ผลการทดสอบที่ดีที่สุดโดยมีค่า MAE เท่ากับ 1138.238, MAPE เท่ากับ 0.073, MSE เท่ากับ 3496491.842, และ R-Squared score เท่ากับ 0.96 ตามมาด้วย Linear Regression Model ที่ให้ผลการทดสอบที่สองดีที่สุดโดยมีค่า MAE เท่ากับ 0.13, MAPE เท่ากับ 0.95, MSE เท่ากับ 0.03, และ R-Squared score เท่ากับ 0.96 แต่โมเดลอื่น ๆ ที่ทดสอบอย่าง Decision Tree Regression, Ridge Regression, และ Lasso มีผลการทดสอบที่แย่กว่า โดยเฉพาะ Lasso ที่ให้ผลการทดสอบที่แย่ที่สุด ดังนั้นในการทำนายราคารถยนต์มือสอง ควรใช้โมเดล Random Forest หรือ Linear Regression Model เพื่อให้ได้ผลการทำนายที่มีความแม่นยำสูงที่สุด

ตารางที่ 2 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์แยกแต่ละแบรนด์จากชุดข้อมูลรายการรถยนต์มือสองของแบรนด์ Ford

Model	MAE	MAPE	MSE	R-square
LinearRegressionModel	0.203872	1.452	0.082	0.91
DecisionTreeRegression	1050.322	0.089	1050.322	0.92
Random forest	863.571	0.074	863.571	0.92
RIDGE	1356.146	0.136	1356.146	0.83
LASSO	1361.892	0.138	1361.892	0.82

จากผลการทดสอบโมเดลตามตารางที่ 2 พบว่า Random Forest และ Decision Tree Regression มีผลการทดสอบที่ดีกว่าทั้ง Ridge และ Lasso Regression โดย Random Forest และ Decision Tree Regression มีค่า MAE, MAPE, MSE และ R-Squared score ที่สูงกว่าและใกล้เคียงกัน ซึ่งแสดงให้เห็นว่าโมเดลสองตัวนี้สามารถทำนายราคารถมือสองในตลาดอย่างแม่นยำได้เป็นอย่างดี ในขณะที่โมเดล Ridge และ Lasso Regression มีผลการทดสอบที่แย่กว่า แสดงให้เห็นว่าโมเดลสองตัวนี้ไม่เหมาะสมในการทำนายราคาของรถมือสองในตลาด สำหรับ Linear Regression มีผลการทดสอบที่ดีเช่นกัน แต่ค่า MAE, MAPE, MSE ต่างจาก Random Forest และ Decision Tree Regression อยู่ในระดับที่ต่ำกว่าเล็กน้อย

ตารางที่ 3 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์แยกแต่ละแบรนด์จากชุดข้อมูลรายการรถยนต์มือสองของแบรนด์ Mercedes-Benz

Model	MAE	MAPE	MSE	R-square
LinearRegressionModel	0.165594	0.79360	0.05485	0.94
DecisionTreeRegression	2039.870	0.087	2039.870	0.94
Random forest	1576.053	0.066	1576.053	0.94
RIDGE	3490.593	0.161	3490.593	0.77
LASSO	3492.440	0.161	3492.440	0.76

จากผลลัพธ์การวิเคราะห์แบบจำลองต่างๆตามตารางที่ 3 ในการทำนายราคารถยนต์มือสองของแบรนด์ Mercedes-Benz พบว่าโมเดล Random Forest ให้ผลลัพธ์ที่ดีที่สุดโดยมีค่า MAE เท่ากับ 1576.053, MAPE เท่ากับ 0.066, MSE เท่ากับ 1576.053, และ R-square เท่ากับ 0.94 ซึ่งแสดงว่าโมเดลสามารถทำนายราคาได้อย่างแม่นยำสูงสุดในกลุ่มโมเดลที่ได้ทดสอบ ตามมาด้วย Linear Regression Model ที่ให้ผลลัพธ์ดีเช่นกัน โดยมีค่า MAE เท่ากับ 0.165594, MAPE เท่ากับ 0.79360, MSE เท่ากับ 0.05485, และ R-square เท่ากับ 0.94 และ Decision Tree Regression ที่ให้ผลลัพธ์ค่อนข้างเป็นที่น่าพอใจด้วยค่า MAE เท่ากับ 2039.870, MAPE เท่ากับ 0.087, MSE เท่ากับ 2039.870, และ R-square เท่ากับ 0.94 ในทางกลับกัน Ridge Regression และ Lasso Regression ให้ผลลัพธ์ที่แย่ลง ด้วยค่า MAE และ MAPE เป็นจำนวนมาก และค่า R-square ต่ำกว่า 0.8

ตารางที่ 4 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์แยกแต่ละแบรนด์จากชุดข้อมูลรายการรถยนต์มือสองของแบรนด์ Toyota

Model	MAE	MAPE	MSE	R-square
LinearRegressionModel	0.145	1.287	0.041	0.95
DecisionTreeRegression	973.042	0.089	973.042	0.96
Random forest	816.251	0.072	816.251	0.96
RIDGE	1132.666	0.122	1132.666	0.92
LASSO	1131.593	0.122	1131.593	0.92

จากผลลัพธ์ที่ได้ตามตารางที่3 จะเห็นว่า แบบจำลอง Decision Tree Regression และ Random Forest ให้ผลลัพธ์ที่ดีกว่าแบบจำลองอื่นๆ โดยมีค่า MAE, MAPE, และ MSE ต่ำสุด และค่า R-square สูงสุด ในขณะที่แบบจำลอง Linear Regression Model, Ridge Regression และ Lasso Regression ให้ผลลัพธ์ที่ค่อนข้างใกล้เคียงกัน โดยแบบจำลอง Linear Regression Model ให้ผลลัพธ์ที่ดีกว่าเล็กน้อย และแบบจำลอง Ridge Regression และ Lasso Regression ให้ผลลัพธ์ที่แย่กว่า แต่ก็ยังคงให้ค่า R-square ที่สูงอยู่ที่ระดับ 0.92

สรุปผลการวิจัย

รถยนต์มือสอง คือ รถยนต์ที่ผ่านการใช้งานมาแล้วจากบุคคลใดบุคคลหนึ่ง อาจมีการใช้งานมาก น้อยหรือหนัก เบา แตกต่างกันออกไป รถยนต์มีส่วนสำคัญในการดำรงชีวิตของมนุษย์เป็นอย่างมาก ตอบสนองความสะดวกสบายในการเดินทางในชีวิตประจำวันไม่ว่าจะเป็นการเดินทางไปทำงาน ประกอบธุรกิจ หรือท่องเที่ยวเป็นต้น เนื่องด้วยเศรษฐกิจในปัจจุบันรถยนต์มือสองหรือรถยนต์ใช้แล้ว เข้ามามีบทบาททดแทนรถยนต์มือหนึ่ง จุดเด่นสำคัญอยู่ที่เรื่องของราคาที่ถูกกว่ารถมือหนึ่งซึ่งนับเป็นอีกทางเลือกหนึ่งที่จะหาซื้อรถยนต์ไว้ในครอบครองได้ในราคาที่เพียงพอกับรายได้ ปัจจุบันรถมือสองมีจำหน่ายอย่างแพร่หลายทั้งจากบริษัท เต็นท์รถ หรือบุคคลทั่วไปที่ต้องการขายรถ ทำให้ผู้ที่กำลังจะซื้อรถมือสองมีตัวเลือกในการเปรียบเทียบเรื่องของราคา รวมไป

ถึงคุณภาพของรุ่นรถที่สนใจได้ แม้จะเป็นรถมือสอง แต่สามารถนำไปขายต่อได้หากต้องการรถยนต์รุ่นที่ใหม่กว่า อีกทั้งรถมือสองโอกาสผ่านการอนุมัติค่อนข้างง่าย เนื่องจากวงเงินในการกู้ซื้อรถไม่สูงมากเท่ารถยนต์มือหนึ่ง ผู้วิจัยได้ใช้ชุดข้อมูลจากKaggle โดยใช้ชุดข้อมูลถยนต์มือสองในตลาดรถยนต์ของสหราชอาณาจักร (UK car market)เพื่อทำนายราคาของรถยนต์มือสองด้วยเทคนิคการเรียนรู้ด้วยเครื่อง และทดสอบแบบจำลองใช้เทคนิค การถดถอยต้นไม้การตัดสินใจ Decision Tree Regression, การถดถอยแบบเชิงเส้น Linear Regression,การถดถอยแบบ Ridge, ลาสโซ่ Lasso เป็นต้น จากการทดลองครั้งนี้แบบจำลองทำนายที่พัฒนามีประสิทธิภาพสูงสุดคือ Random Forest ด้วยค่า MAE 1139.238, MAPE 0.073, MSE 3496491.842, และ R-Squared score 0.96 แบบจำลองต่อไปคือ Linear Regression ด้วยค่า MAE 0.13, MAPE 0.95, MSE 0.030, และ R-Squared score 0.96 ส่วนแบบจำลองอื่น ๆ ที่ทดสอบนั้นมีประสิทธิภาพต่ำกว่า โดย Decision Tree Regression มีค่า MAE 1417.302, MAPE 0.90, MSE 6752453.883, และ R-Squared score 0.96, Ridge Regression มีค่า MAE 2254.940, MAPE 0.180, MSE 13459786.977, และ R-Squared score 0.86, และ Lasso มีค่า MAE 2309.359, MAPE 0.186, MSE 13761927.624, และ R-Squared score 0.85 และการวิเคราะห์แบบจำลองต่างๆโดยวิเคราะห์แยกตามชุดข้อมูลแต่ละแบรนด์ของรายการรถยนต์มือสองจาก 3 แบรนด์ดังนี้ Ford, Mercedes-Benz, Toyota พบว่า DecisionTreeRegression และ Random forest ให้ผลลัพธ์ที่ดีที่สุดในการทำนายราคาโดยในแต่ละแบรนด์อาจจะต่างกันไป ส่วน LinearRegressionModel, RIDGE, และ LASSO ให้ผลลัพธ์ที่น้อยกว่า DecisionTreeRegression และ Random forest ในทุกแบรนด์ ดังนั้นผลลัพธ์การเปรียบเทียบโมเดลต่างๆในการทำนายราคารถยนต์มือสอง จะเห็นได้ว่าโมเดล Random Forest ให้ผลลัพธ์ที่ดีที่สุด โดยมีค่า MAE ที่ต่ำที่สุดและค่า R-square ที่สูงที่สุดในทุกตาราง โมเดล Random Forest จึงเป็นโมเดลที่ดีที่สุดในการทำนายราคารถยนต์มือสองในชุดข้อมูลที่ใช้ในการวิจัยนี้

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนมาตลอดจากผศ.ดร.จันตรี ผลประเสริฐ ที่ได้ให้คำปรึกษาและคำแนะนำในการวิจัยอย่างเต็มที่ ทำให้งานวิจัยเป็นไปได้อย่างมีประสิทธิภาพและคุณภาพ ขอบคุณอย่างสูงที่เสริมสร้างความรู้ความเข้าใจและประสบความสำเร็จในการวิจัยและขอขอบคุณคณะอาจารย์ภาควิชาวิทยาศาสตร์สาขาวิทยาการข้อมูลมหาวิทยาลัยศรีนครินทรวิโรฒ ที่ให้การสนุบสนุนอย่างเต็มที่ ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] C. Jin, "Price Prediction of Used Cars Using Machine Learning," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021, pp. 223-230, doi: 10.1109/ICESIT53460.2021.9696839.

[2] [online] Available: https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes.

[3] https://www.marketingoops.com/exclusive/trending-exclusive/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers/

[4] https://www.statology.org/mape-excel/

[5] https://hackernoon.com/my-notes-on-mae-vs-mse-error-metrics