การทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง

PREDICTION OF USED CAR PRICES USING MACHINE LEARNING

ชัรวรรค์ ปูเตะ

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2565

การทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง

ซัรวรรค์ ปูเตะ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2565

PREDICTION OF USED CAR PRICES USING MACHINE LEARNING

SARWAN PUTEH

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2022

สารนิพนธ์

เรื่อง

การทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง

ของ

ชัรวรรค์ ปูเตะ

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... ประธาน

(ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ

(อาจารย์ ดร.วีระ สอิ้ง)

ชื่อเรื่อง	การทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง
ผู้วิจัย	ชัรวรรค์ ปูเตะ
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2565
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. จันตรี ผลประเสริฐ

วัตถุประสงค์ของงานวิจัยนี้คือการสร้างโมเดลทำนายที่สามารถทำนายราคาขายรถยนต์มือสองได้อย่างแม่นยำ โดยใช้เทคนิคการเรียนรู้ของเครื่องและการศึกษาปัจจัยที่เกี่ยวข้องกับคุณลักษณะต่างๆ ของรถยนต์ เช่น รุ่นรถยนต์, ปีที่ผลิต, ขนาดเครื่องยนต์, ประเภทเชื้อเพลิง, ภาษีทางถนน, ประเภทของเกียร์ และระยะทางที่รถยนต์ได้เคลื่อนที่ไปจนถึงปัจจุบัน ชุดข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลรถยนต์มือสองที่จำหน่ายในตลาดสหราชอาณาจักร ข้อมูลถูกวิเคราะห์และเปรียบเทียบโมเดลต่างๆ ที่รวมถึง Decision Tree Regression, Linear Regression, Ridge Regression, Lasso Regression, และ Random Forest ชุดข้อมูลถูกแบ่งเป็นชุดการฝึกและการทดสอบด้วยอัตราส่วน 80/20 เพื่อประเมินประสิทธิภาพของโมเดล ในการทดลองนี้ โมเดล Random Forest ให้ผลลัพธ์ที่ดีที่สุดด้วยค่าความผิดพลาดเฉลี่ยสัมบูรณ์ (Mean Absolute Error) = £942, ค่าความผิดพลาดเปอร์เซ็นต์สัมบูรณ์ (Mean Absolute Percentage Error) = 6%, และค่าสัมประสิทธิ์ (coefficient of determination) = 0.959 โมเดลถัดมาคือ Decision Tree Regression ที่ให้ผลลัพธ์ที่ดีเป็นอันดับสองด้วยค่าความผิดพลาดเฉลี่ยสัมบูรณ์ = £1,074, ค่าความผิดพลาดเปอร์เซ็นต์สัมบูรณ์ = 7%, และค่าสัมประสิทธิ์ = 0.938 ตามลำดับ โมเดล Linear Regression มีค่าความผิดพลาดเฉลี่ยสัมบูรณ์ = £1,824, ค่าความผิดพลาดเปอร์เซ็นต์สัมบูรณ์ = 14%, และค่าสัมประสิทธิ์ = 0.877 โมเดล Ridge มีผลลัพธ์ที่คล้ายกับ Linear Regression และโมเดลที่ให้ผลลัพธ์ที่น้อยที่สุดคือ Lasso ด้วยค่าความผิดพลาดเฉลี่ยสัมบูรณ์ = £1,848, ค่าความผิดพลาดเปอร์เซ็นต์สัมบูรณ์ = 15%, และค่าสัมประสิทธิ์ = 0.875 งานวิจัยนี้มีประโยชน์ต่อผู้ขาย, ผู้ซื้อ และผู้ผลิตรถยนต์ในตลาดรถยนต์มือสองเนื่องจากมีการทำนายราคาที่แม่นยำ

คำสำคัญ : ราคา, รถยนต์มือสอง, เทคนิคการเรียนรู้ของเครื่อง, การเรียนรู้ของเครื่องแบบสุ่ม

Title	PREDICTION OF USED CAR PRICES USING MACHINE LEARNING
Author	SARWAN PUTEH
Degree	MASTER OF SCIENCE
Academic Year	2022
Thesis Advisor	Assistant Professor Chantri Polprasert , Ph.D.

The objective of this research is to create a predictive model that accurately predicts the prices of used cars using machine learning techniques and studying the factors that affect the prices of used cars from a dataset of used car listings in the UK market. This research can be beneficial to sellers, buyers, and car manufacturers in the used car market by providing accurate price predictions. The research develops predictive models for multiple brands and models of cars using machine learning techniques, considering various car characteristics such as car model, manufacturing year, engine size, fuel type, road tax, gearbox type, and mileage. The data is analyzed and various models are compared and evaluated, including Decision Tree Regression, Linear Regression, Ridge Regression, Lasso Regression, and Random Forest. The dataset is divided into training and testing sets with an 80/20 split for model evaluation. In this experiment, Random Forest performs the best with the following performance measurements: MAE = £942, MAPE = 6%, and R^2 = 0.959. The next best model is Decision Tree Regression with MAE = £1,074, MAPE = 7%, and R^2 = 0.938. Linear Regression follows with MAE = £1,824, MAPE = 14% and R^2 = 0.877. The Ridge model has similar results to Linear Regression, and the least performing model is Lasso with MAE = £1,848, MAPE = 15%, and R^2 = 0.875.

Keyword : Price, Used car, Machine Learning, Random Forest

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จาก ผศ.ดร.จันตรี ผลประเสริฐ อาจารย์ที่ปรึกษาที่ได้สละเวลาอันมีค่าและชี้แนะแนวทางการทำสารนิพนธ์ ให้คำปรึกษา รวมทั้งตรวจสอบแก้ไขข้อบกพร่องต่างๆ อย่างดียิ่งด้วยความเอาใจใส่จนสำเร็จได้ด้วยดีและผู้วิจัยต้องขอบพระคุณ อ.ดร.วีระ สอิ้ง สำหรับคำแนะนำที่มีประโยชน์จนช่วยให้ผลงานวิจัยนี้ได้ผลลัพธ์ที่ดีกว่าที่คาดไว้ ขอขอบพระคุณ ดร.สุทธิพงศ์ ธัชยพงษ์ ผู้ทรงคุณวุฒิจากภายนอกที่ได้เสียสละเวลาอันมีค่า ให้เกียรติมาเป็นประธานสอบปากเปล่าและให้คำแนะนำเติมเต็มให้ผลลัพธ์มีความสมบูรณ์มากยิ่งขึ้น ผู้วิจัยตระหนักถึง ความตั้งใจจริงและความทุ่มเทของอาจารย์ที่ปรึกษาและความช่วยเหลือจาก อ.ดร.วีระ สอิ้ง ผู้วิจัยกราบขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณคณะกรรมการสอบสารนิพนธ์ ได้เมตตาให้ความรู้ แนวคิด และข้อเสนอแนะในการปรับปรุงแก้ไขสารนิพนธ์ฉบับนี้ให้ถูกต้องมีความสมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์และกรรมการบริหารหลักสูตรสาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ให้กับผู้วิจัย รวมถึงบทความในเรื่องที่เกี่ยวข้องกับการทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่องตลอดจนได้อ้างอิงถึงผลงานทางวิชาการซึ่งเป็นประโยชน์ในการเรียบเรียงสารนิพนธ์นี้เป็นอย่างยิ่ง ผู้วิจัยหวังว่าสารนิพนธ์ฉบับนี้คงเป็นประโยชน์ต่อผู้ที่สนใจศึกษาต่อไป ส่วนข้อบกพร่องทั้งหลายผู้วิจัยขอน้อมรับและกราบขออภัยไว้ ณ ที่นี้

ชัรวรรค์ ปูเตะ

สารบัญ

หน้า

บทที่1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

รถยนต์มือสอง คือ ยานพาหนะประเภทรถยนต์ที่เคยถูกใช้งานและถูกขายต่อโดยโอนกรรมสิทธิ์ให้กับผู้ที่ไม่ใช่เจ้าของเดิมของรถยนต์นั้น รถยนต์มือสองอาจถูกขายโดยเจ้าของเดิมโดยตรงหรือผ่านตัวแทนขายรถยนต์มือสอง และรถยนต์มือสองสามารถเป็นที่นิยมในตลาดรถยนต์เนื่องจากมีราคาที่ถูกกว่ารถยนต์ใหม่ นอกจากนี้ยังมีความหลากหลายในรุ่นรถยนต์และสภาพรถที่ถูกใช้งานตามระยะทางและระยะเวลาต่างกันออกไป ทำให้ผู้ซื้อสามารถเลือกได้ตามความต้องการและงบประมาณที่มีอยู่ ทั้งนี้ ส่วนที่สำคัญที่สุดในการพิจารณาเลือกซื้อรถยนต์มือสองนั้นคือราคาที่สมเหตุสมผลและสภาพของรถยนต์ ผู้ซื้อควรตรวจสอบราคาตลาดของรถยนต์รุ่นที่สนใจและเปรียบเทียบกับราคาที่ขายอยู่ในตลาดเพื่อให้ได้ราคาที่เหมาะสม นอกจากประเด็นดังกล่าวผู้ซื้อควรตรวจสอบ อายุของรถยนต์ ประวัติการใช้งาน การดูแลรักษา และประวัติการชนหรืออุบัติเหตุอื่น ๆ เพื่อให้มั่นใจว่ารถยนต์ที่ซื้อมือสองมีความพร้อมใช้งานและไม่มีปัญหาที่อาจเกิดขึ้นในอนาคต รวมทั้งรถยนต์ยังมีประโยชน์ในด้านความสะดวกสบายในการใช้ชีวิต ซึ่งถือเป็นสิ่งอำนวยสะดวกสำหรับการเดินทางไม่ว่าจะเป็นการเดินทางไปทำงาน ใช้งานกับธุรกิจส่วนตัว และท่องเที่ยว เป็นต้น จุดเด่นสำคัญที่ทำให้รถยนต์มือสองเข้ามามีบทบาททดแทนรถยนต์มือหนึ่งได้นั้นคือเรื่องของราคาที่ถูกกว่ารถมือหนึ่งซึ่งนับเป็นอีกทางเลือกหนึ่งที่จะหาซื้อรถยนต์ไว้ใช้งานในราคาที่ถูกกว่ารถยนต์มือหนึ่งซึ่งสอดคล้องกับรายได้ โดยปัจจุบันรถยนต์มือสองได้รับความนิยมอย่างแพร่หลายและมีจำหน่ายทั้งจากบริษัทเต็นท์รถ หรือบุคคลทั่วไปที่ต้องการขายรถของตัวเอง ทำให้ผู้ที่กำลังจะซื้อรถมือสองมีตัวเลือกในการเปรียบเทียบเรื่องของราคา รวมไปถึงคุณภาพของรุ่นรถที่สนใจได้ แม้จะเป็นรถยนต์ที่ถูกใช้งานมาแล้วแต่ก็ยังสามารถนำไปขายต่อได้หากต้องการรถยนต์รุ่นที่ใหม่กว่า อีกทั้งในประเด็นทางด้านสินเชื่อ การกู้ซื้อรถยนต์มือสองมีโอกาสได้รับการอนุมัติสินเชื่อที่ง่ายกว่าเพราะว่าวงเงินในการกู้ซื้อรถยนต์มีมูลค่าไม่สูงเท่ารถยนต์มือหนึ่ง การซื้อรถยนต์มือสองยังมีข้อดีอื่น ๆ ที่สามารถพิจารณาได้คือ 1.คุณภาพรถยนต์: ในบางครั้งอาจพบว่ารถยนต์มือสองที่มีราคาถูกกว่ารถยนต์ใหม่ยังมีคุณภาพที่ดีอยู่ รถยนต์แต่ละคันมีประวัติการใช้งานที่แตกต่างกันไป บางคันอาจมีการดูแลรักษาที่ดี รักษาระยะทางและตรวจเช็คอย่างสม่ำเสมอ ทำให้ผู้ซื้อสามารถหารถยนต์มือสองที่มีคุณภาพดีและยังคงให้ประสบการณ์ขับขี่ที่ดีได้, 2.ค่าเสื่อมราคาต่ำ: รถยนต์ใหม่มักมีการเสื่อมราคาอย่างรวดเร็วในช่วงแรกหลังจากการซื้อ เมื่อรถออกจาก

ตัวแทนจำหน่ายแล้ว ผู้ซื้อจะเสียเงินมากกว่าราคาซื้อเมื่อรถออกจากสต็อก ในขณะที่รถยนต์มือสองอาจมีการเสื่อมราคาช้าลงและมีราคาสูงเมื่อขายต่อในอนาคต นั่นหมายความว่าผู้ซื้ออาจสามารถขายรถยนต์มือสองในอนาคตได้ในราคาที่สูงกว่าที่ซื้อมา, 3.ตรวจสอบประวัติการใช้งาน: การซื้อรถยนต์มือสองให้โอกาสตรวจสอบประวัติการใช้งานของรถยนต์ที่สนใจ เราสามารถตรวจสอบจำนวนกิโลเมตรที่รถยนต์เคยใช้งานไปและประวัติการซ่อมบำรุง อาจมีบันทึกการซ่อมแซมที่เราสามารถตรวจสอบได้ว่ามีปัญหาอะไรบ้างหรือไม่ รวมถึงประวัติการชนหรืออุบัติเหตุอื่น ๆ ที่อาจมีผลกระทบต่อสภาพของรถยนต์ในอนาคต, 4.ระยะเวลาส่งมอบรวดเร็ว: เมื่อเราตัดสินใจซื้อรถยนต์ใหม่ บางครั้งอาจต้องรอให้มีการผลิตและจัดส่งรถ ซึ่งอาจใช้ระยะเวลานาน ในทางกลับกัน การซื้อรถยนต์มือสองอาจช่วยลดระยะเวลานั้นลง โดยเราสามารถซื้อรถยนต์มือสองและรับมอบในเวลาที่สั้นกว่า, 5.การประหยัดเงินในการประกันภัย: รถยนต์มือสองมักมีค่าประกันภัยที่ถูกกว่ารถยนต์ใหม่ ราคาประกันภัยของรถยนต์มือสองจะต่ำลงเนื่องจากมูลค่าของรถลดลงเมื่อเทียบกับรถยนต์ใหม่ ซึ่งอาจช่วยประหยัดเงินในการจ่ายค่าประกันภัยได้, 6.ตัวเลือกทางการเงินที่หลากหลาย: หากเราต้องการที่จะสั่งซื้อรถยนต์ใหม่แต่ไม่มีเงินสดเพียงพอ การซื้อรถยนต์มือสองอาจเป็นตัวเลือกทางการเงินที่ดีกว่า เราสามารถเลือกใช้วิธีการเงินที่หลากหลาย เช่น การจัดไฟแนนซ์หรือสินเชื่อรถยนต์เพื่อส่งเสริมการซื้อรถยนต์มือสองได้, 7.ราคาอะไหล่และการซ่อมบำรุงที่คุ้มค่า: รถยนต์มือสองมีราคาอะไหล่ที่ถูกกว่ารถยนต์ใหม่ อะไหล่สำหรับรถยนต์มือสองอาจมีมากในตลาดและมีราคาที่เป็นธรรม นอกจากนี้ การซ่อมบำรุงรถยนต์มือสองยังมีค่าที่คุ้มค่ากว่ารถยนต์ใหม่ เนื่องจากมีความเรียบง่ายและง่ายต่อการตรวจสอบและซ่อมแซม, 8.รถยนต์มือสองเป็นทางเลือกสำหรับรถยนต์คลาสสิค: หากเราสนใจรถยนต์คลาสสิคหรือรุ่นที่ไม่ผลิตใหม่แล้ว การซื้อรถยนต์มือสองอาจเป็นทางเลือกที่ดีกว่า เนื่องจากเราสามารถหารถยนต์คลาสสิคในสภาพที่ดีและใช้งานได้ตามปกติได้, 9.รูปแบบการชำระเงินที่หลากหลาย: การซื้อรถยนต์มือสองอาจมีรูปแบบการชำระเงินที่หลากหลาย เช่น การชำระเงินด้วยเงินสด การจัดไฟแนนซ์ หรือการแลกเปลี่ยนรถยนต์เก่าของเรากับรถยนต์มือสอง นี้อาจช่วยให้เราสามารถเลือกวิธีการชำระเงินที่เหมาะสมกับสภาพการเงินและความต้องการของเราได้, 10.การรับประกันที่มีอยู่: บางรถยนต์มือสองยังคงมีระยะเวลารับประกันจากผู้ผลิตหรือผู้จำหน่ายอยู่ ซึ่งอาจช่วยให้เรามั่นใจได้ในความพร้อมและคุณภาพของรถยนต์, 11.ความหลากหลายในตลาด: ตลาดรถยนต์มือสองมีความหลากหลายที่มากกว่าตลาดรถยนต์ใหม่ เราสามารถเลือกรถยนต์มือสองที่ตรงกับความต้องการและความสนใจของเราได้มากขึ้น ไม่ว่าจะเป็นรถยนต์ระหว่างประเภท ขนาด สี รายการอุปกรณ์หรือคุณสมบัติเฉพาะ, 12.ตรวจสอบรายละเอียดก่อนซื้อ: ก่อนซื้อรถยนต์มือสอง เราสามารถ

ตรวจสอบรายละเอียดต่างๆ เช่น สภาพภายนอกและภายในของรถยนต์ ประวัติการเข้าศูนย์บริการหรือการซ่อมแซม ระยะทางที่รถยนต์เคยใช้งาน และเอกสารประกันภัย นี้ช่วยให้เราได้รับข้อมูลที่ชัดเจนเพื่อให้ตัดสินใจได้ถูกต้อง, 13.อนุญาติให้ทดสอบขับขี่: หากเราต้องการทราบถึงประสบการณ์ขับขี่ของรถยนต์มือสอง เราอาจขออนุญาตให้ทดสอบขับขี่ก่อนการซื้อ นี้ช่วยให้เราทราบถึงสภาพการขับขี่ ความสบาย และประสบการณ์ที่เราจะได้รับจากการใช้รถยนต์นั้น, 14.ราคาต่อเมื่อถือครองและค่าความเสียหาย: รถยนต์มือสองมีราคาที่ถูกกว่ารถยนต์ใหม่ นอกจากนี้เรายังสามารถต่อราคาหรือขอส่วนลดได้กับผู้ขาย นอกจากนี้ ความเสียหายหรือตำแหน่งของรถยนต์มือสองยังสามารถใช้เป็นตัวต่อราคาได้, 15.ความนิยมและความเป็นที่ยอมรับ: มีรถยนต์มือสองบางรุ่นที่ได้รับความนิยมและความเป็นที่ยอมรับในช่วงเวลานาน รถยนต์เหล่านี้มักมีการทดแทนส่วนประกอบและอะไหล่ที่ง่ายต่อการหา รวมถึงช่างที่มีความชำนาญในการซ่อมบำรุงได้ง่าย, 16.สภาพการใช้งานในช่วงต้นและปลายอายุการใช้งาน: การซื้อรถยนต์มือสองในช่วงต้นของอายุการใช้งานของรถยนต์อาจมีราคาที่ถูกกว่า แต่คุณต้องพิจารณาถึงสภาพของรถยนต์ในระยะเวลาต่อไป การซื้อรถยนต์มือสองในช่วงปลายของอายุการใช้งานอาจทำให้เราเสี่ยงต่อค่าใช้จ่ายในการซ่อมแซมและบำรุงรักษาที่มากขึ้น, 17.การปรับแต่งและการอัพเกรด: หากเราสนใจการปรับแต่งหรือการอัพเกรดรถยนต์ รถยนต์มือสองอาจเป็นทางเลือกที่ดี รถยนต์มือสองที่อยู่ในราคาที่ถูกมากกว่ารถยนต์ใหม่ มักมีพื้นฐานที่แข็งแกร่งและสามารถรองรับการปรับแต่งและอัพเกรดได้, 18.ซื้อจากแหล่งที่น่าเชื่อถือ: การซื้อรถยนต์มือสองควรทำจากแหล่งที่น่าเชื่อถือเช่น ผู้ขายรถยนต์มือสองที่มีชื่อเสียง หรือผ่านการรับประกันจากผู้จำหน่ายรถยนต์มือสองที่เชื่อถือได้

อนึ่ง ในเดือนสิงหาคม 2564 ได้มีผลงานวิจัยด้านการตลาดในประเทศอินเดียของ Ken Research เรื่องการได้รับความนิยมของธุรกิจรถยนต์มือสองไปทั่วโลกตั้งแต่เกิดเหตุการณ์โรคระบาดโคโรนาไวรัส 2019 ซึ่งความนิยมดังกล่าวไม่ได้จำกัดอยู่เฉพาะในประเทศอินเดียหรือทวีปเอเชียเท่านั้น แต่ยังรวมไปถึงกลุ่มประเทศในทวีปยุโรป และสหรัฐอเมริกาอีกด้วย ทั้งนี้ ตามรายงานของ Ken Research ได้ชี้ให้เห็นถึงความสอดคล้องกับกระแสนิยมในตลาดโลกดังกล่าวที่เกิดขึ้นกับประเทศไทย แม้ว่าในสถานการณ์ปกตินั้นการซื้อรถยนต์มือสองจะได้รับความนิยมอยู่ก่อนแล้วในประเทศไทย แต่ตั้งแต่เกิดเหตุการณ์โรคระบาดโคโรนาไวรัส 2019 ขึ้น กระแสความนิยมดังกล่าวกลับปรับตัวเพิ่มขึ้น โดยประเทศไทยมีอัตราการซื้อรถยนต์มือสองเพิ่มขึ้นเป็นจำนวนมากเมื่อเทียบกับตลาดในกลุ่มประเทศในแถบเอเชียตะวันออกเฉียงใต้ ปัจจัยที่ทำให้ตลาดรถยนต์มือสองในปี 2563 มีอัตราเพิ่มสูงขึ้นเนื่องจากสภาวะการเกิดโรคระเบิดนั้นทำให้ประชาชนเกิด

ความระมัดระวังในการใช้ชีวิต และป้องกันตัวจากโรคระบาด โดยการหลีกเลี่ยงการใช้ระบบขนส่งสาธารณะ จึงส่งผลให้ความต้องการรถยนต์ส่วนตัวเพิ่มมากขึ้น

ซึ่งในปี 2564 ก็ปรากฏรายงานการเติบโตของตลาดรถยนต์มือสองที่สอดคล้องกัน โดยมีรายงานจากสมาคมผู้ประกอบการรถยนต์ใช้แล้วระบุว่า มูลค่าที่เกิดในตลาดรถยนต์มือสองแบ่งเป็นการเติบโตของรถยนต์นั่ง เติบโตเพิ่มขึ้น 10-15% โดยประมาณ และ การเติบโตของรถบรรทุกเติบโตเพิ่มขึ้น 1.5-3.5% โดยประมาณ ซึ่งมีมูลค่ารวมประมาณ 80,000 ล้านบาท

ทั้งนี้ ยังมีข้อมูลเน้นย้ำไปในในทิศทางเดียวกันกับผลวิจัยและรายงานข้างต้นเกี่ยวกับโอกาสการเติบโตในทิศทางที่ดีของตลาดรถยนต์มือสองในปี 2565 ซึ่งอ้างอิงจากข้อมูลของกรุงเทพธุรกิจระบุว่า ปัจจัยที่ทำให้รถยนต์มือสองยังเป็นที่ต้องการของผู้บริโภคเมื่อเทียบกับรถยนต์ใหม่ เนื่องจากความได้เปรียบทางด้านราคา รวมทั้งอายุการใช้งานที่น้อยของรถยนต์มือสองในปัจจุบัน ซึ่งทำให้ยังคงได้รับประโยชต์ในด้านการคุ้มครองจากการรับประกันอยู่ จึงส่งผลในเชิงบวกทำให้เกิดความมั่น เกิดความประหยัด และทางเลือกในการซื้อสำหรับกลุ่มลูกค้าเพิ่มมากขึ้น

นอกจากนี้ ข้อเสียที่เคยเกิดขึ้นในอดีตของตลาดรถยนต์มือสอง เช่น ราคาที่ไม่สมเหตุสมผล โดยลูกค้าไม่สามารถประเมินราคาที่ถูกต้องได้ เนื่องจากการถูกดัดแปลงสภาพของรถยนต์มือสองให้อยู่ในสภาพดีและใหม่ โดยแท้จริงแล้วอาจถูกซ่อมแซมหรือกลับเลขไมล์ให้ระยะทางน้อยลงนั้น ซึ่งปัญหาดังกล่าวได้ถูกแก้ไขไปในทิศทางที่ดีมากขึ้นเนื่องจากตลาดรถยนต์มือสองในปัจจุบันเปิดกว้างและเข้าถึงข้อมูลได้ง่ายขึ้น โดยลูกค้าสามารถใช้ประโยชน์จากเว็บไซต์ขายรถยนต์มือสองในการศึกษารายละเอียดเบื้องต้นเกี่ยวกับรุ่น หรือลักษณะของรถยนต์ที่ตนเองสนใจ และสำหรับปัญหาการประเมินราคาและประเมินสภาพรถยนต์ก็สามารถว่าจ้างผู้เชี่ยวชาญในการตรวจสอบรถยนต์เข้าไปตรวจสอบและประเมินสภาพกับผู้ขายโดยตรงได้ จึงเป็นแนวทางที่ทำให้เกิดความเชื่อมั่นกับกลุ่มลูกค้ารถยนต์มือสองเพิ่มมากขึ้น และส่งผลให้ตลาดรถยนต์มือสองเติบโตขึ้นได้ (prakai, 2021)

จากแหล่งข้อมูลอ้างอิงที่ชี้ให้เห็นถึงทิศทางการเติบโตของตลาดรถยนต์มือสองแล้ว งานวิจัยนี้จะเป็นอีกหนึ่งเครื่องมือที่เป็นประโยชน์สำหรับการขับเคลื่อนให้ตลาดรถยนต์มือสองเติบโตอย่างเป็นธรรมต่อผู้ซื้อและผู้ขายมากยิ่งขึ้น เนื่องจากปัญหาที่เคยอ้างถึงเกี่ยวกับการตั้งราคาให้เหมาะสมกับสภาพ ราคาที่เป็นกลาง และพึงพอใจระหว่างผู้ซื้อและผู้ขายนั้นถือเป็นปัจจัยสำคัญในการเลือกซื้อรถยนต์มือสอง ซึ่งโดยทั่วไปแล้วพฤติกรรมในการเลือกซื้อรถยนต์มือสองของฝ่ายผู้ซื้อในปัจจุบันจะหาข้อมูลและเปรียบเทียบราคาของรถยนต์มือสองต่างๆ ผ่านช่องทางออนไลน์หรือแพลต์ฟอร์ม E-commerce ของตลาดรถยนต์มือสองเพื่อประเมินราคาก่อน ใน

ทำนองเดียวกันกับพฤติกรรมสำหรับทางฝ่ายผู้ขายในปัจจุบันก็จำเป็นจะต้องสำรวจราคาในตลาดรถยนต์มือสองก่อนจะตั้งขาย โดยลักษณะงานวิจัยนี้จะใช้แบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ทำนายราคายนต์มือสอง ให้ได้ผลลัพธ์ที่ได้จากการทำนายเป็นราคาที่แม่นยำสอดคล้องกับสภาพรถยนต์มือสอง จึงเป็นการแก้ปัญหา และถือเป็นประโยชน์ต่อทั้งผู้ซื้อและซื้อขาย สำหรับด้านผู้ขายหรือผู้ประกอบการ งานวิจัยนี้จะเข้าไปส่งเสริมด้านการกำหนดราคาขายที่เหมาะสมกับสภาพรถยนต์มือสองและราคาตลาดในขณะนั้น และสำหรับผู้ซื้อหรือลูกค้าก็สามารถใช้งานวิจัยนี้ในการทำนายราคารถยนต์มือสองรุ่นที่ตนเองสนใจได้อย่างแม่นยำและเหมาะสมกับรถยนต์คันดังกล่าว ซึ่งงานวิจัยนี้ได้พัฒนาแบบจำลองการทำนายราคารถยนต์มือสองจากการพิจารณาคุณสมบัติและส่วนประกอบต่างๆ ของรถยนต์นำมาใช้วิเคราะห์ข้อมูลและเปรียบเทียบหลากหลายแบบจำลอง ตัวอย่างเช่น นำข้อมูล ยี่ห้อ รุ่น ปี เลขไมล์ เชื้อเพลิง เกียร์ของรถยนต์มือสองมาเป็นปัจจัยของการทดลองโดยนำมาใช้กับแบบจำลอง การถดถอยแบบเชิงเส้น (Linear Regression), การถดถอยแบบ (Ridge), ลาสโซ่ (Lasso) และการถดถอยต้นไม้การตัดสินใจ (Decision Tree Regression) เป็นต้น ซึ่งเมื่อปรากฏผลลัพธ์ที่แม่นยำที่สุดเปรียบกันระหว่างแบบจำลองดังกล่าวแล้วก็จะเลือกเอาเฉพาะแบบจำลองที่แม่นยำที่สุด เพื่อนำมาใช้เป็นเครื่องการทำนายราคารถยนต์มือสองต่อไป

1.2 **ความมุ่งหมายของงานวิจัย**

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อศึกษาข้อมูลราคารถยนต์มือสองจากตลาดรถยนต์มือสองในสหราชอาณาจักร
2. เพื่อการศึกษาและประยุกต์การใช้เทคนิคการเรียนรู้ของเครื่องมาสร้างแบบจำลองการทำนายราคารถยนต์มือสองโดยใช้คุณสมบัติเช่นระยะทางที่ใช้รถ,ปีรถเป็นต้น
3. เพื่อศึกษาและเปรียบเทียบกลุ่มอัลกอริทึมที่มีผลต่อการทำนายราคารถยนต์
4. หาปัจจัยที่ส่งผลต่อราคารถมือสอง

1.3 ขอบเขตของการวิจัย

1. ชุดข้อมูลในการวิจัยมาจจาก www.kaggle.com ซึ่งเป็นชุดข้อมูลรายการรถยนต์มือสองของตลาดรถยนต์สหราชอาณาจักร
2. นำข้อมูลมาวิเคราะห์เพื่อคาดการณ์ช่วงเวลาที่เหมาะสมในการขายรถยนต์มือสอง เช่น มูลค่าการขายต่อตามอายุการใช้งานและระยะทางการใช้งานเท่าใดที่ราคาขายต่อลดลง
3. ใช้แบบจำลองการเรียนรู้ของเครื่องในการทำนายราคารถยนต์มือสองโดยใช้ 4 เทคนิคได้แก่ การถดถอยแบบเชิงเส้น Linear Regression,การถดถอยต้นไม้การตัดสินใจ Decision Tree Regressor, การถดถอยแบบ Ridge และลาสโซ่ Lasso
4. สำรวจปัจจัยที่มีผลต่อราคารถยนต์มือสองและปัจจัยอื่นๆที่เกี่ยวข้อง
5. สร้างแบบจำลองเพื่อทำนายราคาที่เหมาะสมสำหรับการขายรถยนต์มือสอง
6. วัดผลประสิทธิภาพด้วย R-squared, Mean Absolute Error (MAE), Mean SquareError(MSE)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ศึกษาทฤษฎีและแนวคิดการเรียนรู้ของเครื่อง
2. ได้ศึกษาทฤษฎีและแนวคิดการเกี่ยวกับการวิเคราะห์ข้อมูลเชิงสำรวจ
3. ได้ศึกษาขั้นตอนและเทคนิคที่ใช้ในการพัฒนาแบบจำลองที่เหมาะสมกับข้อมูล
4. ได้ศึกษาเปรียบเทียบแบบจำลองและนำแบบจำลองไปใช้กับงานจริงได้อย่างแม่นยำ
5. เพื่อเป็นแนวทางในการศึกษาปัจจัยที่ส่งผลกระทบต่อราคารถยนต์มือสอง

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

2.1 องค์ความรู้เกี่ยวกับรถยนต์

2.2 องค์ความรู้เกี่ยวกับ Data Science

2.3 องค์ความรู้เกี่ยวกับ Machine Learning

2.4 อัลกอริทึมการเรียนรู้แบบมีผู้สอน(Supervised Machine Learning Algorithms)

2.5 องค์ความรู้เกี่ยวกับ Exploratory data analysis (EDA)

2.6 การเลือกคุณลักษณะ (Feature Selection)

2.7 องค์ความรู้เกี่ยวกับการประเมินแบบจำลอง

2.8 องค์ความรู้เกี่ยวกับการวัดค่าประสิทธิภาพของแบบจำลอง

2.9 เครื่องมือที่ใช้ในการทำวิจัย

2.10 งานวิจัยที่เกี่ยวข้อง

2.1 องค์ความรู้เกี่ยวกับรถยนต์

2.1.1 ข้อมูลทั่วไปของรถยนต์

รถยนต์ หมายถึงยานพาหนะทางบกที่ขับเคลื่อนด้วยพลังงานและถ่ายทอดสู่ล้อ เพื่อพาไปยังจุดหมายปลายทาง ตามพระราชบัญญัติการขนส่ง พ.ศ. 2522 ม. 4 จากอดีตจนถึงปัจจุบันอุตสาหกรรมรถยนต์มีการพัฒนา ผลิตรถยนต์อย่างมากจากรถยนต์ที่ใช้น้ำมันเป็นพลังงานเป็นรถยนต์ไฮบริดที่ใช้พลังงานน้ำมันและหลังงานไฟฟ้าในการขับเคลื่อนโดยหลักรถชนิดนี้จะมีแบตเตอรี่ไว้คอยกักเก็บหลังงานมาไว้เพื่อดึงมาใช้ในการออกตัว ปัจจุบันมีการพัฒนารถยนต์พลังงานไฟฟ้า (EV) ผลิตออกสู่ตลาด รถที่นำพลังงานไฟฟ้ามาใช้เต็มระบบโดยไม่พึ่งน้ำมันเชื้อเพลิง โดยหลักการ คือ จะนำพลังงานไฟฟ้าไปปั่นมอเตอร์ส่งกำลังไปยังล้อและระบบต่างๆ อีกทั้งรถไฟฟ้ายังมีอัตราเร่งที่สูงกว่ารถยนต์ที่ใช้น้ำมันอีกด้วย โดยรถประเภทนี้จะไม่มีการปล่อยมลพิษออกมมาเลยแม้แต่น้อย

2.1.2 **ข้อมูลจำเพาะของรถยนต์**

ข้อมูลจำเพาะของรถยนต์ที่ผู้ซื้อ-ผู้ขาย ควรรู้จะประกอบด้วยดังนี้

2.1.2.1 **ระบบเกียร์**

ระบบเกียร์รถยนต์ปัจจุบันประกอบด้วยเกียร์อัตโนมัติ,เกียร์ธรรมดาและระบบเกียร์รถไฟฟ้า ระบบเกียร์ คือ ตัวกลางที่รับหน้าที่ถ่ายทอดกำลังของเครื่องยนต์ไปสู่เพลาล้อรถยนต์ ซึ่งระบบเกียร์จะมีอัตราทดที่มีการคำนวณออกแบบให้เหมาะสมกับรถยนต์รุ่นนั้น ๆ หากอัตราทนของเกียร์และเพลาไม่ลงตัว คุณจะพบความพิเศษของรถยนต์ ยกตัวอย่างเช่น รถวิ่งไม่ออกตลอดจนรถวิ่งอั้น อะไรแบบนี้เป็นต้น

2.1.2.2 **ข้อมูลทั่วไป**

ข้อมูลทั่วไปของรถยนต์หากนำไปค้นหาตามเว็บ E-commerce ด้านรถยนต์ ข้อมูลส่วนนี้จะบอกลักษณะของรถยนต์เช่น มีจำนวนประตูกี่ประตู,จำนวนที่นั่ง,ชนิดของเบรค,สีของรถยนต์,ยางและล้อรถยนต์เป็นต้น

2.1.2.3 **รายละเอียดเครื่องยนต์**

โดยราละเอียดเครื่องยนต์สามารถดูได้จากคู่มือรถยนต์ซึ่งจะมีพร้อมกับรถยนต์ โดยจะมีข้อมูลดังนี้ ความจุเครื่องยนต์, ระยะช่วงชัก, กำลังสูงสุด, รุ่นเครื่องยนต์, Direct Injection , Aspiration, ขนาดกระบอกสูบ ,อัตราส่วนการอัด, แรงบิดสูงสุด ,ชนิดเครื่องยนต์และประเภทเชื้อเพลิงเช่นเบนซิน ,ดีเซล เป็นต้น

2.1.2.4 **อัตราการสิ้นเปลือง**

อัตราสิ้นเปลืองเป็นตัวบอกอัตราการใช้พลังงานน้ำมัน ตามระยะทาง ซึ่งมีหน่วยเป็นกิโลเมตรต่อลิตร

2.1.2.5 **ขนาดและน้ำหนัก**

ข้อมูลขนาดและน้ำหนักของรถยนต์ประกอบด้วย ความยาว ,ความสูง ความกว้าง, ระยะฐานล้อ, ล้อหน้า, ล้อหลัง, น้ำหนักรถ ,ความจุถังน้ำมัน, ข้อมูลเหล่านี้มีประโยชน์เช่นใช้สำหรับเทียบขนาดและเป็นข้อมูลสำหรับใช้เทียบค่าเดิมที่มากจากโรงงานและค่าที่ได้จากการตรวจสภาพจากสถานตรวจสภาพรถ เมื่อต้องการซื้อรถยนต์มือสอง

2.2 องค์ความรู้เกี่ยวกับวิทยาศาสตร์ข้อมูล

2.2.1 การวางกรอบปัญหา

การทำความเข้าใจและกำหนดกรอบปัญหาเป็นขั้นตอนแรกของวัฏจักรวิทยาศาสตร์ข้อมูล การวางกรอบนี้จะช่วยให้สร้างแบบจำลองที่มีประสิทธิภาพ

2.2.2 เก็บข้อมูล

เป็นขั้นตอนเริ่มต้นสำหรับการทำงานด้านการวิเคราะห์ข้อมูล โดยเป็นขั้นตอนที่จะต้องกำหนดและคัดเลือกข้อมูลที่จำเป็นหรือสำคัญต่อการวิเคราะห์หาคำตอบตามเป้าหมายที่กำหนด โดยข้อมูลที่นำมาใช้อาจได้มาจากหลายๆแหล่ง และมีได้หลายรูปแบบ เช่นข้อมูลรวบรวมมาจากเว็บไซต์ต่างๆด้วยวิธี Web APIs ข้อมูลหรือคลังข้อมูลหรือข้อมูลที่เก็บรวบรวมไว้แล้วและมีการแจกจ่ายให้ดาวน์โหลดไปใช้ได้ เช่น ข้อมูลจาก UCI Machine Learning Repository, Data World และ Kaggle Datasets เป็นต้น เพื่อให้ได้ข้อมูลที่มีประสิทธิภาพนักวิเคราะห์ข้อมูลจะต้องมีความเข้าใจเกี่ยวกับวัตถุประสงค์ ขอบเขตและเป้าหมายของงานที่จะทำการวิเคราะห์ข้อมูลมาก่อน เพื่อที่จะสามารถกำหนดและคัดเลือกข้อมูลได้ถูกต้อง

2.2.3 การทำความสะอาดข้อมูล

ข้อมูลส่วนใหญ่ที่รวบรวมระหว่างขั้นตอนการเก็บรวบรวมจะไม่มีโครงสร้าง ไม่เกี่ยวข้อง และไม่ผ่านการกรอง ข้อมูลที่ไม่ถูกต้องทำให้เกิดผลลัพธ์ที่ไม่ดี ดังนั้นความถูกต้องและประสิทธิภาพของการวิเคราะห์จะขึ้นอยู่กับคุณภาพของข้อมูลเป็นอย่างมากการล้างข้อมูลช่วยขจัดค่าที่ซ้ำกันและค่า Null ข้อมูลที่เสียหาย ประเภทข้อมูลที่ไม่สอดคล้องกัน รายการที่ไม่ถูกต้อง ข้อมูลที่ขาดหายไป และการจัดรูปแบบที่ไม่เหมาะสมขั้นตอนนี้เป็นกระบวนการที่ใช้เวลามากที่สุด แต่การค้นหาและแก้ไขข้อบกพร่องของข้อมูลมีความสำคัญต่อการสร้างแบบจำลองที่มีประสิทธิภาพ

2.2.4 การสำรวจข้อมูล

ขั้นตอนนี้เป็นขั้นตอนของกลั่นกรองข้อมูลที่ได้มา และจัดเตรียมให้พร้อมต่อการนำไปประมวลผลขั้นต่อไป โดยขั้นตอนนี้เพื่อให้ได้ข้อมูลที่มีคุณภาพ นักวิเคราะห์ข้อมูลจำเป็นต้องทำความเข้าใจในทุกๆ ส่วนของข้อมูลที่เก็บรวบรวมมาได้ ประเมินว่าข้อมูลที่ได้มีความครบถ้วนหรือไม่มีความผิดปกติของข้อมูลเกิดขึ้นหรือไม่ จากนั้นก็ทำการจัดเตรียมข้อมูลให้เป็นชุดข้อมูล (Data set) ที่พร้อมสำหรับนำไปสร้างเป็นแบบจำลองวิเคราะห์ ขั้นตอนนี้ถือว่าเป็นขั้นตอนที่ต้องการความละเอียดอย่างมากและใช้ระยะเวลานาน เนื่องจากว่าจะต้องทำการคัดเลือกเอาเฉพาะข้อมูลคุณลักษณะ (Attributes) หรือตัวแปรที่เกี่ยวข้องจริงๆ รวมไปจนกระทั่งปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับวิธีการที่นำมาใช้วิเคราะห์ข้อมูล

2.2.5 การสร้างแบบจำลองวิเคราะห์ข้อมูล

ขั้นตอนนี้เป็นขั้นตอนของการนำข้อมูลที่จัดเตรียมไว้มาวิเคราะห์ด้วยเทคนิคต่างๆ ตามความเหมาะสม โดยมากจะเป็นการวิเคราะห์ด้วยเทคนิคมากกว่าหนึ่งแบบเพื่อประเมินหาเทคนิคที่ให้ค่าการวิเคราะห์ที่เป็นไปตามวัตถุประสงค์และมีประสิทธิภาพมากที่สุด

2.2.6 การนำเสนอผลลัพธ์

ขั้นตอนนี้สามารถทำได้ในทุกๆ ช่วงของกระบวนการวิเคราะห์ข้อมูล โดยเป็นขั้นตอนของการสื่อสารเพื่อสร้างการรับรู้ให้กับกลุ่มคนที่เกี่ยวข้องทั้งนี้อาจจะเป็นหัวหน้างานหรือผู้บริหารได้เรียนรู้และเข้าใจเกี่ยวกับสิ่งที่ได้จากข้อมูล โดยสิ่งสำคัญในขั้นตอนนี้ที่ผู้นำเสนอจะต้องคำนึงถึงก็คือ การสื่อสารและรูปแบบการนำเสนอที่ง่ายต่อความเข้าใจ เพราะส่วนมากแล้วผู้ที่รับฟังข้อมูลมักจะไม่ใช่ผู้เชี่ยวชาญที่มีความเข้าใจในกระบวนการวิเคราะห์ข้อมูลเชิงเทคนิคมากนัก ดังนั้นรูปแบบและวิธีการนำเสนอข้อมูลผลลัพธ์ที่ได้ในแต่ละขั้นตอนการวิเคราะห์ควรจัดทำให้อยู่ในรูปแบบที่ง่ายต่อการทำความเข้าใจของคนทั่วไป

2.2.7 การทดสอบการใช้งาน

ขั้นตอนนี้เป็นขั้นตอนของการทดสอบใช้แบบจำลองการวิเคราะห์ข้อมูลที่ได้ก่อนนำไปใช้งานจริง เพื่อพัฒนาและปรับปรุงแบบจำลองให้เหมาะสมกับสถานการณ์ที่อาจเกิดขึ้นจริงในทางปฏิบัติ

2.2.8 การนำผลลัพธ์ไปใช้

เป็นขั้นตอนหลังจากทำการทดสอบใช้งานแบบจำลองและทำการปรับปรุงให้เหมาะสมเรียบร้อยแล้วก็จะนำแบบจำลองที่ได้มาใช้งานจริงๆโดยอาจพัฒนาให้อยู่ในรูปแบบโปรแกรมหรือพัฒนาให้อยู่ในรูปของแผนหรือกลยุทธ์ทางธุรกิจเป็นต้น

2.3 องค์ความรู้เกี่ยวกับการเรียนรู้ของเครื่อง

เป็นระบบที่สามารถเรียนรู้ได้จากตัวอย่างด้วยตนเองโดยปราศจากการป้อนคำสั่งของโปรแกรมเมอร์ ความก้าวหน้าในครั้งนี้มาพร้อมกับความคิดที่ว่าเครื่องคอมพิวเตอร์สามารถเรียนรู้เพียงแค่จากข้อมูลอย่างเดียวเพื่อที่จะผลิตผลลัพธ์ที่แม่นยำออกมาได้ โดย Machine Learning แบ่งได้ 3 แบบดังนี้

2.3.1 Supervised Learning คือการเรียนรู้โดยมี data มาสอน

2.3.2 Unsupervised Learning คือการเรียนรู้โดยไม่มี data สอน

2.3.3 Reinforcement Learningคือเรียนรู้ตามสภาพแวดล้อม

2.4 อัลกอริทึมการเรียนรู้แบบมีผู้สอน (Supervised Machine Learning Algorithms)

อัลกอริทึมจำเป็นต้องใช้ ข้อมูลในส่วนสำหรับ train (การฝึกข้อมูล) และส่วนที่รับกลับมาเพื่อปรับปรุง (feedback) จากมนุษย์เพื่อที่จะเรียนรู้ความสัมพันธ์ระหว่างข้อมูลที่ถูกป้อนเข้ามาสู่ข้อมูลที่ออกไป ยกตัวอย่างเช่น มีเด็กหัดคนหนึ่งใช้การค่าใช้จ่ายทางการตลาดและการพยากรณ์สภาพอากาศเป็นข้อมูลขาเข้าเพื่อทำนายจำนวนกระป๋องที่จะขายได้ สามารถใช้ การเรียนรู้แบบมีผู้สอนเมื่อผลลัพธ์ของข้อมูลเป็นสิ่งที่รู้อยู่แล้ว อัลกอริทึมนี้ก็จะทำนายข้อมูลใหม่ได้ supervised learning มีอยู่ 2 ประเภทคือ การแบ่งแยกประเภท (Classification) และการถดถอย (Regression)

2.5 **การเรียนรู้แบบไม่มีผู้สอน** (Unsupervised learning)

อัลกอริทึมจะตรวจสอบเฉพาะข้อมูลที่ป้อนเข้ามาเท่านั้นโดยปราศจากการให้ผลลัพธ์ที่จะเกิดขึ้นเช่น การสำรวจข้อมูลประชากรเพื่อหาแบบแผน (pattern) ของข้อมูลนั้น เราสามารถใช้การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) เมื่อไม่ต้องการที่จะรู้ว่า machine แบ่งประเภทได้อย่างไร และเราต้องการอัลกอริทึมนั้นเพื่อหาpattern และแบ่งประเภทข้อมูล

2.7 **การเรียนรู้การเสริมแรง** (Reinforcement Learning)

Reinforcement Learning เป็นแนวทางการเรียนรู้ของ Artificial Intelligence (AI) ซึ่งมีลักษณะที่เหมือนกับการเรียนรู้ของมนุษย์ นั่นคือเป็นการเรียนรู้จากการลองผิดลองถูก และพยายามค้นหาแนวทางรับมือกับปัญหาหนึ่ง ๆ ให้ดีที่สุด ซึ่งนำไปสู่ผลลัพธ์ที่มีประสิทธิภาพ เช่น self-driving car, stock tradingbot แนวทางการเรียนรู้แบบ Reinforcement Learning แตกต่างจาก Supervised Learning อย่างสิ้นเชิง เพราะ Supervised Learning เป็นการเรียนรู้จากข้อมูลที่มีอยู่เพื่อพยากรณ์ข้อมูลที่อยู่นอกขอบเขตที่มี ในขณะที่ Reinforcement Learning เป็นการเรียนรู้โดยอาศัยประสบการณ์จากการลองผิดลองถูกและเรียนรู้ผลดี-ผลเสียของวิธีแก้ปัญหาหนึ่งๆโดยมุ่งหวังให้ได้ผลลัพธ์ที่มีประสิทธิภาพหรือเสาะหาวิธีการที่ดีที่สุดในการแก้ปัญหาหนึ่ง ๆ เช่น self-driving car และ stock trading bot ซึ่งแตกต่างจาก Supervised Learning ที่เน้นการเรียนรู้จากข้อมูลที่มีอยู่เพื่อพยากรณ์ผลลัพธ์ของข้อมูลใหม่

2.8 **องค์ความรู้เกี่ยวกับ** Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) ถือว่าเป็นการวิเคราะห์ข้อมูลที่จำเป็นก่อน ที่จะวิเคราะห์ข้อมูลแบบอื่นๆ เช่นงาน Predictive ลักษณะการทำงานคือ สำรวจข้อมูลในมุมต่างๆ ในทุกๆตัวแปร หรือเปรียบเทียบกันระหว่างตัวแปร วิธีการทำ EDA ก็มีหลากหลายวิธี เช่น Visualization, Statistical Analysis วิเคราะห์ตัวแปล, การทำ clustering เป็นต้น EDA จะไม่มีการตั้งธงหรือสมมุติฐานไว้ ให้ข้อมูลเป็นตัวบอก จึงเหมาะอย่างยิ่งในงานที่หา Insights

EDA คือกระบวนการตรวจสอบ สำรวจข้อมูลเบื้องต้น เป็นการวิเคราะห์ข้อมูลที่จำเป็นก่อนการนำข้อมูลไปใช้ หรือนำไปวิเคราะห์เชิงลึก โดยประโยชน์ของการทำ EDA จะช่วยทำให้เราเข้าใจพื้นฐานเกี่ยวกับข้อมูลชุดนั้น และเป็นการตรวจความผิดพลาดของชุดข้อมูลได้อีกด้วย

2.7 องค์ความรู้เกี่ยวกับการวัดค่าประสิทธิภาพของแบบจำลอง

การเลือกค่าวัดประสิทธิภาพของแบบจำลองจะต้องสอดคล้องกับสิ่งที่ต้องการทราบจาก

แบบจำลองเนื่องจากค่าที่วัดแต่ละค่าจะบ่งบอกประสิทธิภาพของแบบจำลองในแง่มุมที่ต่างกันซึ่งการวัดค่าที่มีประสิทธิภาพของแบบจำลองที่ใช้ในงานวิจัยนี้มีดังนี้

2.7.1 Accuracy

Accuracy คือ การวัดความแม่นยำของการทำนายผล หรือ metric ที่ใช้งานง่ายที่สุด บอกว่าโมเดลเราทำนายถูกทั้งหมดกี่ % ซึ่งมีสูตรว่า accuracy = (TP + TN)/ N ซึ่งTP ย่อมาจาก True Positiveมีความหมายว่าผลบวกจริง และNP ย่อมาจาก True Negative(TN)มีความหมายว่าผลลบจริง ก็คือที่ไม่เจอเป็นจริง โดยค่า accuracy จะมีค่าอยู่ระหว่าง 0-1 ยิ่งเข้าใกล้ 1 แปลว่าโมเดลเราทำนายผลได้ดีมาก

2.7.2 Mean Squared Error (MSE)

ย่อมาจาก Mean Square Error ค่า MSE จะมีการทำให้ค่า Error กลายเป็นบวกก่อน โดยการนำค่า Error มายกกำลังสอง ก่อนที่จะนำค่า Error มาหาค่าเฉลี่ย

2.7.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) คือ ค่าเฉลี่ยของผลต่างสัมบูรณ์ ระหว่างผลเฉลยกับค่าที่เกิดจากการทำนายของแบจำลอง

2.7.4 R-Squared

ค่า R-Squared คือตัวสถิติที่ใช้วัดว่าตัวแบบคณิตศาสตร์ที่ได้นี้มีความสมรูปกับข้อมูลมากน้อยอย่างไร หรือรู้จักกัน ในอีกความหมายหนึ่งว่าเป็น ค่าสัมประสิทธิ์แสดงการตัดสินใจ (Coefficient of Determination) หรือ ค่าสัมประสิทธิ์แสดง การตัดสินใจเชิงซ้อน (Coefficient of Multiple Determination) สำหรับการวิเคราะห์การถดถอยแบบพหุคูณ (Multiple Regression) นิยามของค่า R-Squared คือค่าความผันแปรของตัวแปรตอบสนองที่สามารถอธิบายได้มีอยู่ในตัวแบบเชิงเส้นนี้ กี่เปอร์เซ็นต์ หรือR-Squared = ความผันแปรที่สามารถอธิบายได้ / ความผันแปรทั้งหมด (Explained variation / Total Variation)ค่า R-Squared จะมีค่าอยู่ระหว่าง 0% - 100% โดย 0% แสดงให้เห็นว่า ตัวแบบคณิตศาสาตร์ที่ได้มานั้นไม่สามารถอธิบายความผันแปรของค่าตัว

แปรตอบสนอง ต่างที่กระจายรอบค่าเฉลี่ยได้และ100% แสดงให้เห็นว่า ตัวแบบคณิตศาสาตร์ที่ได้มานั้นสามารถอธิบายความผันแปรของค่าตัวแปรตอบสนอง ต่างที่กระจายรอบค่าเฉลี่ยได้เป็นอย่างดีโดยทั่วไปแล้ว ค่า R-Squared สูงๆ หมายความว่า ตัวแบบคณิตศาสตร์นั้นดี (เหมาะสมกับข้อมูล)

2.8 **การเลือกคุณลักษณะ** (Feature Selection)

Feature Selection ก็ คือ การเลือก Feature เพราะ Feature เรามีเยอะไปหมด ทำให้เวลานำไปใช้งาน เช่นการสร้าง Model ก็อาจจะเป็นอะไรที่สร้างความเสียหายในแง่ของResource ทำให้การเลือก Feature เป็นเรื่องสำคัญมาก

2.9 **เครื่องมือที่ใช้ในการทำวิจัย**

2.9.1 Google Colab

ชื่อเต็มคือ Google Colaboratory เป็นบริการ Software as a Service (Saas) โฮสต์โปรแกรม Jupyter notebook บนCloud จาก Google ซึ่งมีการใช้งานที่สะดวก โดยไม่จำเป็นต้องติดตั้งโปรแกรมใดๆก่อนการใช้งาน สามารถใช้งานได้เพียงแค่มีบัญชี Google drive เพื่อใช้จัดเก็บ ซอร์สโค้ดและชุดข้อมูล โดยภาษา Python เป็นภาษาหลักที่ใช้ในการเขียนและรันงานแต่ละcell ผ่านเว็บ Colab ซึ่งในเวอร์ชั่น Pro มีโปรมีKernelไว้รันโค๊ด

2.9.2 **ชุดคำสั่งไซคิทเลิร์น** (Scikit-learn)

เป็นชุดคำสั่งเสริมหรือไลบารี่ของภาษาไพธอน สำหรับทำงานด้านการเรียนรู้ของเครื่อง จุดเด่นคือฟังก์ชันในการแบ่งประเภทข้อมูล การแบ่งกลุ่มข้อมูล การวิเคราะห์การถดถอยหลายอย่างไม่ว่าจะเป็น ซัพพอร์ตเวกเตอร์แมชชีน (svm) การเรียนรู้ต้นไม้ตัดสินใจ และการแบ่งกลุ่มข้อมูลแบบเคมีน

2.10 งานวิจัยที่เกี่ยวข้อง

บทความวิจัยที่ 1

เรื่อง Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning (J. Varshitha, 2022)

ด้วยการเติบโตทางเศรษฐกิจและความต้องการใช้งานรถยนต์ที่เพิ่มขึ้น มีลูกค้าจำนวนหลายคนที่ต้องการรถยนต์แต่ไม่สามารถซื้อรถที่ผลิตใหม่ด้วยเหตุผลหลายประการ เช่น ราคาสูง ความคุ้มค่าไม่เป็นตามที่ต้องการ ความสามารถทางการเงิน การผลิตที่มีจำนวนจำกัด และอื่นๆ ดังนั้นตลาดรถยนต์ใช้แล้วจึงเพิ่มขึ้นทั่วโลก จะประเมินราคาของรถมือสอง ในโมเดลนี้ใช้การเรียนรู้แบบมีผู้สอน โมเดลโครงข่ายประสาทเทียม และโมเดลการเรียนรู้ด้วยเครื่องแบบสุ่มพัฒนาขึ้นซึ่งสามารถเรียนรู้จากชุดข้อมูลของรถยนต์ที่มีให้ งานวิจัยนี้นำเสนอรูปแบบการทำนายราคารถยนต์มือสองที่มีความผิดพลาดต่ำ น่าเชื่อถือและแม่นยำ ใช้แบบจำลองเชิงเส้นอย่างง่าย. โครงข่ายประสาทเทียม (Artificial Neural Network) สร้างขึ้นโดยใช้อัลกอริทึม Keras Regression คือ Keras Regressor และอัลกอริธึมการเรียนรู้ของเครื่องอื่น ๆ ได้แก่ แบบสุ่ม, Lasso, Ridge, Linear regressions ถูกสร้างขึ้น อัลกอริทึมเหล่านี้ได้รับการทดส อบด้วยชุดข้อมูลรถยนต์ ผลการทดลองดังแสดงในตาราง 1 แสดงให้เห็นว่าแบบจำลองแบบสุ่ม (Random forest) ที่มีค่า Mean Absolute Error (MAE) เท่ากับ 0.74 และค่าความผิดพลาด R-squared เท่ากับ 0. 91 ให้ข้อผิดพลาดน้อยกว่าในอัลกอริทึม

ตาราง 1 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	MAE	R-Squared
Regression	Deep Neural Network	0.76	84%
Regression	Linear Regression	1.15	83%
Regression	Lasso Regression	1.05	87%
Regression	Ridge Regression	1.14	81%
Regression	Random Forest	0.74	91%

บทความวิจัยที่ 2

เรื่อง Prediction of prices for used car by using regression models (N. Monburinon, 2018)

งานวิจัยนี้ ศึกษาเปรียบเทียบประสิทธิภาพการทำนายราคารถยนต์มือสองโดยใช้แบบจำลองการเรียนรู้ด้วยเครื่อง โดยแต่ละรุ่นได้รับการฝึกอบรมโดยใช้ข้อมูลตลาดรถยนต์มือสองที่รวบรวมจากเว็บไซต์อีคอมเมิร์ซของเยอรมัน จากผลการทดลองพบว่า gradient boosted regression trees ที่ให้ประสิทธิภาพที่ดีที่สุดโดยมีค่าเฉลี่ยผิดพลาด แบบสัมบูรณ์เฉลี่ย = 0.28 ตามด้วย random forest regression ด้วย MAE =0.35 และ multiple linear regression ด้วย MAE =0.55 ตามลำดับ ซึ่ง MAE ค่าที่เข้าใกล้ศูนย์แสดงว่าดี

ตาราง 2 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	MAE	MSE
trees /regression	gradient boosted regression trees	0.28	0.28
regression	random forest regression	0.35	0.35
regression	multiple linear regression	0.55	0.55

บทความวิจัยที่ 3

เรื่อง Used Car Price Prediction using Different Machine Learning Algorithms(Bharambe, 2022)

บทความนี้เน้นไปที่การทำงานของอัลกอริธึมการถดถอยแบบต่างๆ สามแบบ ซึ่งใช้ในการทำนายราคารถยนต์มือสอง ในโครงการนี้ โดยใช้คุณลักษณะ อายุของรถยนต์ ประวัติการชน ในการสร้างแบบจำลองการทำนายราคารถยนต์มือสอง เราได้ใช้เทคนิคการเรียนรู้ของเครื่อง 3 แบบ ได้แก่ การถดถอยเชิงเส้น การถดถอยแบบลาสโซ่ และ การถดถอยของ ridge ตามลำดับ เรา

ใช้ไลบรารี Python เพื่อออกแบบ GUI สำหรับโครงการของเราและไลบรารีที่เกี่ยวข้องกับการเรียนรู้ของเครื่องอื่น ๆ เช่น Numpy, Pandas, Sklearn เป็นต้น เราได้คำนวณและเปรียบเทียบความถูกต้องของอัลกอริธึมการเรียนรู้ของเครื่องสามตัว ค่าความแม่นยำของการถดถอยเชิงเส้น การถดถอยแบบ Lasso และ Ridge เท่ากับ 83.65%, 87.09% และ 84.00% ตามลำดับ ดังแสดงในตาราง 3 การถดถอยแบบแลซโซ่จะให้ความแม่นยำสูงสุดในบรรดาอัลกอริธึมทั้งสามแบบ

ตาราง 3 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	Accuracy
Regression	Linear Regression	83.65%
Regression	Lasso Regression	87.09%
Regression	Ridge Regression	84.00%

บทความวิจัยที่ 4

เรื่อง Vehicle Price Classification and Prediction Using Machine Learning in the IoT Smart Manufacturing Era (Al-Turjman, Hussain, Alturjman, และ Altrjman, 2022)

ในบทความนี้ มีการใช้การเรียนรู้ด้วยเครื่องในการคาดการณ์ราคารถยนต์ถือเป็นหนึ่งในหัวข้อการวิจัยที่สำคัญที่สุด_เนื่องจากต้องใช้ความพยายามที่จะสังเกตสภาพของรถยนต์และประสบการณ์การดูรถยนต์ การสร้างแบบจำลองที่คาดการณ์ราคารถยนต์ เราใช้วิธีการเรียนรู้ด้วยเครื่อง สามวิธี 1.โครงข่ายประสาทเทียม ,2.ต้นไม้ตัดสินใจ ,3.ซัพพอร์ตเวกเตอร์ และ 4.การถดถอยเชิงเส้นได้นำไปใช้เพื่อทำงานร่วมกันเป็นกลุ่มในแบบจำลองไฮบริด เทคนิค ML หลายๆ แบบถูกเปรียบเทียบกันเพื่อเปิดเผยว่าวิธีใดเหมาะสม สรุปว่าแบบจำลอง SVM ดีที่สุด มีความแม่นยำถึง 90%

ตาราง 4 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	Accuracy
Regression/classification	Linear Regression (LR)	84%
Regression/classification	Support vector machine (SVM)	90%
Regression/classification	Decision tree(DT)	88%
Regression/classification	Neural Networks	85%

บทความวิจัยที่ 5

เรื่อง Used Car Price Prediction using Machine Learning: A Case Study (M. Hankar, 2022)

ในบทความนี้ใช้การเรียนรู้ของเครื่อง แบบจำลองการวิเคราะห์การถดถอย เพื่อคาดการณ์ราคาขายต่อของรถยนต์มือสองโดยพิจารณาจากปัจจัยหลายประการ เช่น ระยะทาง ประเภทเชื้อเพลิง กำลังทางการเงิน เครื่องยนต์ แบรนด์ รุ่น และปีที่ผลิตรถยนต์ ในแบบจำลองที่ทดสอบทั้งหมด การถดถอยตามลำดับแสดงให้คะแนน R-squared สูง และค่าเฉลี่ยรากที่สอง error ต่ำ 44516.20

ตาราง 5 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	R-squared	RMSE
Regression	MLR	57%	63933.52
Regression	KNNR	70%	51224.96
Regression	RFR	74%	44939.79
Regression	GBR	80%	44516.20
Regression	ANN	67%	549.57.98

จากตารางที่ 5 คำย่อของแบบจำลองต่าวๆดังนี้ multiple linear regression (MLR) , K-nearestneighbors regressor (KNNR) ,random forest regressor (RFR), gradient bosting regressor (GBR), artificial neural network (ANN)

บทความวิจัยที่ 6

เรื่อง Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business (Narayana และ Likhitha, 2021)

วัตถุประสงค์หลักของบทความนี้คือ การสร้างแบบจำลองการคาดการณ์ เช่น กลไกราคายุติธรรมเพื่อทำนายราคาขายรถยนต์ตามคุณลักษณะต่างๆ เช่น รุ่นรถ จำนวนปีที่รถเก่า ประเภทของเชื้อเพลิงที่ใช้ ประเภทของผู้ขาย ประเภทของเกียร์ และจำนวนกิโลเมตรที่รถขับมา จนถึงปัจจุบัน บทความนี้จะช่วยให้ทราบราคาขายรถยนต์มือสองโดยประมาณโดยพิจารณาจาก คุณลักษณะและลดความเสี่ยงของผู้ขายและผู้ซื้อ แบบจำลองที่เสนอใช้อัลกอริธึมการเรียนรู้ด้วย เครื่องและเทคนิคการถดถอยของสถิติ เช่น เส้นตรง ต้นไม้ในการตัดสินใจ และการถดถอยของ ฟอเรสต์แบบสุ่ม เพื่อให้บรรลุงานนี้ ใช้โมเดลการถดถอยของฟอเรสต์แบบสุ่มได้ผลลัพธ์จากการวัด ประสิทธิภาพของแบบจำลอง Mean squared error (MSE) เท่ากับ 0.101

ตาราง 6 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	MSE
Regression	Linear Regression	0.368
Regression/classification	Decision Tree	0.217
Regression/classification	Random Forest	0.101

บทความวิจัยที่ 7

เรื่อง Car Price Prediction using Machine Learning Techniques (Gegic, Isakovic, Keco, และ Masetic, 2019)

การทำนายราคารถยนต์มือสองในบอสเนียและเฮอร์เซโกวีนา เราใช้ 3 เทคนิค (Artificial Neural Network, Support Vector Machine and Random Forest) ข้อมูลที่ใช้ในการทำนายคือรวบรวมจากเว็บ portal autopijaca.ba แบบจำลองได้รับการประเมินโดยใช้ข้อมูลการทดสอบและความความแม่นยำคือ 87.38%

ตาราง 7 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	Accuracy
classification	ANN	86.11%
classification	SVM	83.33%
classification	Random Forest	87.38%

บทความวิจัยที่ 8

เรื่อง Used Car Price Prediction using K-Nearest Neighbor Based Model (K.Samruddhi, 2020)

การทำนายราคารถยนต์มือสองเป็นหนึ่งในพื้นที่ที่สำคัญและน่าสนใจในการวิเคราะห์ ด้วยความต้องการเพิ่มขึ้นในตลาดรถยนต์มือสอง ธุรกิจสำหรับผู้ซื้อและผู้ขายรถยนต์มือสองก็เพิ่มขึ้นเช่นกัน การทำนายที่เชื่อถือได้และแม่นยำต้องการความรู้ของผู้เชี่ยวชาญในสาขานี้ เนื่องจากราคาของรถยนต์ขึ้นอยู่กับปัจจัยหลายประการที่สำคัญ ในงานวิจัยนี้ ผู้วิจัยได้เสนอโมเดลการเรียนรู้ของเครื่องที่มีการนำเสนอข้อมูลรถยนต์มือสองผ่านขั้นตอนการฝึกอบรมแบบดูแลเพื่อวิเคราะห์ราคา โดยใช้ขั้นตอนการเรียนรู้แบบจำกัดของ KNN (K Nearest Neighbor) เราได้ฝึกโมเดลของเราด้วยข้อมูลรถยนต์มือสองที่เก็บรวบรวมมาจากเว็บไซต์ Kaggle ผ่านการทดลองนี้

ข้อมูลได้รับการตรวจสอบด้วยอัตราส่วนการฝึกและทดสอบที่แตกต่างกัน ผลลัพธ์ของโมเดลที่เสนอแสดงให้เห็นว่าความแม่นยำอยู่รอบ 85% และเป็นโมเดลที่ถูกปรับแต่งให้เหมาะสม

ตาราง 8 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	Accuracy
Regression	KNN	85%

บทความวิจัยที่ 9

เรื่อง Car Price Prediction : An Application of Machine Learning(S. Shaprapawad, 2023)

ในการพัฒนาโมเดลเรียนรู้ของเครื่องในการกำหนดราคารถรุ่นใหม่ ผู้วิจัยใช้เทคนิคการลดการเข้ากันของโมเดล (overfitting) และทำให้โมเดลสามารถทำงานได้อย่างทั่วถึง ผู้วิจัยใช้เทคนิคการลดโมเดล (regularization techniques) รวมถึงเทคนิคการปรับค่าพารามิเตอร์ (hyperparameter tuning techniques) เพื่อเอาชนะความท้าทายนี้ โดยพัฒนาโมเดลเชิงเส้น (linear regression), โมเดลลาโส (lasso regression), โมเดลริดจ์ (ridge regression), โมเดลอีลาสติกเน็ต (elastic net regression), โมเดลต้นไม้สุ่ม (random forest), ต้นไม้การตัดสินใจ (decision tree) และเครื่องมือเวกเตอร์สนับสนุน (Support Vector Machine) พร้อมทั้งปรับค่าพารามิเตอร์ให้เหมาะสม วัตถุประสงค์ของบทความนี้คือการสร้างโมเดลที่สามารถทำนายราคารถยนต์มือสองได้อย่างแม่นยำ โดยใช้ข้อมูลต่าง ๆ เช่นเลขไมล์ของรถยนต์ ปีที่ผลิต อากรถ ประเภทเชื้อเพลิงที่ใช้ ขนาดของเครื่องยนต์ เป็นต้น โมเดลที่ดีที่สุดในศึกษานี้คือเครื่องมือเวกเตอร์สนับสนุน (Support Vector Regressor) โดยมีค่า R-Squared ที่ร้อยละ 95.27, ค่าความคลาดเคลื่อนเฉลี่ย (MAE) เท่ากับ 0.142 เมื่อใช้ข้อมูลฝึกอบรม 90% และข้อมูลตรวจสอบ 10%

ตาราง 9 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	MAE	R-squared
Regression	MLR	0.260	8.64%
Regression	Lasso	0.265	8.64%
Regression	Ridge	0.265	8.64%
Regression	E-NET	0.260	86.65%
Regression/Classification	Decision Tree	0.217	90.60%
Regression	Support Vector Regressor	0.142	95.27%

บทความวิจัยที่ 10

เรื่อง Research on used car price prediction based on random forest and LightGBM (Y. Li, 2022)

ตลาดรถยนต์มือสองกำลังขยายตัว ระบบประเมินราคาของตลาดรถยนต์มือสองในประเทศจีนได้เปิดเผยปัญหาที่ไม่ตอบสนองต่อความต้องการของตลาด การทำนายราคารถยนต์มือสองที่แม่นยำสามารถช่วยให้ผู้คนตัดสินใจที่ถูกต้อง ในงานวิจัยนี้เราใช้อัลกอริทึมแบบ Random Forest และ LightGBM เพื่อทำนายราคารถยนต์มือสองและเปรียบเทียบและวิเคราะห์ผลการทำนาย การทดลองพบว่าตัวชี้วัดการประเมินที่เกี่ยวข้องกับโมเดล Random Forest และ LightGBM คือ ค่า MSE ตามลำดับคือ 0.0373 และ 0.0385; ค่า MAE ตามลำดับคือ 0.125 และ 0.117; และ R square ของการทำนายตามลำดับคือ 0.936 และ 0.933 ระหว่างโมเดลการทำนายสองตัว ค่าความคลาดเคลื่อนของโมเดล LightGBM น้อยกว่าและสามารถพิจารณาใช้ในงานวิจัยในอนาคตได้

ตาราง 10 ผลลัพธ์การทำนายราคารถยนต์มือสอง

Problem (Regression/classification)	ML Model	MAE	MSE	R-squared
Regression	RF	0.118	0.039	93.14%
Regression	LightGBM	0.125	0.037	93.55%

สรุปงานวิจัยที่เกี่ยวข้อง

ตาราง 11 สรุปผลลัพธ์การทำนายราคารถยนต์มือสองที่ดีที่สุดของงานวิจัยที่เกี่ยวข้องทั้งหมด

ลำดับบทความวิจัย	ML Model	MAE	MSE	Accuracy	R-squared
1	Random Forest	0.74	-	-	91%
2	gradient boosted regression trees	0.28	0.28	-	-
3	Lasso Regression	-	-	87.09%	-
4	Support vector machine (SVM)	-	-	90%	-
5	gradient bosting regressor	-	-	-	80%
6	Random Forest	-	0.101	-	-
7	Random Forest	-	-	87.38%	-
8	K-nearestneighbors	-	-	85%	-
9	Support Vector Regressor	0.142	-	-	95.27%
10	LightGBM	0.125	0.037	-	93.55%

จากผลการวัดประสิทธิภาพแบบจำลองตามตาราง 11 ผู้วิจัยได้นำผลลัพธ์ของแบบจำลองที่ดีที่สุดในแต่ละงานวิจัยมาแสดงในตาราง จะสังเกตได้ว่างานวิจัยในตารางที่ใช้ตัววัดประสิทธิเดียวกันกับงานวิจัยนี้โดยดูจาก MAE และ R-square ประกอบด้วยงานวิจัยที่ 9 ใช้แบบจำลอง Support Vector Regressor ได้ผลลัพธ์ MAE 0.142 และ R-square 95.27% , งานวิจัยที่ 10 ใช้แบบจำลองLightGBM ได้ผลลัพธ์ MAE 0.125 และ R-square 93.55% และงานวิจัยที่ 11 ใช้แบบจำลอง Random forest ได้ผลลัพธ์ MAE 0.74 , R-square 91% ซึ่งงานวิจัยนี้ใช้แบบจำลอง Random Forest ได้ผลลัพธ์ R-square 95.9% สามารถสรุปได้ว่างานวิจัยนี้ทำนายราคารถยนต์ได้ในระดับดี ซึ่งงานวิจัยนี้แก้ปัญหาที่แตกต่างจากงานวิจัยอื่นคือการเตรียมข้อมูลที่ละเอียดและปรับแต่งแบบจำลองโดยใช้ GridSearch เพื่อให้ได้ผลลัพธ์ดีขึ้น

บทที่ 3
วิธีดำเนินการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

3.1 ทำความเข้าใจปัญหาและความต้องการ

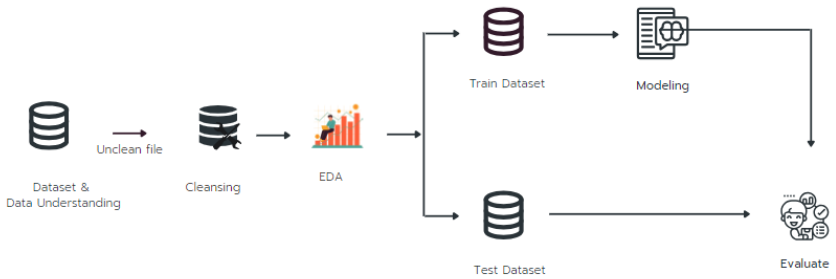
3.2 การทำความเข้าใจข้อมูล

3.3 การเตรียมข้อมูล

3.4 การสร้างแบบจำลอง

3.5 การประเมินแบบจำลอง

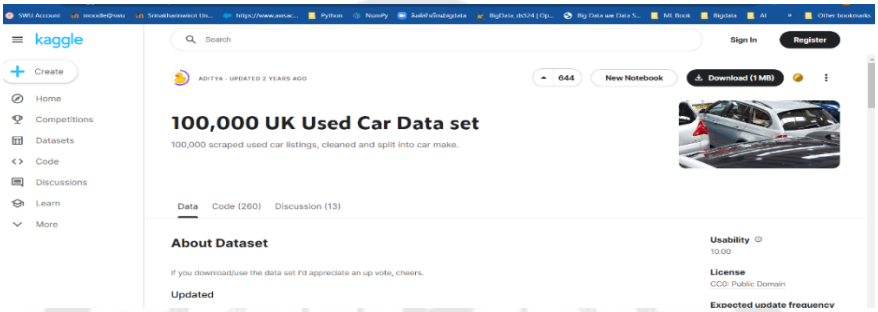
วัตถุประสงค์ในการทำวิจัยคือ การพัฒนาแบบจำลองการทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง มีขั้นตอนแบ่งเป็น 2 ส่วน ส่วนที่หนึ่งเป็นการทำความเข้าใจและเตรียมข้อมูล โดยนำข้อมูลมาผ่านกระบวนการต่างๆ เช่น ทำความสะอาดวิเคราะห์ข้อมูลเชิงสำรวจทำวิศวกรรมข้อมูลและเลือกคุณลักษณะ เพื่อให้ได้ชุดข้อมูลใหม่จากนั้นแบ่งชุดข้อมูลเป็น 2 ชุดคือ ชุดข้อมูลฝึก(Train) และชุดข้อมูลทดสอบ(Test) ในอัตราส่วน 80:20 เพื่อนำไปสร้างแบบจำลองต่อไป ดังภาพประกอบ 1



ภาพประกอบ 1 ภาพกระบวนการทำงานของแบบจำลอง

3.1 ทำความเข้าใจปัญหาและความต้องการ

การทำความเข้าใจปัญหาและกำหนดกรอบปัญหาเพื่อวางแผนการทำงาน ว่าต้องการนำข้อมูลที่มีไปแก้ปัญหาอะไรบ้างซึ่งเป็นขั้นตอนแรกของวัฏจักรวิทยาศาสตร์ข้อมูล โดยมีจุดประสงค์ของงานวิจัยนี้คือ การพัฒนาแบบจำลองการทำนายราคารถยนต์มือสอง โดยศึกษาและทดลองจากตัวอย่างข้อมูล รถยนต์มือสองของสหราชอาณาจักรซึ่งเป็นข้อมูลของผู้ใช้ Kaggle ชื่อว่า ADITYA ดังภาพประกอบ 2



ภาพประกอบ 2 ภาพบันทึกหน้าจอเว็บ Kaggle

3.2 การทำความเข้าใจข้อมูล

ชุดข้อมูลนี้คือข้อมูลรถยนต์มือสองของสหราชอาณาจักรโดยมีรายการรถมือสอง 100,000รายการซึ่งแยกเป็นไฟล์ตามผู้ผลิตรถยนต์แต่ละรายซึ่งเป็นข้อมูลจากผู้ใช้ Kaggle ชื่อว่า ADITY

3.2.1 ประกอบด้วยข้อมูลจำนวน 108540 แถวมีรายละเอียดดังนี้

1.จำนวนแถวข้อมูล 108540 แถว แสดงถึงข้อมูลของแต่ละบุคคลที่ไม่ซ้ำกัน

2.คอลัมน์จำนวน 10 คอลัมน์ ดังตาราง 12 ถึงตาราง 13 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description)

ตาราง 12 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

ตาราง 13 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description)

No	Name	Data type	Description
3	Price	Data type is Float	ราคารถยนต์
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
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

3.2.2 อ่านข้อมูล

จากชุดข้อมูลรถยนต์จากบริษัทผู้ผลิตต่างๆทั้งหมด13ไฟล์ประกอบด้วย audi,bmw,cclass,focus,ford,hyundi,merc,skoda,Toyota, uncleancclass, unclean focus, vauxhall, vw เรามีไฟล์ข้อมูลที่ไม่สะอาด 2 ไฟล์คือ ไฟล์ unclean cclass.csv และunclean focus.csv เราจะทำให้เหมือนกับไฟล์ที่เหลือ

3.3 การเตรียมข้อมูล

การเตรียมข้อมูล แบ่งออกเป็นการทำความสะอาดข้อมูล(Cleansing Data) การจัดรูปแบบข้อมูล (Formatting) และการวิเคราะห์ข้อมูลโดยทำให้ชุดข้อมูลอยู่ในรูปแบบที่สามารถนำไปใช้งานได้

3.3.1 การทำความสะอาดข้อมูล (Cleansing Data)

กระบวนการตรวจสอบ การแก้ไข หรือการลบ เพื่อให้รายการข้อมูลที่ไม่ถูกต้องออกไปจากชุดข้อมูล ตารางหรือฐานข้อมูลการทำความสะอาดข้อมูล เกิดขึ้นเนื่องจาก มีความไม่สอดคล้องของข้อมูล ซึ่งอาจเกิดจากข้อผิดพลาดของการบันทึกข้อมูล การส่งข้อมูล หรือการให้ความหมายของข้อมูลที่จัดเก็บแตกต่างกัน อาทิเช่น พิมพ์ผิด มีการเว้นว่างไม่กรอกข้อมูล กรอกข้อมูลที่ไม่สามารถอ้างอิงในระบบได้ หรือ เป็นตัวเลขที่ไม่มีทางเป็นไปได้ในความเป็นจริง ดังนั้นจึงต้องมีการบูรณาการกับฐานข้อมูลอื่น ๆ ไม่ว่าจะเป็น คลังข้อมูล หรือหลายฐานข้อมูล ซึ่งจะมีโอกาสสูงที่จะเกิด ข้อมูลที่ไม่สะอาด ขึ้นได้

- ไฟล์ที่ไม่สะอาด(unclean) , focus และ cclass ไม่มีคอลัมน์ภาษี(Tax)และ mpg ดังนั้นเราจึงเพิ่มค่าเริ่มต้นเป็น 0 คอลัมน์ภาษี (Tax)ในไฟล์ hyundi คือภาษีสกุลเงินปอนด์ Tax (£) ดังนั้นเราจึงเปลี่ยนเป็นภาษี(Tax)สกุลเงินดอลลาร์และเราจะเพิ่มคอลัมน์ company เป็นคอลัมน์สุดท้าย(ชื่อบริษัท)และจะอ่านข้อมูลที่ดีในข้อมูลดาต้าเฟรมและข้อมูลที่ไม่สะอาดในดาต้าเฟรมที่ไม่สะอาด จากนั้นควรจัดการข้อมูลที่ว่างและข้อมูลที่เป็นค่าNaN โดยใช้เมธอด dropna เพื่อลบแถวที่มี NaN อยู่ ดังภาพประกอบ 3

```
unclean=unclean.replace('nan', np.NaN)
unclean=unclean.replace('NAN', np.NaN)
unclean=unclean.replace('NaN', np.NaN)
unclean=unclean.replace('Unknown', np.NaN)
unclean.head()
```

ภาพประกอบ 3 แสดงถึงการแก้ไขค่าในคอลัมน์ที่แตกต่างกันให้เท่ากัน

- ทำความสะอาด mileage และ mileage2 และรวมกัน ในไฟล์ unclean โดยเปลี่ยนชนิดของคอลัมน์และใส่เลขไมล์(mileage)เป็นค่าสูงสุดของเลขไมล์(mileage)และเลขไมล์2 (mileage2) ดังภาพประกอบ 4

	model	year	price	transmission	mileage	fuel type	engine size	fuel type2	engine size2	company	tax	mpg
7692	Focus	2012.0	4000.0	Manual	107300.0	15	£20	Diesel	1560	unclean focus	0	0
8364	Focus	2017.0	11800.0	Manual	13601.0	NaN	£20	Petrol	999	unclean focus	0	0
3618	C Class	2017.0	23966.0	Automatic	29540.0	NaN	£205	Petrol	2000	unclean cclass	0	0
819	C Class	2019.0	31980.0	Semi-Auto	44.0	NaN	£145	Petrol	1.5	unclean cclass	0	0
2017	C Class	2019.0	20849.0	Semi-Auto	10925.0	NaN	£145	Diesel	2	unclean cclass	0	0
5567	Focus	2016.0	9450.0	Manual	15545.0	NaN	£0	Diesel	1.5	unclean focus	0	0
3225	C Class	2013.0	12000.0	Automatic	37236.0	31	£160	Petrol	1595	unclean cclass	0	0
2616	C Class	2015.0	16789.0	Semi-Auto	20100.0	37	£30	Diesel	2.1	unclean cclass	0	0
8065	Focus	2018.0	14299.0	Manual	7498.0	NaN	£145	Petrol	1	unclean focus	0	0
4266	Focus	2019.0	16998.0	Manual	14206.0	NaN	£150	Petrol	1	unclean focus	0	0

ภาพประกอบ 4 แสดงถึงการรวมกันของคอลัมน์ mileage และ mileage 2

- ทำความสะอาดคอลัมน์ขนาดเครื่องยนต์ (engine size)และคอลัมน์ขนาดเครื่องยนต์2 (engine size2) แล้วรวมเข้าด้วยกันในคอลัมน์ใหม่ขนาดเครื่องยนต์ (engine size) โดยลบค่าสตริงเพื่อเปลี่ยนประเภทของคอลัมน์จากนั้น ใส่ค่า engine Size ใน 'engine size2' ยกเว้นมี nan ใส่ค่าในคอลัมน์ engine Size จะได้ข้อมูลดังภาพประกอบ 5

	engine size	engine size2	engineSize
3712	£30	2.143	2.143
8955	£145	1.500	1.5
8935	1	NaN	1
5746	£145	2.000	2.0
1622	£145	2.100	2.1
8896	1	NaN	1
3575	2.1	NaN	2.1
1522	£300	4.000	4.0
9449	£260	1596.000	1596.0
192	£145	2.000	2.0

ภาพประกอบ 5 การทำความสะอาดข้อมูล การรวมเข้าด้วยกันในคอลัมน์ใหม่ขนาดเครื่องยนต์

- การ drop คอลัมน์เก่า คือ คอลัมน์ engine size 2 เป็นคอลัมน์ engine size ดังภาพประกอบ 6

	model	year	price	transmission	mileage	fuel type	fuel type2	company	tax	mpg	engineSize
2047	C Class	2019.0	27699.0	Semi-Auto	1714.0	NaN	Petrol	unclean cclass	0	0	1.5
4500	Focus	2015.0	9698.0	Manual	30149.0	NaN	Diesel	unclean focus	0	0	1.5
5676	Focus	2018.0	13580.0	Manual	7983.0	NaN	Petrol	unclean focus	0	0	1.0
4133	Focus	2017.0	13000.0	Manual	12456.0	Petrol	NaN	unclean focus	0	0	1.0
8532	Focus	2015.0	8250.0	Manual	62706.0	NaN	Diesel	unclean focus	0	0	1.5

ภาพประกอบ 6 การ drop คอลัมน์เก่าคือ คอลัมน์ engine size 2

- ทำความสะอาดข้อมูลคอลัมน์ ประเภทเชื้อเพลิง (fuel type) และประเภทเชื้อเพลิง2 (fuel type2)จากนั้นรวมคอลัมน์ประเภทเชื้อเพลิง และประเภทเชื้อเพลิง 2 เป็นคอลัมน์ใหม่แล้วลบคอลัมน์ประเภทเชื้อเพลิง 2 คอลัมน์เก่าออก จะได้ผลลัพธ์ดังภาพประกอบ 7

	model	year	price	transmission	mileage	company	tax	mpg	engineSize	fuelType
6240	Focus	2014.0	12661.0	Manual	34457.0	unclean focus	0	0	2.0	Petrol
1470	C Class	2019.0	39000.0	Semi-Auto	2857.0	unclean cclass	0	0	3.0	Petrol
8435	Focus	2016.0	16250.0	Manual	19500.0	unclean focus	0	0	2.0	Petrol
7475	Focus	2019.0	15950.0	Automatic	14190.0	unclean focus	0	0	1.0	Petrol
2857	C Class	2017.0	23749.0	Semi-Auto	12000.0	unclean cclass	0	0	2.1	Diesel
3342	C Class	2016.0	16799.0	Automatic	38785.0	unclean cclass	0	0	2.1	Diesel
2628	C Class	2019.0	29899.0	Semi-Auto	4405.0	unclean cclass	0	0	2.0	Diesel
865	C Class	2019.0	36980.0	Automatic	61.0	unclean cclass	0	0	2.0	Diesel
3918	C Class	2013.0	9995.0	Automatic	69000.0	unclean cclass	0	0	1.6	Petrol
6176	Focus	2013.0	5850.0	Manual	66020.0	unclean focus	0	0	1.6	Diesel

ภาพประกอบ 7 แสดงผลลัพธ์หลังทำความสะอาดข้อมูล

- เรากำลังจะเสร็จสิ้นขั้นตอนการทำความสะอาดข้อมูลแต่ค่า NAN ยังคงอยู่ต้องกำจัดข้อมูลที่มี NAN และ ลบทุกแถวที่มีค่าเป็น null อยู่ในคอลัมน์เนื่องจากไม่มีค่า คำสั่ง inplace=True จะทำให้ค่าใน unclean อันใหม่หลังจาก drop แล้วแทนที่อันเดิมทันที ดังภาพประกอบ 8

```
unclean=unclean.replace('NAN', np.NaN)
```

```
unclean.isnull().sum()
```

```
model           155
year            247
price           155
transmission    155
mileage         210
company           0
tax               0
mpg               0
engineSize      159
fuelType        155
dtype: int64
```

```
unclean.dropna(inplace=True)
```

ภาพประกอบ 8 แสดงถึงการจัดการกับข้อมูลที่มีNAN

- จากนั้นเปลี่ยนชนิดข้อมูล price เป็นint, year เป็นint, mileage เป็นint, tax เป็น int, mpgเป็นfloat และเปลี่ยนชื่อบริษัทจาก unclean focus เป็น focus และ unclean cclass เป็น cclass ดังภาพประกอบ 9

```
unclean.dropna(inplace=True)
unclean = unclean.astype({'price':'int','year':'int','mileage':'int','tax':'in
t','mpg':'float'})
```

```
unclean["company"] = unclean["company"].str.replace('unclean focus', 'focus')
unclean["company"] = unclean["company"].str.replace('unclean cclass', 'cclas
s')
```

ภาพประกอบ 9 แสดงถึงวิธีการเปลี่ยนชนิดข้อมูลและเปลี่ยนชื่อบริษัทผู้ผลิต

- ตรวจดูข้อมูลหลังจากหลังทำความสะอาดข้อมูลเสร็จดังภาพประกอบ 10

```
unclean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9301 entries, 0 to 9609
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   model         9301 non-null   object 
 1   year          9301 non-null   int64  
 2   price         9301 non-null   int64  
 3   transmission  9301 non-null   object 
 4   mileage       9301 non-null   int64  
 5   company       9301 non-null   object 
 6   tax           9301 non-null   int64  
 7   mpg           9301 non-null   float64
 8   engineSize    9301 non-null   float64
 9   fuelType      9301 non-null   object 
dtypes: float64(2), int64(4), object(4)
memory usage: 799.3+ KB
```

ภาพประกอบ 10 แสดงถึงข้อมูลหลังทำความสะอาดของไฟล์ที่ไม่ได้ทำความสะอาด (unclean)

- จากนั้นทำการดูข้อมูลไฟล์ที่สะอาดแล้วของชุดข้อมูล เพื่อตรวจดูจำนวนคอลัมน์ระหว่างทั้งสองไฟล์ว่าเท่ากันไหมคือไฟล์ unclean ชุดข้อมูลที่ยังไม่ได้ทำความสะอาดมาก่อน ซึ่งปัจจุบันได้ทำความสะอาดแล้วกับชุดข้อมูลที่สะอาดมาตั้งแต่ต้นคือชุดข้อมูล data ที่ประกอบด้วยชุดข้อมูลของบริษัทผู้ผลิตดังนี้ audi,bmw,cclass,focus,ford,hyundi,merc,skoda,Toyota,Vauxhall,vw ดังภาพประกอบ 11

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 108540 entries, 0 to 108539
Data columns (total 10 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   model         108540 non-null  object
 1   year          108540 non-null  int64
 2   price         108540 non-null  int64
 3   transmission  108540 non-null  object
 4   mileage       108540 non-null  int64
 5   fuelType      108540 non-null  object
 6   engineSize    108540 non-null  float64
 7   company       108540 non-null  object
 8   tax           108540 non-null  int64
 9   mpg           108540 non-null  float64
dtypes: float64(2), int64(4), object(4)
memory usage: 8.3+ MB
```

ภาพประกอบ 11 แสดงถึงจำนวนข้อมูล,ชื่อคอลัมน์และความจำที่ใช้

- ขั้นตอนสุดท้าย การต่อ Data Frame เมื่อเราได้ข้อมูลที่เข้ากันได้ จะทำการรวบรวมข้อมูลเข้าด้วยกันคือเอา 2 ชุดข้อมูลทั้ง 2 ชุดมาต่อกันแนวนอน ดังภาพประกอบ 12 ถึง 13

```
data = pd.concat([data,unclean], axis=0, ignore_index=True)
```

```
#Shuffle data
data=data.sample(frac = 1)
data.head()
```

ภาพประกอบ 12 แสดงถึงการการต่อ Data Frameทั้ง 2 ชุดข้อมูลเข้าด้วยกัน

	model	year	price	transmission	mileage	fuelType	engineSize	company	tax	mpg
58683	Mokka X	2019	13498	Manual	5895	Petrol	1.4	vauxhall	145	39.2
27255	Fabia	2017	8650	Manual	20378	Petrol	1.0	skoda	145	64.2
46168	Focus	2019	17500	Automatic	13895	Petrol	1.0	ford	150	42.8
6878	Focus	2019	26890	Manual	4000	Petrol	2.3	focus	0	0.0
70986	5 Series	2019	30000	Semi-Auto	1678	Petrol	2.0	bmw	145	50.4

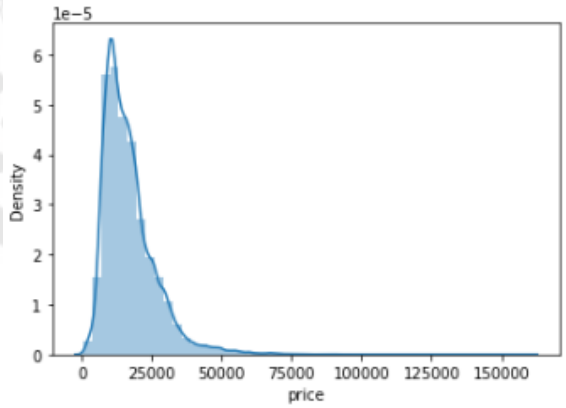
ภาพประกอบ 13 แสดงถึงข้อมูลที่รวมชุดข้อมูลทั้งสองเข้าด้วยกันแล้ว

3.3.2 **การวิเคราะห์ข้อมูลเชิงสำรวจ** (Exploratory Data Analysis)

EDAเป็นการวิเคราะห์ข้อมูลที่จำเป็นก่อนการนำข้อมูลไปใช้หรือนำไปวิเคราะห์เชิงลึกก่อนที่จะเข้าสู่กระบวนการนำข้อมูลไปพัฒนาแบบจำลองในลำดับต่อไปโดยประโยชน์ของการทำ EDA จะช่วยทำให้เราเข้าใจพื้นฐานเกี่ยวกับข้อมูลชุดนั้น และเป็นการตรวจความผิดพลาดของชุดข้อมูลได้อีกด้วย

3.3.2 .1 **การทำวิศวกรรมข้อมูล** (Feature Engineering) **และการวิเคราะห์ข้อมูล**

- เราสามารถพลอตกราฟราคาของชุดข้อมูลด้วย seabornโดยใช้คำสั่ง sns.distplot(data['price']); ซึ่ง Seaborn เป็นไลบรารีที่ใช้ Matplotlib เพื่อพล็อตกราฟซึ่งกราฟนี้ประกอบด้วยแกน X คือราคารถยนต์มือสองสกุลเงินคือปอนด์ และแกน Y คือ ความหนาแน่น จะได้ช่วงความหนาแน่นของรถยนต์ในชุดข้อมูลนี้คือรถยนต์ที่มีราคาโดยประมาณอยู่ในช่วง 16,000 ถึง 19,000 ปอนด์มีความหนาแน่นมากที่สุดดังภาพประกอบ 14

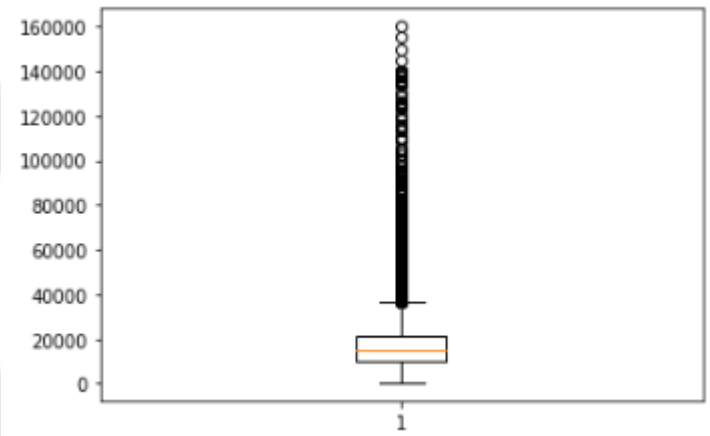


ภาพประกอบ 14 กราฟราคาของชุดข้อมูล

- จากข้อมูลกราฟภาพประกอบ 14 ไม่สามารถระบุค่าที่ชัดเจนได้ ดังนั้นจึง แสดงข้อมูลราคาและใช้ Box plot เพื่อให้ได้อ่านค่าได้ชัดเจนขึ้นดังภาพประกอบ 15 ถึง 16

```
count    117841.000000
mean      16967.226313
std        9660.817139
min         450.000000
25%       10395.000000
50%       14966.000000
75%       20980.000000
max      159999.000000
Name: price, dtype: float64
```

ภาพประกอบ 15 แสดงข้อมูลราคารถยนต์ของชุดข้อมูล



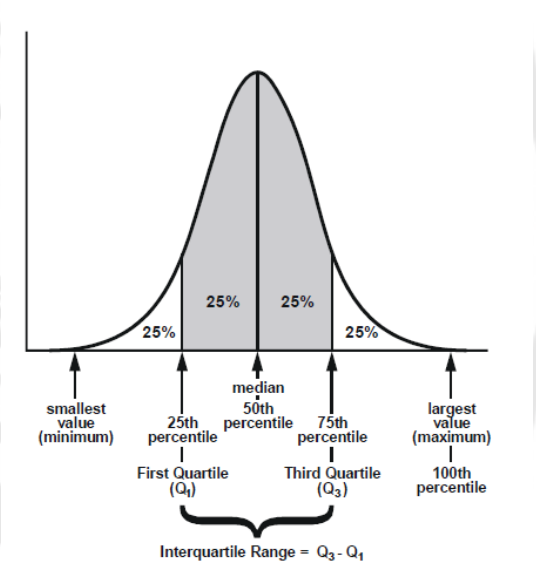
ภาพประกอบ 16 แสดงข้อมูลราคารถยนต์ของชุดข้อมูลด้วย box plot

- จะสังเกตได้ว่าข้อมูลราคารถยนต์มีค่าผิดปกติ (outlier)ซึ่งมีราคารถยนต์ที่สูงเกินไป ดังภาพประกอบ 17

```
count    117841.000000
mean      16967.226313
std        9660.817139
min         450.000000
25%       10395.000000
50%       14966.000000
75%       20980.000000
max      159999.000000
Name: price, dtype: float64
```

ภาพประกอบ 17 ข้อมูลราคารถยนต์ที่มีค่าผิดปกติ

- จากนั้นควรจัดการกับราคารถยนต์ที่มีค่าสูงเกินไป โดยการลบราคารถยนต์ที่สูงเกินไปออก ซึ่งก่อนจะลบราคารถยนต์ที่สูงเกินไป ต้องคำนวณหาค่า ราคารถยนต์ที่สูงเกินไป (data_price_max) โดยหาจากวิธีการคำนวณหาจากสูตร Outlier ดังนี้ Upper Anomaly = Q3+1.5*IQR กำหนดให้ data_price_max คือ Upper Anomaly,Q3 คือ จุดที่มีราคารถยนต์ 75% ของทั้งหมด จากสูตร IQR ย่อมากจาก Interquartile Range (IQR) คือการแบ่งการระจายตัวนั้นออกมาเป็น 4 ช่วง โดยการนำ data ทั้งหมดมาเรียงกันจากน้อยไปหามากและ mark จุดที่ 25% , 50% , 75% , 100% ของจำนวนราคารถยนต์ทั้งหมด ตามลำดับ ในที่นี้เราสนใจที่ค่า 25% และ 75% เป็นหลักเราจะนำทั้งสองจุดมาลบกันเพื่อให้ได้ค่า IQR ดังภาพประกอบ 18



ภาพประกอบ 18 Interquartile Rang(IQR)

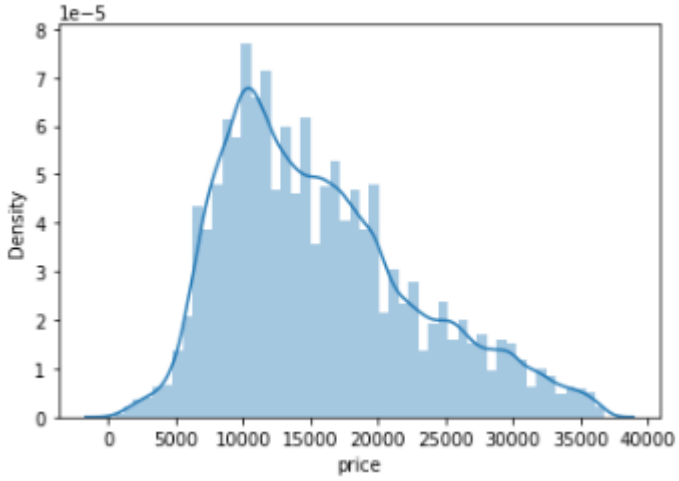
- จากนั้นมาเขียนcodeเพื่อหาค่า Q1 จากเปอร์เซ็นไทล์ราคารถยนต์ที่ 25% และหาค่า Q3 จากเปอร์เซ็นไทล์ราคารถยนต์ที่ 75 % เพื่อนำทั้งสองจุดมาลบกันเพื่อให้ได้ค่า IQR และหาค่าราคารถยนต์ที่สูงเกินไปดังภาพประกอบ 19

```
#remove data over price max
Q1 = np.percentile(data['price'], 25,
                   interpolation = 'midpoint')

Q3 = np.percentile(data['price'], 75,
                   interpolation = 'midpoint')
IQR = Q3 - Q1

data_price_max = Q3+1.5*IQR
data = data.drop(data[data.price>=data_price_max].index)
data.shape
```

```
(113473, 10)
```

ภาพประกอบ 19 การเขียนโค๊ดเพื่อหาค่าและดรอปราคารถยนต์ที่สูงเกินไป

- ทำการพลอตกราฟเพื่อดูราคารถยนต์หลังลบข้อมูลราคารถยนต์ที่สูงเกินไปโดย แกน x คือความหนาแน่น และ แกน Y คือราคารถยนต์ ดังภาพประกอบ 20

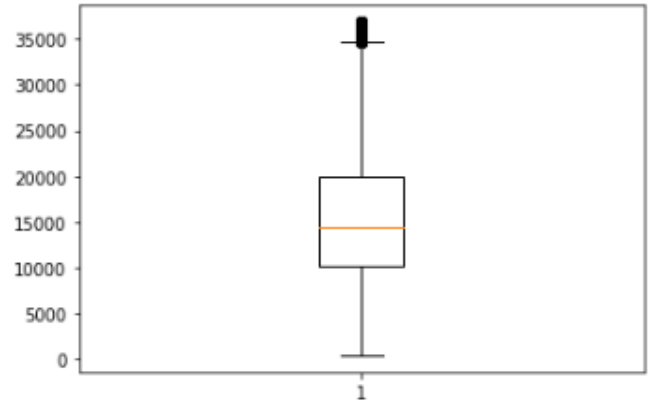


ภาพประกอบ 20 กราฟราคารถยนต์

- ทำการปริ้นแสดงข้อมูล เลขไมล์สะสมของรถยนต์ (mileage) และพลอตกราฟบ็อกพล็อตดังภาพประกอบ 21 ถึง 22

```
count    113473.000000
mean      23603.906207
std       21248.421913
min           1.000000
25%        8179.000000
50%       17918.000000
75%       32840.000000
max      323000.000000
Name: mileage, dtype: float64
```

ภาพประกอบ 21 ข้อมูลของเลขไมล์สะสม (mileage)



ภาพประกอบ 22 กราฟบ็อกพล็อตผลลัพธ์
การแสดงข้อมูลเลขไมล์สะสม (mileage)

- **คอลัมน์ใดที่ส่งผลต่อราคารถยนต์มือสองมากที่สุด?** เราอยากรู้ว่าชุดข้อมูลเรามีความสัมพันธ์กันอย่างไรสามารถหา Correlation ดังภาพประกอบ 22

```
data.corr()['price'].sort_values(ascending=False)
```

```
price         1.000000
year          0.554287
engineSize    0.412974
tax           0.198238
mpg          -0.182732
mileage      -0.470077
Name: price, dtype: float64
```

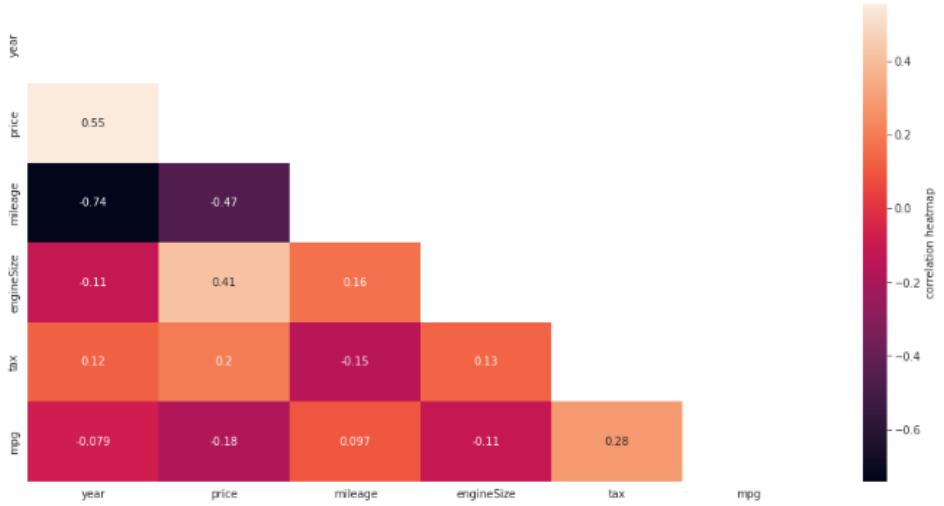
ภาพประกอบ 22 แสดงถึง Correlation ราคารถยนต์กับคอลัมน์อื่นๆ

- แต่ค่าออกมาเป็นตัวเลขอาจจะดูยาก เราสามารถพลอตกราฟสวยๆด้วย seabornสามารถใช้โค้ด ดังภาพประกอบ 23

```
corr  = data.corr()
f = plt.figure(figsize = (16,8))
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("white"):
    ax1 = sns.heatmap(corr, annot=True, mask=mask, cbar_kws={'label': 'correla
tion heatmap'})
```

ภาพประกอบ 23 โค้ดการพลอตกราฟ correlation heatmap

- จากการวิเคราะห์ข้อมูลคอลัมน์อื่นๆที่ส่งผลต่อราคามากที่สุด พบว่าคอลัมน์ ปีรถ (year),ขนาดเครื่องยนต์ (engine Size), เลขไมล์สะสม (mileage) ส่งผลต่อราคารถยนต์มือสองมากที่สุดตามลำดับ Correlation หรือ ค่าสหสัมพันธ์ เป็นการดูทิศทางความสัมพันธ์ระหว่างตัวแปร 2 ตัว โดยมี Correlation Coefficient (r) หรือค่าสัมประสิทธิ์สหสัมพันธ์ เป็นตัวบ่งชี้ถึงความสัมพันธ์นี้ ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์นี้จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 ซึ่งหากมีค่าใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากมีค่าใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันโดยตรงอย่างมาก และหากมีค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกันดังภาพประกอบ 24



ภาพประกอบ 24 Correlation heatmap ด้วย seaborn

- **เราจะวิเคราะห์ข้อมูลเชิงสำรวจว่าบริษัทใดมีราคาเฉลี่ยมากกว่ากัน?**ซึ่งสามารถเขียนโค้ดคิวรี่ราคาเฉลี่ยจากชุดข้อมูลโดยการจัดกลุ่มข้อมูลของริษัทผู้ผลิตใช้คำสั่ง group by เพื่อจัดกลุ่มข้อมูลตามบริษัทผู้ผลิตดังภาพประกอบ 25 ถึง 26

```
d=pd.DataFrame(data.groupby('company').mean().sort_values(by='price',ascending=False)
['price'])
d
```

ภาพประกอบ 25 โค้ดคิวรี่ราคาเฉลี่ยจากชุดข้อมูลโดยการจัดกลุ่มข้อมูลของริษัทผู้ผลิต

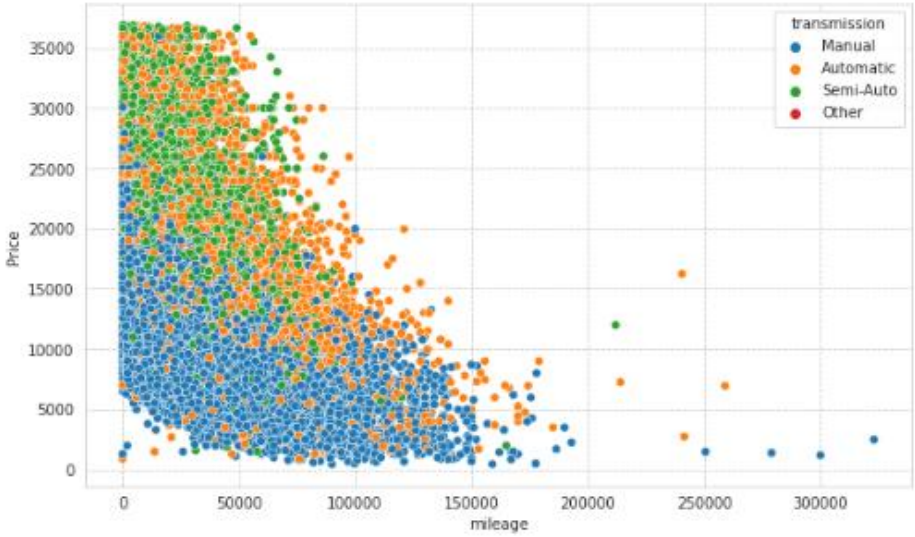
- จากวิเคราะห์ข้อมูลเชิงสำรวจพบว่าบริษัทที่มีราคารถยนต์เฉลี่ยมากที่สุดคือบริษัท cclass รองลงมาคือบริษัท merc ย่อมากจาก merczedezbenz ,บริษัท audi และลำดับสุดท้ายคือ bmw ดังภาพประกอบ 26

Company	Price
cclass	22287
merc	21594
audi	20140
bmw	20048
vw	16277
skoda	14240
focus	13600
hyundi	12681
toyota	12286
ford	12238
vauxhall	10401

ภาพประกอบ 26 แสดงถึงผลลัพธ์การคิวรี่ราคารถยนต์เฉลี่ยจากชุดข้อมูล

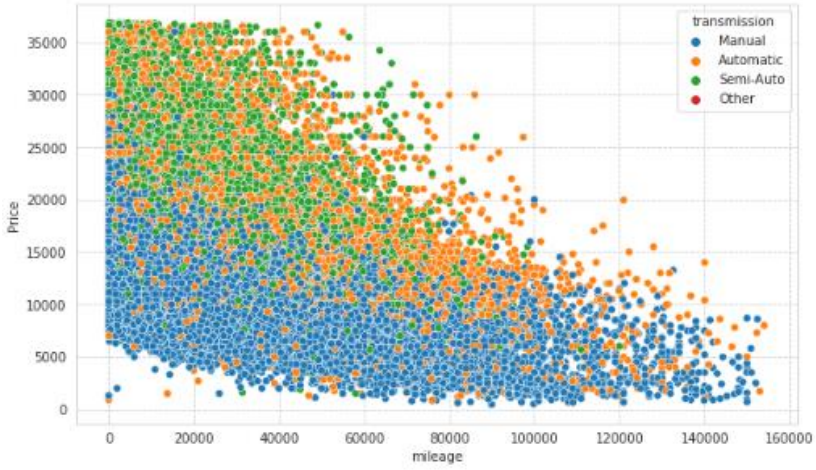
- **แต่ตอนนี้เราต้องการทราบว่าความสัมพันธ์ระหว่างระยะทางสะสม (mileage)และราคามีความสัมพันธ์กันอย่างไร?**โดยการพลอตกราฟใช้ scatterplotให้ label แกน x คือเลขไมล์สะสม (mileage) แกนyคือราคารถยนต์(price)และให้แสดงตามข้อมูลตามเกียร์รถยนต์(Transmission) จะเห็นได้

ว่าระยะทางสะสม(mileage)ตั้งแต่ 0-50,000 กิโลเมตร ราคาจะไม่ได้รับผลกระทบอย่างมีนัยสำคัญ แต่หากระยะทางสะสม(mileage) สูงเกิน 50,000 กิโลเมตร ราคารถยนต์จะมีราคาต่ำกว่าระยะทางที่น้อยกว่า 50,000 กิโลเมตร และรถยนต์เกียร์อัตโนมัติจะมีราคาสูงสุด จากนั้นเป็นรถกึ่งอัตโนมัติจากนั้นเป็นรถเกียร์ธรรมดาดังภาพประกอบ 27



ภาพประกอบ 27 กราฟแสดงความสัมพันธ์ระหว่างระยะทางและราคา

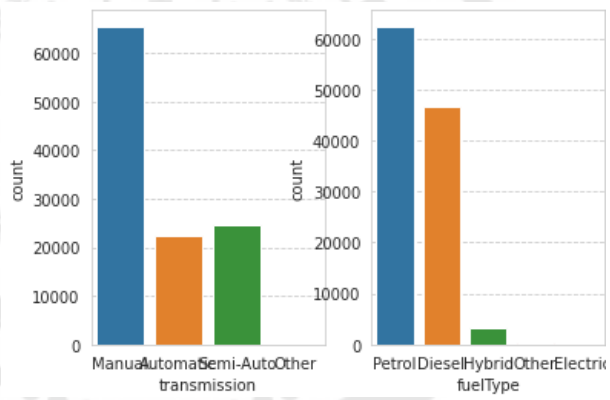
จากภาพประกอบ 27 จะเห็นได้ว่ามีข้อมูลผิดปกติเล็กน้อย จึงทำการลบออก จากนั้นพลอตกราฟอีกครั้งจะได้กราฟดังภาพ 28



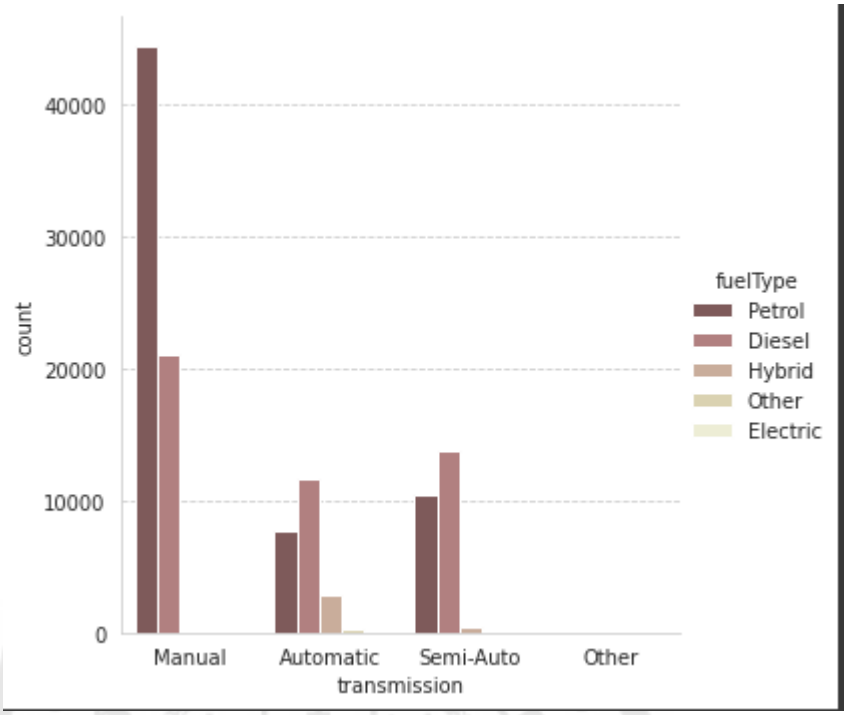
ภาพประกอบ 28 กราฟแสดงความสัมพันธ์ระหว่างระยะทางและราคาที่สมบูรณ์

จากภาพประกอบ 28 ผู้เขียนต้องการที่จะดูข้อมูลช่วงเลขไมล์สะสมที่กระจุกอยู่ให้ชัดเจนขึ้นเลยกำหนดเลขไมล์สะสม0-160000ส่วนช่วงอื่นๆก็ไม่ได้ตัดทิ้งต้องดูภาพประกอบ 27

- **การพล็อตกราฟความสัมพันธ์ระหว่างเกียร์รถ(transmission) กับชนิดเชื้อเพลิงมีความสัมพันธ์อย่างไร?** จะเห็นได้ว่ากราฟฝั่งซ้ายของภาพประกอบ 29 จากชุดข้อมูลจำนวนการใช้รถเกียร์ธรรมดามากกว่าเกียร์อื่นๆ รองลงมาคือเกียร์ semi-auto และสุดท้ายคือเกียร์ออโต้ หากสังเกตกราฟทางขวามือจำนวนการใช้ชนิดเชื้อเพลิง petrol มากที่สุด,รองลงมาคือชนิดเชื้อเพลิงดีเซลและสุดท้ายชนิดเชื้อเพลิง hybrid ต่อมาความความสัมพันธ์ระหว่างเกียร์รถ(transmission) กับชนิดเชื้อเพลิงมีความสัมพันธ์โดยรถเกียร์ธรรมดาManual ใช้น้ำมันpetrol, รถเกียร์ automatic ใช้น้ำมันดีเซลและรถเกียร์semi-autoใช้น้ำมันดีเซลดังภาพประกอบ 30

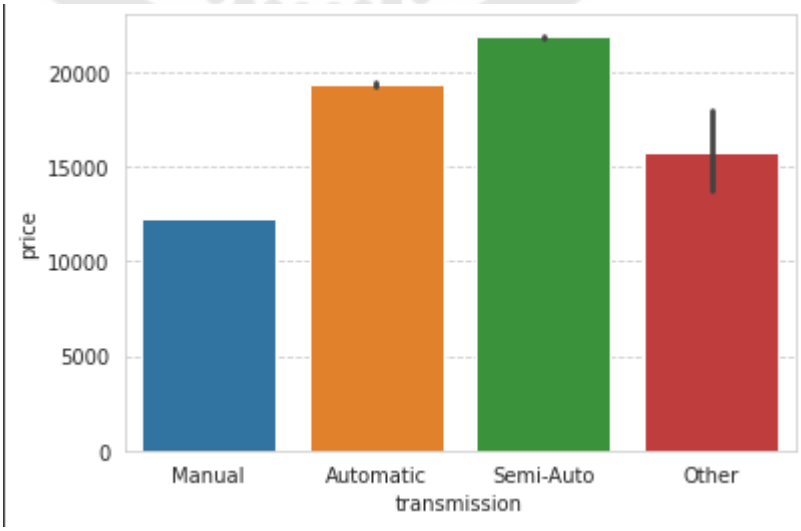


ภาพประกอบ 29 แสดงกราฟจำนวนการใช้งานเกียร์รถยนต์และกราฟแสดงจำนวนการใช้ชนิดเชื้อเพลิง



ภาพประกอบ 30 กราฟแสดงความสัมพันธ์เกียร์ชนิดต่างๆกับจำนวนการใช้งาน

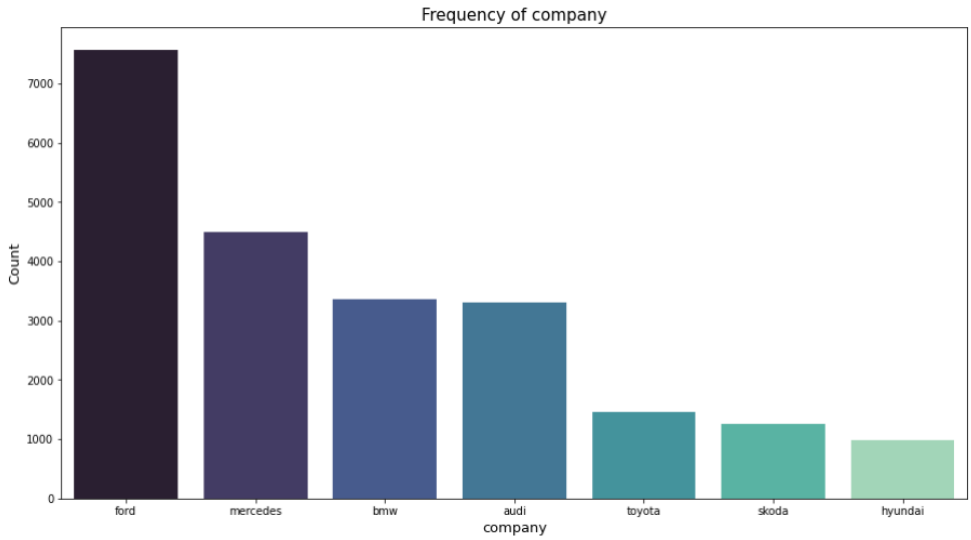
- จากนั้นทำการพลอตกราฟเพื่อแสดงความสัมพันธ์ระหว่างเกียร์รถยนต์ชนิดต่างๆ กับราคารถยนต์มือสอง จะเห็นได้ว่ารถยนต์เกียร์ Semi-Auto มีราคาที่สูงกว่า รองลงมาคือรถยนต์เกียร์ออโต้ รองลงมาคือเกียร์อื่นๆและสุดท้ายคือรถยนต์เกียร์ธรรมดา (Manual) ราคาถูกที่สุดดังภาพประกอบ 31



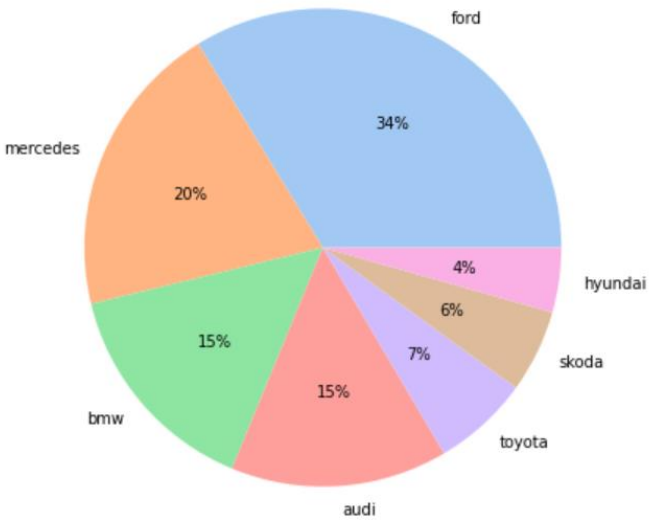
ภาพประกอบ 31 กราฟแสดงความสัมพันธ์เกียร์ชนิดต่างๆกับราคารถยนต์

3.3.2 .2 การวิเคราะห์ข้อมูลเชิงสำรวจปัจจัยอื่นๆที่เกี่ยวข้องกับรถยนต์มือสอง

- การตรวจสอบบริษัทที่ได้รับความนิยมสูงสุดคือบริษัทอะไร? ใช้seabornในการพลอตกราฟจะได้ว่าบริษัท Ford มีจำนวนการใช้งานมากที่สุดซึ่งได้รับความนิยมสูงสุดของชุดข้อมูลนี้ดังภาพประกอบ 32 ถึงภาพประกอบ 33



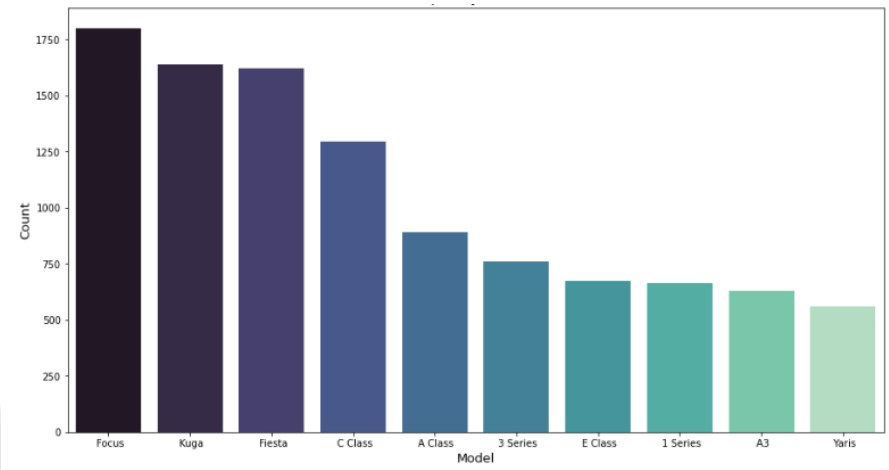
ภาพประกอบ 32 แสดงข้อมูลบริษัทที่ได้รับความนิยมสูงสุด



ภาพประกอบ 33 บริษัทที่ได้รับความนิยมสูงสุดเป็นเปอร์เซ็น

จากภาพประกอบ 33 บริษัทรถยนต์ที่โดดเด่นที่สุดในตลาดรถยนต์ในสหราชอาณาจักรคือฟอร์ด และบริษัทรถยนต์ที่น้อยที่สุดคือ hyundai

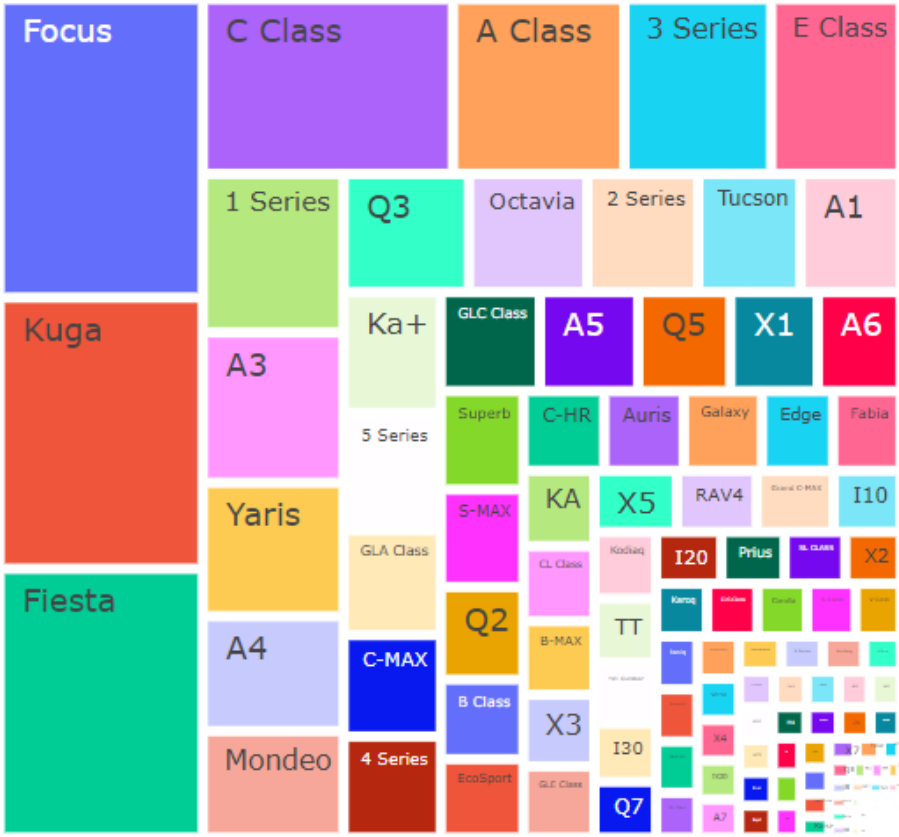
- ตรวจสอบรุ่นยอดนิยมสำหรับรถแต่ละรุ่น โดยการพลอตกราฟ ให้แกน y คือจำนวนการใช้งานรถยนต์ แกน x คือรถยนต์แต่ละรุ่น ดังภาพประกอบ 34



ภาพประกอบ 34 กราฟความนิยมการใช้งานรถยนต์แต่ละรุ่น

จากภาพประกอบ 34 แสดงถึงความนิยมการใช้รถรุ่นต่างๆของชุดข้อมูลนี้โดยประกอบด้วยรถยนต์รุ่น Focus, Kuga, Fiesta จากแบรนด์ Ford และ Cclass, AClass, EClassจากแบรนด์ Benz และรุ่น 1 Series,3 Series จากแบรนด์ BMW , รถยนต์รุ่นA 3 จากแบรนด์ Audi และสุดท้ายรถยนต์รุ่น Yaris จากแบรนด์ Toyota

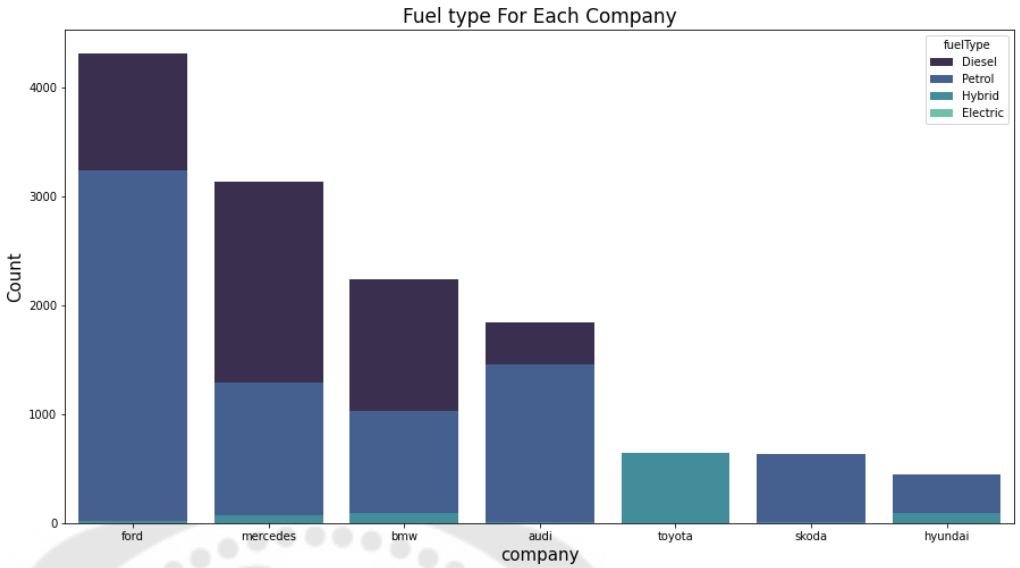
- นำมาพลอตกราฟแบบ Tree maps เพื่อสามารถดูว่ารถยนต์รุ่นที่ได้รับความนิยมสูงสุดในตลาดรถยนต์ในสหราชอาณาจักรได้มากขึ้น ดังภาพประกอบ 35



ภาพประกอบ 35 ตรวจสอบรุ่นรถยนต์ยอดนิยม

จากภาพประกอบ 35 จะเห็นได้ว่ารุ่น Focus,Kuga,Fiesta จากแบรนด์ Ford และ Cclass,AClass,EClassจากแบรนด์ Benz และรุ่น 1 Series,3 Series จากแบรนด์ BMW , รถยนต์รุ่น A 3 จากแบรนด์ Audi และสุดท้ายรถยนต์รุ่นYaris จากแบรนด์ Toyota ซึ่งผลลัพธ์การพลอตกราฟแบบTreemapsนี้สอดคล้องกับผลลัพธ์ดังภาพประกอบ 34 แต่การพลอตกราฟแบบ Tree maps จะเห็นรุ่นรถยนต์ได้มากกว่า

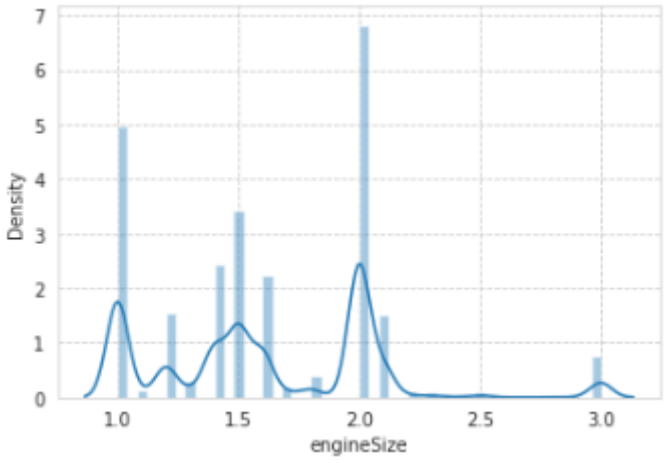
- ตรวจสอบประเภทเชื้อเพลิงส่วนใหญ่ที่ใช้สำหรับผู้ผลิตรถยนต์แต่ละราย ดังประกอบ 36



ภาพประกอบ 36 กราฟแสดงชนิดเชื้อเพลิงยอดนิยมของบริษัทผู้ผลิต

จากภาพประกอบ36 เป็นกราฟเปรียบเทียบการใช้เชื้อเพลิงต่างๆโดยวิเคราะห์แยกแต่ละแบรนด์เช่น รถยนต์แบรนด์Ford มีการใช้เชื้อเพลิง Petrol มากกว่าเชื้อเพลิงDieselและเชื้อเพลิงอื่นๆ ,รถยนต์แบรนด์ mercedes มีการใช้เชื้อเพลิง Diesel มากกว่าเชื้อเพลิงอื่นๆเป็นต้น

- จากนั้นมาดูขนาดเครื่องยนต์กันว่า ในชุดข้อมูลนี้ขนาเครื่องยนต์ไหนมีความนิยมมากกว่ากัน ดังภาพประกอบ 37



ภาพประกอบ 37 กราฟความนิยมการใช้งานขนาดเครื่องยนต์จากชุดข้อมูลนี้

3.4 การสร้างแบบจำลอง

ขั้นตอนนี้เป็นขั้นตอนของการนำข้อมูลที่จัดเตรียมไว้มาวิเคราะห์ด้วยเทคนิคต่างๆ ตามความเหมาะสม โดยมากจะเป็นการวิเคราะห์ด้วยเทคนิคมากกว่าหนึ่งแบบเพื่อประเมินหาเทคนิคที่ให้ค่าการวิเคราะห์ที่เป็น ไปตามวัตถุประสงค์และมีประสิทธิภาพมากที่สุด จากการผ่านขั้นตอนการเตรียมข้อมูลจะได้ชุดข้อมูลจำนวน 110,061 แถวและ 12 คอลัมน์ จากนั้นทำการแบ่งข้อมูล (split data) ออกเป็น 2 ส่วนคือ Training Set 80% และ Test set 20%

3.4.1 การแบ่งข้อมูลสำหรับฝึกและทดสอบ (Train and Test Split)

การแบ่งข้อมูลออกเป็น 2 ส่วนคือ Train และ Test โดย Train คือชุดข้อมูลการฝึกสำหรับสร้างแบบจำลองมีจำนวน 93551 แถว, 11 คอลัมน์ และ Test คือชุดข้อมูลสำหรับทดสอบประสิทธิภาพของแบบจำลองมีจำนวน 16510 แถว, 11 คอลัมน์ จากนั้นแสดงจำนวนแถวและจำนวนคอลัมน์ของ training set และ Train set ดังภาพประกอบ 38

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, ra
ndom_state=42)
print("data is ",data.shape)
print("X is ",X.shape)
print("X_train is ",X_train.shape)
print("y_train is ",y_train.shape)
print("X_test is ",X_test.shape)
print("y_test is ",y_test.shape)
```

```
data is  (110061, 12)
X is  (110061, 11)
X_train is  (93551, 11)
y_train is  (93551,)
X_test is  (16510, 11)
y_test is  (16510,)
```

ภาพประกอบ 38 การ split data

3.4.2 Regression for car prices เราจะทำการนำเข้า แบบจำลองจาก Scikit learn

แบบจำลองที่นำเข้ามาใช้คือ Linear Regression,RandomForestRegressor ,Ridge และ DecisionTreeRegressor และนำเข้า metrics เพื่อวัดผลข้อมูลประกอบโดย mean_absolute_error mean_squared_error และmean_squared_error ดังปภาพประกอบ 39 ถึง ภาพประกอบ 40

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import median_absolute_error
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import Ridge
from sklearn.tree import DecisionTreeRegressor
```

ภาพประกอบ 39 การนำเข้าแบบจำลองและนำเข้า metrics

- ขั้นตอนที่สำคัญที่สุด แต่สั้นที่สุด นั่นคือการฝึกโมเดลด้วย Linear regression algorithm ดังภาพประกอบ 40

```
LinearRegressionModel = LinearRegression(fit_intercept=True, copy_X=True,n_j
obs=-1)
forest = RandomForestRegressor(criterion = 'mse',random_state = 1,n_jobs = -
1)
rr = Ridge(alpha=0.0001)
dt = DecisionTreeRegressor(random_state = 100,max_depth=2 )
model=forest
model.fit(X_train, y_train)
```

ภาพประกอบ 40 แสดงการฝึกโมเดล

- รายละเอียดการคำนวณและแสดงข้อมูลคะแนนจากการฝึกโมเดล (Train) และทดสอบ (Test) ดังภาพประกอบ 41

```
#Calculating Details
print('Linear model Train Score is : ' , model.score(X_train, y_train))
print('Linear model Test Score is : ' , model.score(X_test, y_test))
print('----------------------------------------------------')
```

```
Linear model Train Score is :  0.9940897077050586
Linear model Test Score is :  0.9624784047380534
----------------------------------------------------
```

ภาพประกอบ 41 แสดงการคำนวณหาค่า R-squared

- ทำการ Predict value จาก x_test และ x_train หลังจากนั้นทำการแสดงข้อมูลจากการ Predict value กับข้อมูลจริงดังภาพประกอบ 42

```
y_pred = model.predict(X_test)
yt_pred = model.predict(X_train)
print('Predicted Value for Linear Regression is : \n' ,np.around(y_pred[:1
0],3))
print('real Value for Linear Regression is : \n' ,np.around(np.array(y_test
[:10]),3))
#----------------------------------------------------
```

```
Predicted Value for Linear Regression is :
 [ 0.071 -1.221  0.154 -0.215 -0.532  0.238 -0.333 -0.271 -1.255 -0.237]
real Value for Linear Regression is :
 [ 0.042 -1.231  0.254 -0.418 -0.383  1.102 -0.312 -0.312 -1.09  -0.253]
```

ภาพประกอบ 42 แสดงข้อมูลการ Predict ค่า

- จากนั้นทำการคำนวณค่า metrics ต่างๆสำหรับวัดผลดังภาพประกอบ 43 และผลลัพธ์ตาราง 13

```
#Calculating Mean Absolute Error
MAEValue = mean_absolute_error(y_test, y_pred, multioutput='uniform_averag
e') # it can be raw_values
print('Mean Absolute Error Value is : ', MAEValue)
#----------------------------------------------------
#Calculating Mean Squared Error
MSEValue = mean_squared_error(y_test, y_pred, multioutput='uniform_average')
# it can be raw_values
print('Mean Squared Error Value is : ', MSEValue)
#----------------------------------------------------
#Calculating Median Squared Error
MdSEValue = median_absolute_error(y_test, y_pred)
print('Median Squared Error Value is : ', MdSEValue )
```

```
Mean Absolute Error Value is :  0.1314113340228776
Mean Squared Error Value is :  0.03747610893274977
Median Squared Error Value is :  0.08876607441820165
```

ภาพประกอบ 43 ค่าMAE, MSE, Medain Square Error

3.5 การประเมินผลแบบจำลอง

ขั้นตอนนี้เป็นขั้นตอนของการทดสอบใช้แบบจำลองการวิเคราะห์ข้อมูลที่ได้ก่อนนำไปใช้งานจริง โดยการใช้องค์ความรู้ในการวัดค่าประสิทธิภาพของแบบจำลองจะได้ผลลัพธ์ดังตาราง 14

ตาราง 14 การประเมินแบบจำลองโดยใช้อัลกอริทึมการถดถอยเชิงเส้น Linear regression

Model	MAE	MAPE	R-square
LinearRegressionModel	1,827.58 £	15%	0.87

สำหรับการประเมินประสิทธิภาพแบบจำลองผลตัวแปรที่เป็นข้อมูลเชิงปริมาณ ใช้สูตรในการวัดคือ MAE, MAPE,MSE และ R-Square

MAE (Mean Absolute Error) เป็นค่าเฉลี่ยของความคลาดเคลื่อนในการทำนายของโมเดล คือค่าสัมบูรณ์ของความแตกต่างระหว่างค่าที่ทำนายได้กับค่าจริง โดยจะมีหน่วยเดียวกับตัวแปรที่กำลังทำนาย ซึ่งเป็นค่าบวกที่ยิ่งน้อยแสดงถึงการทำนายที่แม่นยำมากขึ้น หาก MAE มีค่าเป็น 0 แสดงว่าการทำนายเป็นไปตามค่าจริงอย่างแม่นยำที่สุด ในงานวิจัยนี้ใช้หน่วยของตัวแปร Price คือ ปอนด์ซึ่งเป็นสกุลเงินของสหราชอาณาจักร

สูตรคำนวณ MAE คือ

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \qquad (3.1)$$

โดยที่ y_i คือค่าจริง (actual values) และ $\hat{y_i}$ คือการทำนาย (predicted values)

MAPE (Mean Absolute Percentage Error) เป็นค่าเฉลี่ยของค่าสัมบูรณ์ของร้อยละของความต่างระหว่างค่าที่ทำนายได้กับค่าจริง ค่า MAPE จะแสดงให้เห็นถึงอัตราการคลาดเคลื่อนของการทำนายในรูปของร้อยละ ค่า MAPE มีความเหมาะสมในกรณีที่ค่าตัวแปรที่ทำนายมีขนาดไม่เท่ากัน เช่น การทำนายราคาของรถยนต์มือสองที่มีราคาสูงและต่ำต่างกันเป็นต้น

สูตรคำนวณ MAPE คือ

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y_i}|}{y_i} \times 100\% \qquad (3.2)$$

โดยที่y_iคือค่าจริง (actual values) และ $\widehat{y_i}$ คือการทำนาย (predicted values)

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

สูตรของ MSE คือ $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad (3.3)$$

เมื่อ

n = จำนวนข้อมูลในชุดข้อมูล

y_i= ค่าเป้าหมายของข้อมูลคือค่าที่จริงหรือค่าที่ต้องการทำนาย

$\widehat{y_i}$ = ค่าที่โมเดลทำนายขึ้นมา

R-Square (Coefficient of Determination) เป็นค่าวัดประสิทธิภาพของโมเดล โดยค่า R-Square เป็นค่าที่ใช้วัดความสัมพันธ์ของค่าที่ทำนายได้กับค่าจริง ค่า R^2 อยู่ในช่วงระหว่าง 0 ถึง 1 โดย R^2 = 1 หมายความว่าการทำนายเป็นไปตามค่าจริงอย่างแม่นยำที่สุด ส่วน R^2 = 0 หมายความว่าการทำนายไม่ดีเลย หรือไม่สามารถอธิบายความเปลี่ยนแปลงของตัวแปรตามได้เลย ค่า R^2 นับได้จากสัดส่วนของความคลาดเคลื่อนที่อธิบายได้ของตัวแปรตามที่ได้จากการทำนาย สังเกตว่า R^2 มีความเหมาะสมในกรณีที่ค่าตัวแปรที่ทำนายมีการกระจายแบบเส้นตรง

สูตรของ R-Square คือ

$$R^2 = 1 - \frac{SSR}{SST} \quad (3.4)$$

เมื่อ SSR = ผลรวมของค่า residuals ยกกำลังสอง หรือ $(y_i - \widehat{y_i})^2$ ผลรวมของความคลาดเคลื่อนระหว่างการทำนายและค่าเฉลี่ยของค่า

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

y_i= ค่าเป้าหมายของข้อมูลคือค่าที่จริงหรือค่าที่ต้องการทำนาย

$\widehat{y_i}$ = ค่าที่โมเดลทำนายขึ้นมา$\bar{y}$= ค่าเฉลี่ยของค่าตัวแปรตาม

บทที่ 4 การทดลอง

ในงานวิจัยการทำนายราคารถยนต์มือสอง โดยใช้ชุดข้อมูลรายการรถยนต์มือสองของตลาดสหราชอาณาจักร ด้วยเทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามกระบวนการต่างๆ ตลอดจนวัดประสิทธิภาพ เพื่อให้บรรลุจุดประสงค์ของการวิจัยที่ได้กำหนดไว้ดังนี้

4.1 การปรับจูนแบบจำลอง

4.2 ผลลัพธ์จากการวัดประสิทธิภาพแบบจำลอง

4.3 ผลลัพธ์ของคุณลักษณะสำคัญที่มีผลต่อการทำนาย

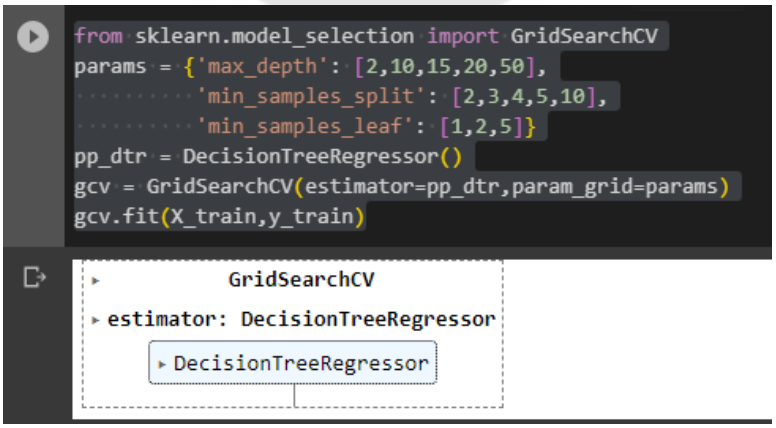
4.4 ผลลัพธ์จากการศึกษาปัจจัยที่มีผลต่อราคารถยนต์มือสอง

4.5 ผลลัพธ์การวิเคราะห์ข้อมูลโดยแยกแต่ละแบรนด์

4.6 การวัดความสำคัญของแต่ละคุณสมบัติในการทำนายผลของแบบจำลอง

4.1 การปรับจูนแบบจำลอง

การปรับแต่งแบบจำลองต้นไม้ (Decision Tree) โดยใช้เทคนิค Grid Search Grid Search เป็นเทคนิคที่ใช้ในการค้นหาค่าพารามิเตอร์ที่เหมาะสมสำหรับแบบจำลอง โดยการกำหนดช่วงค่าที่สนใจให้กับแต่ละพารามิเตอร์ และทดลองความเป็นไปได้ทุกคอมบินเนชั่นของพารามิเตอร์ดังกล่าว จากนั้นวัดประสิทธิภาพของแบบจำลองในแต่ละคอมบินเนชั่น และเลือกคอมบินเนชั่นที่มี ผลลัพธ์ที่ดีที่สุดสำหรับแบบจำลองดังภาพประกอบ 44



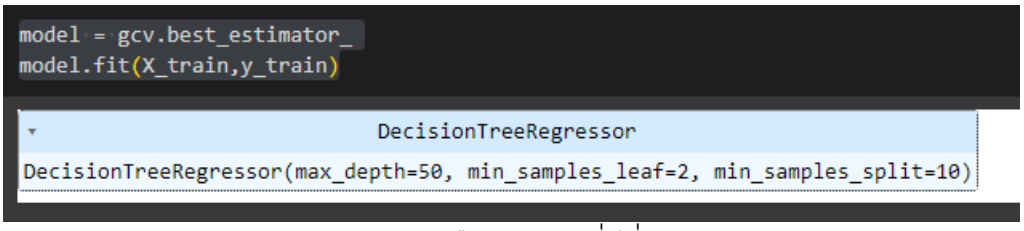
ภาพประกอบ 44 การใช้ Grid Search CV สำหรับค้นหาค่าพารามิเตอร์ที่เหมาะสม

จากภาพประกอบ 44 from sklearn.model_selection import GridSearchCV: เรา import GridSearchCV จาก sklearn.model_selection เพื่อใช้ในการทำ Grid Search สำหรับค้นหาค่าพารามิเตอร์ที่เหมาะสมสำหรับแบบจำลอง สามารถอธิบายโค๊ดดังนี้

- params: เป็นพารามิเตอร์ที่เรากำหนดให้ Grid Search ทดลองค้นหา ซึ่งประกอบด้วย max_depth, min_samples_split, และ min_samples_leaf ที่เป็นค่าที่เราสนใจในการปรับแต่งแบบจำลอง
- pp_dtr = DecisionTreeRegressor():เราสร้างตัวแบบ Decision Tree Regressor เพื่อใช้เป็นตัวแบบหลักที่จะถูกปรับแต่ง
- gcv = GridSearchCV(estimator=pp_dtr, param_grid=params): เราสร้างอ็อบเจ็กต์ GridSearchCV โดยระบุตัวแบบหลักที่จะปรับแต่ง (estimator) เป็น pp_dtr และกำหนดพารามิเตอร์ที่จะทดลอง (param_grid) เป็น params
- gcv.fit(X_train, y_train): เริ่มกระบวนการ Grid Search โดยใช้ข้อมูลการฝึกอบรม (X_train) และเป้าหมายการฝึกอบรม (y_train) โมเดลจะถูกปรับแต่งด้วยค่าพารามิเตอร์ที่ต่างกันในแต่ละรอบของ Grid Search เพื่อหาค่าพารมิเตอร์ที่ให้ประสิทธิภาพที่ดีที่สุด

ดังนั้นเมื่อโค้ดนี้ทำงานจะได้ผลลัพธ์ของ Grid Search ซึ่งเก็บไว้ในอ็อบเจ็กต์ gcv และเราสามารถเข้าถึงผลลัพธ์เหล่านี้ได้ เช่น gcv.best_params_ หรือ gcv.best_score_ เพื่อดูค่าพารามิเตอร์ที่ดีที่สุดและคะแนนประสิทธิภาพที่ดีที่สุดของแบบจำลองที่ปรับแต่งแล้ว

เลือกตัวแบบที่ดีที่สุดจาก Grid Search และทำการฝึกอบรมโมเดลด้วยข้อมูลการฝึกอบรม เพื่อให้ได้โมเดลที่ปรับแต่งแล้วและพร้อมที่จะใช้งานกับข้อมูลใหม่ดังภาพ 45



ภาพประกอบ 45 ภาพโค๊ดการเลือกตัวแบบที่ดีที่สุดจาก Grid Search

จากภาพประกอบ 45 สามารถอธิบายได้ว่า model = gcv.best_estimator_ เลือกตัวแบบที่ดีที่สุดที่ได้จาก Grid Search และเก็บไว้ในตัวแปร model และคำสั่งนี้ model.fit(X_train, y_train) จะทำการฝึกอบรมโมเดลด้วยข้อมูลการฝึกอบรม (X_train,y_train) โดยใช้ตัวแบบที่ถูกเลือกจาก Grid Search (gcv.best_estimator_) เมื่อการฝึกอบรมเสร็จสิ้นโมเดลจะได้เรียนรู้และปรับตัวให้เหมาะสมกับข้อมูลการฝึกอบรม

โค้ดนี้จะประเมินความแม่นยำของโมเดลในการฝึกอบรมโดยใช้เทคนิค k-fold cross-validation และแสดงผลค่าเฉลี่ยของคะแนนความแม่นยำในการฝึกอบรมคือ Train ScoreและTest Score ดังภาพประกอบ 46

```
train_scores_dtr = cross_val_score(model, X_train, y_train, cv=10)
print('Train Scores:',train_scores.mean())

Train Scores: 0.9372887386163115

[74] test_scores_dtr = cross_val_score(model, X_test, y_test, cv=10)
     print('Test  Scores:',test_scores.mean())

     Test  Scores: 0.9005517477999604
```

ภาพประกอบ 46 การประเมินความแม่นยำของโมเดลโดยใช้เทคนิค k-fold cross-validation

จากภาพประกอบ 46 สามารถอธิบายโค๊ดได้ว่า train_scores_dtr = cross_val_score (model, X_train, y_train, cv=10): เราใช้ cross_val_score เพื่อประเมินค่าความแม่นยำของโมเดลในการฝึกอบรม โดยใช้โมเดลที่ถูกเลือกจาก Grid Search (model) และใช้ข้อมูลการฝึกอบรม (X_train, y_train) ในกระบวนการ cross-validation โดยกำหนด cv=10 ให้ใช้เทคนิค k-fold cross-validation และคืนค่าคะแนนความแม่นยำของแต่ละรอบ cross-validation ใน train_scores_dtr print('TrainScores:',train_scores.mean()):เราปริ้นค่าเฉลี่ยของคะแนนความแม่นยำของแบบจำลองในการฝึกอบรม โดยใช้ train_scores.mean() โดยที่ train_scores เป็นตัวแปรที่เก็บคะแนนความแม่นยำที่ได้จากcross_val_scoreซึ่งเรานับเฉพาะค่าเฉลี่ยเท่านั้น

4.2 ผลลัพธ์การวัดประสิทธิภาพของแบบจำลอง

เมื่อทำการนำข้อมูลที่จัดเตรียมไว้มาวิเคราะห์ด้วยเทคนิคต่างๆ และทำการสร้างแบบจำลองการถดถอยแบบเชิงเส้น (Linear Regression),การถดถอยต้นไม้การตัดสิน ใจ Decision Tree Regressor,การถดถอยแบบสุ่ม (Random Forest) เพื่อเปรียบเทียบและฝึกทำนายข้อมูลจากนั้นทำการวัดประสิทธิภาพของแบบจำลองโดยใช้ตัวชี้วัดดังนี้

1. Mean Absolute Error (MAE)
2. Mean Absolute Percentage Error (MAPE)
3. R-square

ซึ่งได้ผลการทดลองดังนี้

ตาราง 15 สรุปผลการทดลองของแต่ละแบบจำลองโดยวิเคราะห์จากแบรนด์ทั้งหมดในชุดข้อมูลนี้

Model	MAE	MAPE	R-square
LinearRegressionModel	£1,824	14%	0.877
DecisionTreeRegression	£1,074	7%	0.938
Random forest	£ 942	6%	0.959
RIDGE	£1,825	14%	0.877
LASSO	£1,848	15%	0.875

จากผลการทดสอบโมเดลที่แตกต่างกันตามตารางที่ 15 พบว่าโมเดล โมเดล Linear Regression Model มีค่า MAE ที่ 1,824 £ และ MAPE ที่ 14% ซึ่งหมายความว่าค่าเฉลี่ยของความคลาดเคลื่อนของโมเดลจากค่าจริงมีค่าประมาณ 1,824 £ หรือ 14% ของค่าจริง และมีค่า R-square เท่ากับ 0.877 ซึ่งหมายถึงโมเดลสามารถอธิบายตัวแปรตามได้ร้อยละ 87.7 โมเดล Decision Tree Regression มีค่า MAE ที่ 1,074.08 £ และ MAPE ที่ 7% ซึ่งหมายความว่าค่าเฉลี่ยของความคลาดเคลื่อนของโมเดลจากค่าจริงมีค่าประมาณ 1,074.08 £ หรือ 7% ของค่าจริง และมีค่า R-square เท่ากับ 0.938 ซึ่งหมายถึงโมเดลสามารถอธิบายตัวแปรตามได้ร้อยละ 93.8 โมเดล Random Forest มีค่า MAE ที่ 942 £ และ MAPE ที่ 6% ซึ่งหมายความว่าค่าเฉลี่ยของความคลาดเคลื่อนของโมเดลจากค่าจริงมีค่าประมาณ 942 £ หรือ 6% ของค่าจริง และมีค่า

R-square เท่ากับ 0.959 ซึ่งหมายถึงโมเดลสามารถอธิบายตัวแปรตามได้ร้อยละ 95.9 โมเดล RIDGE มีค่า MAE ที่ 1,824 £ และ MAPE ที่ 14% ค่า MAE ของโมเดล RIDGE มีค่าที่ 1,824 £ ซึ่งหมายความว่าค่าเฉลี่ยของความคลาดเคลื่อนของโมเดลจากค่าจริงมีค่าประมาณ 1,824 £ หรือ 14% ของค่าจริง นอกจากนี้โมเดล RIDGE ยังมีค่า R-square เท่ากับ 0.877 ซึ่งหมายถึงโมเดลสามารถอธิบายตัวแปรตามได้ร้อยละ 87.7 โมเดล LASSO มีค่า MAE ที่ 1,848 £ และค่า MAPE เท่ากับ 0.15 ซึ่งหมายความว่าค่าเฉลี่ยของความคลาดเคลื่อนของโมเดลจากค่าจริงมีค่าประมาณ 1,848 £ หรือ 15% ของค่าจริง นอกจากนี้โมเดล LASSO ยังมีค่า R-square เท่ากับ 0.875 ซึ่งหมายถึงโมเดลสามารถอธิบายตัวแปรตามได้ร้อยละ 87.5 โดยทั่วไปแล้ว เมื่อเราวิเคราะห์ผลลัพธ์ของโมเดล ส่วนใหญ่เราคาดหวังว่าค่า MAE และ MAPE จะต่ำ และค่า R-square สูงซึ่งในกรณีนี้โมเดล Random Forest มีค่า MAE และ MAPE ที่ต่ำที่สุด และมีค่า R-square สูงที่สุดซึ่งแสดงให้เห็นว่าทุกโมเดลสามารถอธิบายตัวแปรตามได้ในระดับที่ดีและมีความแม่นยำสูงในการทำนายผล ดังนั้นนอกจากแบบจำลอง RandomForest มีค่า MAE และ MAPE ที่ต่ำที่สุด จะได้ว่าค่าที่ใช้วัดความสัมพันธ์ของค่าที่ทำนายได้กับค่าจริง ค่า R-squared อยู่ในช่วงระหว่าง 0 ถึง 1 โดย R-squared = 0.959 หมายความว่าการทำนายเป็นไปตามค่าจริงอย่างแม่นยำที่สุด

ตาราง 16 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์แยกแต่ละแบรนด์จากชุดข้อมูลรายการรถยนต์มือสองของแบรนด์ Ford

Model	MAE	MAPE	R-square
Linear Regression Model	£1,356.23	13%	0.830
Decision Tree Regression	£1,053.98	9%	0.865
Random forest	£861.88	7%	0.915
RIDGE	£1,356.14	13%	0.830
LASSO	£1,361.90	14%	0.829

จากผลการทดสอบโมเดลตามตารางที่ 16 เราสามารถสรุปได้ว่าโมเดล Random forest มีประสิทธิภาพที่ดีที่สุดในการทำนาย โดยมีค่า MAE ที่ต่ำที่สุด และค่า MAPE ที่ต่ำที่สุด ในขณะที่ค่า R-square เท่ากับ 0.915 แสดงถึงความสามารถในการอธิบายความเปลี่ยนแปลงของตัวแปรตามได้ดี

ตาราง 17 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์แยกแต่ละแบรนด์จากชุดข้อมูลรายการรถยนต์มือสองของแบรนด์ Mercedes-Benz

Model	MAE	MAPE	R-square
LinearRegressionModel	£3,371.62	17%	0.776
DecisionTreeRegression	£1,948.21	8%	0.909
Random forest	£1,522.73	6%	0.951
RIDGE	£3,366.27	17%	0.778
LASSO	£3,370.43	17%	0.776

จากผลการทดสอบโมเดลตามตารางที่ 17 พบว่าโมเดล Random forest มีประสิทธิภาพที่ดีที่สุดในการทำนาย โดยมีค่า MAE ที่ต่ำที่สุดเทียบกับโมเดลอื่น ๆ รวมถึงค่า MAPE ที่ต่ำที่สุด นอกจากนี้ โมเดล Random forest ยังมีค่า R-square เท่ากับ 0.951 ซึ่งแสดงถึงความสามารถในการอธิบายความเปลี่ยนแปลงของตัวแปรตามได้ อย่างไรก็ตาม ควรพิจารณาเพิ่มเติมเกี่ยวกับข้อมูลและเป้าหมายเพื่อให้การเลือกโมเดลเป็นไปอย่างเหมาะสมที่สุด

ตาราง 18 การเปรียบเทียบผลลัพธ์แบบจำลองต่างๆ โดยวิเคราะห์แยกแต่ละแบรนด์จากชุดข้อมูลรายการรถยนต์มือสองของแบรนด์ Toyota

Model	MAE	MAPE	R-square
LinearRegressionModel	£1,122.80	12%	0.830
DecisionTreeRegression	£973.04	8%	0.865
Random forest	£816.25	7%	0.915
RIDGE	£1,132.67	12%	0.830
LASSO	£1,131.59	12%	0.829

จากผลลัพธ์ที่ได้ตามตารางที่ 18 จะเห็นว่า แบบจำลอง Random forest มีประสิทธิภาพที่ดีที่สุดในการทำนาย โดยมีค่า MAE และ MAPE ที่ต่ำที่สุด และค่า R-square ที่สูงที่สุดที่ 0.915 ซึ่งหมายความว่าโมเดลนี้มีความสามารถในการอธิบายความเปลี่ยนแปลงของตัวแปรตามได้ดีที่สุดในกลุ่มของโมเดลที่พิจารณา

4.3 ผลลัพธ์ของคุณลักษณะสำคัญที่มีผลต่อการทำนาย

เมื่อการสร้างแบบจำลองทั้ง 3 แบบและวัดประสิทธิภาพเรียบร้อยแล้ว จากนั้นทำการหาคุณลักษณะสำคัญที่มีผลต่อราคารถยนต์มือสองด้วยการหา correlationจะได้ผลลัพธ์ดังตาราง 19

ตาราง 19 แสดงถึงค่าถึงความสัมพันธ์อย่างแท้จริงระหว่างคู่ของตัวแปร

feature	ค่า Correlation
price	1
year	0.55
engineSize	0.41
tax	0.19
mpg	-0.18
mileage	-0.47

จากตาราง 19 จะเห็นได้ ปีที่จดทะเบียน (year),ขนาดเครื่องยนต์ (engineSize) ,ระยะทางสะสม (mileage) เป็นคุณลักษณะที่มีผลต่อราคารถยนต์มือสองมากที่สุด

4.4 ผลลัพธ์จากการศึกษาปัจจัยที่มีผลต่อราคารถยนต์มือสอง

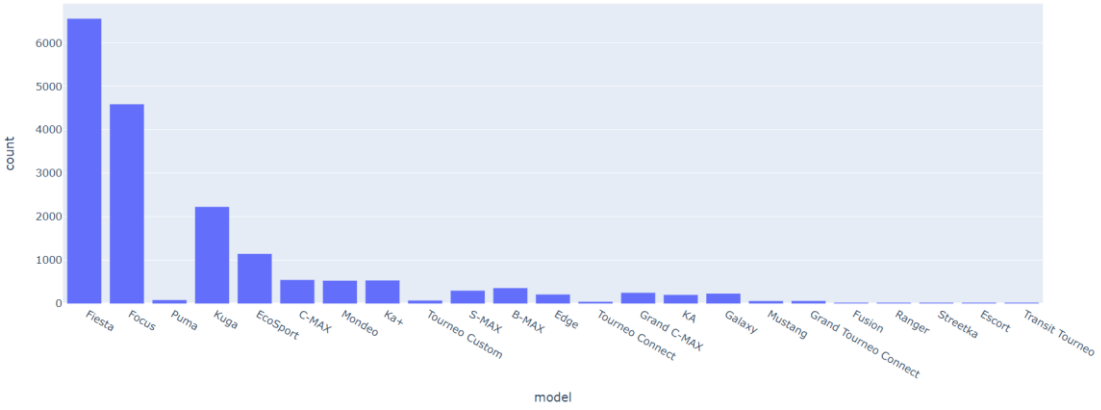
จากการสำรวจในบทที่ 3 จะเห็นได้ว่าปัจจัยที่มีผลต่อราคารถยนต์มือสองมีดังนี้

1. ยี่ห้อและโมเดลของรถยนต์ ยี่ห้อและโมเดลของรถยนต์เป็นตัวชี้วัดสำคัญในการกำหนดราคาขายรถยนต์มือสอง รถยนต์ที่มียี่ห้อและโมเดลที่นิยมและมีความนิยมสูงอาจมีราคาสูงกว่ารถยนต์ที่ไม่ได้รับความนิยมเท่านั้น
2. อายุของรถยนต์ รถยนต์ที่มีอายุมากกว่าจะมีค่าเสื่อมราคามากกว่ารถยนต์ที่อายุน้อยกว่า ซึ่งอาจทำให้ราคาขายลดลง
3. ระยะทางที่เคยใช้งาน การใช้งานระยะไกลหรือการขับรถบ่อยๆ อาจส่งผลต่อความสมบูรณ์ของรถยนต์และส่งผลให้ราคาลดลง

4.5 ผลลัพธ์การวิเคราะห์ข้อมูลโดยแยกแต่ละแบรนด์

การวิเคราะห์ข้อมูลเพื่อหาปัจจัยที่ผลต่อราคารถยนต์มือสอง

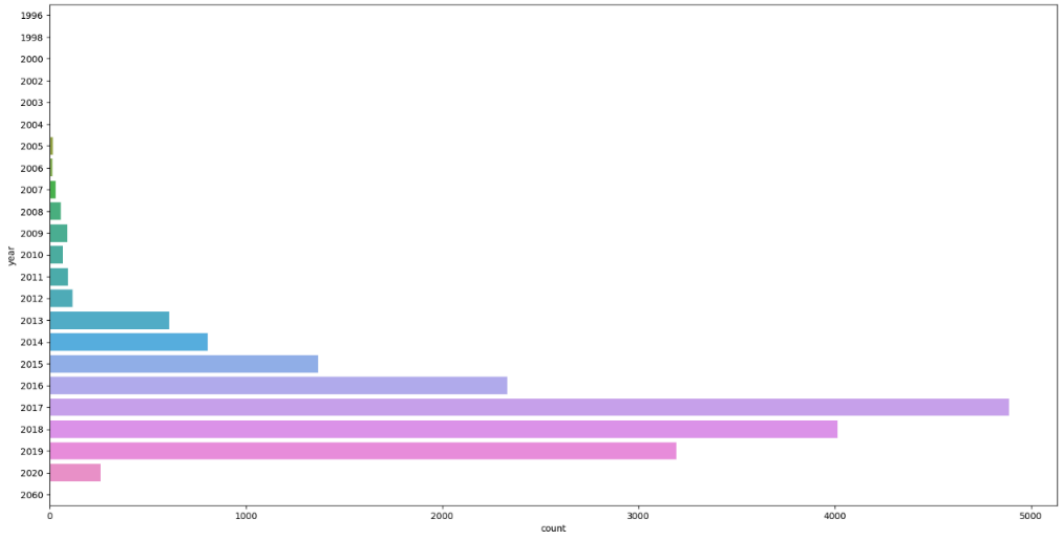
- ผลลัพธ์การวิเคราะห์ข้อมูลความนิยมการใช้งานรถยนต์รุ่นต่างๆของแบรนด์Ford ดังภาพ 44



ภาพประกอบ 47 กราฟแท่งแสดงข้อมูลรุ่นยอดนิยมของแบรนด์ฟอร์ด

จากภาพประกอบ 47 พบว่ารถยนต์รุ่น Fiesta เป็นรถยนต์รุ่นที่ได้รับความนิยมสูงสุดของแบรนด์ฟอร์ดในชุดข้อมูลนี้ รองลงมาคือรถยนต์รุ่น Focus และถัดมาคือรถยนต์รุ่น Kuga ในขณะเดียวกันรถยนต์รุ่น Mustang,Grand Tourneo connect,Fusion,Ranger,Streetka,Escort และ Transit Tourneo จำนวนการใช้งานไม่มาก

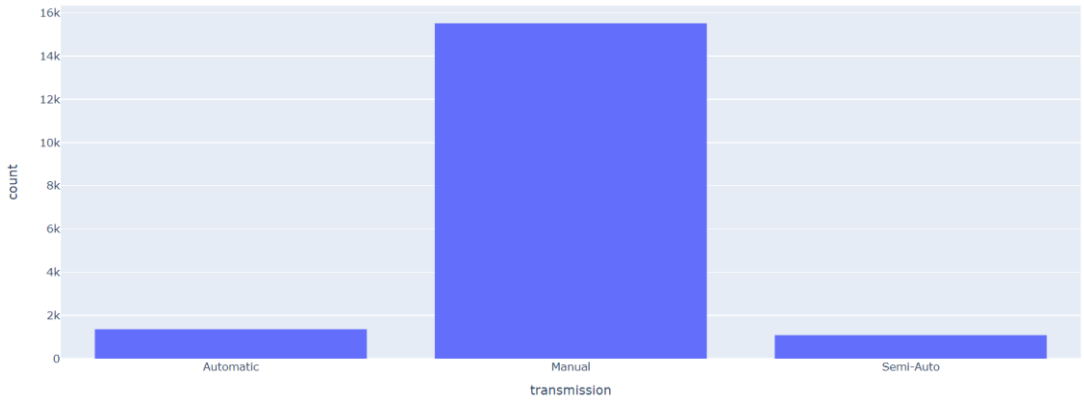
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานรถยนต์มือสองตามปีที่ผลิตของรถยนต์แบรนด์Ford ดังภาพ 48



ภาพประกอบ 48 กราฟจำนวนการใช้งานรถยนต์มือสองตามปีที่ผลิตของรถยนต์

จากภาพประกอบ 48 พบว่ารถยนต์แบรนด์ฟอร์ด ปี 2017 มีจำนวนการใช้งานมากที่สุดในชุดข้อมูลนี้ รองลงมาคือปี 2018 ,ปี 2019 และตามด้วยปี 2016 รถยนต์รุ่นปีเก่าๆจะมีจำนวนผู้ใช้น้อยตามลำดับจนถึงไม่มีคนใช้งานเลย หากเป็นรถยนต์ปีที่เก่ามากๆ

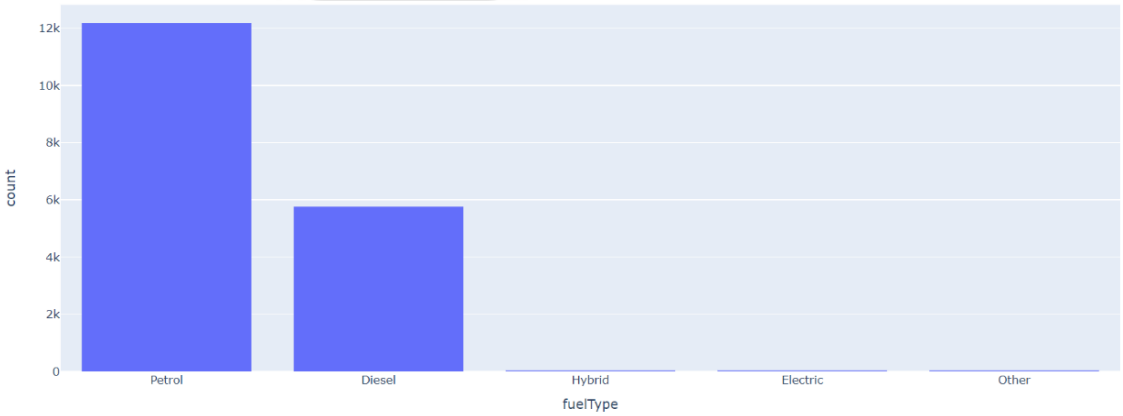
- ผลลัพธ์การวิเคราะห์จำนวนการใช้รถยนต์แบรนด์ฟอร์ด โดยแบ่งตามชนิดเกียร์คือ Automatic,Manual และ Semi-Auto ภาพประกอบ 49



ภาพประกอบ 49 การวิเคราะห์จำนวนการใช้เกียร์รถยนต์ชนิดต่างๆแบรนด์ฟอร์ด

จากภาพประกอบ 49 จะเห็นได้ว่ารถยนต์แบรนด์ฟอร์ดมีผู้ใช้งานเกียร์ชนิด Manual มากที่สุดรองลงมาคือเกียร์ชนิด Automatic และสุดท้ายคือเกียร์ชนิด Semi-Auto

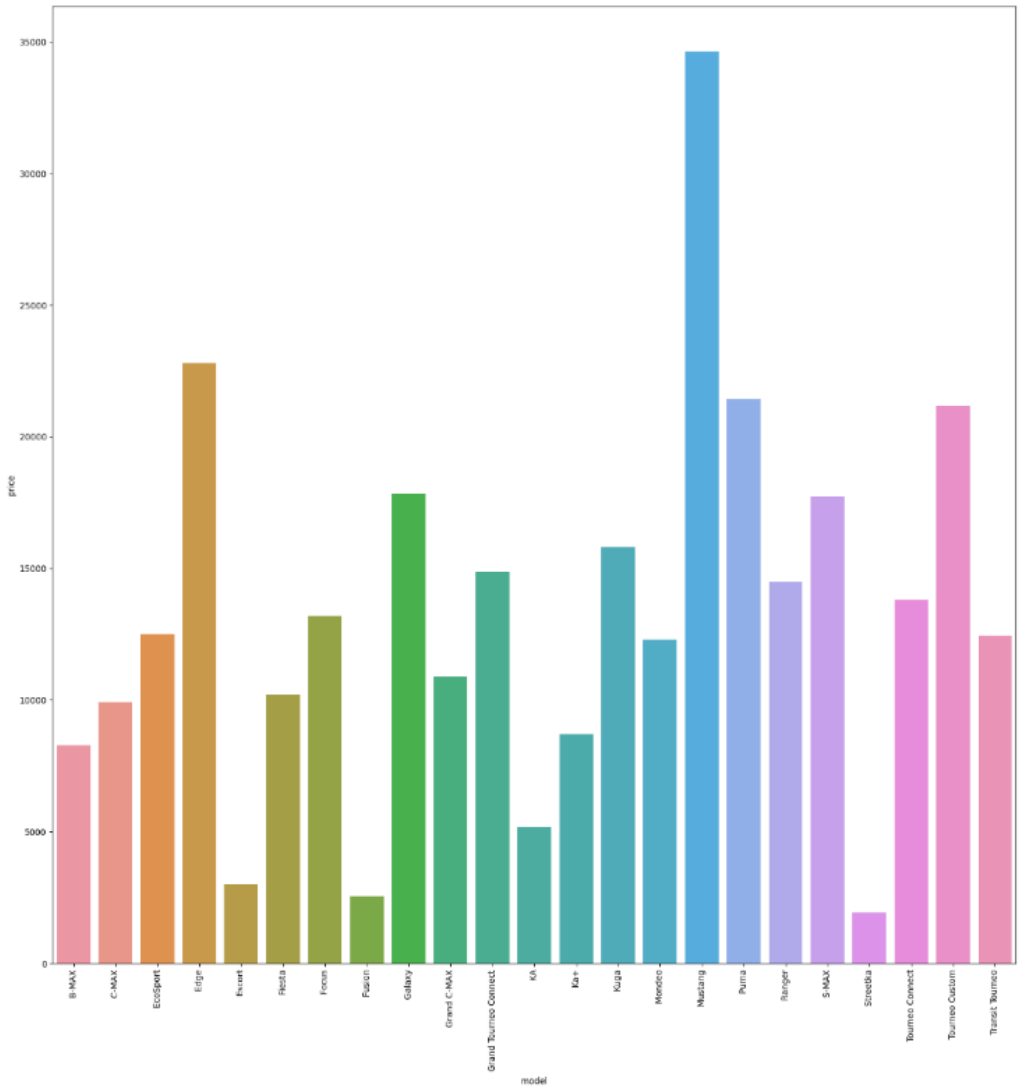
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานเชื้อเพลิงชนิดต่างๆของรถยนต์มือสองแบรนด์Ford ดังภาพประกอบ 50



ภาพประกอบ 50 การวิเคราะห์จำนวนการใช้เชื้อเพลิงชนิดต่างๆแบรนด์ฟอร์ด

จากภาพประกอบ 50 จะเห็นได้ว่ารถยนต์แบรนด์ฟอร์ด มีผู้ใช้งานรถยนต์ชนิดเชื้อเพลิง Petrol มากที่สุด รองลงมาคือ รถยนต์ชนิดเชื้อเพลิง Dieselและเชื้อเพลิงอื่นๆมีจำนวนผู้ใช้งานในระดับเดียวกัน

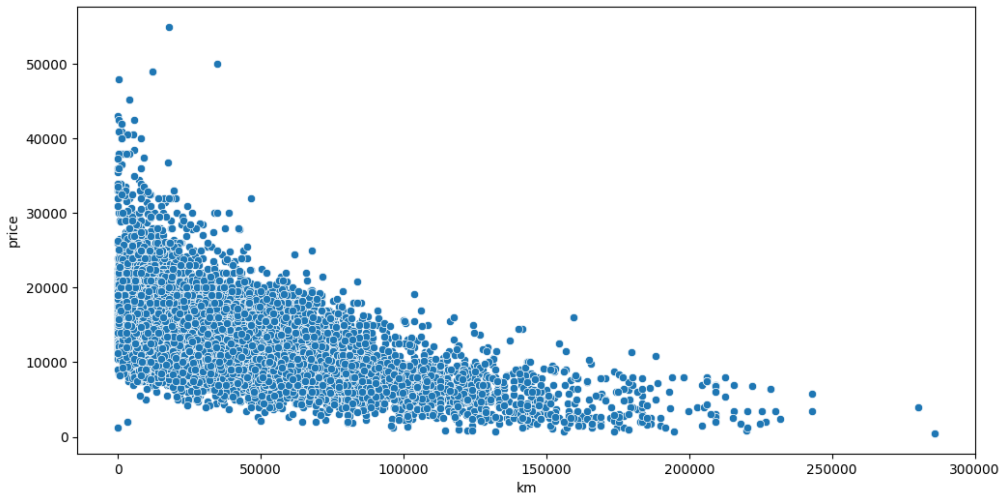
ผลลัพธ์การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคาดังภาพประกอบ 51



ภาพประกอบ 51 กราฟแท่งแสดงราคารถรุ่นต่างๆ

จากภาพประกอบ 51 ในแนวแกน x จะเป็นรุ่นรถยนต์และในแนวแกน Y คือราคารถยนต์แบรนด์ฟอรด์รุ่นต่างๆในชุดข้อมูลนี้ จะเห็นได้ว่ารถยนต์รุ่น Mustang มีราคาสูงใกล้จะถึงราคา 35,000 ปอนด์ และรองลงมาคือช่วงราคา 20,000 – 25,000 ประกอบด้วยรุ่นรถยนต์ Edge,puma และ Toumeo custom และในทางกลับกันรถยนต์ที่มีราคาต่ำกว่า 5,000 ปอนด์ ประกอบด้วย Escort,Fusion และ Streetka

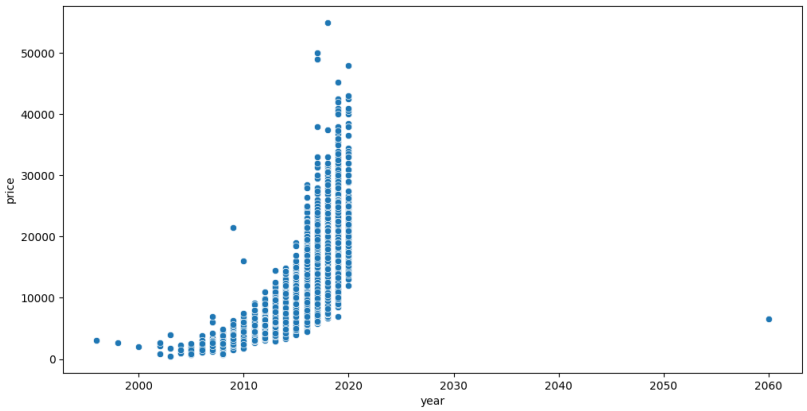
- ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้แบรนด์ Ford ดังภาพประกอบ 52 ในหน้าถัดไป



ภาพประกอบ 52 ภาพความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้

จากภาพประกอบ 52 จะเห็นได้ว่าจำนวนระยะทางที่ใช้น้อย ราคารถยนต์จะสูงในทางตรงกันข้ามจำนวนระยะทางที่ใช้มาก ราคาก็จะต่ำลง

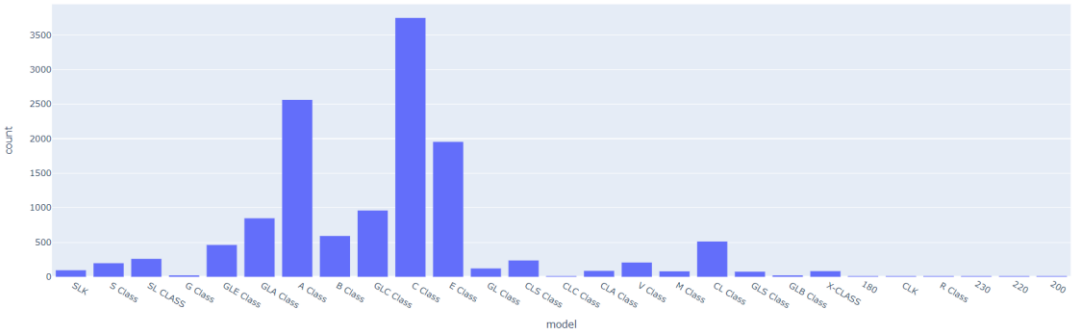
- ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับรถปีต่างๆแบรนด์ Ford ดังภาพประกอบ 53



ภาพประกอบ 53 ภาพความสัมพันธ์ระหว่างราคากับรถปีต่างๆแบรนด์ Ford

จากภาพประกอบ 53 จะเห็นได้ว่ารถปีใหม่ๆจะมีราคาที่สูงถึง 50,000 ปอนด์และรถปีเก่าจะมีราคาที่ต่ำสังเกตจากกราฟรถยนต์ปีที่ต่ำกว่าปี 2010 ราคาจะต่ำลงเรื่อยๆ

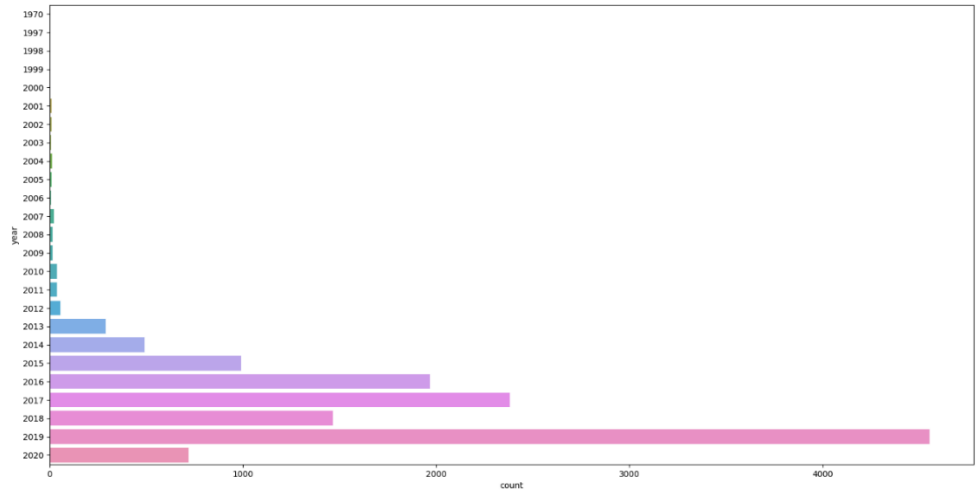
- ผลลัพธ์การวิเคราะห์ข้อมูลความนิยมการใช้งานรถยนต์รุ่นต่างๆของแบรนด์ Mercedez-benz ดังภาพประกอบ 54



ภาพประกอบ 54 กราฟแท่งแสดงข้อมูลรุ่นยอดนิยมของแบรนด์ฟอร์ด

จากภาพประกอบ 54 พบว่ารถยนต์รุ่น C classเป็นรถยนต์รุ่นที่ได้รับความนิยมสูงสุดของแบรนด์ Mercedez ในชุดข้อมูลนี้ รองลงมาคือรถยนต์รุ่น A-Class และถัดมาคือรถยนต์รุ่น E-Class

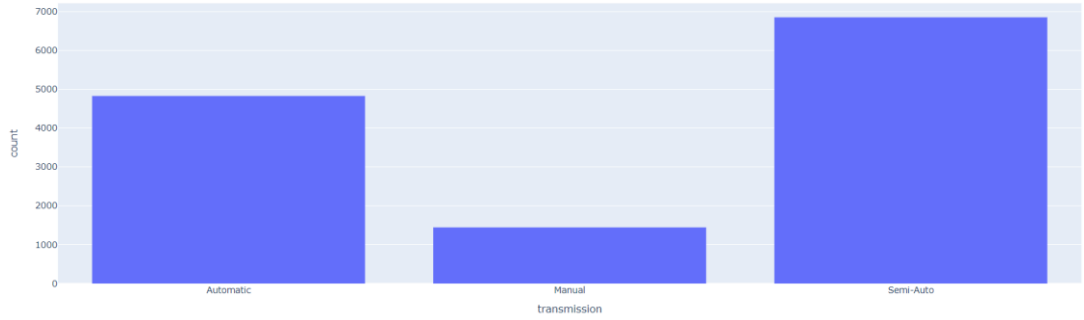
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานรถยนต์มือสองตามรุ่นปีที่ผลิตของรถยนต์แบรนด์ Mercedezbenz ภาพประกอบ 55



ภาพประกอบ 55 กราฟรถยนต์มือสองตามรุ่นปีที่ผลิตของรถยนต์แบรนด์ Mercedezbenz

จากภาพประกอบ 55 พบว่ารถยนต์แบรนด์ Mercedez-benz ปี 2019 มีจำนวนการใช้งานมากที่สุดในชุดข้อมูลนี้ รองลงมาคือปี 2017 รองลงมาคือปี 2018 และตามด้วยรถยนต์รุ่นปีเก่าๆจะมีจำนวนผู้ใช้น้อยตามลำดับจนถึงไม่มีคนใช้งานเลย หากเป็นรถยนต์ปีที่เก่ามากๆ

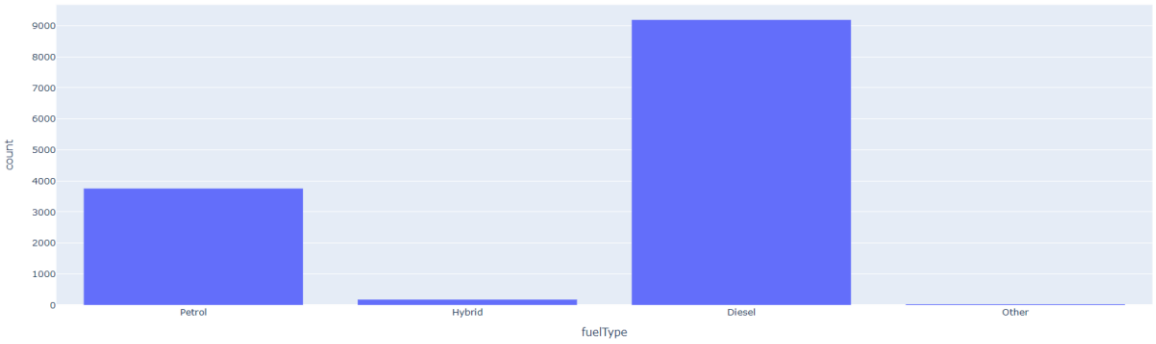
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานเกียร์ชนิดต่างๆของรถยนต์มือสองแบรนด์ Mercedezbenz ดังภาพประกอบ 56



ภาพประกอบ 56 การวิเคราะห์จำนวนการใช้เกียร์รถยนต์ชนิดต่างๆ

จากภาพประกอบ 56 จะเห็นได้ว่ารถยนต์แบรนด์ Mercedezbenz มีผู้ใช้งานเกียร์ชนิด Semi-Auto มากที่สุดรองลงมาคือเกียร์ชนิด Automatic และสุดท้ายคือเกียร์ชนิด Manual

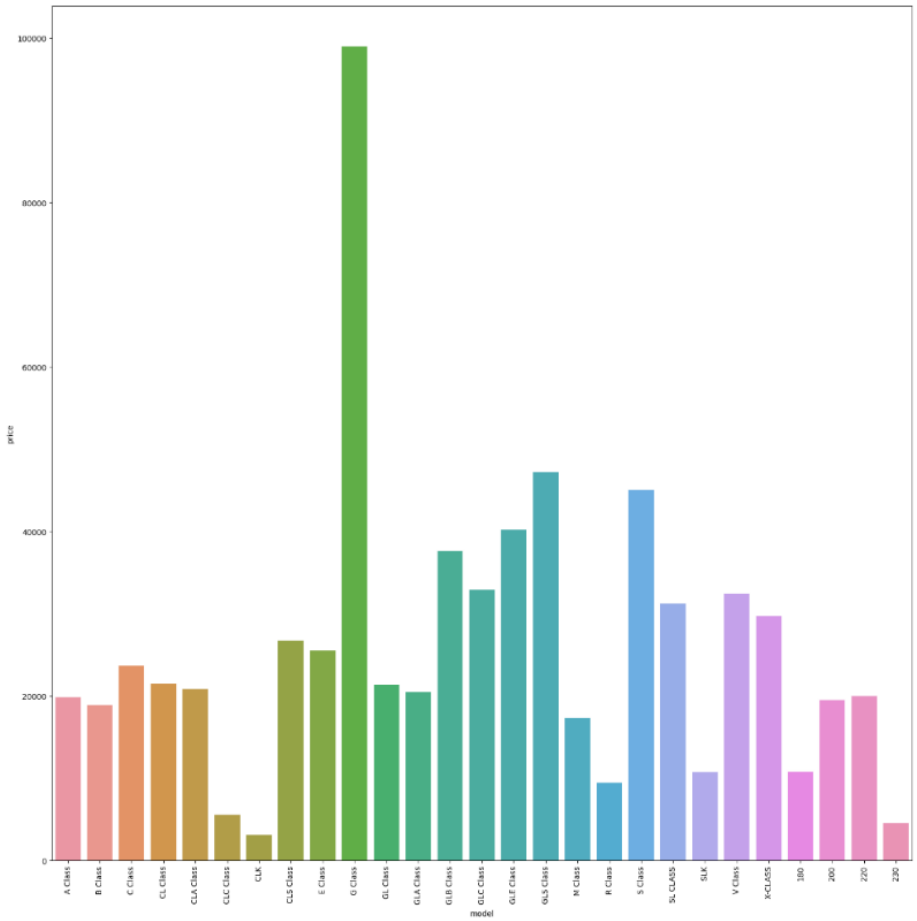
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานเชื้อเพลิงชนิดต่างๆของรถยนต์มือสองแบรนด์ Mercedezbenz ภาพประกอบ 57



ภาพประกอบ 57 การวิเคราะห์จำนวนการเชื้อเพลิงชนิดต่างๆ

จากภาพประกอบ 57 จะเห็นได้ว่ารถยนต์แบรนด์ Mercedezbenz มีผู้ใช้งานรถยนต์ชนิดเชื้อเพลิง Diesel มากที่สุด รองลงมาคือ รถยนต์ชนิดเชื้อเพลิง Petrol และตามด้วย Hybrid และเชื้อเพลิงอื่นๆ

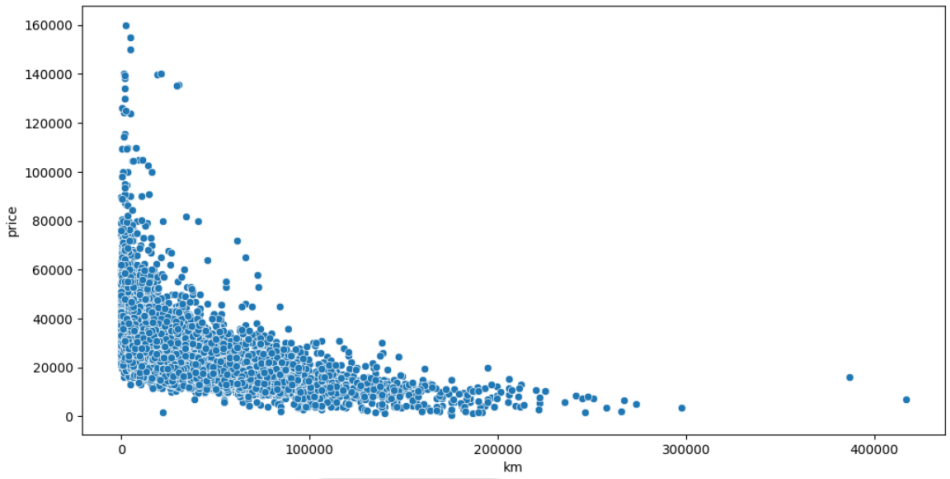
- ผลลัพธ์การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคาดังภาพประกอบ 58



ภาพประกอบ 58 การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคา

จากภาพประกอบ 58 ในแนวแกน x จะเป็นรุ่นรถยนต์และในแนวแกน Y คือราคารถยนต์แบรนด์ Mercedez รุ่นต่างๆในชุดข้อมูลนี้ จะเห็นได้ว่ารถยนต์รุ่น G Class มีราคาสูงใกล้จะถึงราคา 100,000 ปอนด์ และรองลงมาคือช่วงราคา 40,000 – 60,000 ประกอบด้วยรุ่นรถยนต์ gle class,gls class และ S class และในทางกลับกันรถยนต์ที่มีราคาต่ำกว่า 20,000 ปอนด์ ประกอบด้วย B Class,CLC Class,CLK,N Class,R Class,SLK,180 และ 230

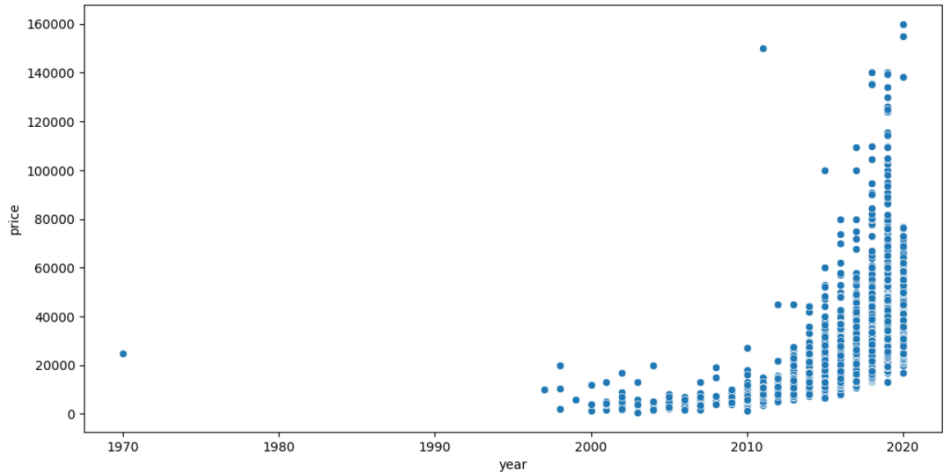
- ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้แบรนด์ Mercedezbenz ดังรูปภาพ 59



ภาพประกอบ 59 ภาพความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้

จากภาพประกอบ 59 จะเห็นได้ว่าจำนวนระยะทางที่ใช้น้อย ราคารถยนต์จะสูงในทางตรงกันข้ามจำนวนระยะทางที่ใช้มาก ราคาก็จะต่ำลง

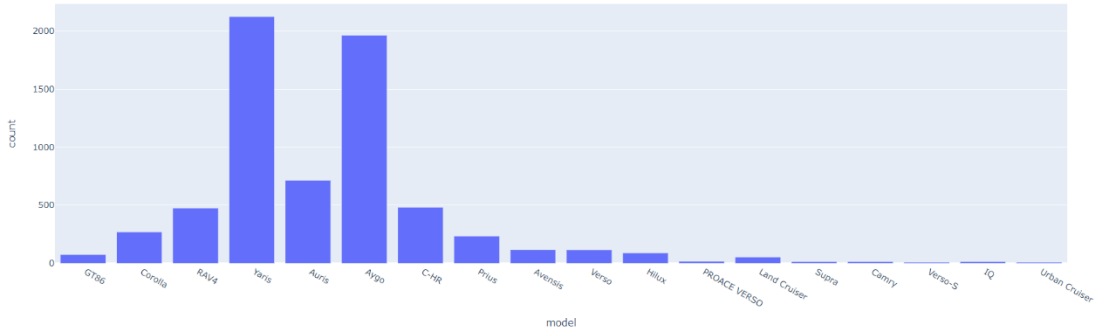
- ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับรถปีต่างๆแบรนด์ Mercedezbenz ดังรูปภาพ 60



ภาพประกอบ 60 ภาพความสัมพันธ์ระหว่างราคากับรถปีต่างๆแบรนด์ Mercedezbenz

จากภาพประกอบ 60 จะเห็นได้ว่ารถปีใหม่ๆจะมีราคาที่สูงถึง 160,000 ปอนด์และรถปีเก่าจะมีราคาที่ต่ำสังเกตจากกราฟรถยนต์ปีที่ต่ำกว่าปี 2000 แทบจะไม่มีรถยนต์กระจุกอยู่ในช่วงปี 2000-1970

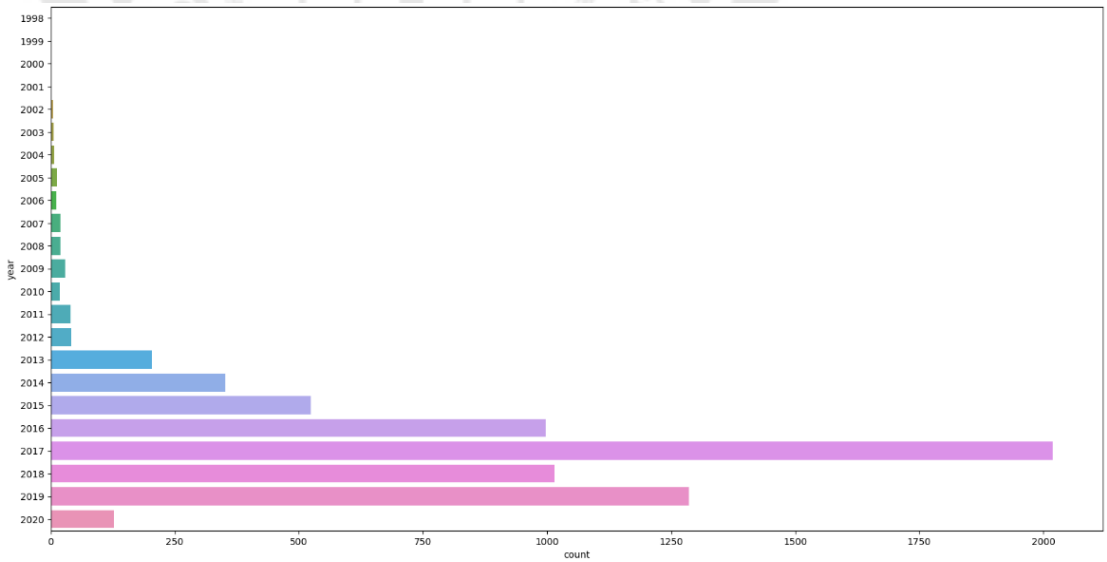
- ผลลัพธ์การวิเคราะห์ข้อมูลความนิยมการใช้งานรถยนต์รุ่นต่างๆของแบรนด์ Toyota



ภาพประกอบ 61 การวิเคราะห์จำนวนการใช้รถยนต์รุ่นๆต่างของแบรนด์Toyota

จากภาพประกอบ 61จะเห็นได้ว่ารถยนต์แบรนด์ Toyota รุ่นที่ได้รับความนิยมสูงสุดคือรุ่น Toyota Yaris ,Aygo และรุ่นที่มีจำนวนคนใช้งานในระดับปานกลางที่อยู่ในช่วง 500-1000 คน คือ Auris และรุ่นที่คนใช้งานน้อยสุดในชุดข้อมุลนี้คือ Verso-S แล ะUrban Cruiser

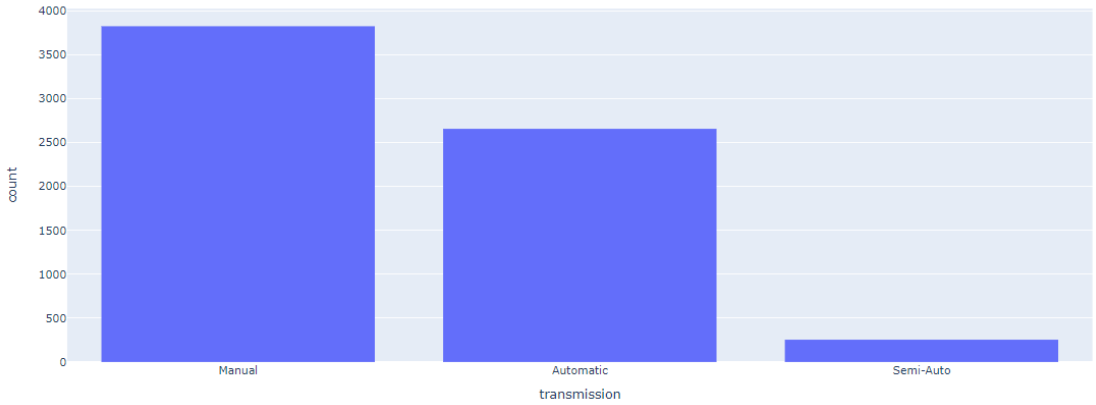
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานรถยนต์มือสองตามรุ่นปีที่ผลิตของรถยนต์แบรนด์ Toyota ดังภาพประกอบ 62



ภาพประกอบ 62 กราฟรถยนต์มือสองตามรุ่นปีที่ผลิตของรถยนต์แบรนด์ Toyota

จากภาพประกอบ 62 พบว่ารถยนต์แบรนด์ Toyota ปี 2017 มีจำนวนการใช้งานมากที่สุดในชุดข้อมูลนี้ รองลงมาคือปี 2019 รองลงมาคือปี 2018 และปี2016ตามด้วยรถยนต์รุ่นปีเก่าๆจะมีจำนวนผู้ใช้น้อยตามลำดับจนถึงไม่มีคนใช้งานเลย หากเป็นรถยนต์ปีที่เก่ามากๆ แต่ทั้งนี้ก็ขึ้นอยู่ปัจจัยอื่นๆด้วย

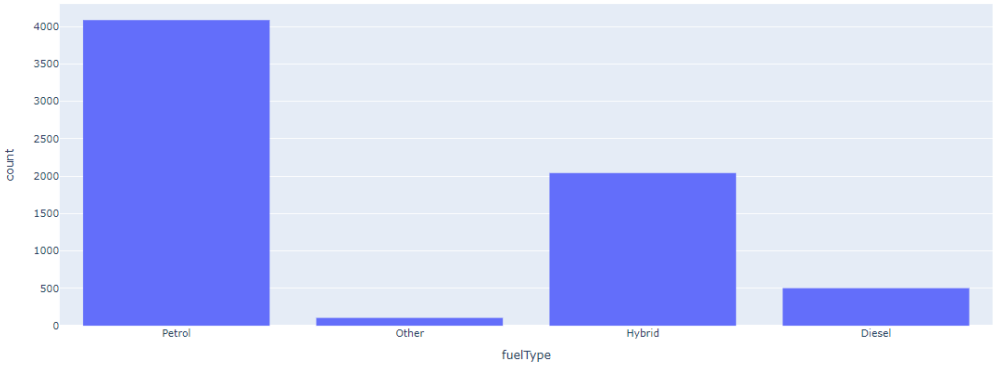
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานเกียร์ชนิดต่างๆของรถยนต์มือสองแบรนด์ Toyota ดังภาพประกอบ 63



ภาพประกอบ 63 กราฟจำนวนการใช้เชื้อเพลิงชนิดต่างๆของรถยนต์แบรนด์ Toyota

จากภาพประกอบ 63 จะเห็นได้ว่ารถยนต์แบรนด์ Toyota มีผู้ใช้งานเกียร์ชนิดManual มากที่สุดรองลงมาคือเกียร์ชนิด Automaticและสุดท้ายคือเกียร์ชนิด Semi-Auto

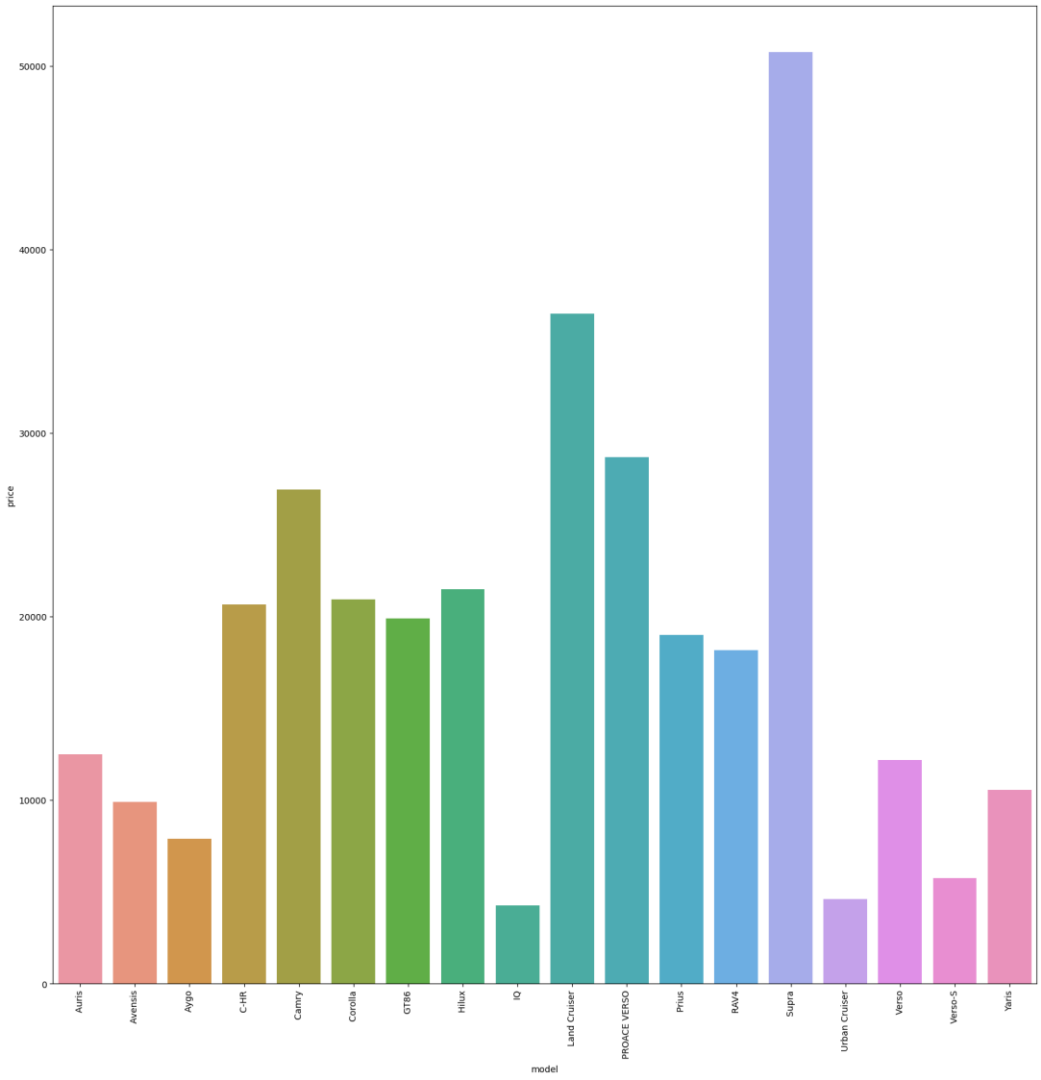
- ผลลัพธ์การวิเคราะห์จำนวนการใช้งานเชื้อเพลิงชนิดต่างๆของรถยนต์มือสองแบรนด์ Toyota ดังภาพประกอบ 64



ภาพประกอบ 64 การวิเคราะห์จำนวนการเชื้อเพลิงชนิดต่างๆ

จากภาพประกอบ 64 จะเห็นได้ว่ารถยนต์แบรนด์ Toyota มีผู้ใช้งานรถยนต์ชนิดเชื้อเพลิงPetrol มากที่สุด รองลงมาคือ รถยนต์ชนิดเชื้อเพลิง Hybrid และตามด้วย Dieselและเชื้อเพลิงอื่นๆ

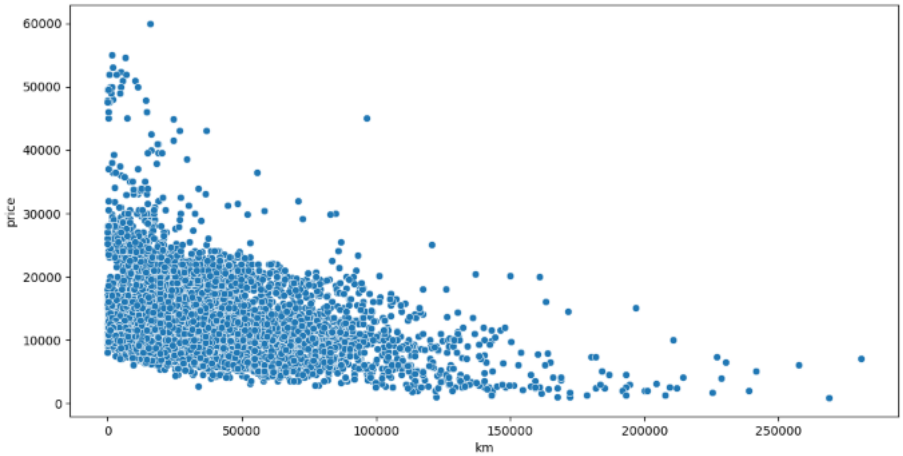
- ผลลัพธ์การวิเคราะห์ข้อมูลระหว่างรถยนต์รุ่นต่างๆกับราคาดังภาพประกอบ 65



ภาพประกอบ 65 ภาพจำนวนการใช้งานรถโดยแบ่งตามรุ่นรถแบรนด์toyota

จากภาพประกอบ 65 ในแนวแกน x จะเป็นรุ่นรถยนต์และในแนวแกน Y คือราคารถยนต์แบรนด์Toyotaรุ่นต่างๆในชุดข้อมูลนี้ จะเห็นได้ว่ารถยนต์รุ่น Supraมีราคาสูงใกล้จะถึงราคา 50,000 ปอนด์ และรองลงมาคือช่วงราคา 30,000 – 40,000 คือรถยนต์รุ่น Land Cruiser

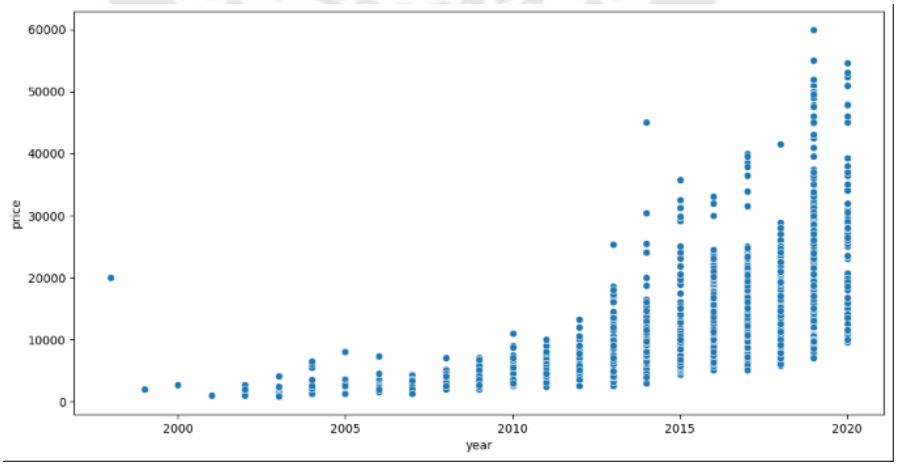
- ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทางที่ใช้แบรนด์ Toyota ดังภาพประกอบ 66



ภาพประกอบ 66 การวิเคราะห์ความสัมพันธ์ระหว่างราคากับจำนวนระยะทาง

จากภาพประกอบแบรนด์Toyota รถยนต์ที่มีการใช้งานอยู่ที่100,000-50,000 กิโลยังสามารถขายต่อที่ราคาค่อนข้างดี

- ผลลัพธ์การวิเคราะห์ความสัมพันธ์ระหว่างราคากับรถปีต่างๆแบรนด์ Toyota ดังภาพประกอบ 68

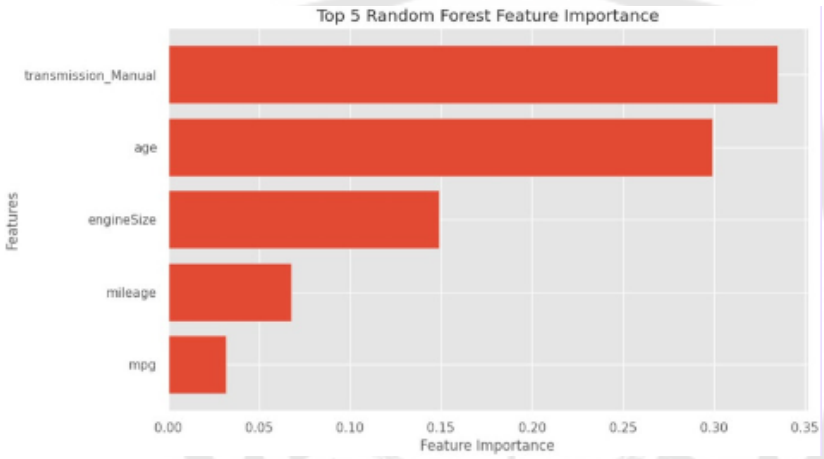


ภาพประกอบ 68 ภาพความสัมพันธ์ระหว่างราคากับรถปีต่างๆแบรนด์ Toyotaจาก

จากภาพประกอบ 68 จะเห็นได้ว่ารถปีใหม่ๆจะมีราคาที่สูงถึง 60,000 ปอนด์และรถปีเก่าจะมีราคาที่ต่ำ

4.6 การวัดความสำคัญของแต่ละคุณสมบัติในการทำนายผลของแบบจำลอง

Feature importance คือการวัดความสำคัญของแต่ละคุณสมบัติในการทำนายผลของแบบจำลองหรือคุณสมบัติใดที่มีผลกระทบให้กับผลลัพธ์ของแบบจำลองมากที่สุด เมื่อแบบจำลองถูกสร้างและทำนายข้อมูล ค่า feature importance จะบ่งบอกถึงอัตราส่วนของผลกระทบในการเปลี่ยนแปลงของคุณสมบัติต่อผลลัพธ์ การที่คุณสมบัติมีค่า feature importance มากนั้นคือคุณสมบัตินั้นมีบทบาทที่สำคัญในการตัดสินใจของแบบจำลอง และมีอิทธิพลต่อผลลัพธ์ในรูปแบบที่มีความชัดเจน ในงานวิจัยนี้ผู้วิจัยนำ Feature importance มาใช้กับแบบจำลอง Random Forestและได้ผลลัพธ์มาพล็อตกราฟดังภาพประกอบ 69



ภาพประกอบ 69 กราฟ 5 ตัวแปรที่สำคัญที่สุด

จากภาพประกอบ 69 การพล็อตกราฟ Feature Importance เฉพาะ 5 ตัวแปรที่สำคัญที่สุดจากชุดข้อมูลที่สร้างด้วย Random Forest Regression พบว่า transmission_Manual เป็นคุณสมบัติที่มีความสำคัญมากที่สุดในการทำนายราคารถยนต์มือสอง ซึ่งเป็นระบบเกียร์ที่คนส่วนใหญ่นิยมใช้ในชุดข้อมูลนี้ การเรียนรู้คุณสมบัตินี้เป็นสิ่งสำคัญในการคัดเลือกและสร้างแบบจำลองที่มีประสิทธิภาพในการทำนายราคารถยนต์ ค่า feature importance ที่สูงให้เข้าใจว่าคุณสมบัตินี้มีผลกระทบสูงในการทำนายราคาและการตัดสินใจของแบบจำลองนอกจากนี้ยังมีคุณสมบัติอื่นๆที่มีความสำคัญมากตามลำดับคือ age, engineSize, mileage, และ mpg ซึ่งคุณสมบัตินี้ทั้งหมดเป็นตัวบ่งบอกถึงคุณลักษณะของรถยนต์ที่มีผลต่อราคา การใช้คุณสมบัตินี้ในกระบวนการสร้างแบบจำลองอาจช่วยเพิ่มความแม่นยำและเพื่อให้บรรลุประสิทธิภาพในการทำนายที่ดีที่สุด

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยการทำนายราคารถยนต์มือสอง โดยใช้ชุดข้อมูล 100,000 UK Used Car คือชุดข้อมูลรายการรถยนต์มือสองของตลาดรถยนต์ในสหราชอาณาจักร ด้วยเทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้วัดประสิทธิภาพของแต่ละแบบจำลองเพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผลการวิจัย

5.3 ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

รถยนต์มือสอง คือ ยานพาหนะประเภทรถยนต์ที่เคยถูกใช้งานและถูกขายต่อโดยโอนกรรมสิทธิ์ให้กับผู้ที่ไม่ใช่เจ้าของเดิมของรถยนต์นั้น รถยนต์มือสองอาจถูกขายโดยเจ้าของเดิมโดยตรงหรือผ่านตัวแทนขายรถยนต์มือสอง และรถยนต์มือสองสามารถเป็นที่นิยมในตลาดรถยนต์เนื่องจากมีราคาที่ถูกกว่ารถยนต์ใหม่ จากอดีตจนถึงปัจจุบันการเติบโตของตลาดรถยนต์มือสองมากขึ้นและการเพิ่มขึ้นของธุรกิจขายรถยนต์มือสองที่มีช่องทางจำหน่ายรถยนต์มือสอง ที่หลากหลายมากขึ้น ตัวอย่างเช่น สถานประกอบการขายรถยนต์มือสอง และช่องทางออนไลน์ทำให้ผู้ซื้อมีทางเลือกเพิ่มขึ้น เมื่อมีทางเลือกมากขึ้นการตัดสินใจในการซื้อหรือขายต่อก็จะยากขึ้น เพราะว่าการเปรียบเทียบราคา เพื่อใช้ในการประเมินราคาและการตัดสินใจซื้อหรือขายก็จะยากขึ้นเนื่องจากมีคุณสมบัติของรถยนต์มือสองต่างๆที่แตกต่างกันสำหรับวิเคราะห์และประเมินราคารถยนต์มือสองที่เหมาะสมกับสภาพ งานวิจัยนี้จึงนำคุณสมบัติต่างๆที่เกี่ยวข้องกับราคารถยนต์มาวิเคราะห์หาปัจจัยที่สำคัญต่อราคารถยนต์และพัฒนาแบบจำลองการเรียนรู้ของเครื่องเพื่อเปรียบเทียบหาแบบจำลองที่สอดคล้องกับข้อมูลและมีความแม่นยำที่สุด งานวิจัยนี้นำชุดข้อมูลจากเว็บไซต์ Kaggle ใช้ชุดข้อมูลตลาดรถยนต์มือสองในสหราชอาณาจักร ใช้เทคนิคการเรียนรู้ของเครื่องประกอบโดยเทคนิคดังนี้ การถดถอยแบบเชิงเส้น (Linear Regression), การถดถอยแบบ (Ridge), ลาสโซ่ (Lasso) ,การถดถอยต้นไม้การตัดสินใจ (Decision Tree Regression) และการถดถอยแบบป่าสุ่ม Random Forest เป็นต้น โดยนำผลลัพธ์ที่ได้จากการฝึกแบบจำลองและทดสอบแบบจำลองมาวัดประสิทธิภาพและนำแบบจำลองที่มีผลลัพธ์ที่ดีที่สุดมาใช้

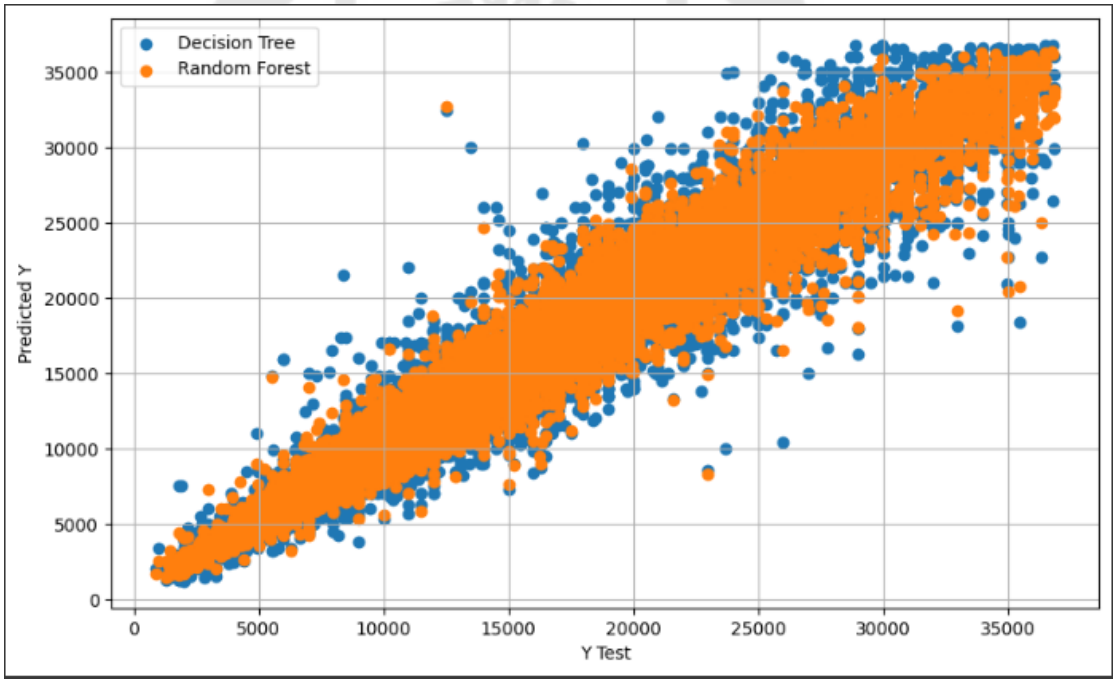
ผลวิจัยพบว่าแบบจำลอง Random Forest ให้ผลลัพธ์ที่ดีที่สุดในการทำนายราคารถยนต์มือสองสำหรับแบรนด์ Ford, Mercedes-Benz, และ Toyota โดยมีค่า MAE และ MAPE ที่ต่ำสุดและค่า R-square ที่สูงสุด เป็นไปตามคาดหวังในการทำนายที่แม่นยำและเชื่อถือได้ การใช้โมเดล Random Forest เป็นทางเลือกที่ดีในการทำนายราคารถยนต์มือสองสำหรับทั้งสามแบรนด์ที่ศึกษา นอกจากนี้ โมเดล Random Forest ยังมีความสามารถในการอธิบายความเปลี่ยนแปลงของราคาได้ดี อย่างไรก็ตาม โมเดล Ridge Regression และ Lasso Regression ให้ผลลัพธ์ที่ใกล้เคียงกัน แต่ค่า MAE มีค่าสูงกว่า Random Forest, Linear Regression และ Decision Tree ดังนั้น ผลลัพธ์การเปรียบเทียบแบบจำลองต่างๆ ในการทำนายราคารถยนต์มือสองแสดงให้เห็นว่าแบบจำลอง Random Forest เป็นแบบจำลองที่ดีที่สุดในการทำนายราคารถยนต์มือสองในชุดข้อมูลนี้และสอดคล้องกับราคารถยนต์ที่เหมาะสมในการทำนายราคารถยนต์มือสองสำหรับทุกแบรนด์ในชุดข้อมูลนี้

5.2 อภิปรายผลการวิจัย

เมื่อวิเคราะห์ถึงความเป็นไปได้ในการนำแบบจำลองLinear Regression, Decision Tree Regression, และ Random Forest เหล่านี้ไปใช้งานจริง สามารถสรุปได้ดังนี้Linear Regressionเป็นแบบจำลองที่เหมาะสำหรับงานที่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรต้นและตัวแปรตาม โดยสมมติฐานว่าความสัมพันธ์เป็นเส้นตรง การใช้ Linear Regression ในการทำนายราคารถยนต์มือสองอาจมีประสิทธิภาพที่ดีเมื่อข้อมูลตัวแปรต้นและตัวแปรตามมีความสัมพันธ์เชิงเส้นและไม่ซับซ้อนมากนัก อีกทั้งต้องพิจารณาด้วยว่า Linear Regression สามารถตอบสนองความต้องการในการทำนายราคาอย่างเพียงพอหรือไม่, Decision Tree Regression: แบบจำลอง Decision Tree Regression เหมาะสำหรับงานที่มีความซับซ้อนและความสัมพันธ์ที่ซับซ้อนระหว่างตัวแปรต้นและตัวแปรตาม โดยแบบจำลองนี้สามารถจับความสัมพันธ์ที่ซับซ้อนได้ดีกว่า Linear Regression และมีความยืดหยุ่นในการปรับแต่งพารามิเตอร์ เมื่อมีการใช้งาน Decision Tree Regression ในการทำนายราคารถยนต์มือสอง อาจได้ผลลัพธ์ที่ดีเมื่อต้องการการทำนายที่ซับซ้อนและความสามารถในการจับความสัมพันธ์ที่ซับซ้อน, Random Forest: แบบจำลอง Random Forest เป็นการนำหลาย Decision Tree มาสร้างและให้คำตอบโดยอิสระจากแต่ละ Decision Tree แล้วจัดเรียงผลลัพธ์เพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำและเสถียรภาพสูงกว่า

Decision Tree เดี่ยว การใช้งาน Random Forest ในการทำนายราคารถยนต์มือสองอาจมีประสิทธิภาพที่ดีเนื่องจากการคำนวณผลลัพธ์โดยใช้หลาย Decision Tree ทำให้การทำนายมีความเสถียรภาพและความแม่นยำสูงขึ้น, Ridge Regression ใช้การเพิ่มค่า Regularization Term เข้าไปในฟังก์ชันความสูญเสีย ซึ่งมีรูปแบบของ L2 Regularization จุดประสงค์ของ Ridge Regression คือการลดความเกี่ยวข้องกับข้อมูลที่มีความสัมพันธ์ซับซ้อน และช่วยป้องกันการเกิด Overfitting ในโมเดล Linear Regression, Lasso Regression ใช้การเพิ่มค่า Regularization Term เข้าไปในฟังก์ชันความสูญเสีย ซึ่งมีรูปแบบของ L1 Regularization โดยมีการเพิ่มค่าสัมประสิทธิ์ของพารามิเตอร์ที่ใช้ในโมเดล ซึ่งส่งผลให้บางพารามิเตอร์มีค่าเป็นศูนย์ และลดความเกี่ยวข้องกับข้อมูลจริงออกไป จุดประสงค์ของ Lasso Regression คือการลดความเกี่ยวข้องกับข้อมูลที่ไม่สำคัญ และช่วยในการเลือกคุณลักษณะที่สำคัญสำหรับโมเดล Linear Regressionทั้ง Ridge Regression และ Lasso Regression สามารถช่วยลดความเกี่ยวข้องกับข้อมูลที่ซับซ้อนและช่วยป้องกันการเกิด Overfitting ในโมเดล Linear Regression โดยการเพิ่มค่า Regularization Term เข้าไปในการคำนวณฟังก์ชันความสูญเสีย การเลือกใช้ Ridge Regression หรือ Lasso Regression ขึ้นอยู่กับลักษณะของข้อมูลและความต้องการของงาน โดยทั่วไป Ridge Regression มักใช้เมื่อต้องการโมเดลที่เสถียรและนิยมใช้ในข้อมูลที่มีความซับซ้อน ในขณะที่ Lasso Regression มักใช้เมื่อต้องการโมเดลที่เข้าใจง่ายและสามารถเลือกคุณลักษณะที่สำคัญได้ในข้อมูลที่มีขนาดใหญ่โดยทั่วไปแล้ว เมื่อมีการทำนายราคารถยนต์มือสอง ควรพิจารณาประสิทธิภาพและความเหมาะสมของแต่ละโมเดล โดยการทดลองและประเมินผลทางสถิติ เช่น ค่า MAE, MAPE, และ R-square เพื่อเลือกโมเดลที่ให้ผลลัพธ์ที่ดีที่สุดตามความต้องการและลักษณะของข้อมูลเมื่อมีการทำนายราคารถยนต์มือสอง ควรพิจารณาประสิทธิภาพและความเหมาะสมของแต่ละโมเดลโดยใช้เครื่องมือทางสถิติ เพื่อเลือกโมเดลที่ให้ผลลัพธ์ที่ดีที่สุดตามความต้องการและลักษณะของข้อมูล ซึ่งการประเมินผลสามารถทำได้โดยใช้ค่า MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), และ R-square (Coefficient of Determination) ซึ่งมีความหมายดังนี้ MAE (Mean Absolute Error): ค่า MAE ใช้วัดความคลาดเคลื่อนเฉลี่ยของค่าทำนายจากโมเดลกับค่าจริง มีหน่วยเดียวกับตัวแปรตามที่ทำนาย เช่น ในกรณีนี้คือ ปอนด์ (£) ค่า MAE ที่น้อยที่สุดแสดงถึงโมเดลที่ให้ค่าทำนายที่เข้าใกล้ค่าจริงมากที่สุด ซึ่งอาจจะแสดงให้เห็นถึงความแม่นยำของโมเดลในการทำนายราคารถยนต์มือสอง, MAPE (Mean Absolute Percentage Error): ค่า MAPE ใช้วัดความคลาดเคลื่อนเฉลี่ยของค่าทำนายจากโมเดลเป็นเปอร์เซ็นต์ แสดงถึงความคลาดเคลื่อนของการทำนายเมื่อเปรียบเทียบกับค่าจริง ค่า

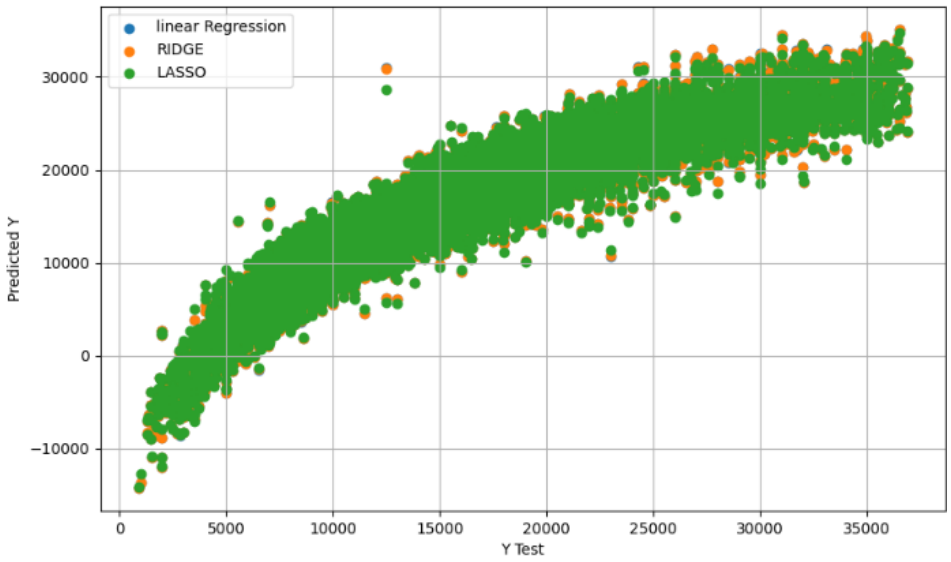
MAPE ที่น้อยที่สุดแสดงถึงความแม่นยำของโมเดลในการทำนายราคารถยนต์มือสอง โดยค่า MAPE ที่น้อยกว่า 10% ถือว่ามีประสิทธิภาพสูง, R-square (Coefficient of Determination): ค่า R-square ใช้วัดความสัมพันธ์ระหว่างค่าทำนายจากโมเดลกับค่าจริง มีค่าอยู่ในช่วง 0-1 โดยค่าที่ใกล้เคียง 1 แสดงถึงความสัมพันธ์ที่ดีของโมเดลในการทำนาย ค่า R-square ที่สูงแสดงให้เห็นถึงความแม่นยำของโมเดลในการทำนายราคารถยนต์มือสอง ดังนั้น เมื่อทำการประเมินผลโดยใช้ MAE, MAPE, และ R-square ค่าที่น้อยที่สุดหรือสูงที่สุดในแต่ละตัวแปรจะแสดงให้เห็นถึงโมเดลที่ให้ผลลัพธ์ที่ดีที่สุดตามความต้องการและลักษณะของข้อมูลที่มีในงานที่กำลังจะใช้โมเดลในการทำนายราคารถยนต์มือสอง ทั้งนี้ผมพลอตกราฟ Scatter เพื่ออธิบายผลลัพธ์ที่ได้ดังภาพประกอบ 70



ภาพประกอบ 70 การพล็อตกราฟ Scatter Plot เปรียบเทียบ 2 แบบจำลอง

จากภาพการพล็อตกราฟ Scatter สำหรับ Decision Tree Regression และ Random Forest Regression เพื่อวิเคราะห์ผลลัพธ์การทำนายราคารถยนต์มือสองนี้ แกน x แทนค่าราคาจริง (Actual Price) ซึ่งมากจากค่า Y Test และแกน y แทนค่าที่ทำนายราคารถยนต์มือสองได้ (Predicted Price) จากแต่ละแบบจำลอง จะสังเกตได้ว่าโดยจุดที่แตกต่างกันบนกราฟแสดงถึงความคลาดเคลื่อนของค่าทำนายของแต่ละแบบจำลองเมื่อเทียบกับค่าจริง หากจุดกระจายมีการกระจายตัวเป็นระเบียบและติดกับเส้นทแยงมุมหรือเส้นตรง แสดงว่าโมเดลทำนาย

ได้ใกล้เคียงกับค่าจริง เนื่องจากจุดทำนายตกลงมาบริเวณเส้นตรงหรือเส้นทแยงมุม อย่างไรก็ตาม หากจุดกระจายมีการกระจายตัวอย่างมากและห่างจากเส้นตรงหรือเส้นทแยงมุม แสดงว่าโมเดลมีความคลาดเคลื่อนในการทำนายและอาจให้ผลลัพธ์ที่ไม่ถูกต้องมากขึ้น ดังนั้นสังเกตได้ว่า Decision Tree Regression และ Random Forest มีค่าทำนายที่แม่นยำทั้งคู่โดยการพิจารณาจุดกระจายและความเข้ากันได้ของค่าราคาจริงและค่าที่ทำนายได้ หากจะทำไปใช้ในการทำนายราคารถยนต์ควรในงานจริงควรพิจจารณาเช่น Decision Tree Regression: หาก Scatter Plot แสดงถึงจุดกระจายที่มีความสอดคล้องกับเส้นทแยงมุมหรือเส้นตรง และมีค่าทำนายที่ใกล้เคียงกับค่าราคาจริง โมเดล Decision Tree Regression อาจเป็นตัวเลือกที่ดี เนื่องจากมีความสามารถในการจับความซับซ้อนของข้อมูลและเหมาะสมกับราคาที่มีลักษณะซับซ้อนและมีความสัมพันธ์ที่ซับซ้อนหาก Random Forest Regression: Scatter Plot แสดงถึงจุดกระจายที่มีความสอดคล้องกับเส้นทแยงมุมหรือเส้นตรง และมีค่าทำนายที่ใกล้เคียงกับค่าราคาจริงอย่างมาก โมเดล Random Forest Regression อาจเป็นตัวเลือกที่ดี เนื่องจากสามารถประมวลผลข้อมูลแบบประมาณการและลดความเอียงของโมเดลได้



ภาพประกอบ 71 การพล็อตกราฟ Scatter Plot เปรียบเทียบแบบจำลอง Linear Regression, Ridge และ Lasso

จากภาพกราฟ Scatter Plot นี้ เราได้ทำการพล็อตค่าที่ทำนายได้ (Predicted Y) จากโมเดล Linear Regression, Ridge, และ Lasso ในแกน y และค่าราคาจริง (Y Test) ในแกน x จุดที่ปรากฏบนกราฟแสดงค่าราคาจริงและค่าที่ทำนายได้จากแต่ละโมเดล โดยในกรณีนี้ เราได้พล็อตจุดสำหรับ Linear Regression, Ridge, และ Lasso ในกราฟ Scatter Plot นี้ เราสามารถ

เปรียบเทียบความแม่นยำของโมเดล Linear Regression, Ridge, และ Lasso ในการทำนายราคา โดยการพิจารณาจุดกระจายและความเข้ากันได้ของค่าราคาจริงและค่าที่ทำนายได้ หากเราสังเกตจุดกระจายที่มีความสอดคล้องกับเส้นทแยงมุมหรือเส้นตรง และค่าทำนายใกล้เคียงกับค่าราคาจริง จะแสดงว่าโมเดลทำนายได้แม่นยำและใกล้เคียงกับค่าจริง จากรูปสามารถเห็นได้ว่าโมเดล Linear Regression มีความแม่นยำในการทำนายราคารถยนต์มือสองค่าราคาที่ใกล้เคียงกับค่าจริง ในขณะที่โมเดล Ridge และ Lasso อาจมีความแม่นยำที่ดีกว่า โดยทั่วไปแล้ว Ridge และ Lasso ช่วยลดความเกินแบบจำลองที่อาจเกิดขึ้นจากการ Overfitting

ดังนั้น หากเราพิจารณาจากกราฟ Scatter Plot ในงานทำนายราคารถยนต์ สามารถสรุปได้ว่าโมเดล Linear Regression, Ridge, และ Lasso มีความแม่นยำในการทำนายราคาต่างกัน โดยอาจพิจารณาแบบจำลองที่มีการกระจายและความสอดคล้องกับเส้นทแยงมุมมากที่สุดเป็นแบบจำลองที่เหมาะสมสำหรับงานทำนายราคารถยนต์

5.3 **ข้อเสนอแนะ**

- อาจเพิ่ม feature engineering เพิ่มเติมเพื่อช่วยในการทำนายราคาของรถยนต์มือสองได้แม่นยำขึ้น
- ควรพิจารณาใช้งานแบบจำลองอื่นๆ เช่น Gradient Boosting, Support Vector Machine เพื่อเปรียบเทียบผลลัพธ์และหาวิธีที่ให้ผลลัพธ์แม่นยำสูงยิ่งขึ้น
- ศึกษาปัจจัยอื่นๆเช่นปัจจัยทางภูมิศาสตร์
- ศึกษาเรื่องการวัดประสิทธิภาพด้วยตัวชี้วัดต่างๆเพื่อนำไปใช้กับงานจริง

บรรณานุกรม

Uncategorized References

Al-Turjman, F., Hussain, A. A., Alturjman, S., และ Altrjman, C. (2022). Vehicle Price Classification and Prediction Using Machine Learning in the IoT Smart Manufacturing Era. *Sustainability, 14*(15), 9147.

Bharambe, P., Bhargav Bagul, Shreyas Dandekar. (2022). Used Car Price Prediction using Different Machine Learning Algorithms,. *Ijraset, 10*(4), 773-778

Gegic, E., Isakovic, B., Keco, D., และ Masetic, Z. (2019). Car price prediction using machine learning techniques. *TEM Journal, 8*(1), 113-118.

J. Varshitha, K. J. a. C. L. (2022). Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning. *IEEE*, 1-4.

K.Samruddhi, D. R. A. K. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. *International Journal of Innovative Research in Applied Sciences and Engineering, 4*(2).

M. Hankar, M. B. a. A. B.-H. (2022). Used Car Price Prediction using Machine Learning: A Case Study,. *IEEE*, 1-4.

N. Monburinon, P. C., T. Kaewkiriya, S. Rungpheung, . (2018). Prediction of prices for used car by using regression models,. *IEEE*, 115-119.

Narayana, C. V., และ Likhitha, C. L. (2021). Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business. *IEEE*, 1680-1687.

prakai. (2021). วิเคราะห์ตลาด ‘รถยนต์มือ 2’ อีกหนึ่งธุรกิจที่เติบโตแรง บูมเพราะความแพนิคในช่วงโควิด.

S. Shaprapawad, P. B. a. N. K. (2023). Car Price Prediction:An Application of Machine Learning. *IEEE*, 242-248.

Y. Li, Y. L. a. Y. L. (2022). Research on used car price prediction based on random forest and LightGBM. *IEEE*, 539-543.

ประวัติผู้เขียน