พัชระ สุวรรณบดินทร์ (6636058)

พัฒนาโมเดลที่ช่วยประเมินราคาที่เหมาะสมของ laptop ที่วางขายอยู่ในช่วงเดือนตุลาคม 2023

Introduction/Problems: จากช่วงนี้คนมักจับจ่ายใช้สอย online มากขึ้นรวมถึงการเลือกซื้ออุปกรณ์อิเล็กทรอนิกส์อย่าง laptop จากผู้ค้าออนไลน์ต่างๆ เราจึงสนใจถึงปัญหาในการป้องกันการโดนโก่งราคา จากผู้ค้าออนไลน์ โดยได้รวบรวมข้อมูลราคาตลาดของ laptop มาจาก amazon.com ในช่วงปี 2023 [1] เพื่อนำมาวิเคราะห์ และทำ model ที่ช่วยประเมินราคาที่เหมาะสมของ laptop นั้นจาก spec ของ laptop

Method:

1. Data Source: Laptop Prices Dataset - October 2023 [1]
2. Background about data: เป็นข้อมูล ของโน๊ตบุ๊คที่เก็บรวบรวมมาจาก website Amazon.com ในช่วงเดือนตุลาคม 2566 โดยมีข้อมูล รายละเอียดสินค้า, ราคา, คะแนนความพอใจ ที่มีขนาดประมาณ 4500 records กับ 14 column

COLUMN NAME	MISSING VALUE	DATA TYPE	DESCRIPTION	DATA SAMPLE
brand	0%	object	ยี่ห้อ	Lenovo, acer, HP
model	26%	object	รุ่น	MacBook Pro, CB315-3HT, ROG Flow Z13
screen_size	1%	object	ขนาดของหน้าจอ (นิ้ว)	15.66 Inches, 16 Inches
color	13%	object	สี	Blue, Silver, Midnight
harddisk	13%	object	Hard disk ที่ติดตั้งมา	1000, 512, 1 TB, 1.92 TB, 32 MB
cpu	2%	object	หน่วยประมวลผล	3. 3, 1.9, 3 GHz, 2133 MHz , 3200 MH

ram	1%	object	หน่วยความจำ	16 GB, 64 MB
OS	1%	object	ระบบประมวลผล	Windows 11 Home, Mac OS
special_features	54%	object	ส่วนเสริมอื่นๆ	Backlit Keyboard
graphics	1%	object	ประเภทการ์ดจอ	Integrated, Dedicated
graphics_coprocessor	42%	object	การ์ดจอ	NVIDIA GeForce RTX 4070, Intel
cpu_speed	66%	object	ความเร็วหน่วยประมวลผล	9120, 1.8
rating	51%	float64	คะแนนจากผู้ใช้งาน 0 ถึง 5	5, 4.8
price	0%	object	ราคา	$1,599.00, $999.99

3. Preprocess:
    a. จากที่เห็นได้ชัดด้วยการดูตัวอย่างข้อมูลเลย คือทำการแปลงข้อมูลประเภท numerical ให้อยู่ใน format เดียวกันเช่น screen_size, harddisk, ram, price และแปลง ข้อมูล categorical ให้อยู่ใน format เดียวกันเช่น OS
    b. จัดการกับ missing value โดยจัดการกับตัว feature ที่มี missing value ค่อนข้างเยอะโดยจะมีการ drop features ทิ้งได้แก่ model, color, special_features, graphics_coprocessor, cpu_speed, rating และเติม harddisk ด้วยค่ากลาง (median) และพวก categorical columns ด้วยการเติม mode และลบข้อมูล records ที่ ไม่มี price ทิ้ง
    c. Visualize ข้อมูลดูความสัมพันธ์ข้อมูลต่างๆกับราคาเช่น ยี่ห้อกับราคา (fig.1) ราคากับขนาดหน้าจอ (fig2), ราคากับ ram (fig3) และอื่นๆ
    d. Transform ข้อมูลประเภท numerical ด้วย Standardize และ encode ข้อมูลประเภท categorical ด้วย one-hot encoding
    e. ทำ train-test split โดยแบ่ง train 80 test 20

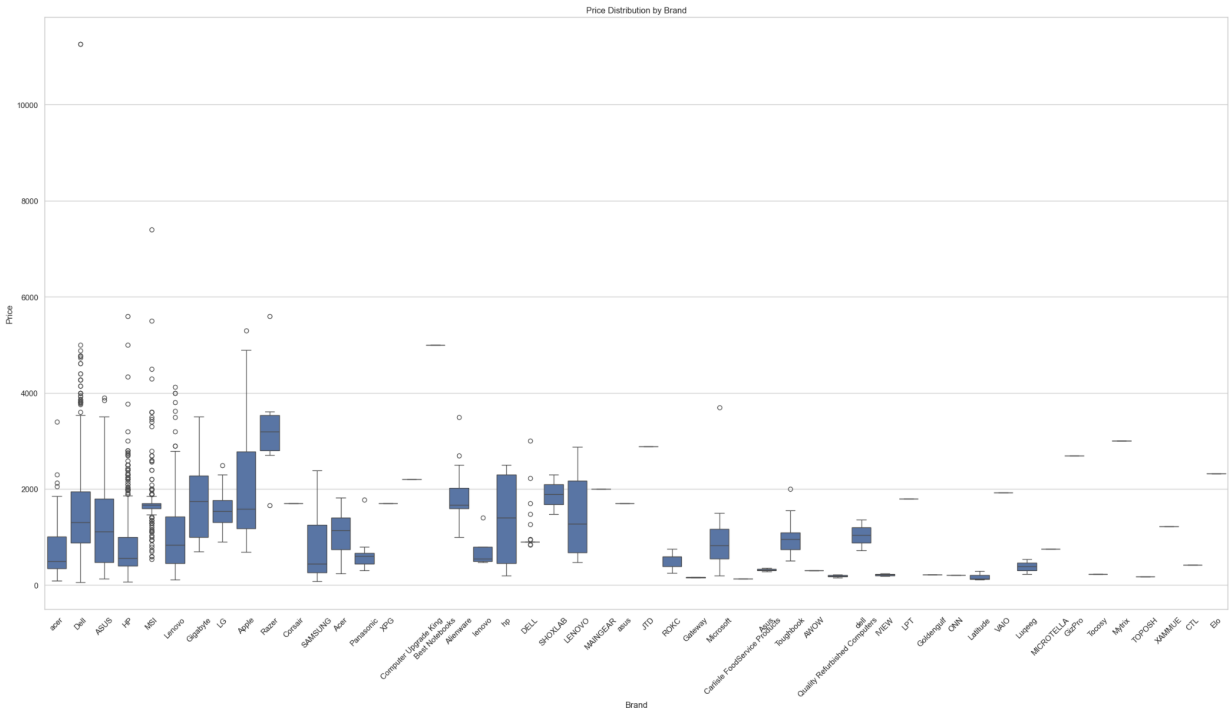

Fig.1 Price distribution by brand

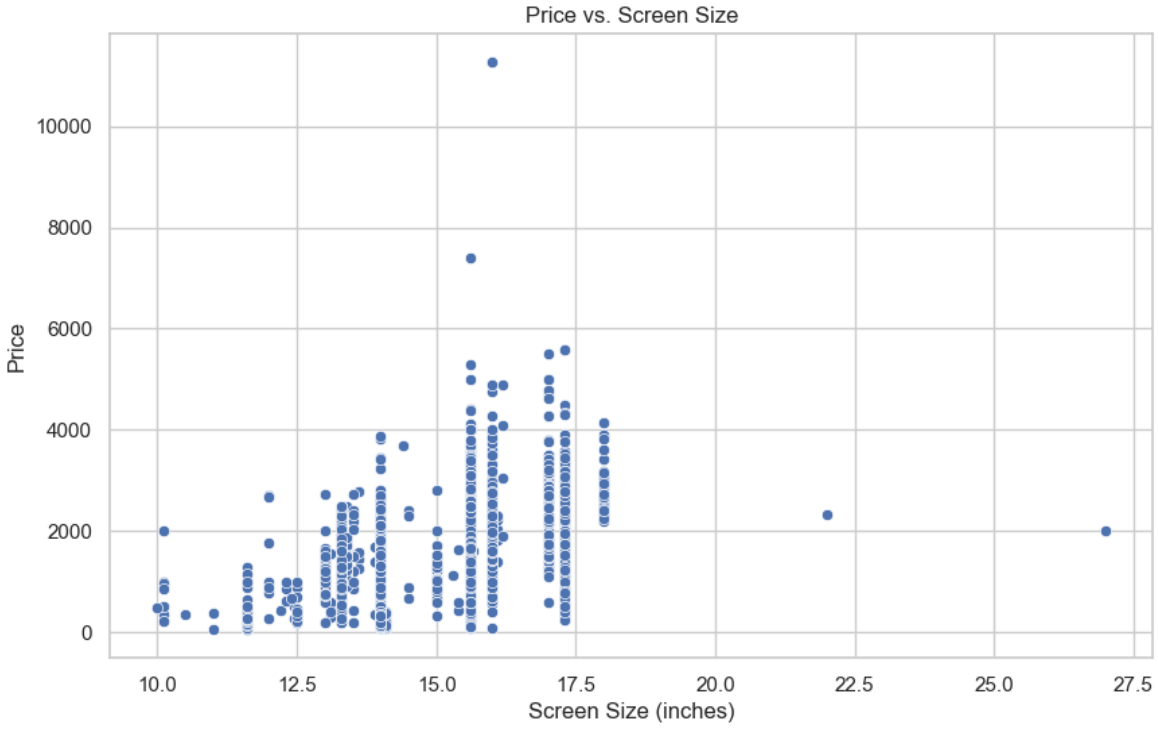

Fig.2 ความสัมพันธ์ระหว่างยี่ห้อกับราคา

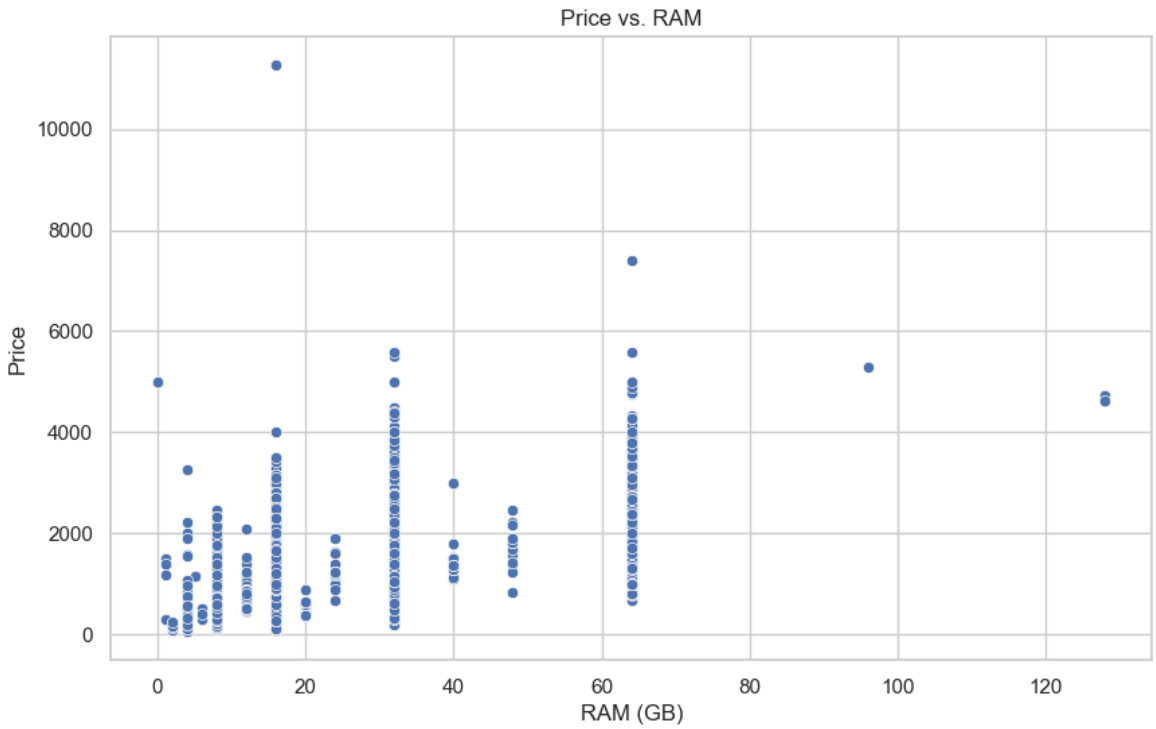

Fig.3 ความสัมพันธ์ระหว่างหน่วยความจำกับราคา

Experimental Results: จากการปรับข้อมูล และจัดการเตรียมข้อมูลแล้วจะเหลือข้อมูลอยู่ทั้งหมด 4441 records โดยแบ่งเป็น train 3552 และ test 889 และนำมาทดสอบกับ model 3 แบบและวัดผล root mean squared error และ R2 ได้ผลดังนี้

1. Linear Regression: RMSE = 478.557, R2 Score = 0.697
2. Random Forest: RMSE = 471.966, R2 Score = 0.706
3. Gradient Boosting: RMSE = 443.282, R2 Score = 0.740

จากการทดลอง 3 วิธีได้ผล R2 ของ Gradient Boosting ดีที่สุด จึงนำ Gradient Boosting มาใช้เพื่อทำนายราคาประเมินของ laptop หลังจากที่ทดสอบการทำนายแล้วจึงลองปรับประสิทธิภาพ model ด้วยการใช้ Randomized search on hyper parameters เพื่อช่วยหา parameters เพื่อปรับประสิทธิภาพให้ดีขึ้น จึงได้ผล RMSE = 422.760, R2 Score = 0.764 ที่ดีขึ้น และได้ผลเทียบการทำนายราคากับราคาจริงดังภาพ (Fig.4) จึงนำไปต่อยอดโดยการทำ restAPI (Fig.5) ที่ request ด้วย model input และตอบกลับเป็น ราคาประเมิน กับ label ที่จะบอกว่าราคาสมเหตุสมผลหรือไม่ ซึ่งถ้าราคาซื้อขายไม่ต่างกับราคาประเมินเกิน 15% ก็จะถือว่าราคาสมเหตุสมผล

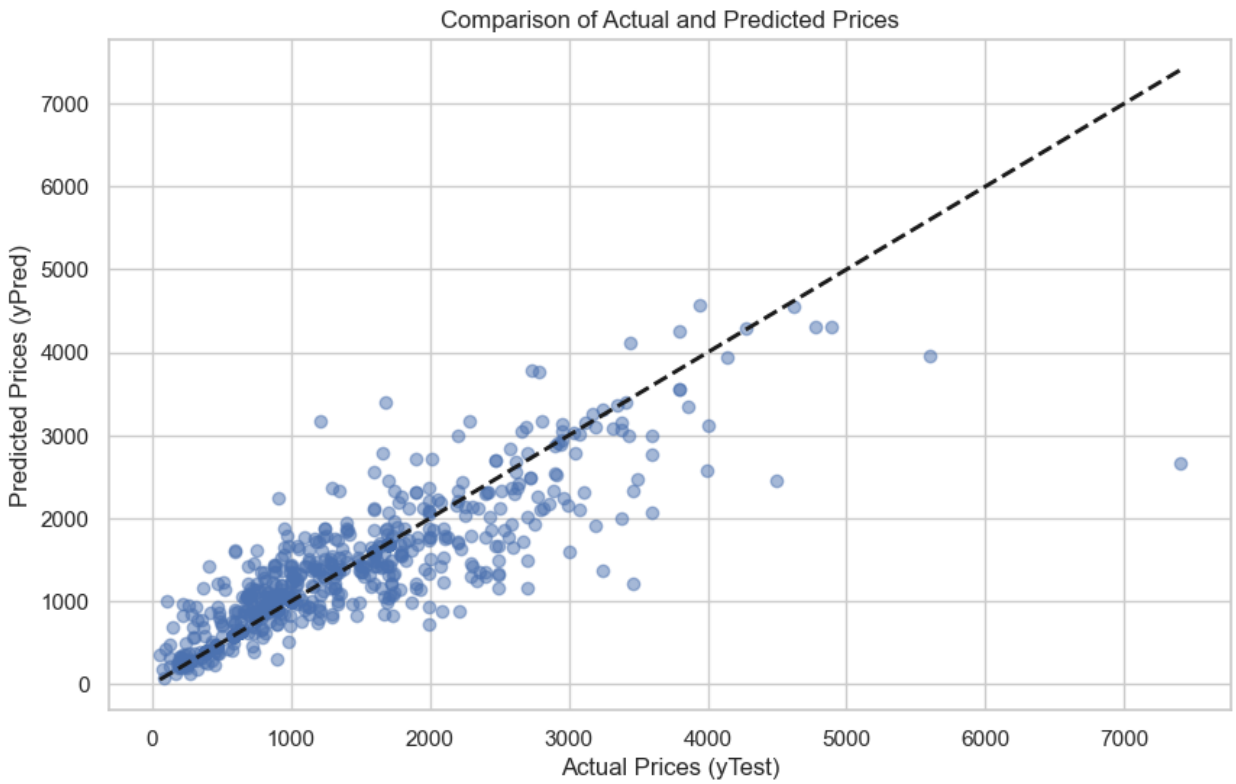

Fig.4 กราฟเปรียบเทียบราคาประเมินกับราคาจริงของข้อมูล test

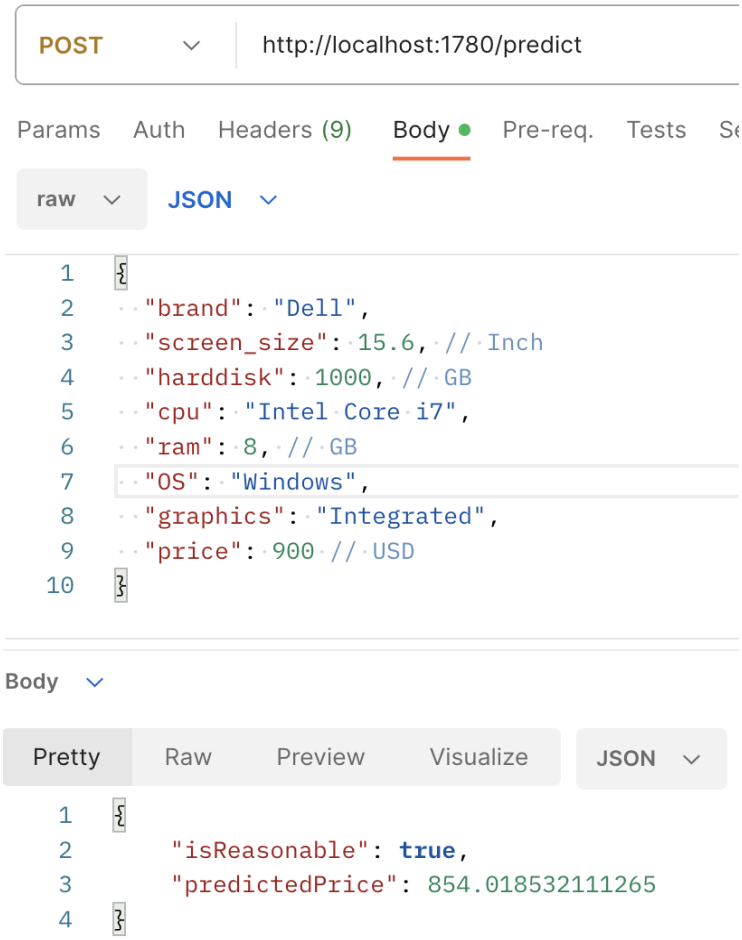

Fig.5 ตัวอย่างการใช้งาน API

Conclusion: จากการทำการทดลอง เริ่มด้วยการทำการจัดการกับชุดข้อมูลบางส่วน ต่อด้วยการ visualize ข้อมูลเพื่อดูความสัมพันธ์ แล้วนำไปทำการเตรียมข้อมูลเพื่อเทรน model ด้วยเทคนิค Gradient Boosting ให้ผลลัพท์เป็นไปในทางที่น่าพอใจ จึงได้นำผลลัพท์ของการทดลอง ไปต่อยอดทำเป็น restAPI เพื่อออกมาใช้งานได้ง่ายขึ้น

Ref:

[1]:https://www.kaggle.com/datasets/talhabarkaatahmad/laptop-prices-dataset-october-2023/data