



หลอกขายรถมือสองราคาถูก

พฤติกรรม

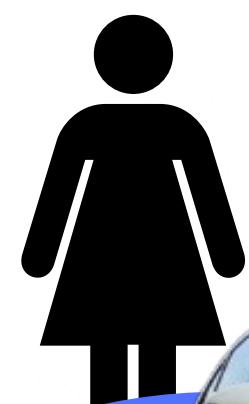
จักรกพ
มอเตอร์
จ.มหาสารคาม

โพสต์ขาย
รถมือสอง
ทางเฟซบุ๊ก

ราคาถูก
กว่าตลาด
2-3
หมื่น

ว่าง
ต้องรอนาน
20-25 วัน

Which one is **MORE** Expensive?



เมย, 29
คนกรุงเทพ



The Luxury Sedan



บอย, 32
คนอุบลราชธานี



The 4WD Pickup

BENZ E350e 2.0 AVANTGARDE

- สีขาว 2000 cc ปี 2018
- เลขไมล์ 88,139
- ทะเบียน กรุงเทพมหานคร
- Hybrid Fuel

Ford Ranger Raptor Double Cab 2.0

- สีดำ 2000 cc ปี 2018
- เลขไมล์ 76,888
- ทะเบียน กรุงเทพมหานคร
- Diesel Fuel

THB 1,190,000



BENZ E350e 2.0 AVANTGARDE

- สีขาว 2000 cc ปี 2018
- เลขไมล์ 88,139
- จ. กรุงเทพมหานคร
- Hybrid Fuel

**Would the model guess
better than you?**

This project showcased how a data science framework could accurately predict the used cars price



Price Revealed !!

THB 1,250,000



Only a 60,000 THB difference.



Ford Ranger Raptor Double Cab 2.0

- สีดำ 2000 cc ปี 2018
- เลขไมล์ 76,888
- จ. อุบลราชธานี
- Diesel Fuel

Final Project



Used Car Price Evaluation Tool

Group เอ็น ไก่กอด

2603498 DATA SCI PRAC



The image shows a promotional graphic for the Roddonjai used car price evaluation tool. It features a smiling man in an orange jacket holding a tablet that displays the Roddonjai app's user interface. The app screen shows several cars and a prominent banner with the text "คัดมาแต่รถโฉนดๆ ก็รอดูนิจ". Below the tablet, there is a list of three key features: 1. ตรวจสอบสภาพโดยคนกลาง มาตรฐานสากลสูงสุด 274 จุด 2. มีราคากลางมาตรฐาน ช่วยตัดสินใจในการซื้อ 3. จัดไฟแนนซ์ได้ทุกคัน โดย ทีมบีดีเพอร์ฟ. At the bottom, there is a call-to-action button with the text "เว็บไซต์ซื้อขายรถมือสอง" and the website URL "Roddonjai.com" with a magnifying glass icon.

- ✓ ตรวจสอบสภาพโดยคนกลาง มาตรฐานสากลสูงสุด 274 จุด
- ✓ มีราคากลางมาตรฐาน ช่วยตัดสินใจในการซื้อ
- ✓ จัดไฟแนนซ์ได้ทุกคัน โดย ทีมบีดีเพอร์ฟ

เว็บไซต์ซื้อขายรถมือสอง

Roddonjai.com 

BUSINESS PROBLEM FACTORS



ความไม่สมมาตรของข้อมูล

ผู้ซื้อและผู้ขายรถยนต์มือสองมีข้อมูลไม่เท่าเทียมกัน ทำให้เกิดการเอาเปรียบในการกำหนดราคา โดยเฉพาะกับส่วนที่ไม่สามารถประเมินได้ด้วยตาเปล่า



ความยากในการกำหนดราคา

ผู้ขายรถยนต์มือสองประสบปัญหาในการกำหนดราคาที่เหมาะสม เนื่องจากปัจจัยหลายประการ ที่ส่งผลต่อ มูลค่ารถ ทำให้การตั้งราคาอาจสูงหรือต่ำเกินไป ส่งผลให้ขายยากหรือขาดทุน



ความต้องการตลาดที่รวดเร็ว

ตลาดรถยนต์มือสองมีการเปลี่ยนแปลงอย่างรวดเร็วตามความนิยมและสภาพเศรษฐกิจ ทำให้ผู้ซื้อและผู้ขายต้องการเครื่องมือที่ทันสมัยในการประเมินราคากลางๆ ตามสภาพตลาดปัจจุบัน

OUR PROPOSED SOLUTION

Regression Predictive Algorithm เพื่อประเมินราคาของรถมือสองใน RODDONJAI จากปัจจัยต่างๆ

Impact of Car Price Evaluation Tool



พิจารณาประกอบการตัดสินใจตั้งราคา
รถยนต์ของตนเองตามเป้าหมายได้



เปรียบเทียบราคารถยนต์ที่ตนเองสนใจ
กับราคามาตรฐานของรถยนต์รุ่นนั้น

Dataset

ข้อมูลรถยนต์มือสอง จากเว็บไซต์ Roddonjai ที่
เป็นข้อมูลทุกตัวแวร์จากการทำ Web Scraping

10829 Rows

ข้อมูลเฉพาะเกี่ยวกับรถยนต์	
ยี่ห้อ	Toyota
รุ่น	Camry
รุ่นย่อย	Camry 2.0 G (MY18)
ประเภท	SUV
ประเภทย่อย	รถเก๋งขนาดใหญ่
สี	สีขาว
เกียร์	เกียร์ดอโนเบติ
เลนส์หน้า	157,069 กม.
จำพวกกันน้ำ	5
ประเภทเครื่องยนต์	Benzine
ขนาดเครื่องยนต์	2000 CC

DATA DICTIONARY

18 Features

6
numerical

12
categorical

Numerical Data

- **carYear:** ปีของรถยนต์
- **carCC:** ความจุกระบอกสูบของรถยนต์
- **carPrice:** ราคาขายปัจจุบัน
- **numberOfSeat:** จำนวนที่นั่งของรถยนต์
- **mileage:** ระยะทางที่รถวิ่ง
- **mileageAvg:** ระยะทางเฉลี่ยที่รถวิ่งต่อปี

Categorical Data

- **id:** รหัสประจำรถ
- **carUrl:** ลิงก์ไปยังหน้ารายละเอียดรถ
- **carBrand:** ยี่ห้อรถ
- **carModel:** รุ่นรถยนต์
- **carSubModel:** รุ่นย่อยของรถ
- **carGear:** เกียร์ของรถยนต์ (A = Auto, M = Manual)
- **carType:** ประเภทรถ
- **carSubSegment:** กลุ่มย่อยของรถ
- **carColor:** สีของรถ
- **fuel:** ประเภทเชื้อเพลิง
- **licensePlateProvince:** จังหวัดของป้ายทะเบียน
- **repairHistory:** ประวัติการซ่อม

Data Wrangling

no duplicated rows

- total = 10829 Rows

Drop irrelevant features : carUrl, id

- As these features not contribute to predictive modeling or insights.

Missing Values Checking

fuel	0.16 %
repairHistory	91.21 %

► **impute**

► **drop column**

id	fuel	repairHistory
CAR202311140025	diesel	เครื่องยนต์เดิมไม่เคยชนหนัก
CAR202310300172	Hybrid	
CAR202307130109	unknown	ไม่ลับอยมาก

Data Preprocessing

- Missing values
- Anomalies

► Imputation อิงจากข้อมูลจริงของ carSubModel แทนที่จะค้น



Data Preprocessing

carColor

ขาว
สีขาว
สีขาว

text correction
→

white

ฟ้า - ดำ
น้ำเงิน - ขาว
เขียว - เหลือง

combine as
one class
→

multiple_colors

mileageAvg

Extract numeric values from the string

เฉลี่ย 24,770 กม./ปี
เฉลี่ย 8,607 กม./ปี



24,770
8,607

licensePlateProvince

- มีข้อมูล จ.กรุงเทพมหานคร **50.22%**

77 จังหวัด

→ แบ่งตาม **region (6 classes)**
ex. Northern, Central, Southern
BangkokAndSurroundingAreas

isBangkok

Bangkok	▶	1
OtherProvince	▶	0

carYear

แปลงเป็น

carAge

คำนวณอายุรถเทียบจากปีปัจจุบัน : 2025 - carYear

Categorical Encoding

Features with no ordinal relationship

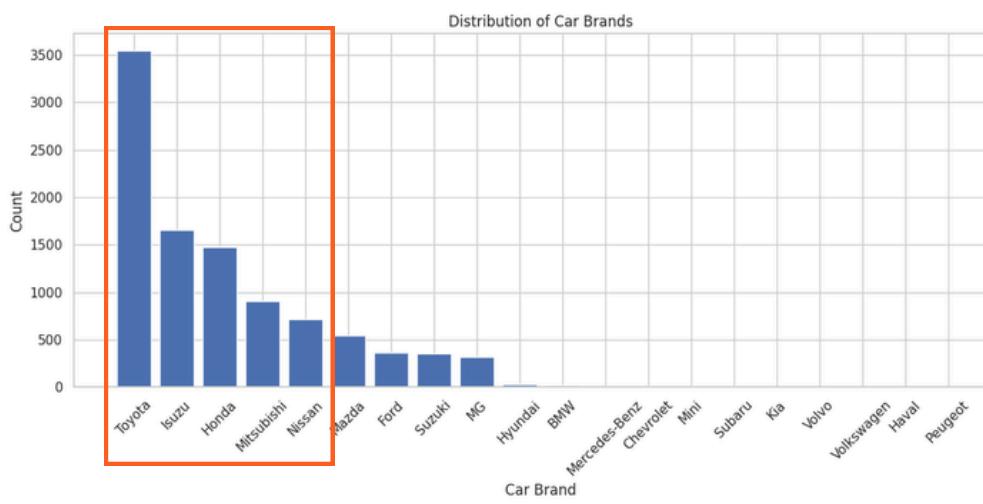
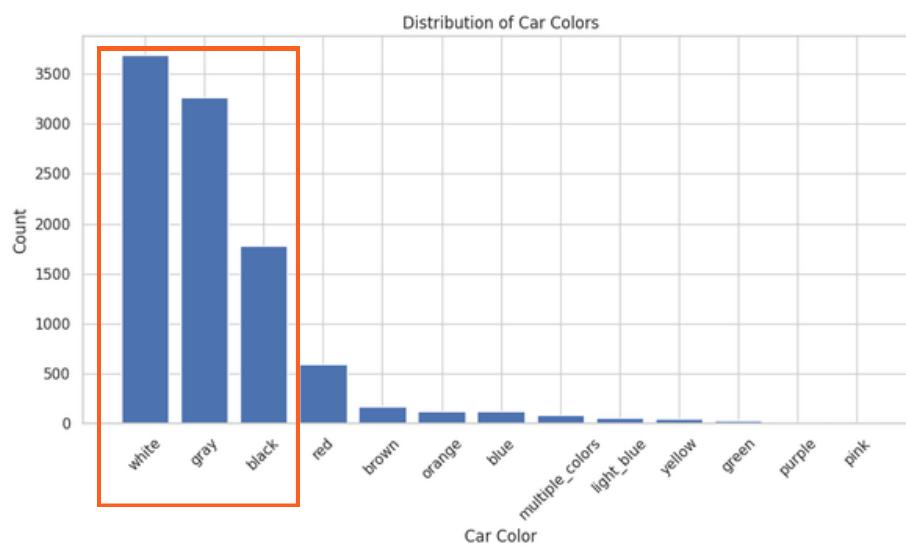


- **carSubSegment**
- **carGear**
- Region
- carType

Some categories rarely occur, which can make the model overly complex



- **carColor** : kept only top 3 colors
- **carBrand** : kept only top 5 carBrand



One hot Encoding



*done after splitting data to avoid data leakage

Categorical Encoding (cont.)

Target encoding



- **carModel** : encode with its average carPrice from the training data

	carModel	carModel_encoded
5236	Hilux Revo	517641.375000
1934	Hilux Revo	517641.375000
6461	HR-V	598458.582904
7520	Commuter	762791.002236
7245	Vios	376890.038568

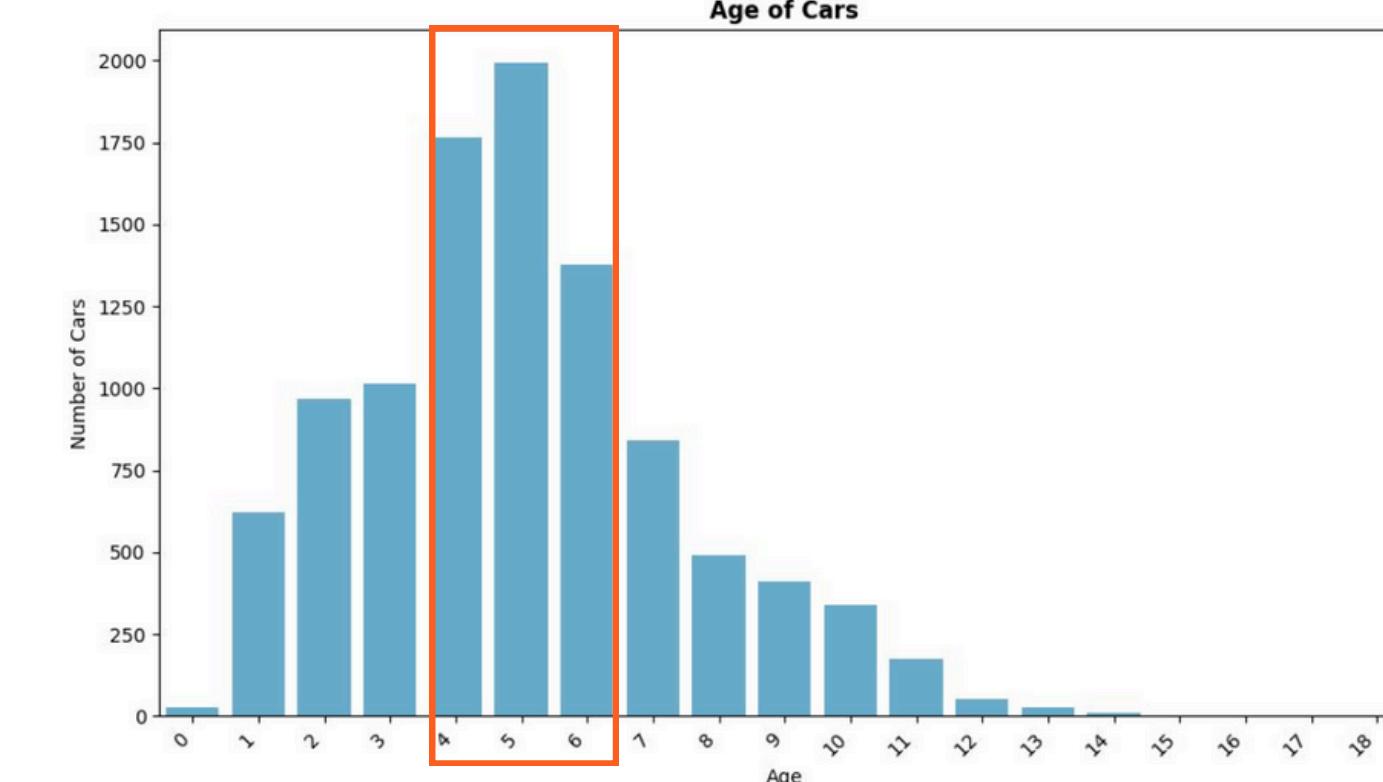
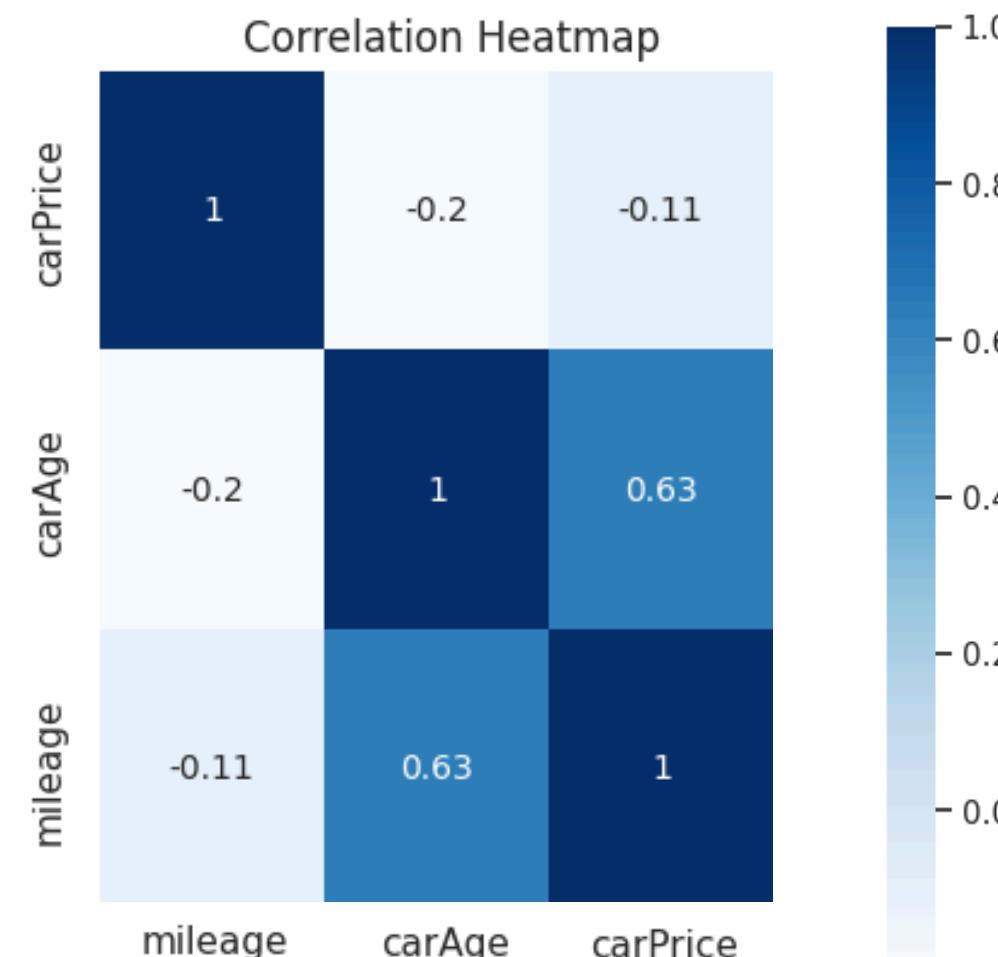
Numerical Encoding

Standard Scaling

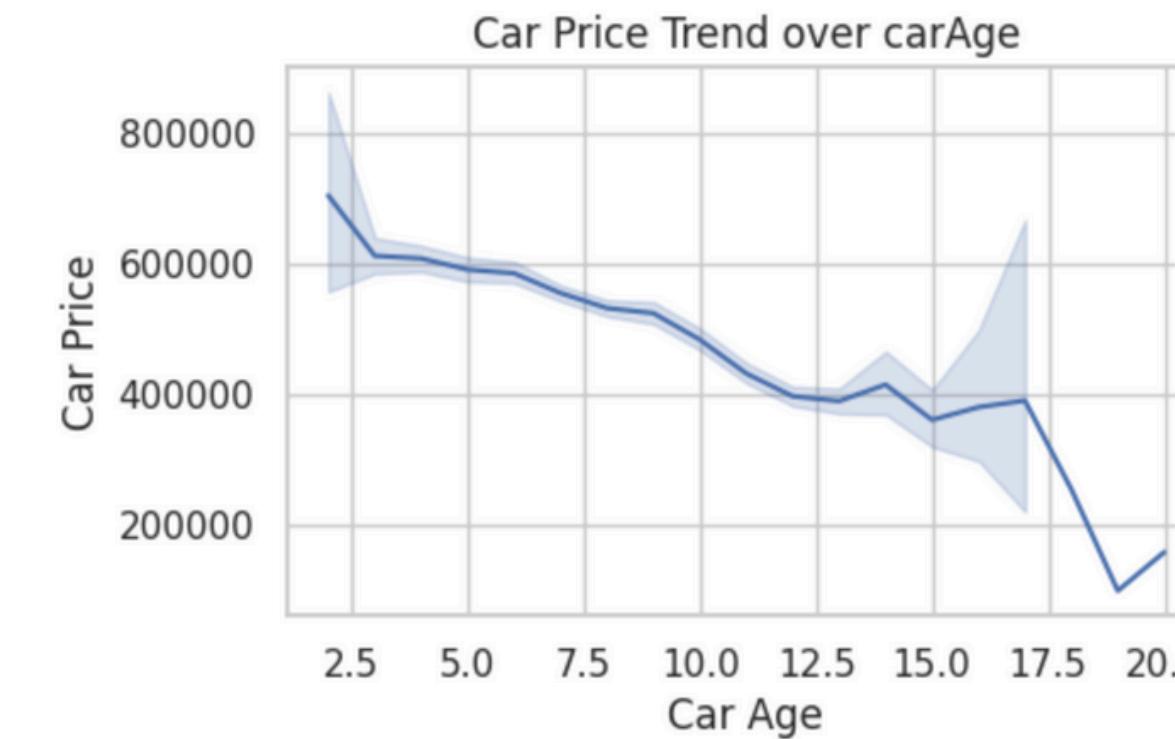


- **mileage**
- **numberOfSeat**
- **carAge**
- **carCC**

Key Insights from EDA

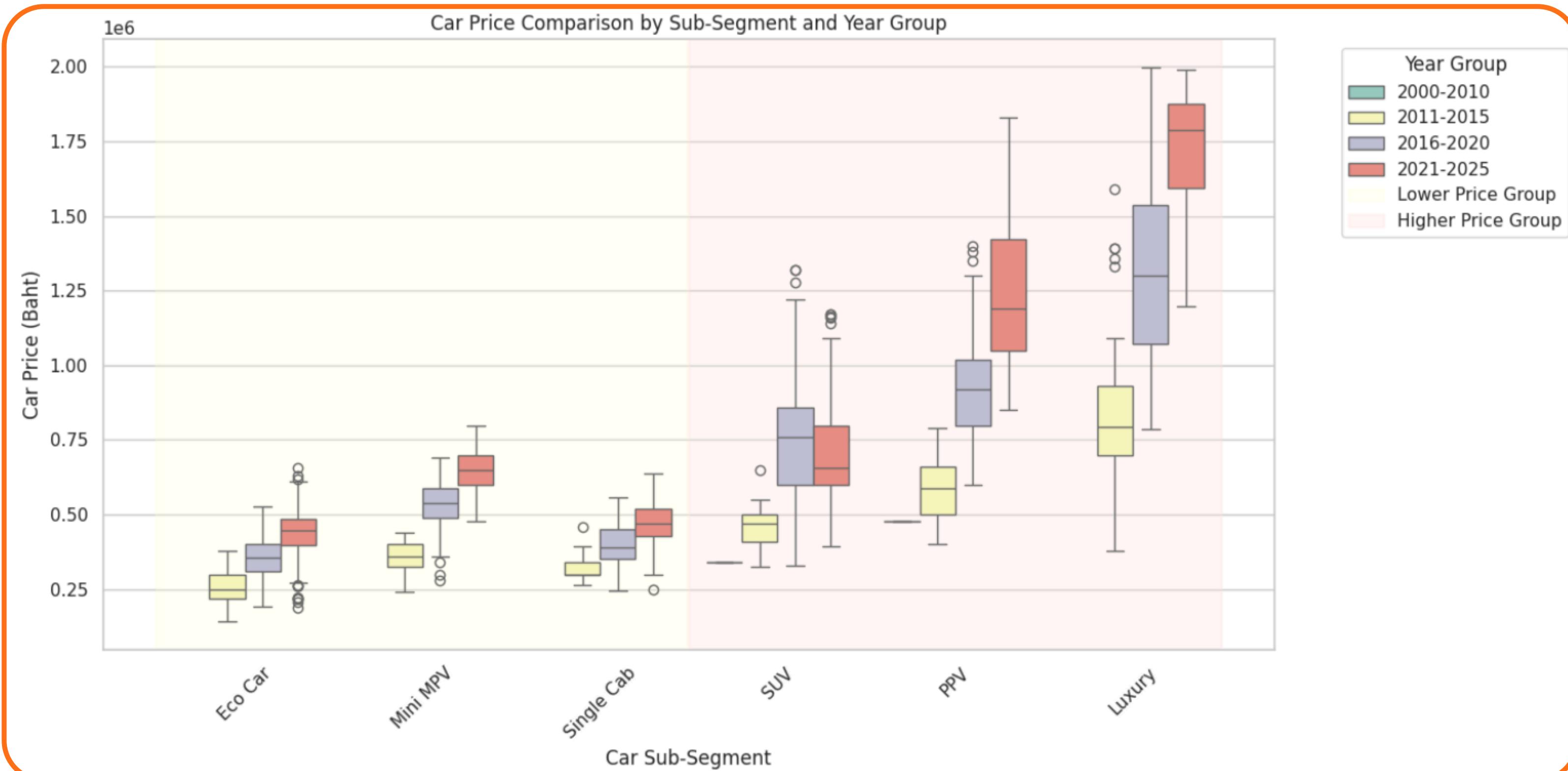


รถมือสองส่วนใหญ่มีอายุอยู่ในช่วง 4-6 ปี



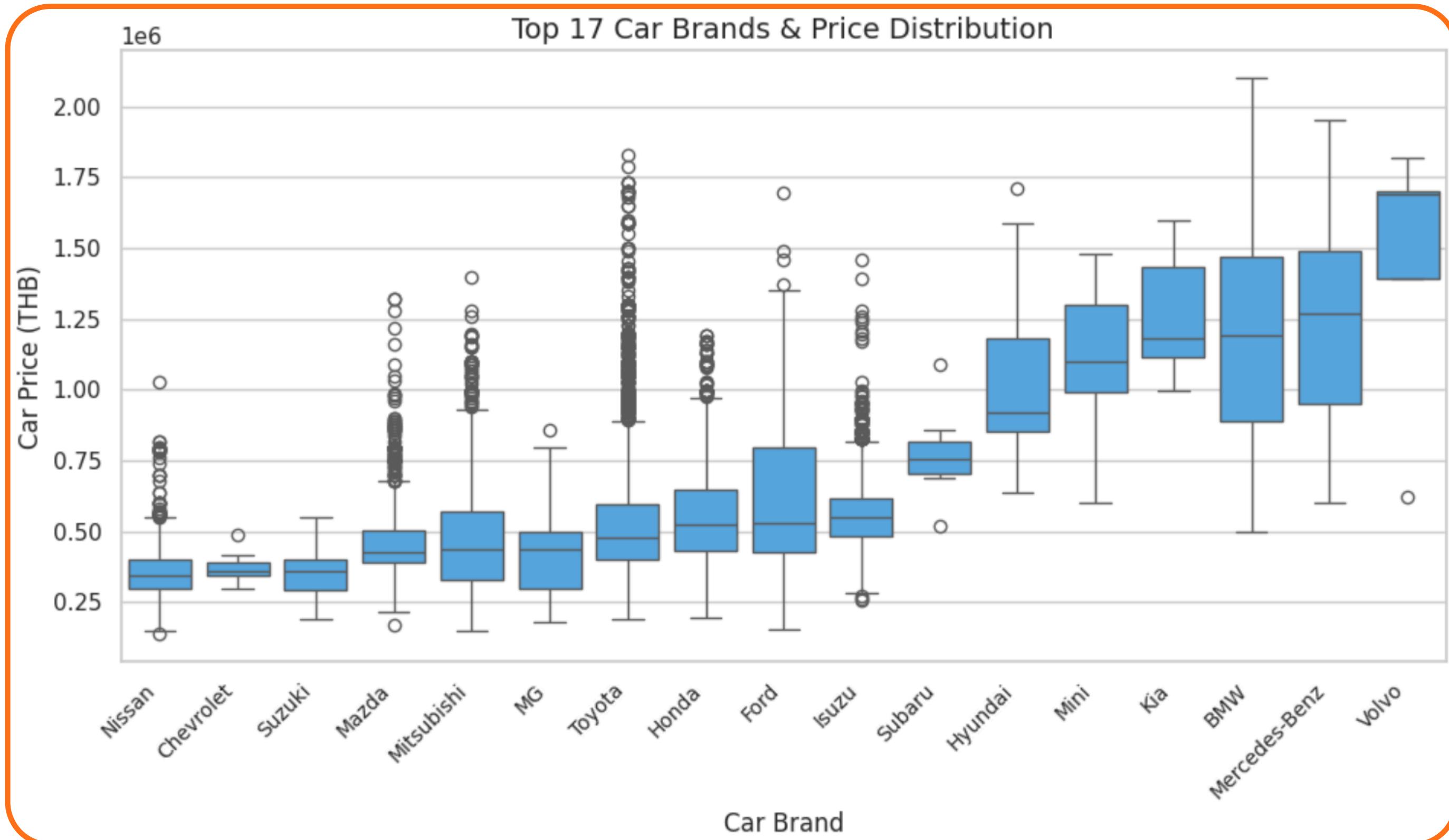
รถที่มีอายุการใช้งานมาก ยิ่งมีค่าเลขไม้ ula กว่าขึ้น ส่งผลให้ราคารถยนต์มีมูลค่าต่ำลง

Older Cars Generally Have Lower Prices Across All Car Sub-Segments



รถรุ่นใหม่ในช่วงปี 2021–2025 มีระดับราคาสูงกว่ารุ่นก่อนหน้าอย่างชัดเจน โดยเฉพาะกลุ่ม SUV, PPV และ Luxury ที่มีราคากลางค่อนข้างกว้างและเพดานราคาสูง ในขณะที่กลุ่ม Eco Car, Mini MPV และ Single Cab มีช่วงราคาคงที่และ การแข่งขันด้านราคาชัดเจน

Car Price Range Distribution by Car Brand



ข้อมูลนี้ช่วยสนับสนุนการแบ่งขันของแต่ละแบรนด์ในกลุ่มเป้าหมายที่ต่างกัน และสามารถนำไปใช้เพื่อวางแผนกลยุทธ์การตลาด
ตั้งราคาขาย หรือจัดหมวดหมู่สินค้าให้ตรงกับกลุ่มลูกค้าได้อย่างมีประสิทธิภาพ

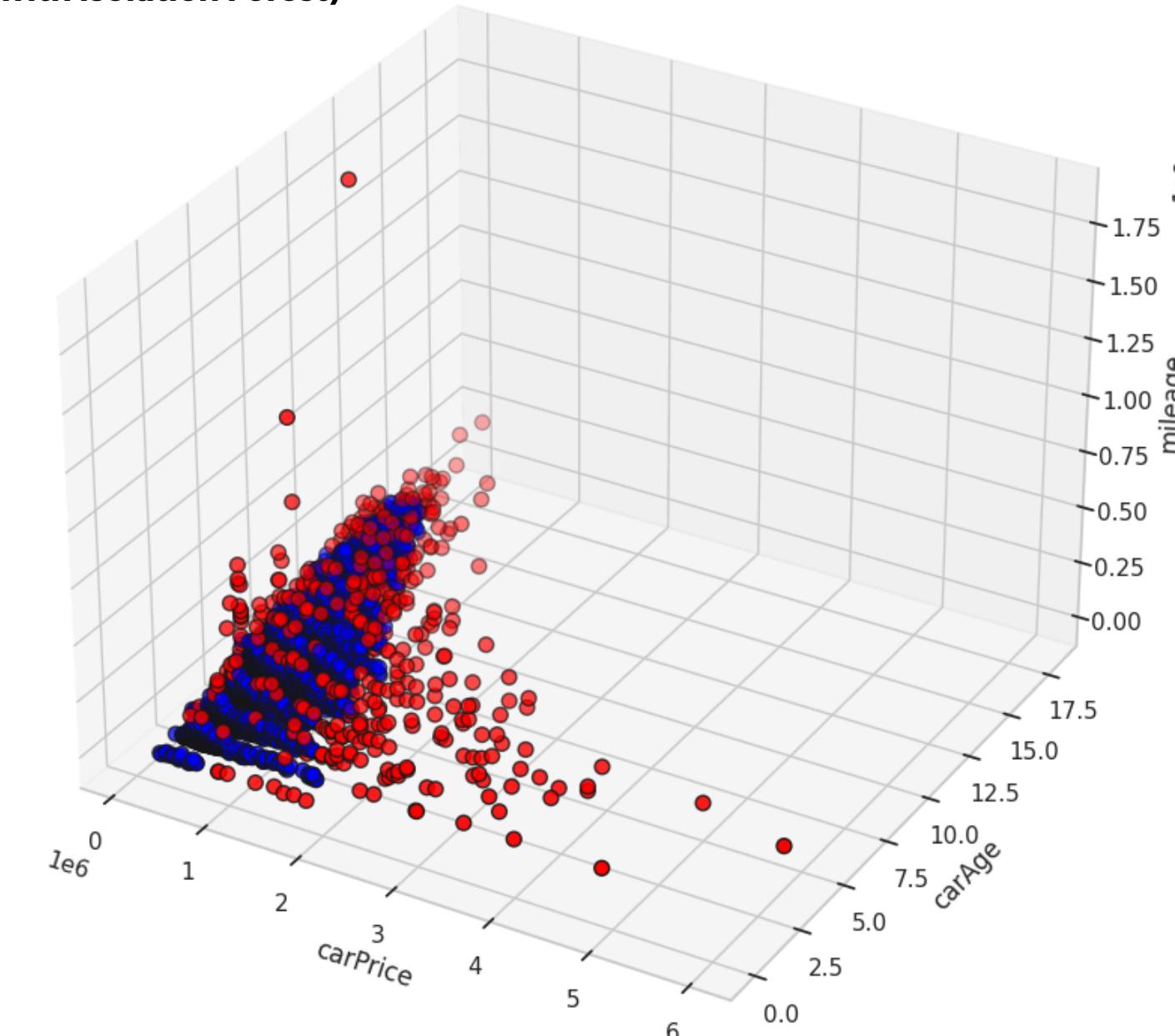
Impact of Outliers on Used Car Price Evaluation

รถกลุ่มที่มีค่าที่สูงหรือต่ำผิดปกติใน carPrice, carAge, mileage
ซึ่งเป็นส่วนน้อย อาจจะไม่สะท้อนถูกต้องของราคารถที่แท้จริง
ซึ่งอาจส่งผลให้โมเดลคลาดเคลื่อนในการประเมินราคารถส่วนใหญ่

Outliers ส่วนใหญ่ที่ถูกตัดออกไปเป็น Luxury Car ซึ่งมีราคาที่สูง
โดยกว่าร้อยละ 80% แม้จะมีอายุรถ หรือ mileage ใกล้เคียงกับรถทั่วไป

ผลลัพธ์: โมเดลมีความแม่นยำมากกว่าการไม่นำ outliers เหล่านี้ออก

Outlier in carPrice, carAge and mileage
(with Isolation Forest)



10,829 rows → 8,847 rows
outlier 1,593 rows (~15%)

STEPS ON MODELLING & EVALUATION

Regression Algorithms

Support Vector Regression

Single model

Random Forest Regression

Ensemble tree-based

CatBoost Regression

Boosting tree-based

XGBoost Regression

Boosting tree-based

Comparing Approaches with Gridsearch Parameters Tuning

1. Baseline Features

ใช้ feature ตั้งต้น สำหรับเป็น benchmark สำหรับการเปรียบเทียบ

2. Baseline Features + Feature Engineering

ตัดแต่ง features เพื่อเพิ่มประสิทธิภาพของโมเดล

3. Baseline Features + Feature Engineering + Outlier Handling

เปรียบเทียบการใช้ IQR , DBSCAN และ Isolation Forest

เพื่อตัดค่า outlier ที่ส่งผลต่อมodel

Evaluation Metric : MAPE

Why MAPE?

- บอกได้ว่าโมเดลคาดไปกี่เปอร์เซ็นต์ของราคารถโดยเฉลี่ย

ราคากลางของรถยนต์มือสองมีช่วงกว้าง
ตั้งแต่หลักแสนไปจนถึงหลักล้านบาท

Ex. ความพิดพลาด 50,000 บาท
จะมีค่ามากสำหรับรถราคา 300,000 บาท (16.7%)
แต่น้อยสำหรับรถราคา 2,000,000 บาท (2.5%)

Key Successful Criteria

Case Study

การทำนายราคารถยนต์มือสองด้วยการเรียนรู้ของเครื่อง
ชั้ววรรค ปูเตะ จันตรี ผลประเสริฐ

The study defines success based on the criteria:^{*}

MAE < 50,000 THB

MAPE <= 6-8%

R2 >= 0.8

MODEL SUMMARY

best approach on each algorithms

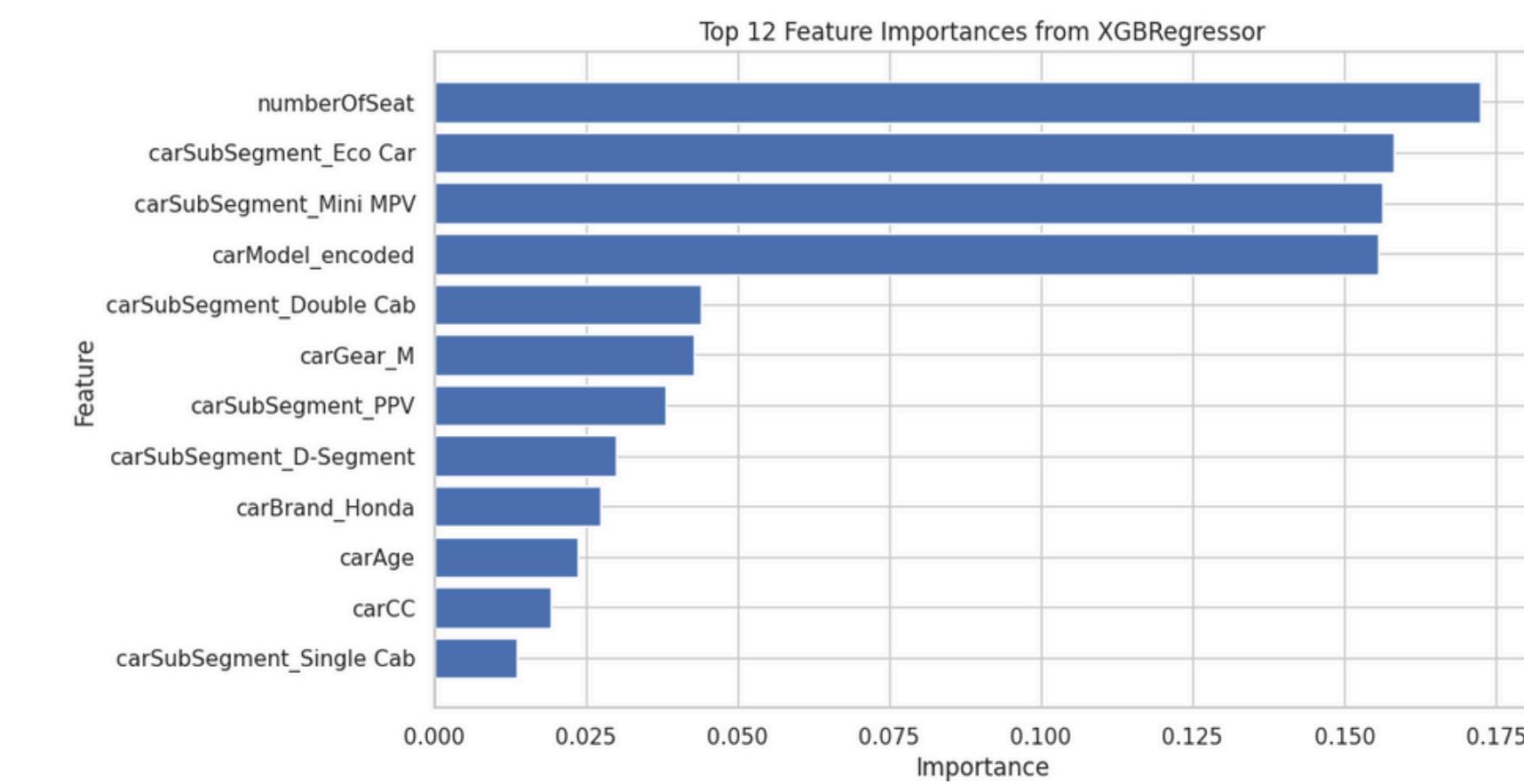
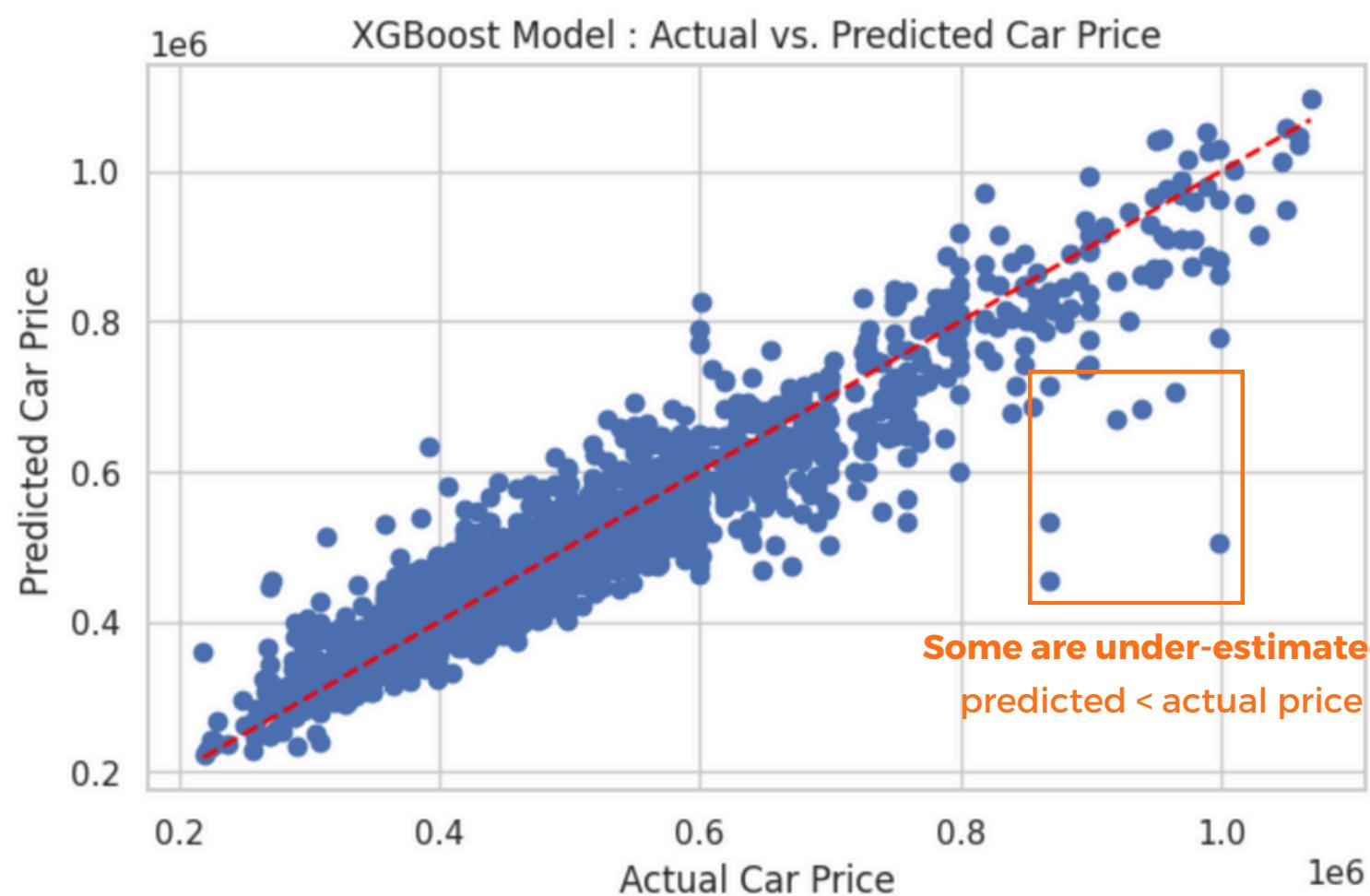
		MAE < 50,000 THB	MAPE <= 6-8%	R2 > 0.8
Support Vector Regression	train	45530.8158	0.0920	0.8465
	test	48171.6614	0.0955	0.8043
RandomForest Regressor	train	35234.8539	0.0716	0.9115
	test	42160.8954	0.0843	0.8593
CatBoost Regressor	train	43307.3063	0.0857	0.8505
	test	45619.1141	0.0894	0.8268
XGBoost Regressor	train	34956.3906	0.0709	0.9131
	test	39842.4297	0.0794	0.8661

FINAL DECISION : BEST MODEL

The model's **average prediction error** is **39,842.43 THB per car**, and averagely **7.94% of the actual price**, indicating that the model is accurate and meets the predefined criteria.



XGBoost Regression	MAE	MAPE	R2
training set	34956.3906	7.09%	0.9131
test set	39842.4297	7.94%	0.8661



KEY TAKEAWAYS



Limited representation of luxury cars

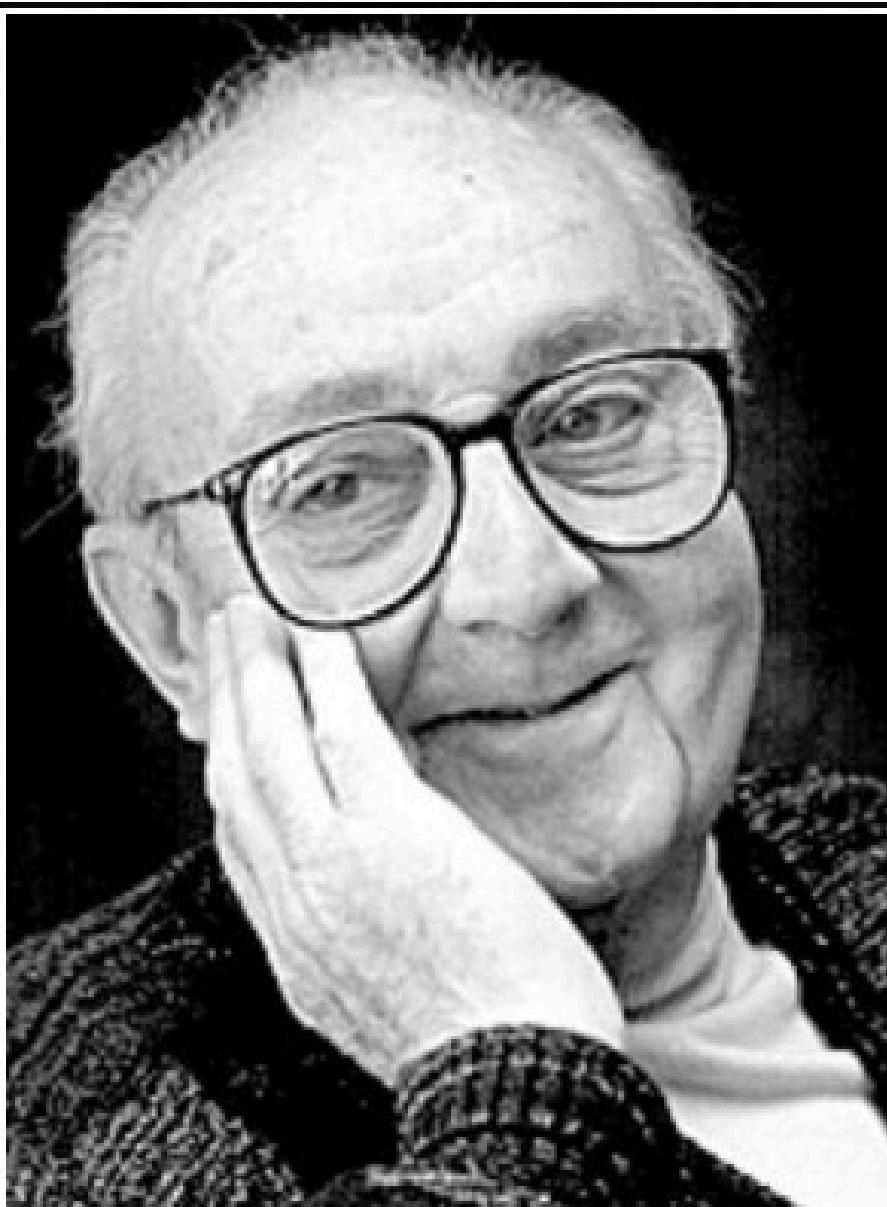
เนื่องจาก luxury car ในเว็บไซต์มีจำนวนน้อย แม้จะมีอายุการใช้งาน
หรือเลขไมล์ใกล้เคียงกับรถยนต์ก็จะไป แต่ก็มีราคาที่สูงกว่ามาก ดังนั้น
การนำข้อมูลเหล่านี้ออก จึงช่วยให้ไมเดลมีเสถียรภาพและให้ผลลัพธ์ที่
แม่นยำในการประเมินราคารถส่วนใหญ่ในเว็บไซต์ Roddonjai



Further Suggestion : เพิ่มข้อมูลด้านสภาพรถ

หากมีข้อมูลที่สมบูรณ์เกี่ยวกับการตรวจสภาพรถ เช่น
ประวัติการซ่อมบำรุง, อุบัติเหตุ, หรือ การตกน้ำ ข้อมูลเหล่านี้
จะสามารถช่วยให้ไมเดลสามารถประเมินราคาได้แม่นยำ
และสมเหตุสมผลมากขึ้น

Roddonjai Presentation



Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

— George E. P. Box —

AZ QUOTES

By : Group N

Group N

Appendix

BACKGROUND INFORMATION



Toyota Fortuner 2020
Fortuner 2.4 Legender (MY20)
บุกบุรี

● เข้าชม 1,047

⚡ รถ 5 ดาว ไม่ถึง 5 ปี

🔥 ราคาดี **1,119,000.-**

ผ่อนกับ **DRIVE** เริ่มต้นเดือนละ ₩19,034 ⓘ

83,161 กม.
เฉลี่ย 16,632 กม./ปี

★★★★★

2ND HAND

เปลี่ยนเกียบ

“รถดอนใจ” เป็นแพลตฟอร์มออนไลน์ที่ช่วยให้การซื้อและขายรถยนต์มือสองเป็นเรื่องง่ายขึ้น ด้วยระบบที่ใช้ AI และการวิเคราะห์ข้อมูล เพื่อประเมินราคารถยนต์อย่างแม่นยำ ผู้ใช้งานสามารถ ตั้งราคาที่เหมาะสม ตามสภาพรถและแนวโน้มตลาด แพลตฟอร์มนี้ยังมี bluebook ที่ช่วยอ้างอิงราคากลางของรถมือสองได้ เปิดโอกาสให้เจ้าของรถสามารถขายได้อย่างมั่นใจ โดยไม่ต้องพึ่งพาคนกลาง และช่วยให้ผู้ซื้อสามารถเลือกซื้อรถที่มีการตรวจสอบมาตรฐานจากผู้เชี่ยวชาญ

HYPERPARAMETER TUNING

Random Forest

```
# Define the Random Forest Regressor
rf_regressor = RandomForestRegressor(random_state=42)

# Define the hyperparameter grid
param_dist = [
    'n_estimators': [200], # Number of trees in the forest
    'max_features': ['sqrt'], # Number of features to consider at each split
    'max_depth': [14], # Maximum depth of the tree
    'min_samples_split': [15], # Minimum number of samples required to split a node
    'min_samples_leaf': [1], # Minimum number of samples required to be at a leaf node
    'bootstrap': [False] # Whether bootstrap samples are used when building trees
]
```

ลด Overfitting:

- จำกัดความลึกของต้นไม้ (`max_depth`)
- ป้องกันต้นไม้แตก node บ่อยเกินไป (`min_samples_split, min_samples_leaf`)
- จำกัดจำนวน feature ที่ใช้ในแต่ละต้นไม้ (`max_features`)

ลด Underfitting:

- เพิ่มจำนวนต้นไม้ (`n_estimators`)
- ทดลองใช้หรือไม่ใช้ `bootstrap`

Handle Data Imbalance:

- ใช้ `min_samples_leaf` และ `min_samples_split`
- ปรับ `bootstrap` เพื่อลดความลำเอียงในการเรียนรู้ข้อมูล



HYPERPARAMETER TUNING

XgBoost

```
# Define the XGBoost Regressor
xgb_regressor = xgb.XGBRegressor(random_state=42)

# Define the hyperparameter grid for XGBoost
param_dist = {
    'n_estimators': [320], # Number of trees in the forest
    'max_depth': [4], # Maximum depth of each tree
    'learning_rate': [0.1], # Step size shrinking
    'subsample': [0.75], # Fraction of samples to be used for fitting each tree
    'colsample_bytree': [0.7], # Fraction of features to be used for fitting each tree
    'gamma': [0], # Minimum loss reduction required to make a further partition
    'min_child_weight': [1], # Minimum sum of instance weight (hessian) needed in a child
    'booster': ['gbtree'], # Type of boosting model ('gbtree' for tree-based model)
    'alpha': [0.5,1], # L1 regularization (lasso)
    'lambda': [1,2,3], # L2 regularization (ridge)
}
```

ลดโอกาส overfitting:

- ลด max_depth=4 (จาก 6 → 4)
- ลด learning_rate=0.1 (จาก 0.3 → 0.1)
- เพิ่ม n_estimators=320 (จาก 100 → 320)

เพิ่ม regularization

- เพิ่ม alpha=[0.5,1] (L1 regularization)
- เพิ่ม lambda=[1,2,3] (L2 regularization)

จัดการ underfitting

- ปรับ colsample_bytree=0.7 (จาก 1 → 0.7)
- ปรับ subsample=0.75 (จาก 1 → 0.75)



HYPERPARAMETER TUNING

CatBoost

```
# Define the parameter grid for GridSearchCV or RandomizedSearchCV
param_grid = {
    'iterations': [300,500,700],
    'depth': [4, 6, 8],
    'l2_leaf_reg': [1, 3, 5],
    'border_count': [32]
}

# For RandomizedSearchCV (random sampling):
grid_search = RandomizedSearchCV(estimator=cat_regressor, param_distributions=param_grid, scoring='neg_mean_squared_error', n_iter=10, cv=5)

# Fit the GridSearchCV or RandomizedSearchCV object to the training data
grid_search.fit(X_cat_train, y_cat_train, eval_set=(X_cat_val, y_cat_val), early_stopping_rounds=100)
```

ลด Overfitting:

- จำกัดความลึกของต้นไม้ → `depth = [4, 6, 8]`
- เพิ่มค่า Regularization → `l2_leaf_reg = [1, 3, 5]`
- จำกัดจำนวน bin → `border_count = [32]`

ลด Underfitting:

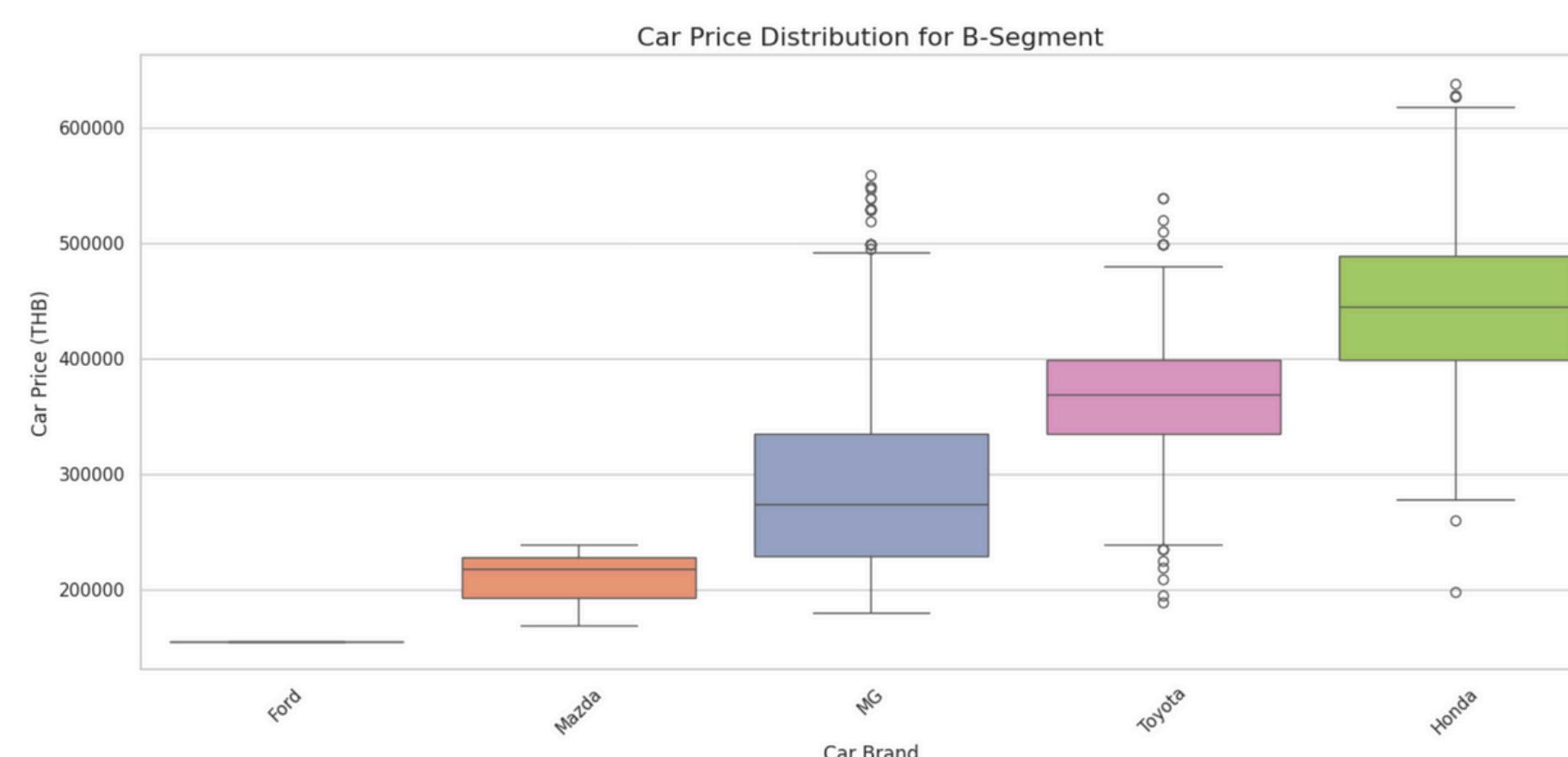
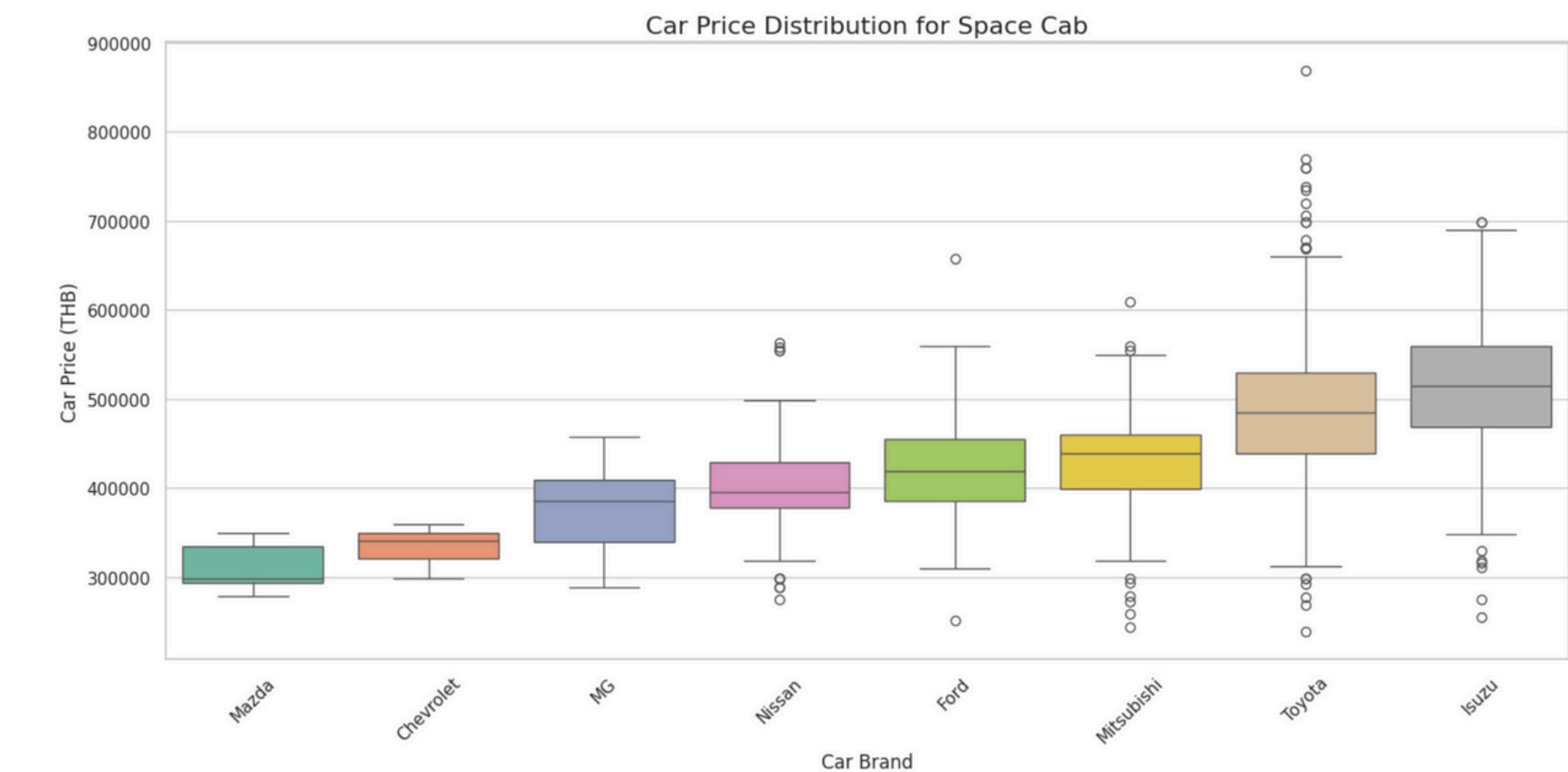
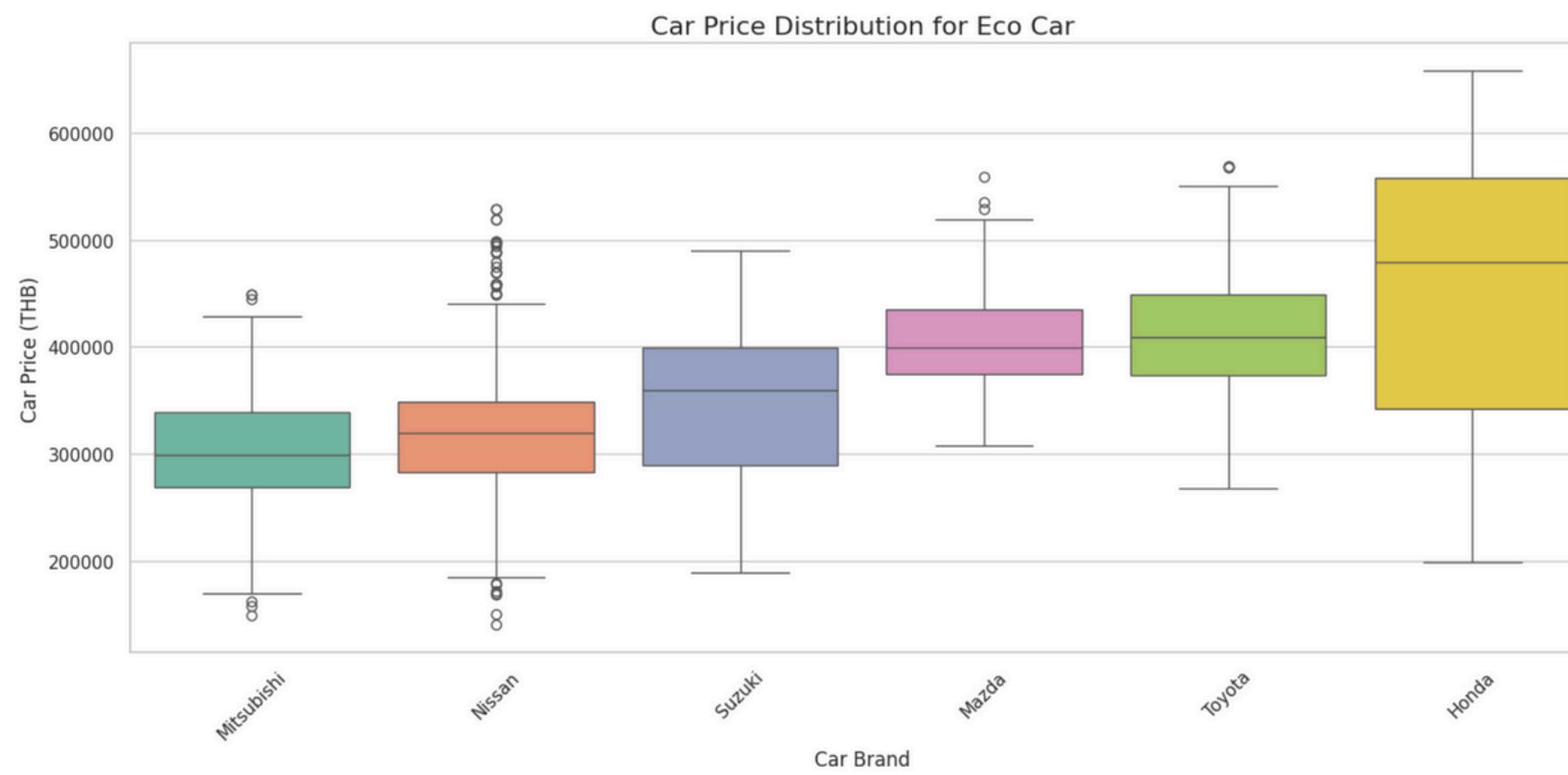
- เพิ่มจำนวนต้นไม้ → `iterations = [300, 500, 700]`

Handle Data Imbalance:

- ใช้ Early Stopping → `early_stopping_rounds=100`



กราฟแสดงการเปรียบเทียบระหว่างแบรนด์รถ (Car Brand) และกลุ่มรถยอดนิยม 5 กลุ่ม (Sub-Segment)



กราฟแสดงการเปรียบเทียบระหว่างแบรนด์รถ (Car Brand) และกลุ่มรถยอดนิยม 5 กลุ่ม (Sub-Segment)

