Capstone Project 3

Pengembangan dan Evaluasi Model Regresi Machine Learning untuk Prediksi Harga Mobil Bekas di Arab Saudi

-Aulia Aorama

OUTLINE

Business Understanding

business context, business problem, steakholder, tujuan, evaluasi model

Data Understanding & Preparation

EDA

Distribusi Price, Hubungan data numerikal dan kategorik dengan Price

Modeling

Cross validation, Hyper tunning

Model Evaluation

Perbandingan sebelum & sesudah tuning, Analisis error

Feature Importance & SHAP Analysis

Kesimpulan

Rekomendasi

Rekomendasi bisnis, Rekomendasi pengembangan model lanjutan

BUSINESS CONTEXT

Pasar mobil bekas di Arab Saudi b**erkembang pesat seiring pertumbuhan industri otomotif dan meningkatnya kebutuhan mobilitas**. **Persaingan yang semakin kompetitif dan transparansi harga** melalui platform digital membuat dinamika pasar menjadi lebih kompleks.

BUSINESS PROBLEM

Penetapan harga mobil bekas dipengaruhi oleh banyak faktor seperti **usia, mileage, mesin, dan merek, sehingga sering bersifat subjektif.**

Hal ini menimbulkan **risiko overpricing maupun underpricing**, sehingga dibutuhkan **model prediksi berbasis data** untuk menghasilkan **estimasi harga yang lebih objektif dan konsisten**.

?

1. Bagaimana membangun model prediksi harga yang akurat?
2. Faktor apa yang paling memengaruhi harga?
3. Bagaimana meminimalkan kesalahan prediksi?

TUJUAN

Tujuan Bisnis

- Mengurangi ketidakpastian harga
- Membantu penetapan harga yang lebih konsisten

Tujuan Teknis

- Membangun model regresi prediksi harga
- Mengidentifikasi fitur paling berpengaruh
- Menghasilkan model dengan error terukur

EVALUASI MODEL

- **MSE** → tingkat kesalahan utama (semakin kecil semakin baik)
- **RMSE** → error dalam satuan harga (SAR)
- **R^2** → kemampuan menjelaskan variasi harga
- **MAPE** → error dalam persentase

STEAKHOLDER

- Platform jual beli mobil
- Penjual (individual/dealer)
- Pembeli
- Pricing analyst / decision maker

DATA UNDERSTANDING

Column Name	Jumlah Unique	Deskripsi
Type	347	Model/tipe kendaraan
Region	27	Lokasi kendaraan dijual
Make	58	Merek kendaraan
Gear_Type	2	Jenis transmisi
Origin	4	Asal kendaraan
Options	3	Tingkat kelengkapan fitur

Column Name	Jumlah Unique	Deskripsi
Year	50	Tahun produksi
Engine_Size	71	Kapasitas mesin
Mileage	1716	Jarak tempuh kendaraan
Negotiable	2	Status harga bisa dinegosiasikan
Price	467	Harga jual mobil bekas

01 Data

Jumlah data: 5624 data
jumlah kolom: 11
jumlah data numerik: 6
jumlah data kategorik: 7

02 Duplikat, Missing Value, Outlier Extream

Duplikat: 4 data
Missing Value: 0 data
Outlier Extream: 2 data (mileage)

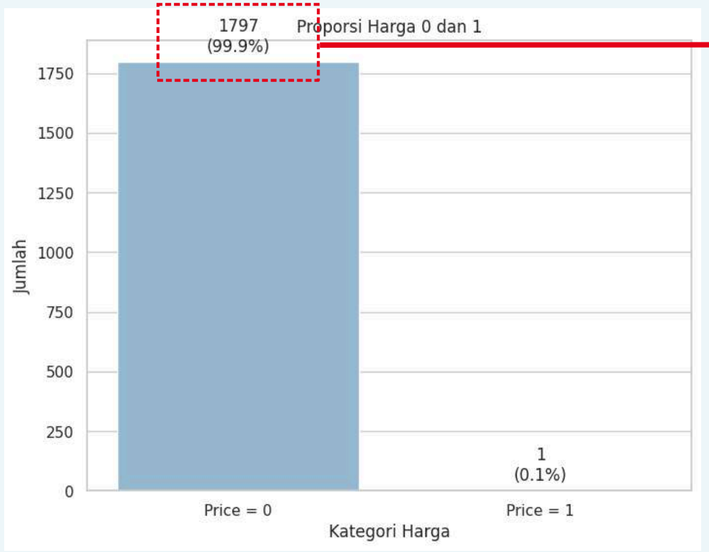
03 Fitur Tambahan

Car_age= umur mobil yang diambil dari tahun 2026 - year

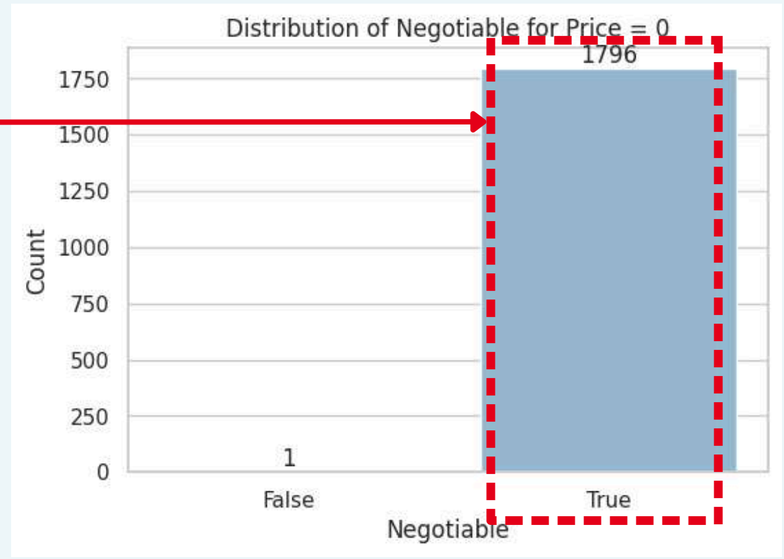
DISTRIBUSI TARGET (PRICE)

Insight

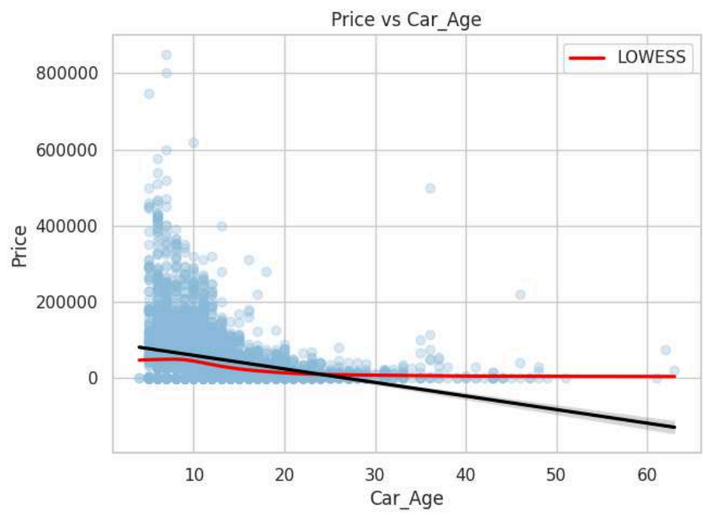
- pola **right-skewed** dengan **konsentrasi tinggi pada segmen harga rendah hingga menengah dan sedikit kendaraan premium berharga sangat tinggi**.
- Struktur ini menyebabkan **model lebih stabil pada segmen menengah, namun cenderung mengalami deviasi lebih besar pada kendaraan ekstrem**.

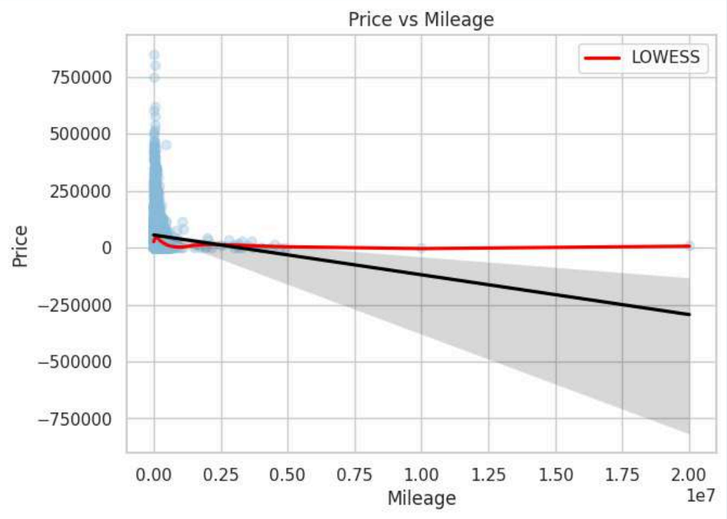


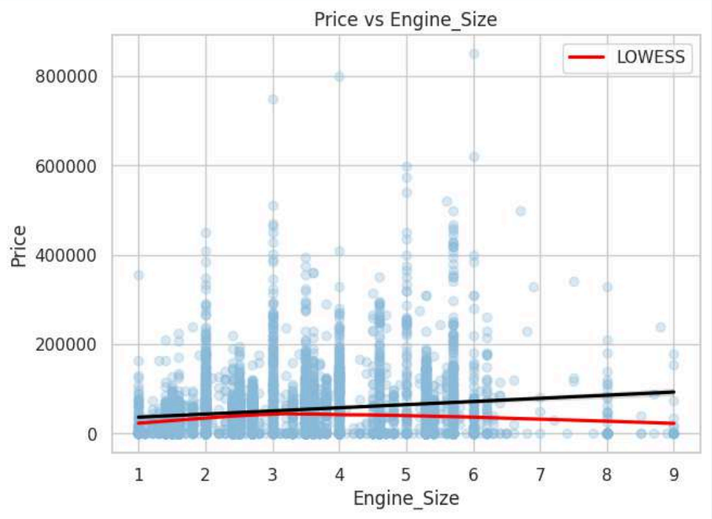
- **Price = 0 (99,9%) memiliki status Negotiable = True**
- nilai 0 **bukan harga aktual**, melainkan placeholder untuk **harga yang dapat dinegosiasikan.**
- Price = 0 diperlakukan sebagai **missing value** dan didrop.



NUMERIK VS PRICE







Car_age

- Semakin **tua mobil → harga semakin turun** (depresiasi).
- Setelah usia tertentu, harga cenderung stabil di level rendah.
- Terdapat outlier (mobil tua dengan harga tinggi), kemungkinan karena merek premium atau kondisi khusus.

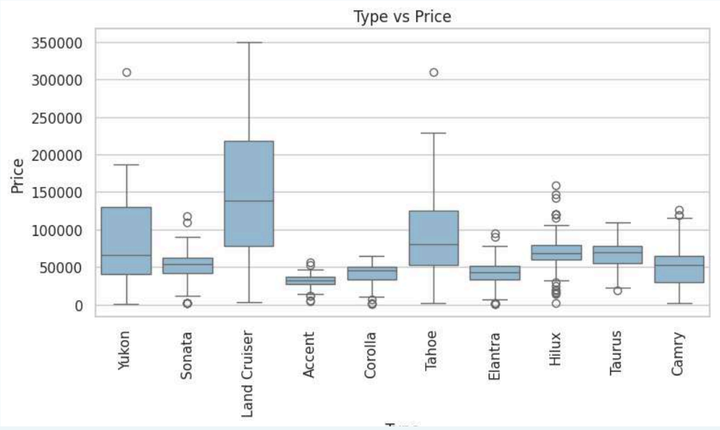
Mileage

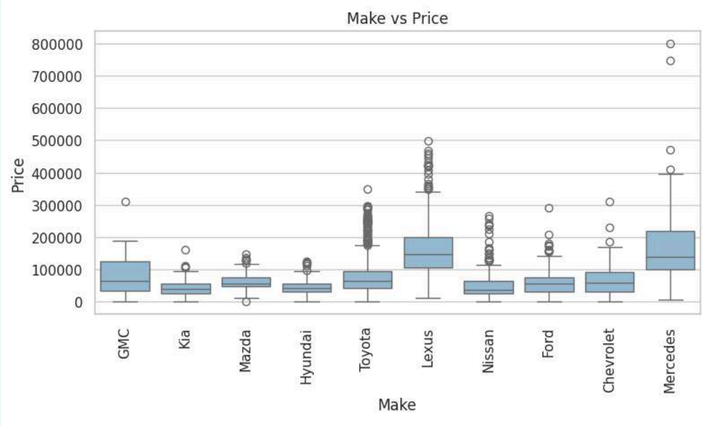
- Mileage **rendah → harga relatif lebih tinggi.**
- Mileage **tinggi → harga cenderung turun**.
- Distribusi sangat skewed (banyak mileage kecil).

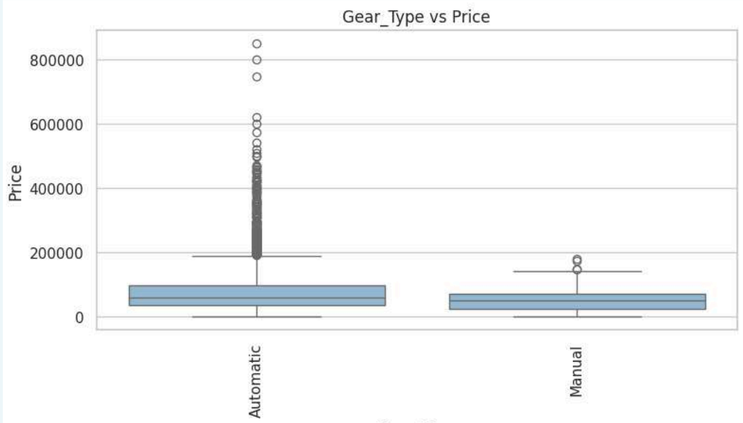
Engine_Size

- Semakin **besar engine size → harga cenderung meningkat**.
- Namun, hubungan tidak linear kuat.
- Variasi harga cukup tinggi pada setiap level engine size.

KATEGORI VS PRICE







10 Type

- SUV dan kendaraan premium (misalnya **Land Cruiser, Tahoe**) memiliki **median harga tertinggi**.
- Sedan kompak (**Accent, Corolla, Elantra**) memiliki **median harga lebih rendah**.
- **Variasi harga dalam satu tipe cukup besar** → faktor lain turut memengaruhi (usia, kondisi, merek).

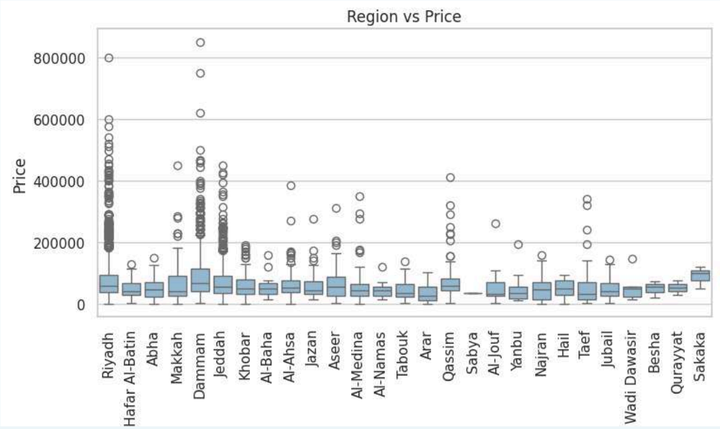
10 Make

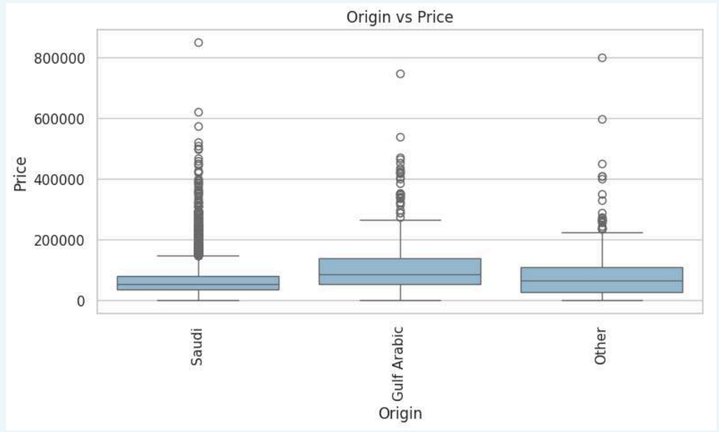
- Merek **premium** (Mercedes, Lexus) menunjukkan **median dan rentang harga lebih tinggi**.
- Merek **mass-market** (Hyundai, Kia, Nissan) berada di segmen **harga menengah–rendah.**

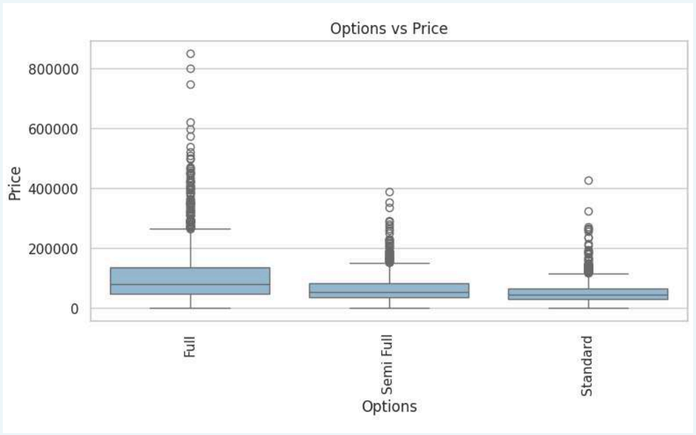
Gear_Type

- **Automatic** memiliki **median harga lebih tinggi** dibanding Manual.
- **Manual** cenderung berada di segmen **harga bawah–menengah**.

KATEGORI VS PRICE







Region

- Kota besar seperti **Riyadh dan Jeddah** menunjukkan lebih **banyak outlier harga tinggi.**
- Lokasi memengaruhi distribusi listing, tetapi **bukan faktor utama penentu harga** dibanding fitur kendaraan.

Origin

- Mobil dengan origin **Gulf Arabic** memiliki median **harga lebih tinggi**.
- Origin **Saudi dan Other** berada di **level menengah.**
- Asal kendaraan memiliki **pengaruh terhadap persepsi nilai dan harga pasar**.

Options

- Varian **Full** option memiliki median **harga tertinggi.**
- **Semi Full** berada di **tengah**.
- **Standard** memiliki harga **paling rendah**.
- Perbedaan antar **kategori cukup konsisten.**

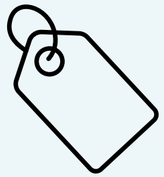
DATA CLEANING STRATEGY

01

Handaling Duplikat

Drop Duplikat (4 data)

02



Handaling price = 0 dan 1

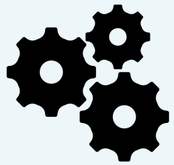
Drop data price = 0 dan price = 1 (1797 data)

03

Handaling outlier extream Mileage

Drop data outlier extream (1 data)

04



Handaling Engine_size > 8L

Drop data engine_size > 8 L (9 data)

05

Handaling Origin = “unknown”

Drop data Origin = “unknown” (15 data)

06

Handaling Type = “ceros,furniture, other” dan Make = “other”

Drop data (77 data)

07

Handaling Standarisasi Nama data (Type)

- D’Max → D-MAX
- Carenz → Carens
- Montero2 → Montero
- Copper→ Cooper
- Van R → Van
- Land Cruiser 70" dan "Land Cruiser Pickup" → Land Cruiser"

Data Preporation

STRATEGI DATA PREPROCESSING

Transformer Car_age

01

Membuat transformer untuk mengubah **Year** menjadi **Car_Age** dan menghapus kolom Year agar tidak terjadi multicollinearity

Robust Scaler

02

Robust scaler untuk data numerik (**Mileage, car_age, Engine_size**)

OneHotEncoder

03

Onehotencoder pada data kategori yang kecil dari 5 yaitu **gear_type, origin, options, negotiable**

BinaryEncoder

04

Binaryencoder pada data kategori yang lebih dari 5 yaitu **Type, make, region**

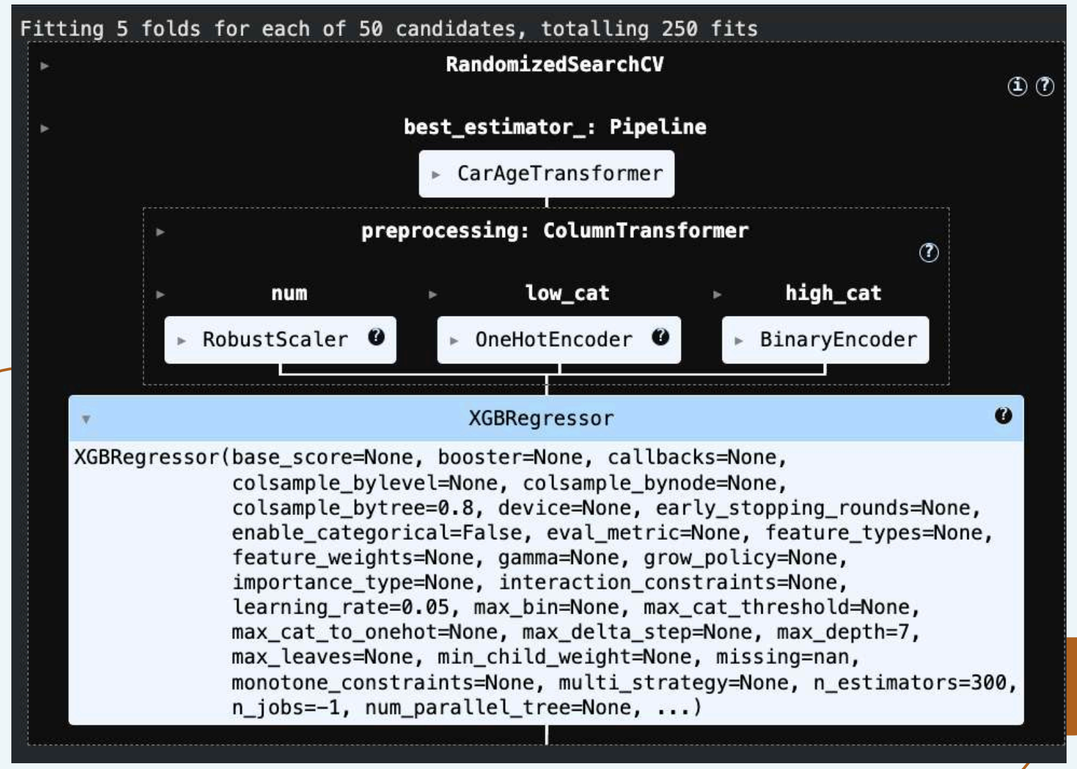
CROSS VALIDATION

CV = 5

	Model	MSE_score	MSE_mean	MSE_std	RMSE_score	RMSE_mean	RMSE_std	MAE_score	MAE_mean	MAE_std	MAPE_score	MAPE_mean	MAPE_std
0	XGBoost	[1597354624.0, 1549300736.0, 1937184000.0, 1062728704.0, 1254188416.0]	1.480151e+09	3.008980e+08	[39966.92, 39361.16, 44013.45, 32599.52, 35414.52]	38271.11	3933.59	[21301.63, 19285.26, 20850.62, 19674.01, 20332.33]	20288.77	738.70	[1.6434, 1.2252, 1.1327, 1.4054, 0.6324]	1.2078	0.3364
1	Stacking	[1938709895.24, 1631297573.82, 2112701306.48, 1194172914.71, 1158482599.93]	1.607073e+09	3.841744e+08	[44030.78, 40389.32, 45964.13, 34556.81, 34036.49]	39795.51	4836.37	[21927.56, 20853.39, 22581.85, 21387.76, 21427.19]	21635.55	582.64	[1.463, 1.2841, 1.0419, 1.2856, 0.6766]	1.1502	0.2721
2	RandomForest	[2234401199.64, 1576918310.25, 2196204483.63, 1194304639.9, 1372261339.45]	1.714818e+09	4.263790e+08	[47269.45, 39710.43, 46863.68, 34558.71, 37044.05]	41089.26	5146.88	[23078.44, 19596.16, 21781.6, 18919.56, 20766.09]	20828.37	1492.46	[1.8042, 1.3273, 1.3218, 1.4555, 0.6762]	1.3170	0.3653
3	Bagging	[2073033697.43, 1724053126.89, 2410622115.44, 1294281223.48, 1244936795.18]	1.749385e+09	4.481545e+08	[45530.58, 41521.72, 49098.09, 35976.12, 35283.66]	41482.03	5350.35	[23386.77, 21521.32, 24015.64, 22442.11, 21892.22]	22651.61	927.27	[1.5193, 1.3344, 1.1158, 1.3209, 0.7042]	1.1989	0.2784
4	KNN	[2149751348.86, 1749061001.32, 2367343795.09, 1380653564.42, 1249914563.84]	1.779345e+09	4.296771e+08	[46365.41, 41821.78, 48655.36, 37157.15, 35354.13]	41870.76	5117.02	[23588.39, 22190.07, 24533.07, 23126.12, 22128.3]	23113.19	901.58	[1.5706, 1.3845, 1.1123, 1.358, 0.6996]	1.2250	0.3005
5	Voting	[2329598059.94, 1932755998.08, 2428514149.7, 1303119636.1, 1487698515.98]	1.896337e+09	4.452294e+08	[48265.91, 43963.12, 49279.96, 36098.75, 38570.7]	43235.69	5197.38	[25044.81, 23213.03, 25200.44, 22402.0, 23992.58]	23970.57	1067.83	[1.5898, 1.3757, 1.1849, 1.408, 0.7822]	1.2681	0.2748
6	GradientBoosting	[2072099052.0, 2071156030.85, 2361262475.53, 1332961615.6, 1755527263.2]	1.918601e+09	3.499573e+08	[45520.31, 45509.96, 48592.82, 36509.75, 41899.01]	43606.37	4133.49	[23825.92, 24536.27, 25309.57, 22656.76, 25101.54]	24286.02	963.49	[1.6791, 1.442, 1.405, 1.3779, 0.7271]	1.3262	0.3180
7	Ridge	[4029555889.59, 3171759522.89, 4074540132.75, 2379640243.7, 3012139872.7]	3.333527e+09	6.438812e+08	[63478.78, 56318.38, 63832.12, 48781.56, 54882.97]	57458.76	5658.44	[36869.85, 34151.86, 36550.41, 33912.72, 36541.29]	35605.23	1291.95	[1.9299, 1.5995, 1.438, 1.7105, 1.1393]	1.5635	0.2656
8	Lasso	[4028982485.24, 3170652023.09, 4075173411.55, 2384374696.32, 3018095423.88]	3.335456e+09	6.419649e+08	[63474.27, 56308.54, 63837.08, 48830.06, 54937.2]	57477.43	5639.21	[36878.27, 34159.57, 36562.46, 33935.37, 36584.3]	35623.99	1293.99	[1.9299, 1.5997, 1.4383, 1.7108, 1.1406]	1.5638	0.2653
9	LinearRegression	[4028982484.05, 3170651940.96, 4075173447.31, 2384374951.94, 3018095677.63]	3.335456e+09	6.419648e+08	[63474.27, 56308.54, 63837.09, 48830.06, 54937.2]	57477.43	5639.21	[36878.27, 34159.57, 36562.46, 33935.38, 36584.3]	35623.99	1293.99	[1.9299, 1.5997, 1.4383, 1.7108, 1.1406]	1.5638	0.2653
10	DecisionTree	[3811368422.81, 3154445138.57, 4119547888.13, 3270825716.71, 2863984625.82]	3.444034e+09	4.564151e+08	[61736.28, 56164.45, 64183.7, 57191.13, 53516.21]	58558.35	3866.97	[29213.6, 28305.8, 29622.29, 27504.72, 28169.34]	28563.15	759.80	[1.6022, 1.4288, 1.2861, 1.4124, 0.7953]	1.3049	0.2740
11	AdaBoost	[5526174205.47, 3840413416.33, 5392333745.01, 3236649174.85, 3362426757.56]	4.271599e+09	9.913265e+08	[74338.24, 61971.07, 73432.51, 56891.56, 57986.44]	64923.96	7515.22	[60993.3, 46494.9, 58390.21, 48280.56, 48456.7]	52523.13	5950.48	[3.7264, 2.5211, 2.8073, 2.7556, 1.7291]	2.7079	0.6391

Model berbasis ensemble, khususnya **XGBoost**, memberikan performa terbaik dengan nilai **MSE, RMSE, dan MAE terendah**.
Hal ini mengindikasikan bahwa hubungan antar **fitur dan harga bersifat non-linear dan kompleks**, sehingga **model linear kurang optimal untuk kasus ini.**

HYPERPARAMETER TUNNING



Best Params XGBoost:

modeling__subsample': **0.8**,
'modeling__reg_lambda': **5**,
'modeling__reg_alpha': **0**,
'modeling__n_estimators': **300**,
'modeling__max_depth': **7,**
'modeling__learning_rate': **0.05**,
'modeling__colsample_bytree': **0.8**

MSE, RMSE,MAE,MAPE

Best MSE: **1269420236.8**
Best RMSE: **35229.840625**
Best MAE: **18143.17734375**
Best MAPE: **1.160978364944458**

Hyperparameter tuning melalui RandomizedSearchCV menunjukkan peningkatan performa pada cross-validation.

BAGAIMANA XGBOOST BEKERJA?

Bagaimana Cara XGBoost Bekerja?

1. **Model** membangun pohon **pertama** untuk memprediksi **harga**
2. Menghitung selisih (error) dengan **harga asli**
3. Pohon berikutnya fokus **memperbaiki** error tersebut
4. Proses diulang berkali-kali
5. Hasil akhir = **kombinasi** semua pohon

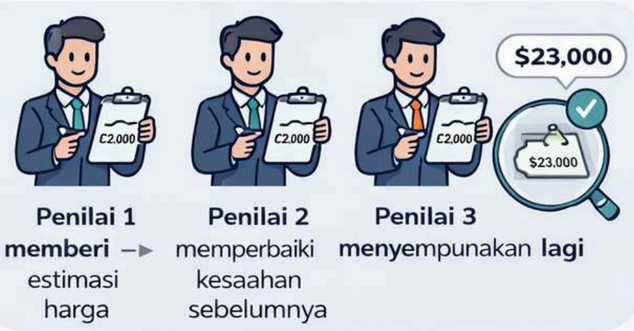
Tahap cara turn XGBoost

Analogi: Tim Appraisal Mobil

1. **Penilai 1** memberi estimasi harga
2. **Penilai 2** memperbaiki kesalahan seblumnya
3. **Penilai 3** menyempurnakan **lagi**
4. Hasil akhir = estimasi **kolektif** yang lebih **akurat**

Intinya:

Setiap model baru belajar dari **kesalahan model sebelumnya.**



KENAPA XGBOOST COCOK?

Karakteristik Data	Alasan
Hubungan non-linear	Tree menangkap pola kompleks
Banyak kategori	Tree handle encoding dengan baik
Interaksi fitur	Otomatis menangkap kombinasi
Banyak outlier	Lebih robust dibanding linear

Intinya

Dataset harga mobil bersifat kompleks dan non-linear, sehingga model boosting lebih efektif dibanding regresi linear.

METRIK EVALUASI

Metric	Before Tuning	After Tuning	Change (%)
MSE	1.167607e+09	9.035434e+08	22.62%
RMSE	3.417027e+04	3.005900e+04	12.03%
MAE	1.876034e+04	1.701851e+04	9.28%
MAPE	7.814325e-01	7.743088e-01	0.91%
R^2	0.772447	0.823910	6.66%

- **Error turun signifikan** (MSE -22%, RMSE -12%)
- **Prediksi lebih dekat ke harga aktual** (MAE -9%)
- **Kemampuan jelaskan variasi harga meningkat** (R^2 naik ke 82%)
- **Model lebih stabil & minim risiko salah harga ekstrem**

EROR ANALYSIS

- **Rata-rata harga mobil: ~79.700**
- **Rata-rata error (MAE): ~17.000**
- **Relative error: ~21%**
- **Model menjelaskan 82% variasi harga (R^2 = 0.82)**

Pola

- **Error terbesar** terjadi pada mobil **harga tinggi**
- Model cenderung **underpredict mobil premium**
- Prediksi **lebih stabil pada segmen harga menengah**

Implikasi ke Bisnis

- Risiko **undervalue mobil mahal** → potensi kehilangan margin
- Kesalahan **ekstrem sudah berkurang** (hasil tuning efektif)
- Model **cukup andal untuk pricing umum**, tetapi segmen **premium perlu perhatian khusus**

Feature Important

FEATURE IMPORTANCE

- **Tingkat Kelengkapan Fitur (Options – Full)**

→ Faktor paling berpengaruh terhadap harga

- **Usia Mobil (Car_Age)**

→ Semakin tua, harga semakin turun

- **Kapasitas Mesin (Engine_Size)**

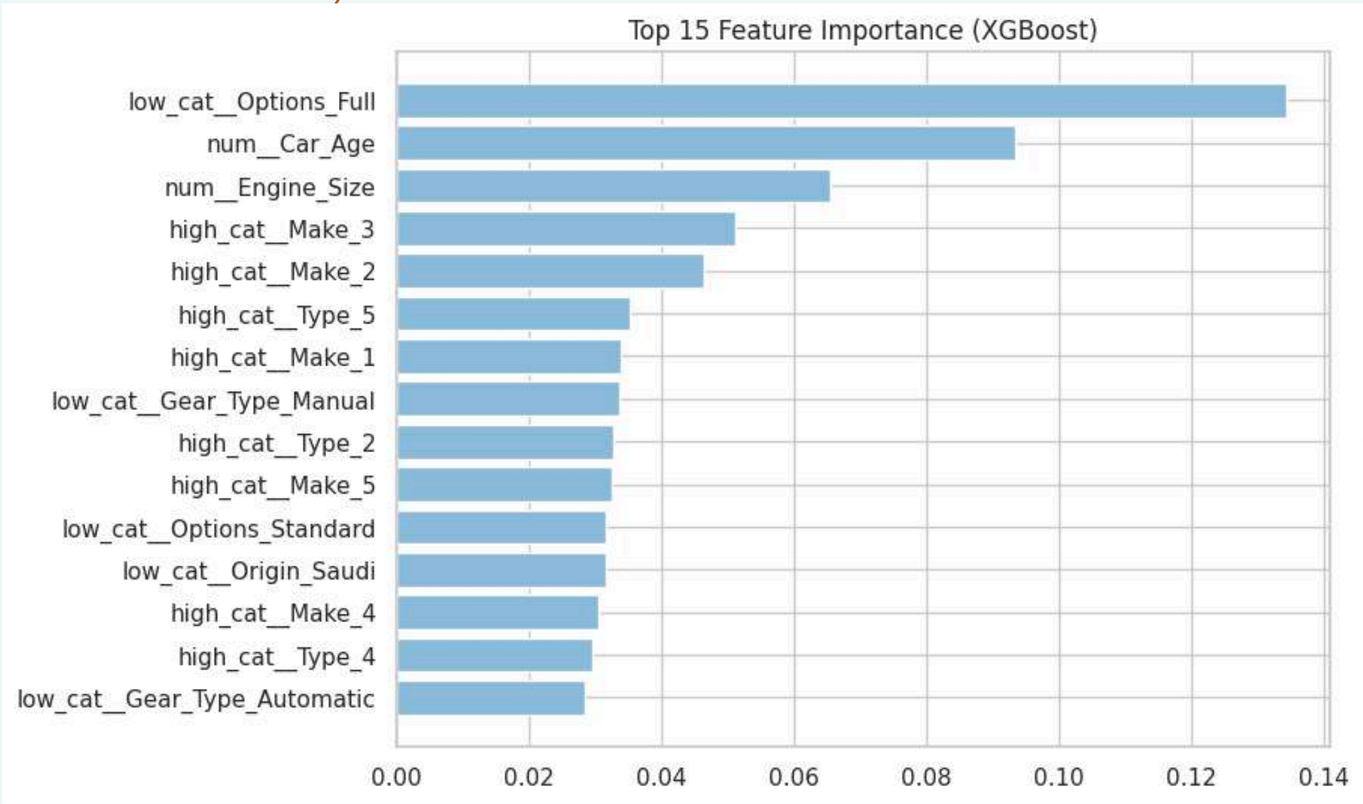
→ Mesin lebih besar cenderung lebih mahal

- **Brand & Tipe Kendaraan**

→ Make dan Type tertentu memberi premium harga

- **Transmisi & Asal Kendaraan**

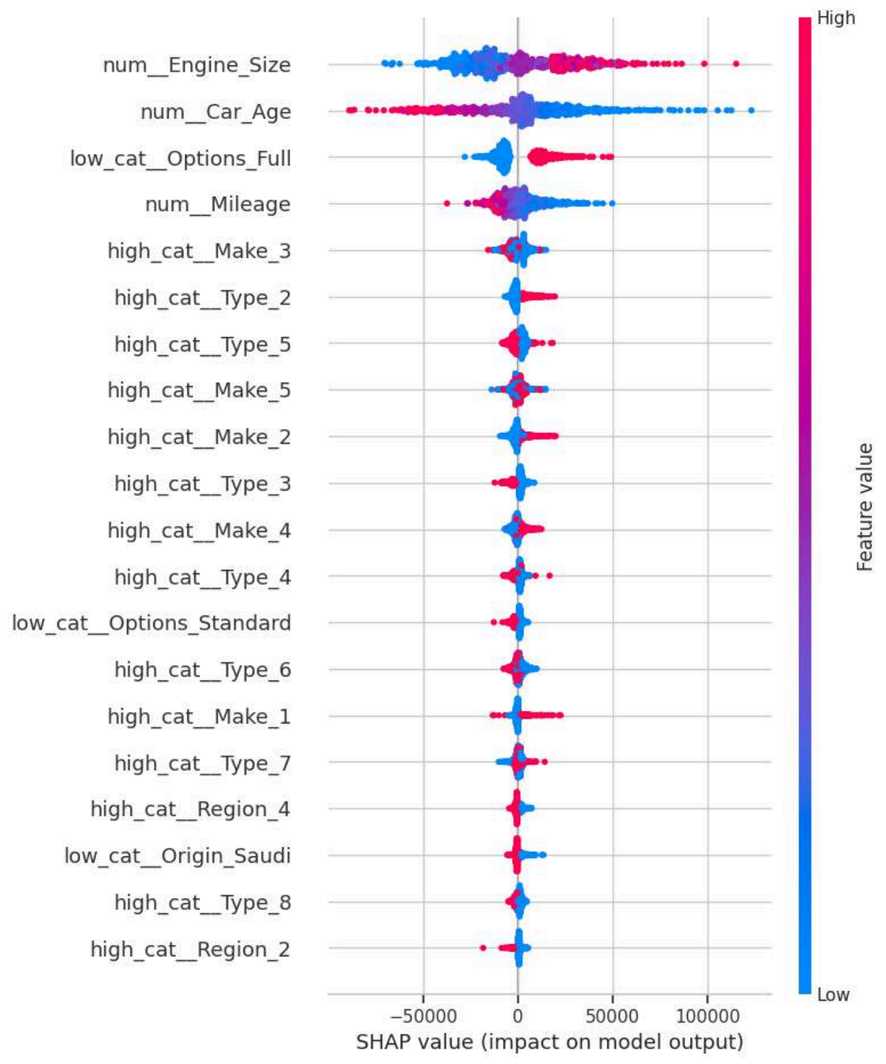
→ Gear Type dan Origin turut memengaruhi harga



INSIGHT BISNIS

- Fitur dan kelengkapan lebih dominan dibanding lokasi
- Faktor teknis kendaraan lebih kuat dari faktor regional
- Brand positioning berpengaruh signifikan terhadap pricing

SHAP VALUE



1. **Engine Size**
   - Nilai besar (merah) → menaikkan harga
   - Nilai kecil (biru) → menurunkan harga
   - Mesin besar = premium
2. **Car Age**
   - Usia tinggi → menurunkan harga
   - Mobil lebih baru → menaikkan harga
   - Depresiasi sangat berpengaruh
3. **Full Options**
   - Full features → dorong harga naik signifikan
4. **Mileage**
   - Mileage tinggi → menekan harga
   - Mileage rendah → meningkatkan nilai

Implikasi Bisnis

- Faktor teknis kendaraan (mesin & usia) paling menentukan
- Fitur lengkap memberi premium harga
- Model menangkap pola depresiasi dengan baik
- Pricing premium terutama didorong oleh mesin & brand

KESIMPULAN

Performa Model

- XGBoost terpilih sebagai model terbaik
- MSE turun 22% setelah tuning
- R^2 meningkat menjadi 82%
- Rata-rata error sekitar 17 ribu SAR (~21%)

Faktor Penentu Harga

- Kelengkapan fitur (Full Options)
- Usia kendaraan
- Kapasitas mesin
- Brand & tipe kendaraan

Model menangkap pola pasar yang logis:
Mobil lebih baru, mesin besar, dan fitur lengkap → harga lebih tinggi.

Implikasi Bisnis

- Membantu estimasi harga lebih objektif
- Mengurangi risiko overpricing & underpricing
- Perlu perhatian khusus pada segmen harga ekstrem (premium & sangat murah)

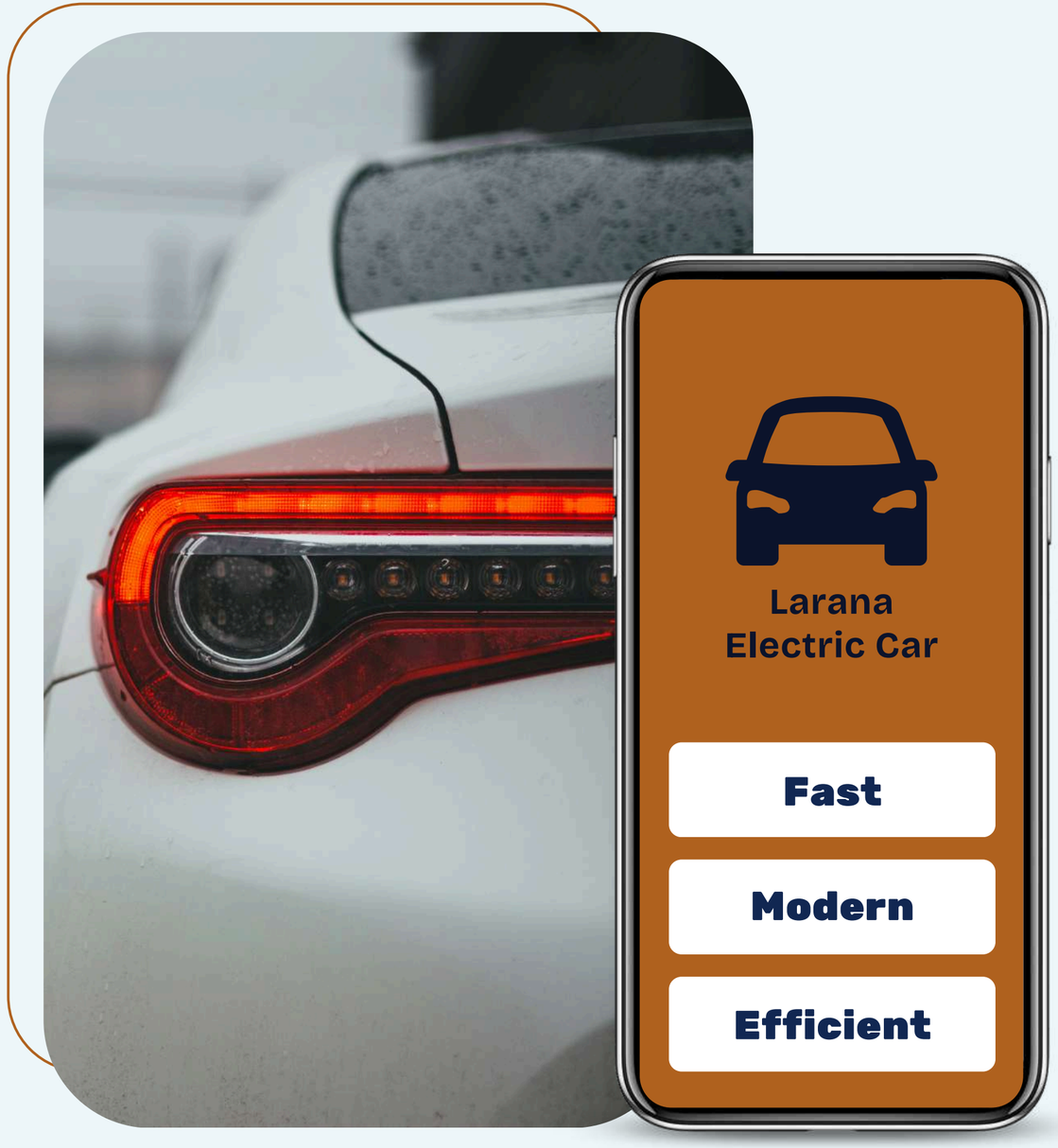
REKOMENDASI PENGEMBANGAN & IMPLEMENTASI MODEL

Rekomendasi Bisnis

- Integrasikan model ke platform untuk estimasi harga instan
- Gunakan sebagai acuan objektif untuk mencegah overpricing & underpricing
- Fokus pada faktor utama: fitur, usia, dan ukuran mesin
- Gunakan model sebagai referensi awal untuk segmen premium (dengan validasi tambahan)

Rekomendasi Teknis

- Tingkatkan akurasi pada harga ekstrem (tambah data & fitur eksternal)
- Gunakan segmented modeling (low-end, mid-range, premium)
- Lakukan retraining berkala agar tetap relevan
- Terapkan monitoring performa untuk deteksi model drift



Capstone Project 3

Thank You!

www.linkedin.com/in/aulia-aorama
aoramaaulia@gmail.com
Capstone Project 3