

# Saudi Arabia Used Car Price Prediction

*Dataset from syarah.com*

Aldino Dian



# Latar Belakang



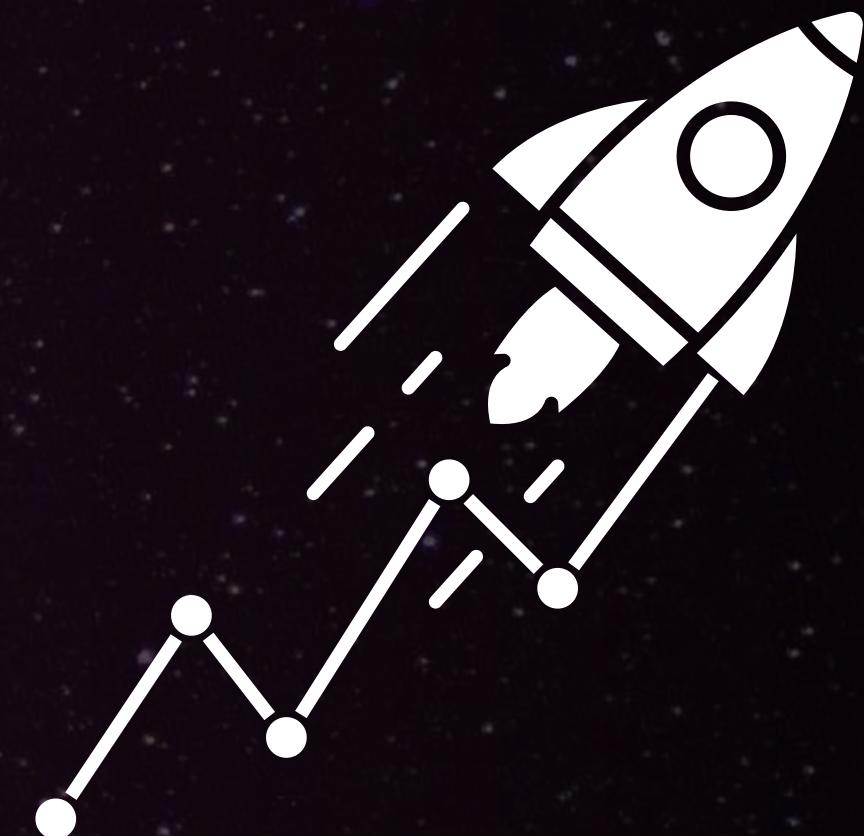
## NILAI PASAR

USD 9.6B (2024) -> USD 16.8B (2033)



## FAKTOR PENDORONG :

- HARGA MOBIL BARU
- PPN
- KEBUTUHAN SOLUSI MURAH



# Tujuan

*MENCIPTAKAN ALAT PREDIKSI HARGA MOBIL BEKAS BERBASIS DATA YANG AKURAT, RELEVAN, DAN MENGUNTUNGKAN SEMUA PIHAK DALAM EKOSISTEM JUAL-BELI MOBIL.*

- MEMBANTU PENJUAL MENETAPKAN HARGA MOBIL BEKAS YANG TEPAT DAN KOMPETITIF
- MENINGKATKAN KEPERCAYAAN PEMBELI
- MENDUKUNG PLATFORM JUAL-BELI SEPERTI SYARAH.COM

# Analytic Approach

MENGANALISIS DATA UNTUK MENEMUKN POLA HUBUNGAN ANTARA **FITUR** KENDARAAN (SEPERTI MEREK, TAHUN, MESIN, MILEAGE, DLL) DENGAN **HARGA MOBIL BEKAS.**

## Metric Evaluation

RMSE

MAE

MAPE

# Pemahaman Data



→ JUMLAH BARIS : **5.624**

| Fitur       | Tipe Data | Deskripsi   |
|-------------|-----------|---|
| Type        | Object    | Jenis mobil bekas   |
| Region      | Object    | Wilayah tempat mobil ditawarkan                                       |
| Make        | Object    | Nama perusahaan atau merek mobil                                      |
| Gear_Type   | Object    | Jenis transmisi mobil (otomatis/manual)                               |
| Origin      | Object    | Asal mobil (lokal atau impor)   |
| Options     | Object    | Fitur tambahan pada mobil (seperti Standard, Semi Full, Full)         |
| Year        | Integer   | Tahun pembuatan mobil   |
| Engine_Size | Float     | Kapasitas mesin mobil dalam liter                                     |
| Mileage     | Integer   | Jarak tempuh mobil (dalam kilometer)                                  |
| Negotiable  | Boolean   | Menunjukkan apakah harga dapat dinegosiasikan ( True jika harga = 0 ) |
| Price       | Integer   | Harga mobil bekas (dalam satuan mata uang lokal, biasanya SAR)        |

# EDA & PreProcessing Data

## DISTRIBUSI HARGA:

HARGA MOBIL BEKAS MEMILIKI DISTRIBUSI RIGHT-SKEWED, DENGAN MAYORITAS HARGA BERADA DI KISARAN RENDAH DAN SEBAGIAN KECIL MOBIL DENGAN HARGA SANGAT TINGGI (OUTLIER).

## KORELASI FITUR NUMERIK:

- YEAR → KORELASI POSITIF TERHADAP HARGA
- MILEAGE → KORELASI NEGATIF
- ARTINYA, MOBIL BARU DENGAN JARAK TEMPUH RENDAH CENDERUNG LEBIH MAHAL.

## PREPROCESSING STEP

PENGECEKKAN DATA

PENANGANAN MISSING VALUE DAN DUPLIKASI

KONVERSI TIPE DATA

PENANGANAN OUTLIER

FEATURE ENGINEERING

# Feature Engineering

*DILAKUKAN SEJUMLAH TRANSFORMASI DATA UNTUK MEMPERSIAPKAN DATASET AGAR OPTIMAL DIGUNAKAN DALAM PROSES PELATIHAN MODEL MACHINE LEARNING.*

*BEBERAPA LANGKAH UTAMA YANG DILAKUKAN MELIPUTI:*

**SCALING**

**FEATURE  
SELECTION**

**ENCODING**

**IRRELEVANT  
DATA**

# Modelling Step

1. DATA DIBAGI MENJADI DATA LATIH DAN DATA UJI.
2. BENCHMARK DILAKUKAN PADA 5 MODEL: LINEAR REGRESSION, KNN, DECISION TREE, RANDOM FOREST, DAN XGBOOST.
3. EVALUASI MENGGUNAKAN METRIK RMSE, MAE, DAN MAPE.
4. XGBOOST MENJADI MODEL TERBAIK DARI BENCHMARKING AWAL, LALU DIBANDINGKAN LAGI DENGAN MODEL BOOSTING LAINNYA: ADA\_BOOST, LIGHTGBM, DAN CATBOOST.
5. CATBOOST UNGGUL PADA DATA UJI, KARENA KEMAMPUANNYA MENANGANI BANYAK FITUR KATEGORIKAL SECARA LANGSUNG.
6. HYPERPARAMETER TUNING MENGGUNAKAN RANDOMIZEDSEARCHCV (20 ITERASI DARI 432 KOMBINASI), DIPILIH UNTUK EFISIENSI WAKTU DAN RESOURCE.

# Modelling Step

1. DATA DIBAGI MENJADI DATA LATIH DAN DATA UJI.
2. BENCHMARK DILAKUKAN PADA 5 MODEL: LINEAR REGRESSION, KNN, DECISION TREE, RANDOM FOREST, DAN XGBOOST.
3. EVALUASI MENGGUNAKAN METRIK RMSE, MAE, DAN MAPE.
4. XGBOOST MENJADI MODEL TERBAIK DARI BENCHMARKING AWAL, LALU DIBANDINGKAN LAGI DENGAN MODEL BOOSTING LAINNYA: ADA\_BOOST, LIGHTGBM, DAN CATBOOST.
5. CATBOOST UNGGUL PADA DATA UJI, KARENA KEMAMPUANNYA MENANGANI BANYAK FITUR KATEGORIKAL SECARA LANGSUNG.
6. HYPERPARAMETER TUNING MENGGUNAKAN RANDOMIZEDSEARCHCV (20 ITERASI DARI 432 KOMBINASI), DIPILIH UNTUK EFISIENSI WAKTU DAN RESOURCE.

# Modelling Step

1. DATA DIBAGI MENJADI DATA LATIH DAN DATA UJI.
2. BENCHMARK DILAKUKAN PADA 5 MODEL: LINEAR REGRESSION, KNN, DECISION TREE, RANDOM FOREST, DAN XGBOOST.
3. EVALUASI MENGGUNAKAN METRIK RMSE, MAE, DAN MAPE.
4. XGBOOST MENJADI MODEL TERBAIK DARI BENCHMARKING AWAL, LALU DIBANDINGKAN LAGI DENGAN MODEL BOOSTING LAINNYA: ADA\_BOOST, LIGHTGBM, DAN CATBOOST.
5. CATBOOST UNGGUL PADA DATA UJI, KARENA KEMAMPUANNYA MENANGANI BANYAK FITUR KATEGORIKAL SECARA LANGSUNG.
6. HYPERPARAMETER TUNING MENGGUNAKAN RANDOMIZEDSEARCHCV (20 ITERASI DARI 432 KOMBINASI), DIPILIH UNTUK EFISIENSI WAKTU DAN RESOURCE.

# Modelling Step

1. DATA DIBAGI MENJADI DATA LATIH DAN DATA UJI.
2. BENCHMARK DILAKUKAN PADA 5 MODEL: LINEAR REGRESSION, KNN, DECISION TREE, RANDOM FOREST, DAN XGBOOST.
3. EVALUASI MENGGUNAKAN METRIK RMSE, MAE, DAN MAPE.
4. XGBOOST MENJADI MODEL TERBAIK DARI BENCHMARKING AWAL, LALU DIBANDINGKAN LAGI DENGAN MODEL BOOSTING LAINNYA: ADA\_BOOST, LIGHTGBM, DAN CATBOOST.
5. CATBOOST UNGGUL PADA DATA UJI, KARENA KEMAMPUANNYA MENANGANI BANYAK FITUR KATEGORIKAL SECARA LANGSUNG.
6. HYPERPARAMETER TUNING MENGGUNAKAN RANDOMIZEDSEARCHCV (20 ITERASI DARI 432 KOMBINASI), DIPILIH UNTUK EFISIENSI WAKTU DAN RESOURCE.

# Modelling Step

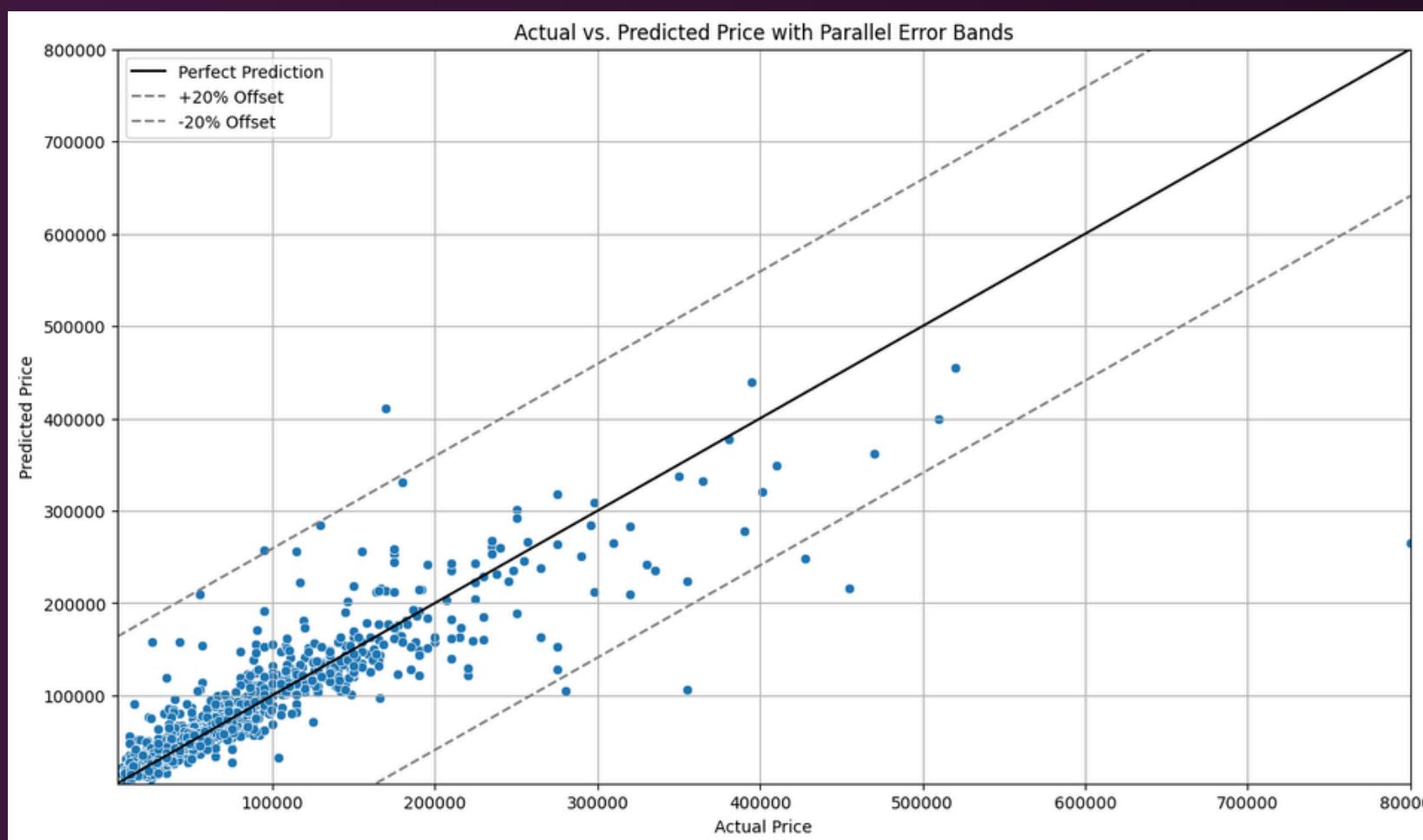
1. DATA DIBAGI MENJADI DATA LATIH DAN DATA UJI.
2. BENCHMARK DILAKUKAN PADA 5 MODEL: LINEAR REGRESSION, KNN, DECISION TREE, RANDOM FOREST, DAN XGBOOST.
3. EVALUASI MENGGUNAKAN METRIK RMSE, MAE, DAN MAPE.
4. XGBOOST MENJADI MODEL TERBAIK DARI BENCHMARKING AWAL, LALU DIBANDINGKAN LAGI DENGAN MODEL BOOSTING LAINNYA: ADABOOST, LIGHTGBM, DAN CATBOOST.
5. CATBOOST UNGGUL PADA DATA UJI, KARENA KEMAMPUANNYA MENANGANI BANYAK FITUR KATEGORIKAL SECARA LANGSUNG.
6. HYPERPARAMETER TUNING MENGGUNAKAN RANDOMIZEDSEARCHCV (20 ITERASI DARI 432 KOMBINASI), DIPILIH UNTUK EFISIENSI WAKTU DAN RESOURCE.

# Modelling Step

1. DATA DIBAGI MENJADI DATA LATIH DAN DATA UJI.
2. BENCHMARK DILAKUKAN PADA 5 MODEL: LINEAR REGRESSION, KNN, DECISION TREE, RANDOM FOREST, DAN XGBOOST.
3. EVALUASI MENGGUNAKAN METRIK RMSE, MAE, DAN MAPE.
4. XGBOOST MENJADI MODEL TERBAIK DARI BENCHMARKING AWAL, LALU DIBANDINGKAN LAGI DENGAN MODEL BOOSTING LAINNYA: ADABOOST, LIGHTGBM, DAN CATBOOST.
5. CATBOOST UNGGUL PADA DATA UJI, KARENA KEMAMPUANNYA MENANGANI BANYAK FITUR KATEGORIKAL SECARA LANGSUNG.
6. HYPERPARAMETER TUNING MENGGUNAKAN RANDOMIZEDSEARCHCV (20 ITERASI DARI 432 KOMBINASI), DIPILIH UNTUK EFISIENSI WAKTU DAN RESOURCE.

# Analysis Evaluasi

## Actual vs Predicted Price



\*Offset error 20%

## Trend Model :

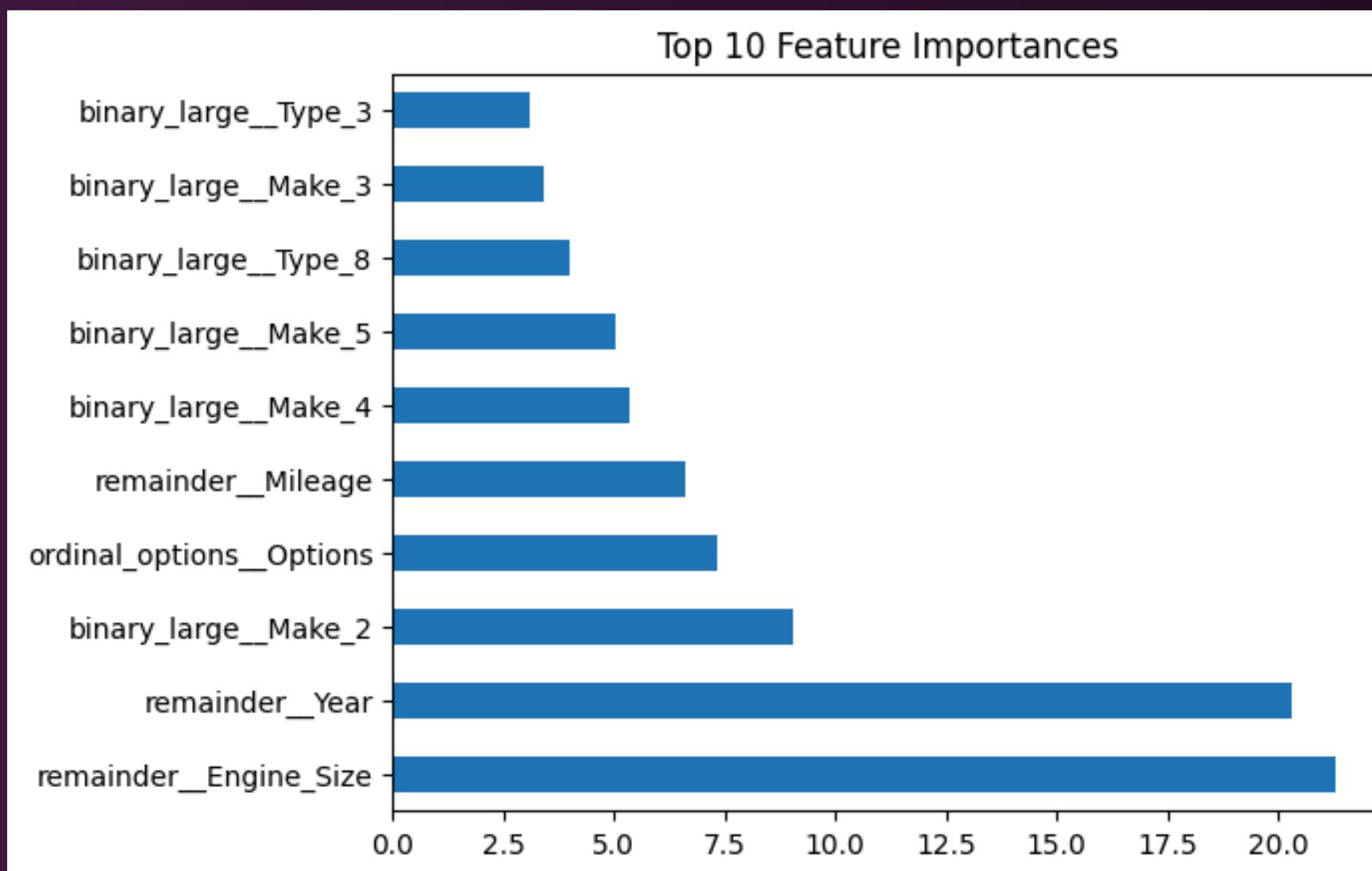
- Titik-titik mengikuti pola linear positif → saat harga aktual naik, prediksi juga naik.
- Ini menunjukkan model sudah menangkap hubungan antara fitur dan harga.

## Penyebaran Error :

- Harga tinggi → Beberapa titik keluar dari batas  $\pm 20\%$ , pada range harga ( $>300.000$  SAR), menunjukkan penurunan akurasi pada harga mahal.(Kalau diliat dari dataset, mobil dengan harga tinggi tidak banyak sehingga model susah pelajari hal ini)
- Harga rendah → prediksi cukup rapat.

# Analysis Evaluasi

## Top 10 Feature Importances



- **Engine\_Size** (*remainder\_Engine\_Size*) → Fitur paling berpengaruh. Artinya, ukuran mesin mobil adalah indikator kuat dalam menentukan harga mobil bekas.
- **Year** (*remainder\_Year*) → Semakin baru tahun produksinya, biasanya harga semakin tinggi. Ini juga masuk akal secara bisnis.
- **Options** (*ordinal\_options\_Options*) dan **Mileage** juga cukup penting: → Banyak fitur tambahan (Options) dan jarak tempuh (Mileage) sering dikaitkan langsung dengan kondisi dan daya tarik mobil.
- **Make** (*binary\_large\_Make\_2*) → Beberapa produsen mobil juga menjadi salah satu feature terpenting dalam menentukan harga mobil bekas.

# Kesimpulan

Berdasarkan analisis feature importance dari **model CatBoost**, fitur-fitur yang paling berpengaruh terhadap prediksi harga mobil bekas di Arab Saudi adalah **Engine Size, Year, Make(hasil encoding), Options** dan **Mileage**.

**Model ini berguna untuk :**

- **Memprediksi harga wajar mobil bekas**, membantu pengguna dalam proses jual-beli (baik dealer maupun individu).
- **Memberi dasar keputusan lebih objektif**, sehingga mengurangi risiko undervaluation atau overpricing.
- Dengan menampilkan rentang error ( $\pm 20\%$ ) dan visualisasi prediksi vs. aktual, pengguna bisa memahami akurasi model secara visual, ini bisa **mendatangkan customer** karena customer **bisa lebih percaya atau memiliki trust** terhadap stakeholder yang menggunakan fitur ini (prediksi harga mobil).

# Evaluasi Performa Model

## **RMSE 33,449.61 SAR**

Rata-rata kesalahan prediksi model adalah sekitar **33 ribu Riyal**, dan karena ini Root Mean Squared Error, maka kesalahan besar lebih diperberat. Ini menunjukkan ada beberapa prediksi yang jauh meleset (outlier), terutama pada harga mobil mahal.

## **MAE 16,113.83 SAR**

Rata-rata selisih absolut antara harga aktual dan prediksi adalah sekitar **16 ribu Riyal**. Ini adalah ukuran kesalahan yang lebih stabil dan tidak terlalu sensitif terhadap outlier seperti RMSE.

## **MAPE 24.61%**

Secara rata-rata, prediksi model meleset sebesar **24.61%** dari harga aslinya. Misalnya jika harga mobil sebenarnya SAR 100,000, maka prediksi model bisa saja berada di kisaran SAR 75,000 – 125,000.

Pada case ini MAPE cukup dekat dengan rata-rata selisih harga dari harga mobil di syarah.com dan competitornya, yakni 10-30%. Meski begitu MAPE masih diatas nilai tengah (20%), sehingga model masih terbuka untuk diperbaiki lagi hingga bisa sampai 20% atau kurang dari. Model catboost ini dapat dituning lebih lanjut misalnya pada parameternya, ataupun perbaikan pada data dengan cara menambah fitur yang bisa membantu prediksi atau membersihkan data dari anomali berdasarkan domain knowledge yang bisa dipelajari ataupun dikordinasikan dengan expertise.

# Rekomendasi

**Lakukan A/B Testing Model,** bisa dilakukan setelah melakukan tuning yang berbeda terhadap model ataupun melakukan perbaikan data atau feature engineering lebih lanjut, baru setelahnya dibuat perbandingan melalui metrics RMSE, MAE, ataupun MAPEnya.

**Perluas Fitur Saat Pengumpulan Data,** seperti kondisi mobil, accident history, riwayat servis, tangan ke berapa, warna, dan lainnya yang bisa ditanyakan kepada expertise dibidang jual beli mobil.

**Eksplorasi Kegunaan Lain Model Ini,** misalnya estimasi depresiasi nilai mobil per tahun, rekomendasi harga pasang iklan, pendekripsi harga tidak wajar (ini bisa mencegah salah listing ataupun penipuan).

Untuk estimasi kasual/awal oleh pengguna umum, model sudah cukup layak digunakan. Namun untuk penggunaan komersial skala besar (seperti dealer resmi atau integrasi langsung ke e-commerce mobil), **model sebaiknya dilatih ulang** dengan fitur tambahan, **tuningnya disempurnakan**, diuji di data baru secara berkala. Agar ketika nantinya ingin menggunakan fitur prediksi harga pada stakeholder, alih-alih mendapatkan trust customer, model yang prediksinya belum benar-benar akurat (dalam konteks ini masih 24% dari 10-30%) dikhawatirkan dapat menjadi backfire. Dimana user malah akan jadi ragu dikarenakan stakeholder menggunakan sistem prediksi harga yang tidak akurat, dampaknya bisa kehilangan user ataupun kehilangan trust user.

**Karena kesalahan terbesar terjadi pada mobil dengan harga tinggi (>300,000 SAR), maka solusinya antara lain :**

- Menambah sampel pada segmen mobil mewah/premium.
- Tambah fitur: misalnya fitur spesifik untuk mobil mahal/premium seperti "imported", "accident history", atau "luxury package".
- Evaluasi apakah harga tersebut terlalu dipengaruhi oleh faktor yang belum terekam di dataset (missing features).



# Thank You