

Model Prediksi Harga Mobil Bekas di DKI Jakarta

Kelompok 1 IF-B

Ahmad Zainur Fadli
123230049
Praktikum IF-C

Muhammad Raihan S.
123230072
Praktikum IF-E

Ahmad Habib Hamidi
123230077
Praktikum IF-C



Latar Belakang dan Tujuan?

Latar Belakang

Pasar mobil bekas di Jakarta memiliki variasi harga yang sangat tinggi dan kompleks. Seringkali terjadi kesulitan dalam menentukan harga wajar akibat subjektivitas penjual serta banyaknya faktor teknis kendaraan yang mempengaruhi nilai jual, sehingga memicu ketidakpastian bagi pelaku pasar.

Tujuan

Penelitian ini bertujuan membangun model Machine Learning untuk memprediksi harga mobil secara objektif berdasarkan data historis, serta mengimplementasikannya ke dalam aplikasi web sederhana agar pengguna dapat mengetahui estimasi harga pasar yang akurat dan transparan secara instan.



Dataset Yang digunakan

Dataset didapatkan dari hasil scraping yang dilakukan pada halaman website Carmudi pada bagian mobil bekas untuk daerah DKI Jakarta

Data Dictionary

Merk

Nama merek mobil (Contoh: Toyota)

Engine CC

Kapasitas mesin (cc)

Fuel Type

Jenis bahan bakar

Model

Nama model mobil

Transmission

Jenis transmisi (Contoh: manual, otomatis)

Jarak Tempuh

Jarak Tempuh (km)

Tahun Produksi

Tahun pembuatan

Seat Capacity

Kapasitas penumpang (orang)

Harga

Harga jual (Rp)

Kualitas Data

Missing Value

Variable	Jumlah	Persentase
Merk	5	0.04 %
Model	5	0.04 %
Tahun Produksi	314	2.45 %
Jarak Tempuh	314	2.45 %
Engine CC	554	4.33 %
Transmission	554	2.45 %
Seat Capacity	314	2.45 %
Harga	337	2.63 %
Fuel Type	319	2.49 %

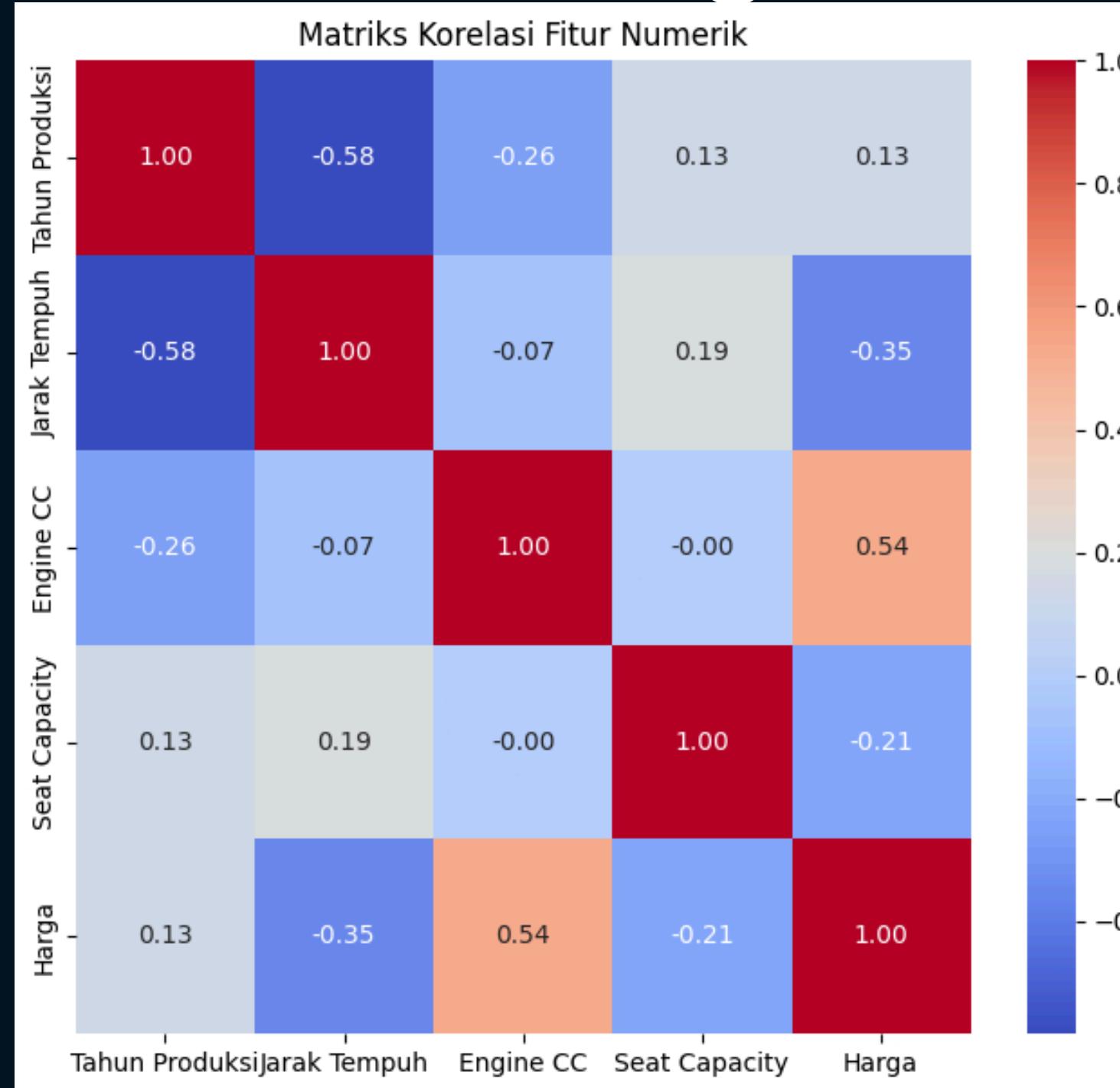
Outlier

Variable	Jumlah	Persen tase	Batas Bawah	Batas Atas
Harga	1171	9.15 %	-462000 000.0	1.282000 e+09
Engine CC	400	3.13 %	10.5	3.974500 e+03
Jarak Tempuh	274	2.14 %	-57500.0	1.425000 e+05
Tahun Produksi	180	1.41 %	2007	2.031000 e+03
Seat Capacity	30	0.23 %	2.0	1.00000e +01

Duplikasi

1936 Data

EDA Berinsight



Positif

Kolom "Tahun Produksi" dan "Engine CC" memiliki nilai positif dengan fitur harga, yang mana kolom ini mempengaruhi kenaikan harga, jadi semakin tinggi kolom "Tahun Produksi" dan "Engine CC" maka semakin tinggi Harga yang diberikan

Negatif

- Kolom "Jarak Tempuh" dan "Seat Capacity" memiliki nilai negatif dengan harga, yang mana kolom ini mempengaruhi penurunan fitur harga, jadi semakin tinggi kolom "Jarak Tempuh" dan "Seat Capacity" maka semakin rendah harga yang diberikan
- Tahun Produksi vs. Jarak Tempuh: Ini adalah korelasi negatif yang kuat. Secara logis, mobil yang lebih baru (Tahun Produksi tinggi) cenderung memiliki Jarak Tempuh yang lebih rendah. Ini adalah hubungan yang wajar di dunia nyata.

1

Menghapus Nilai Null & Duplicated

Dikarenakan data sudah cukup banyak sehingga jika data null dan duplicated dihapus tidak akan berpengaruh secara signifikan ke model

3

Mengubah Jumlah Fitur Model

Fitur model mengandung lebih dari 1 kata, agar tidak terjadi data leakage (dalam kasus ini fitur model terlalu dominan dalam penentuan prediksi harga)

2

Feature Engineering

Membuat fitur umur mobil dari tahun saat ini - fitur tahun produksi

4

Encoding

- Kolom transmission & Fuel Type menggunakan label encoding: Low Cardinality (sedikit kategori)
- Kolom merk & model menggunakan frequency encoding: High Cardinality (kategori sangat banyak), Frekuensi mengandung informasi penting



Model



Algoritma yang dipakai

- Linear Regression, model linear sederhana sebagai pembanding dasar
- Random Forest, model non-linear yang kuat dan menangani variabel kompleks
- LightGBM, model boosting yang cepat dan akurat



Train-test split

- Dataset dibagi menjadi 80% data latih dan 20% data uji
- Pembagian dilakukan menggunakan `train_test_split` dengan `random_state=42` untuk memastikan hasil konsisten



Hyperparameter

- Menggunakan `RandomizedSearchCV` untuk mencari kombinasi hiperparameter terbaik pada Random Forest
- Menggunakan 20 iterasi dan 5-fold cross validation

Evaluasi Model

Linear Regression

MAE (Juta Rp) : 292.35
RMSE (Juta Rp) : 620.45
R-squared : 0.45

Random Forest

MAE (Juta Rp) : 96.30
RMSE (Juta Rp) : 346.34
R-squared : 0.83

LightGBM

MAE (Juta Rp) : 105.10
RMSE (Juta Rp) : 304.46
R-squared : 0.87

Alasan Menggunakan Random Forest

Random Forest dipilih sebagai model utama karena bersifat lebih stabil, lebih mudah dijelaskan kepada stakeholder, dan memiliki performa baseline yang sudah sangat baik tanpa tuning yang kompleks.

Selain itu, Random Forest lebih robust terhadap outlier dan noise yang umum ditemukan pada data harga mobil bekas.

Model ini juga lebih mudah diinterpretasikan sehingga memudahkan dalam menjelaskan faktor-faktor yang mempengaruhi harga kepada pihak bisnis.

Walaupun LightGBM memiliki skor R^2 yang sedikit lebih tinggi, perbedaannya tidak signifikan secara praktikal. Random Forest memberikan trade-off terbaik antara performa, stabilitas, dan interpretabilitas sehingga lebih cocok untuk kasus bisnis prediksi harga mobil bekas.



Masalah & Batasan



Batasan Model

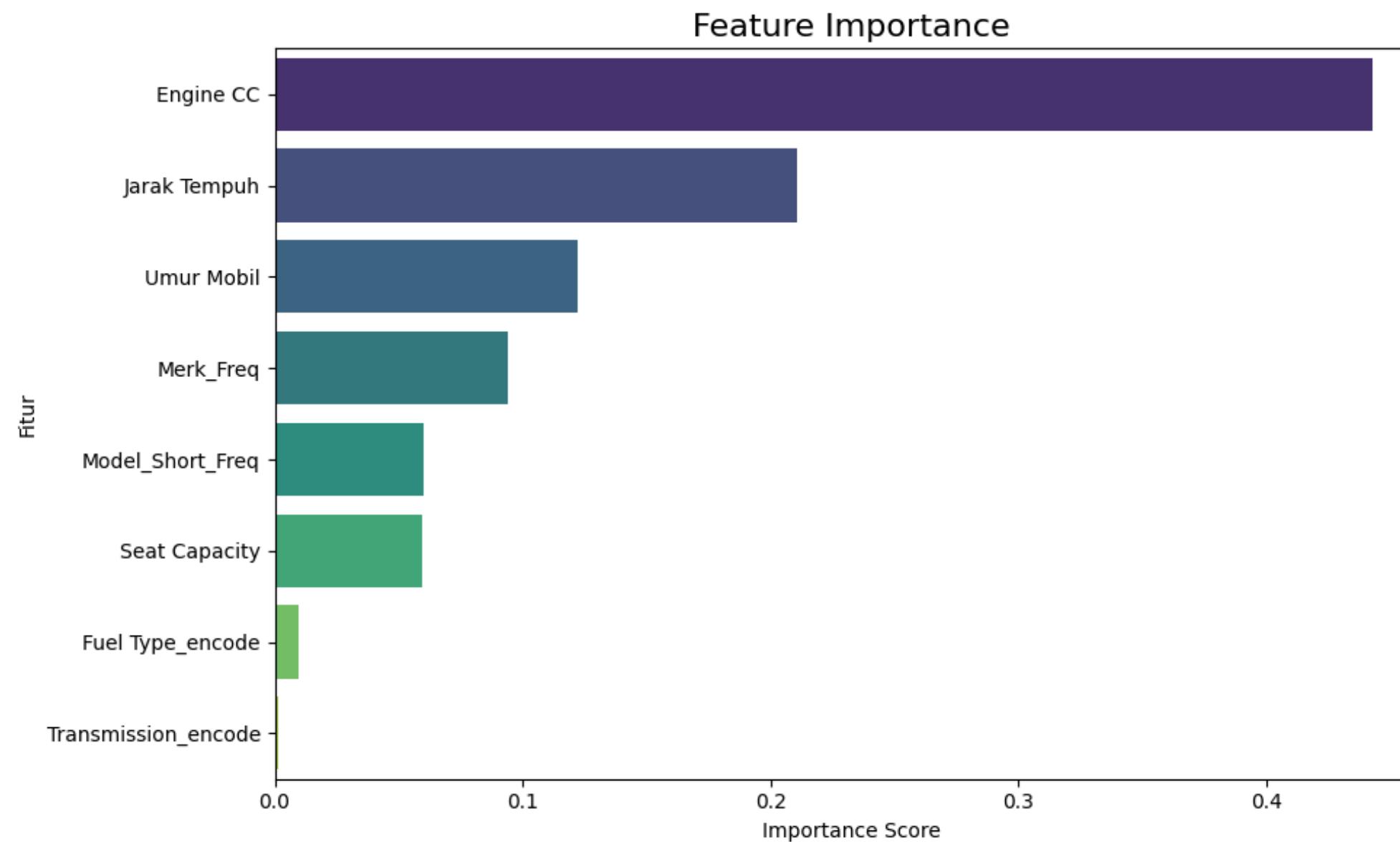
Meskipun performanya baik, model masih memiliki batasan. MAE sebesar ± 96 juta menunjukkan bahwa terdapat error yang cukup besar pada beberapa tipe mobil tertentu, terutama pada mobil premium atau mobil yang jarang muncul pada dataset. Model juga tidak mempertimbangkan faktor non-numerik seperti kondisi fisik kendaraan, riwayat servis, atau harga pasar regional sehingga prediksi dapat berbeda dari harga aktual di lapangan. Selain itu, model sangat bergantung pada kualitas dataset; apabila data tidak lengkap atau tidak representatif, prediksi bisa kurang akurat.

[HOME](#)[ABOUT US](#)[MORE](#)

Hubungan Model dengan Masalah

- Masalah awal: Harga mobil bekas sulit diprediksi secara konsisten karena banyak faktor.
- Tujuan project: Membangun model untuk memprediksi harga mobil secara lebih akurat.
- Hubungan model: Model Random Forest digunakan karena mampu menangkap pola non-linear pada faktor-faktor seperti tahun, kilometer, merk, dan tipe. Model ini secara langsung menjawab masalah awal yaitu kebutuhan prediksi harga yang lebih akurat dan konsisten

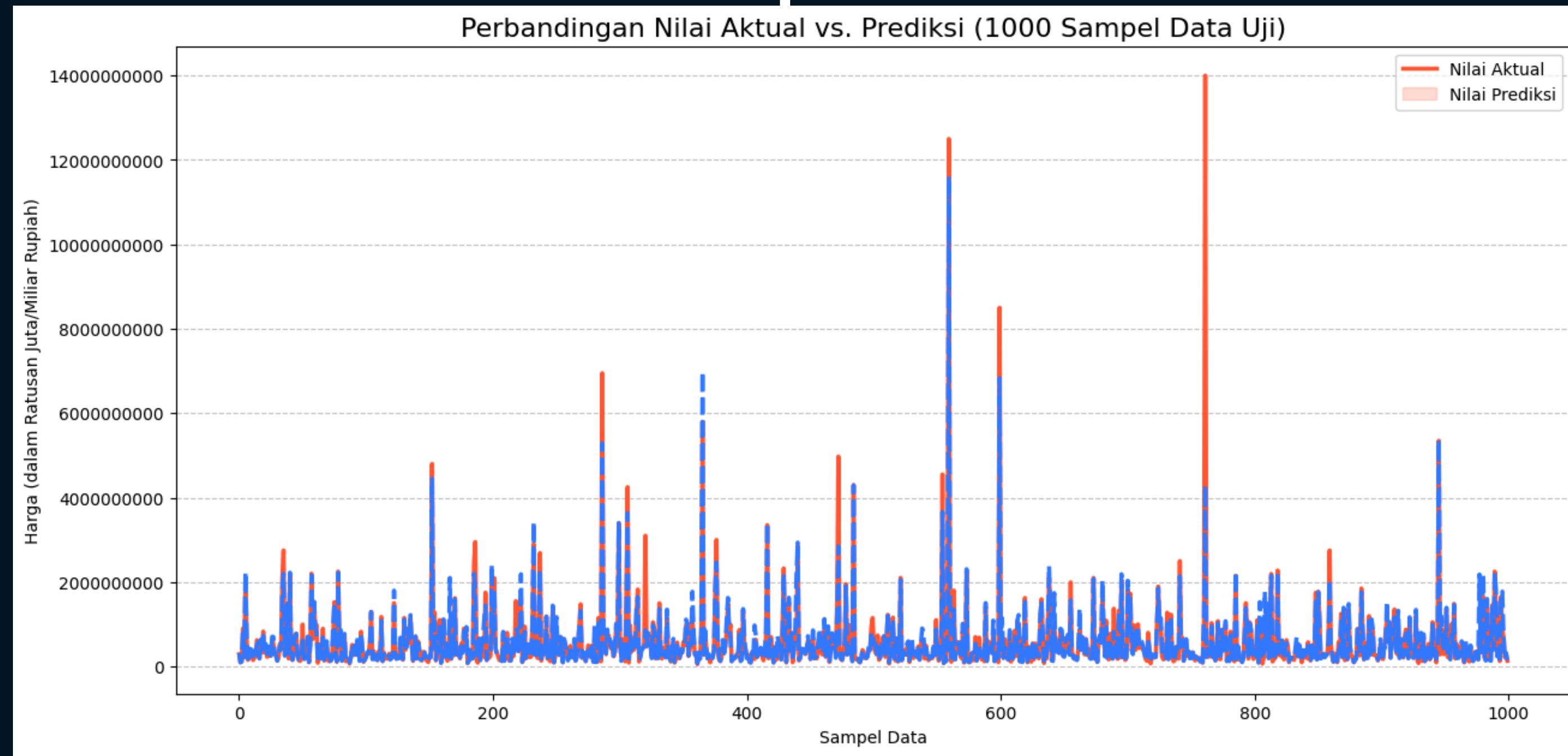
Visualisasi Pendukung Hasil Model

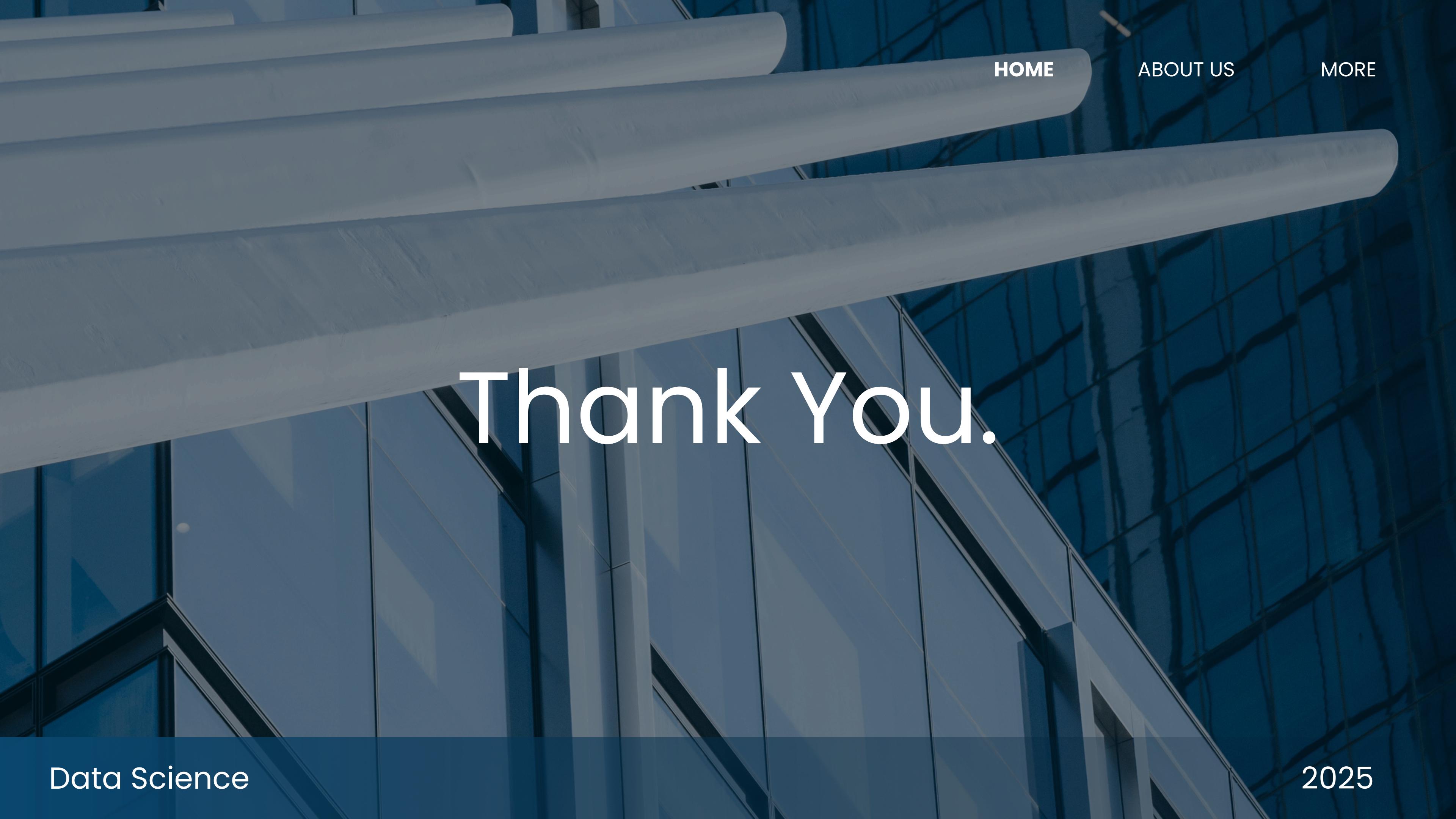


Feature Importance

Engine CC 0.4430
Jarak Tempuh 0.2105
Umur Mobil 0.1221
Merk_Freq 0.0938
Model_Short_Freq 0.0600
Seat Capacity 0.0593
Fuel Type_encode 0.0097
Transmission_encode 0.0016

Visualisasi & Output Model



A dark, semi-transparent background image of a modern architectural structure. It features large, curved, light-colored panels that wrap around the building's facade. The building has a grid of windows, some of which are illuminated from within, creating a warm glow against the cool blue tones of the exterior. The overall composition is dynamic, with the curves of the building's design creating a sense of movement.

HOME

ABOUT US

MORE

Thank You.

Data Analyst and Data Business Analyst



Data Analyst

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam.



Data Business Analyst

Lore ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam.



Tools Commonly Used

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Core Skills

Problem-solving

Critical thinking

Data interpretation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Role in Data Collection and Preparation

How data business analysts gather

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.



Data Visualization and Reporting

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam.

