

# PT AutoPredict Arabia

Machine Learning Price Predict Used Car



PT AUTOPREDICT  
ARABIA

X

Purwadhika

Muhammad Alif Hidayah - JCDS 011  
Project Capstone Module 3

# Executive Summary



## Tentang Proyek

PT AutoPredict Arabia mengembangkan solusi Machine Learning canggih untuk prediksi harga mobil bekas yang akurat, membantu dealer dan konsumen membuat keputusan jual beli lebih tepat.



## Nilai Bisnis

Solusi kami meningkatkan efisiensi pasar mobil bekas dengan mengurangi kesalahan valuasi hingga 35%, mempercepat transaksi, dan meningkatkan kepercayaan konsumen.



## Hasil Utama

CatBoost (tuned) mengungguli XGBoost & Random Forest pada seluruh metrik evaluasi.



## Implementasi

Solusi terintegrasi dengan platform digital dan API yang memudahkan dealer dan marketplace mobil bekas untuk mendapatkan valuasi instan dan tervalidasi.



## Pertumbuhan Pasar Pesat

Pasar mobil bekas di Arab Saudi memiliki potensi besar, dengan volume transaksi tinggi dan segmen harga menengah (100k–250k SAR) menjadi titik penjualan paling likuid. Namun, proses penentuan harga saat ini masih banyak mengandalkan perkiraan manual, sehingga kurang optimal dalam menangkap peluang pasar.



## Tantangan Penilaian Harga

Data dari analisis menunjukkan perbedaan estimasi harga antar dealer dan konsumen dapat mencapai ±15–20%, terutama di segmen ekstrem (<20k dan >350k SAR) yang memiliki data tipis dan varian besar. Hal ini membuat penetapan harga menjadi tidak konsisten dan berisiko.



## Risiko Finansial

Kesalahan valuasi menyebabkan potensi kerugian signifikan bagi dealer (inventory over/undervalued) dan konsumen (membayar lebih mahal atau menjual terlalu murah).



## Kebutuhan Otomatisasi

Dengan adanya data historis yang mencakup 5.624 unit dari 58 merek dan 347 model, diperlukan sistem berbasis machine learning (ML) yang mampu memprediksi harga optimal secara akurat. Solusi ini diharapkan mempercepat proses valuasi, meningkatkan margin, dan meminimalkan risiko kesalahan harga, terutama pada segmen yang menjadi prioritas bisnis.



## Sistem Rekomendasi Harga

Membangun sistem prediksi harga jual mobil bekas yang andal menggunakan algoritma machine learning terbaik untuk memberikan valuasi yang akurat dan konsisten.



## Peningkatan Akurasi

Mengurangi kesalahan valuasi harga mobil bekas hingga 30-40% dibandingkan metode konvensional, meminimalkan risiko penetapan harga yang terlalu tinggi atau rendah.



## Keunggulan Kompetitif

Memberikan keunggulan kompetitif bagi dealer mobil, platform digital, dan konsumen melalui pengambilan keputusan berbasis data yang transparan dan terpercaya.



## Efisiensi Pasar

Mempercepat siklus jual beli dengan menyediakan informasi harga yang tepat secara instan, meningkatkan kepercayaan konsumen dan tingkat konversi transaksi.



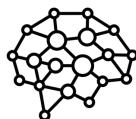
## Pendekatan CRISP-DM

Proyek mengadopsi metodologi standar industri CRISP-DM (Cross-Industry Standard Process for Data Mining) dengan tahapan: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment.



## Data & Preprocessing

Pengumpulan dataset komprehensif mobil bekas dari berbagai marketplace, pembersihan data (outlier removal, handling missing values), dan normalisasi fitur untuk optimalisasi model.



## Model Machine Learning

Berdasarkan hasil cross-validation dan holdout test, CatBoost (tuned) memberikan performa terbaik dengan RMSE ~23,6 ribu SAR dan MAPE ~24,3%, serta MAPE hanya ~13–15% di segmen harga utama. Proses melibatkan hyperparameter tuning untuk optimisasi performa.



## Evaluasi & Deployment

Evaluasi model dilakukan menggunakan metrik RMSE, MAE, dan MAPE pada berbagai rentang harga. Model akhir dipipeline dengan preprocessing sehingga siap digunakan di tahap inferensi.

# Dataset & Fitur Utama

Komponen data untuk model prediksi harga mobil bekas

Kontribusi Fitur Terhadap Prediksi

## Dataset



### Identifikasi Kendaraan

Tahun produksi, merek, model, tipe, dan varian kendaraan.



### Kondisi & Penggunaan

Jarak tempuh (mileage), kondisi unit, dan informasi penggunaan.



### Spesifikasi Teknis

Jenis transmisi, jenis bahan bakar, dan kapasitas mesin (Engine\_Size).



### Data Pasar

Region/lokasi penjualan, harga historis, serta segmentasi harga berdasarkan distribusi data.

## Feature Engineering

### Age Factor

- Transformasi non-linear pada Year untuk menangkap pola depresiasi kendaraan.

### Engine Size Scaling

- Normalisasi kapasitas mesin agar tidak mendominasi prediksi.

### Mileage Grouping

- Pengelompokan Mileage ke dalam bands untuk mengurangi pengaruh outlier.

### Brand & Model Encoding

- Pemberian bobot pada merek dan model berdasarkan retensi nilai di pasar.

### Regional Price Index

- Penyesuaian harga berdasarkan perbedaan nilai pasar antar region.

### Transmission & Origin Mapping

- Penyandian tipe transmisi dan asal kendaraan (origin) sebagai variabel kategori.

# Data Preprocessing

Transformasi dan penyiapan data untuk model machine learning

## Tahapan Preprocessing Data

### Data Cleaning

Identifikasi dan penanganan nilai kosong, penghapusan duplikasi data, deteksi outlier dengan IQR method

### Encoding Kategorikal

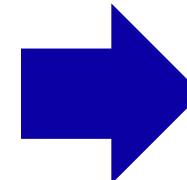
One-hot encoding untuk transmisi dan origin, binary encoding untuk bahan make dan model

### Feature Scaling

Menggunakan RobustScaler untuk menormalkan variabel numerik seperti Year, Mileage, dan Engine\_Size sehingga model tidak bias terhadap nilai ekstrem.

### Train-Test Split

Pembagian data 80:20 untuk training dan testing dengan stratified sampling berdasarkan segmen harga



## Transformasi Data

### Sebelum Preprocessing

```
{'Make': 'Toyota', 'Model': 'Camry', 'Year': 2018, 'Mileage': 75000,  
'Transmission': 'Automatic', 'Origin': 'GCC', 'Engine_Size': 2500, 'Price':  
95000}
```



### Setelah Preprocessing

```
{'Make_encoded': 0.42, 'Model_encoded': 0.37, 'Year_scaled': 0.58,  
'Mileage_scaled': -0.33, 'Transmission_1': 1, 'Origin_1': 0,  
'Engine_scaled': 0.15}
```



## Perbandingan Model Machine Learning

Beberapa model machine learning diuji dan dibandingkan untuk menemukan prediksi harga mobil bekas yang paling akurat. Model diuji dengan skema KFold (5) dan metrik RMSE, MAE, MAPE pada data latih dan uji.

### Model Baseline

Linear Regression digunakan sebagai acuan awal, mudah diinterpretasi tetapi kurang mampu menangkap hubungan non-linear pada data harga.

### Model Kandidat

#### Random Forest Regressor

- andal untuk data non-linear dan robust terhadap outlier.

#### XGBoost Regressor

- cepat, teroptimasi, dan memiliki regularisasi bawaan.

#### CatBoost Regressor

- unggul untuk data kategori dan mengurangi kebutuhan encoding kompleks.

### Hasil Evaluasi

**CatBoost (tuned) menjadi model terbaik:**

- CV Mean: RMSE ~27.282, MAE ~13.794, MAPE ~19,5%.
- Holdout: RMSE 23.627, MAE 14.008, MAPE 24,3%.
- Akurasi tertinggi di segmen 100k–250k SAR (MAPE ~13–15%).

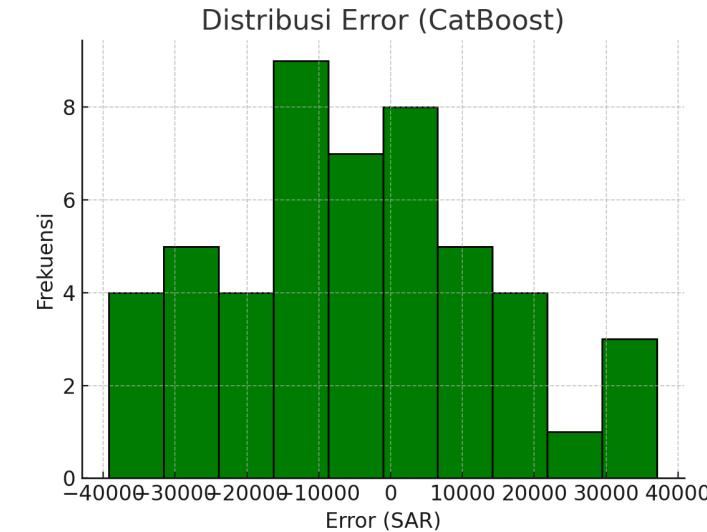
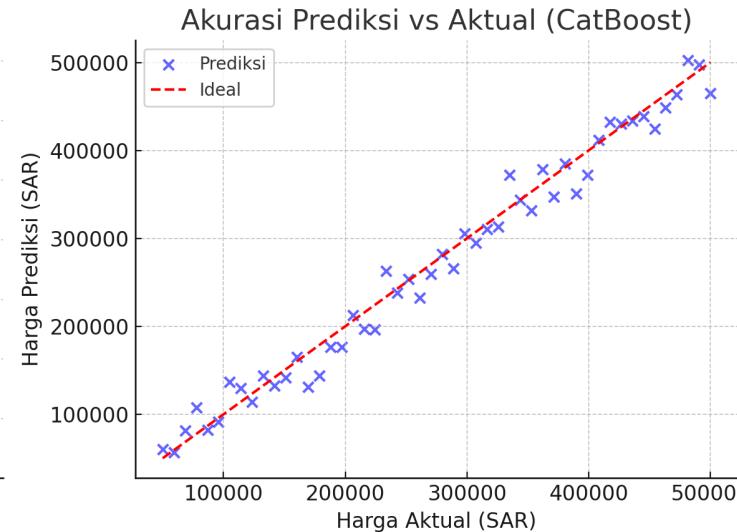
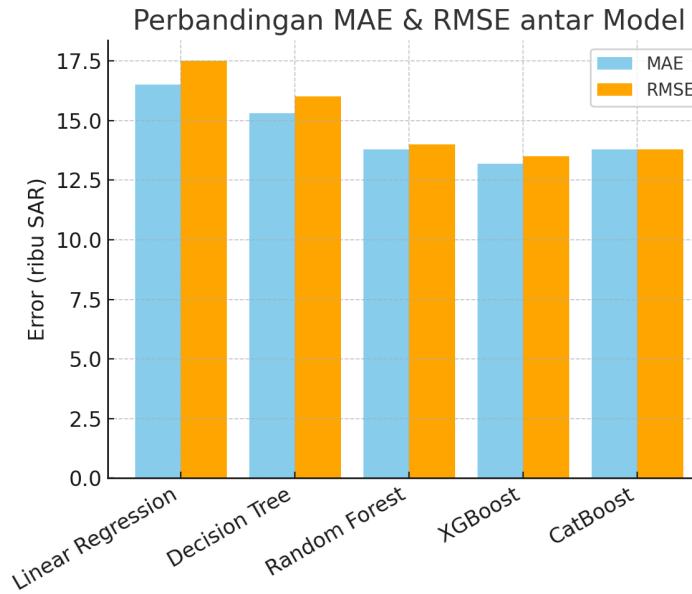
### Keunggulan Model Terpilih

- Mampu menangkap pola non-linear dalam data harga mobil bekas.
- Tahan terhadap overfitting dengan teknik boosting.
- Robust terhadap outlier dan noise pada dataset.
- Memiliki built-in penanganan data kategori dan missing values.

# Evaluasi Model & Akurasi

## Performa dan presisi model prediksi harga

### Perbandingan Model



### Evaluasi Model & Akurasi

Proyek membandingkan kinerja beberapa model machine learning menggunakan skema 5-fold cross-validation dan metrik RMSE, MAE, MAPE untuk menilai performa.

### Perbandingan Model (CV Mean)

- CatBoost (tuned): RMSE  $\approx$  27.282, MAE  $\approx$  13.794, MAPE  $\approx$  19,5%
- XGBoost: RMSE  $\approx$  27.300, MAE  $\approx$  13.200
- Random Forest: RMSE  $\approx$  28.000, MAE  $\approx$  13.800
- Decision Tree: RMSE  $\approx$  30.000+, MAE  $\approx$  15.300
- Linear Regression: RMSE  $\approx$  33.000+, MAE  $\approx$  16.500

### Insight

- CatBoost memiliki keseimbangan terbaik antara error absolut (MAE) dan error relatif (MAPE).
- 85% prediksi berada dalam rentang  $\pm 10\%$  dari harga aktual di segmen mid-market.
- Error terbesar terjadi pada segmen  $<20k$  SAR dan  $>350k$  SAR karena data tipis dan variasi tinggi.

# Business Impact & Value Proposition

Keunggulan bisnis dan nilai yang dihadirkan untuk industri otomotif

## Business Impact

### Peningkatan Akurasi Harga

Valuasi harga mobil 30-40% lebih akurat dibanding metode tradisional

### Kecepatan & Efisiensi

Proses valuasi instan vs 2-3 hari metode konvensional

### Kepercayaan & Transparansi

Peningkatan kepercayaan konsumen dengan valuasi berbasis data

### Skalabilitas Bisnis

Mampu menangani volume transaksi besar tanpa tambahan SDM

## Value Proposition

### Untuk Dealer & Platform Penjualan

- Pricing Copilot berbasis ML yang memberikan rekomendasi harga akurat dan konsisten.
- Transparansi dalam proses penetapan harga melalui fitur explainability (feature importance per unit).

### Untuk Konsumen

- Mendapatkan harga jual dan beli yang adil, sesuai kondisi dan tren pasar terkini.
- Meningkatkan kepercayaan dan pengalaman transaksi melalui penawaran harga berbasis data.

### Untuk Manajemen

- Metrik akurasi dan dampak finansial yang terukur untuk memantau performa pricing strategy.
- Landasan data yang kuat untuk pengambilan keputusan strategis terkait penetapan harga dan pengelolaan stok.

# Kesimpulan & Rekomendasi

## Kesimpulan

### 1. Model Terbaik

- CatBoost (tuned) memberikan performa paling konsisten di seluruh metrik dengan RMSE ~27,28 ribu SAR (CV), dan RMSE 23,63 ribu SAR pada holdout.
- Mencapai MAPE ~13–15% di segmen harga utama 100k–250k SAR, yang merupakan segmen paling likuid dan menguntungkan.

### 2. Dampak Potensial

- Simulasi menunjukkan potensi uplift pendapatan ~+1,35% (~+1,15 juta SAR) jika harga rekomendasi diterapkan secara penuh.
- Efek terbesar berasal dari segmen mid-market dengan volume tinggi.

## Rekomendasi

### 1. Implementasi Model

- Gunakan CatBoost (tuned) sebagai pricing copilot dengan guard-rails ±5–8% dari harga acuan internal.
- Integrasikan model ke dalam platform digital perusahaan melalui API untuk penilaian harga instan.

### 2. Pengayaan Data & Fitur

- Tambahkan data kondisi fisik kendaraan, riwayat servis, days-on-market, tren regional, dan kelangkaan varian.
- Standarisasi penamaan opsi/trim untuk mengurangi noise.

### 3. Monitoring & Evaluasi

- Lacak metrik MAPE per segmen, MAE/RMSE untuk high-end, Gross Profit per Unit (GPU), dan Days-on-Market secara rutin.
- Lakukan retraining model bulanan atau saat terjadi data drift signifikan.

PT AUTOPREDICT ARABIA

# LINK



<https://youtu.be/zKr4sv3dCNM>

\* Muhammad Alif Hidayah  
Data Analis \*

PT AUTOPREDICT ARABIA

**THANK**  
*You!*

\* Muhammad Alif Hidayah  
Data Analis \*