Study Case: Saudi Arabia Used Cars

Capstone Project Module 3 Purwadhika

Gian Habli Maulana

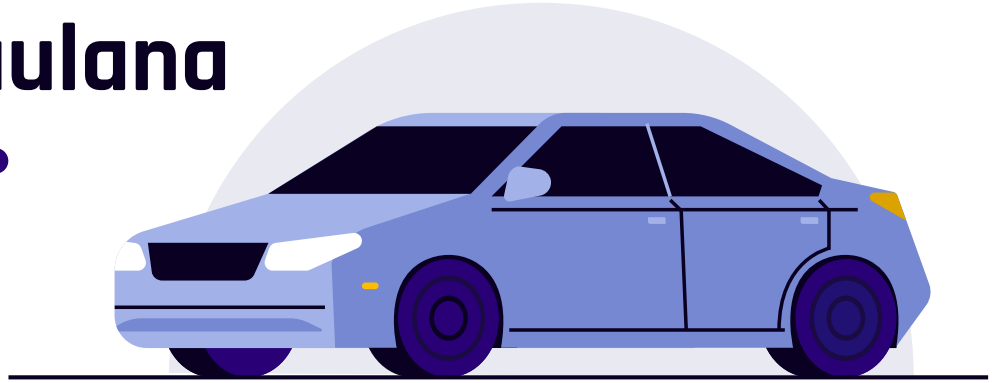
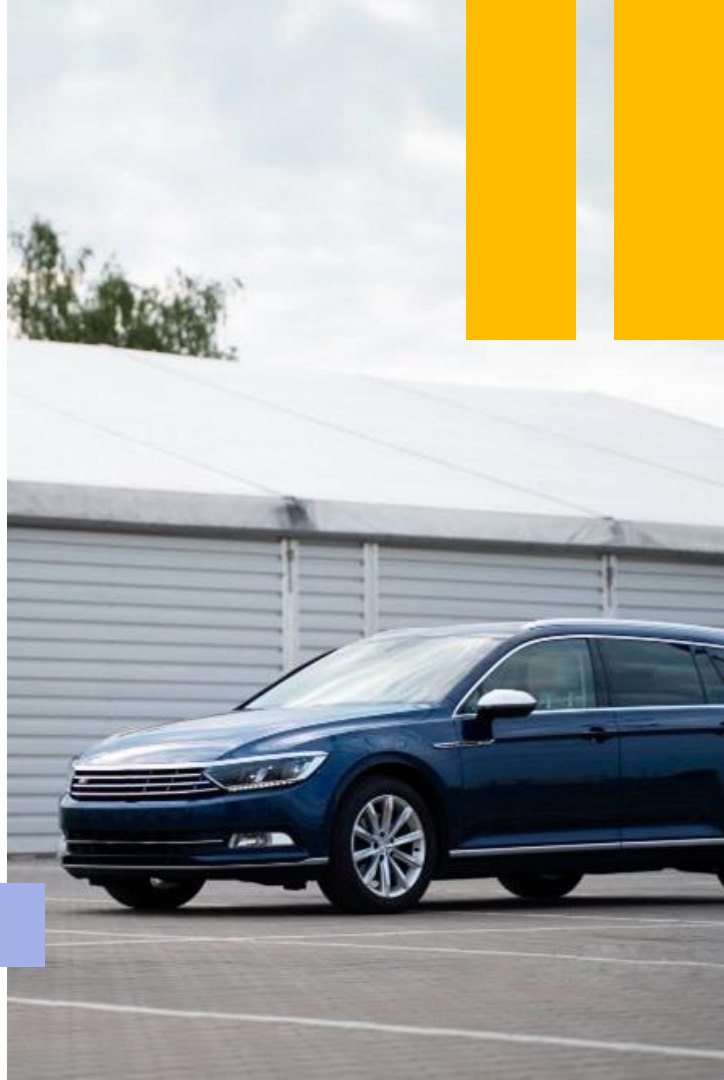


Table of contents

01.

Business Understanding



01. BUSINESS UNDERSTANDING

Context

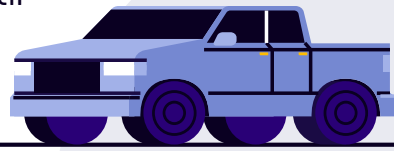
1. Seiring dengan pertumbuhan populasi dan urbanisasi di Arab Saudi, kebutuhan akan mobil sebagai alat transportasi semakin meningkat.

2. Dalam beberapa tahun terakhir, pasar mobil bekas di Arab Saudi telah berkembang pesat akibat meningkatnya permintaan akan kendaraan yang terjangkau.

3. Calon pembeli mobil di Arab Saudi cenderung mencari penawaran terbaik dengan mempertimbangkan berbagai faktor seperti merek, model, tahun pembuatan, kondisi, dan harga.

4. Namun, karena adanya keragaman dalam spesifikasi dan kondisi mobil, calon pembeli sering kali kesulitan untuk menilai harga yang adil untuk kendaraan yang mereka minati.

5. Oleh karena itu, perusahaan yang beroperasi di sektor otomotif, baik itu platform jual beli mobil atau dealer, perlu memahami faktor-faktor yang mempengaruhi harga mobil bekas.



01. BUSINESS UNDERSTANDING

Problem Statement

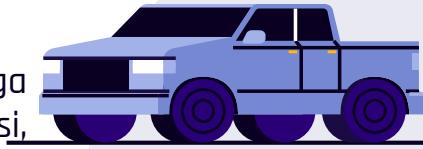
1. Dalam pasar mobil bekas di Arab Saudi, salah satu tantangan utama adalah menentukan harga jual yang kompetitif dan akurat bagi kendaraan yang akan dipasarkan.

2. Data yang tidak konsisten dan beragam tentang harga mobil bekas mengakibatkan kesulitan bagi pembeli dan penjual dalam mengambil keputusan yang tepat.

3. Konsumen sering kali bingung apakah harga yang mereka tawarkan atau diterima sesuai dengan tren pasar saat ini.

4. Selain itu, penjual membutuhkan panduan untuk menetapkan harga yang wajar agar bisa bersaing di pasar, tanpa merugikan diri mereka sendiri.

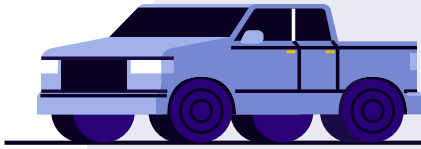
5. Oleh karena itu, perlu dikembangkan model machine learning untuk memprediksi harga mobil bekas yang didasarkan pada berbagai fitur seperti merek, model, tahun produksi, tipe mesin, kilometer yang telah ditempuh, dan kondisi mobil.



01. BUSINESS UNDERSTANDING

Goals

1. Mengembangkan Model Prediksi Harga
2. Meningkatkan Efisiensi Penentuan Harga
3. Meningkatkan Kepercayaan Pasar
4. Menarik Pelanggan Potensial



01. BUSINESS UNDERSTANDING

Key Stakeholders

1. Pemilik Bisnis (Dealers / Penjual Mobil):

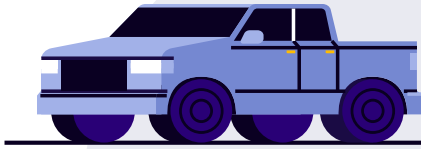
Memanfaatkan hasil prediksi untuk menentukan harga jual mobil bekas yang kompetitif.

2. Konsumen:

Individu yang mencari informasi tentang harga mobil bekas dan diharapkan terbantu melalui rekomendasi harga yang akurat.

3. Pemasar

Menggunakan data dan prediksi untuk strategi pemasaran dan pemahaman pasar.



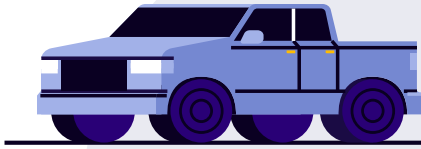
01. BUSINESS UNDERSTANDING

Analytic Approach

Prediksi Nilai Price Menggunakan Model Regresi

1. Dalam membuat model yang dapat memprediksi harga sebuah mobil bekas, pertama-tama dianalisis terlebih dahulu spesifikasi apa saja yang dapat berpengaruh terhadap harga tersebut.

2. Kemudian dibuat berbagai macam model regresi yang bertujuan untuk menentukan harga mobil bekas, dan melalui model-model tersebut ditentukan model terbaik yang memberikan evaluation metrics terbaik yang akan digunakan sebagai final model.



01. BUSINESS UNDERSTANDING

Evaluation Metric

Beberapa Evaluation Metric yang digunakan dalam model ini diantaranya adalah sebagai berikut.

1. MAE (Mean Absolute Error)
2. MAPE (Mean Absolute Percentage Error)
3. RMSE(Root Mean Squared Error)
4. R^2 (R-squared)

02.

Data Understanding

02. Data Understanding

Overview Dataset:

Sumber:
Website Kaggle
https://www.kaggle.com/datasets/reemalruqi/used-cars-in-saudi-arabia

Dataset:
1. Dataset Utama : data_saudi_used_cars.csv

Dimensi:
1. Baris : 5.624
2. Kolom : 11

02. Data Understanding

Overview Dataset:

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Type          5624 non-null   object
 1   Region        5624 non-null   object
 2   Make          5624 non-null   object
 3   Gear_Type     5624 non-null   object
 4   Origin        5624 non-null   object
 5   Options       5624 non-null   object
 6   Year          5624 non-null   int64
 7   Engine_Size   5624 non-null   float64
 8   Mileage       5624 non-null   int64
 9   Negotiable    5624 non-null   bool
 10  Price         5624 non-null   int64
```

Features	Description
`Type`	Tipe mobil bekas.
`Region`	Wilayah tempat mobil bekas ditawarkan untuk dijual.
`Make`	Nama perusahaan.
`Gear_Type`	Ukuran tipe gear mobil bekas.
	1. Automatic
	2. Manual
`Origin`	Asal mobil bekas.
	1. Saudi
	2. Gulf
	3. Other
`Options`	Pilihan mobil bekas.
	1. Full Options
	2. Semi-Full
	3. Standard
`Year`	Tahun pembuatan.
`Engine_Size`	Ukuran mesin mobil bekas.
`Mileage`	Jarak tempuh mobil bekas.
`Negotiable`	Benar, jika harga 0 berarti bisa nego.
`Price`	Harga mobil bekas.

02. Data Understanding

Overview Dataset:

	Type	Region	Make	Gear_Type	Origin	Options	Year	Engine_Size	Mileage	Negotiable	Price
0	Corolla	Abha	Toyota	Manual	Saudi	Standard	2013	1.4	421000	True	0
1	Yukon	Riyadh	GMC	Automatic	Saudi	Full	2014	8.0	80000	False	120000
2	Range Rover	Riyadh	Land Rover	Automatic	Gulf Arabic	Full	2015	5.0	140000	False	260000
3	Optima	Hafar Al-Batin	Kia	Automatic	Saudi	Semi Full	2015	2.4	220000	False	42000
4	FJ	Riyadh	Toyota	Automatic	Saudi	Full	2020	4.0	49000	True	0

03.

Data Cleansing and Preprocessing

Data Cleansing

1. **Handling Duplicate Data**

```
[79] # Drop duplicated rows

     df.drop_duplicates(inplace=True)
```

2. **Handling Handling Column 'Origin' with 'Unknown' Value**

```
[81] df['Origin'] = df['Origin'].replace(['Unknown'],'Other')
```

3. **Handling Column 'Price' with '0' Value**

```
[85] df = df[df['Price'] != 0]
```

4. **Drop Unnecessary Columns**

```
[86] df.drop(columns = ['Negotiable'], inplace=True)
```

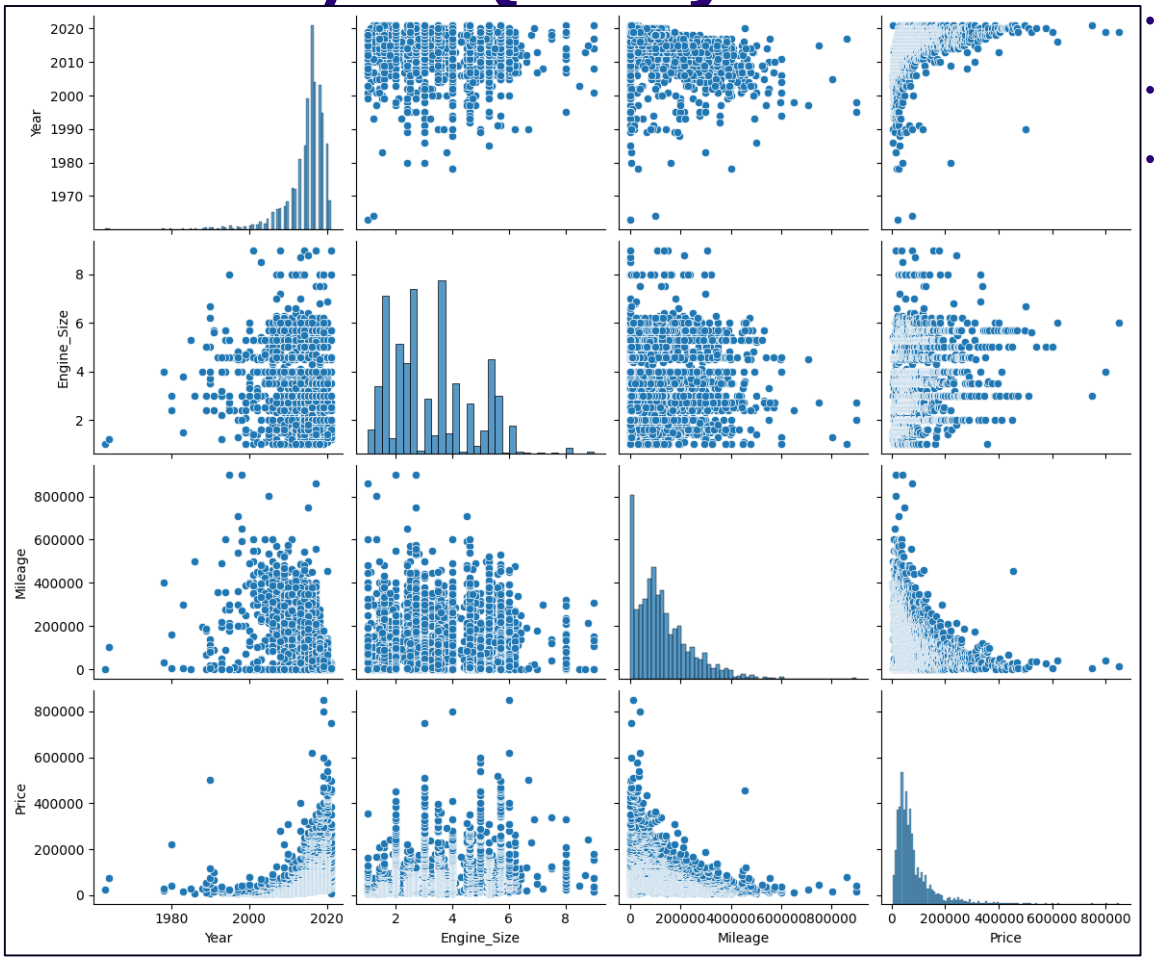
5. **Handling Outliers**

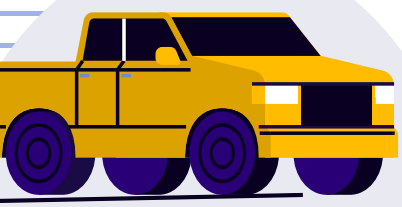
```
[100] # Drop data based on column Price
      df = df[(df['Price'] >= 5000)]
```

```
# Drop data based on column Mileage
df = df[(df['Mileage'] <= 1000000)]
```

Exploratory Data Analysis (EDA)

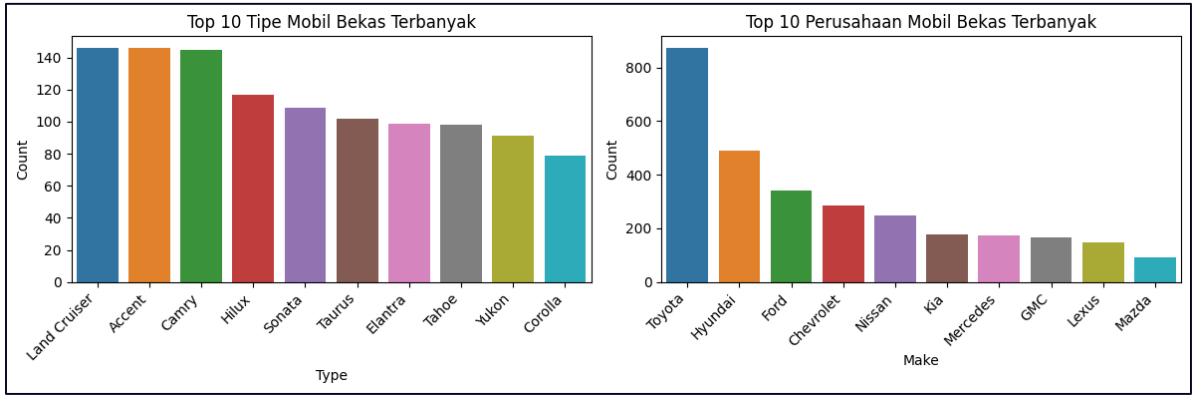
1. **Numerical Variabel**

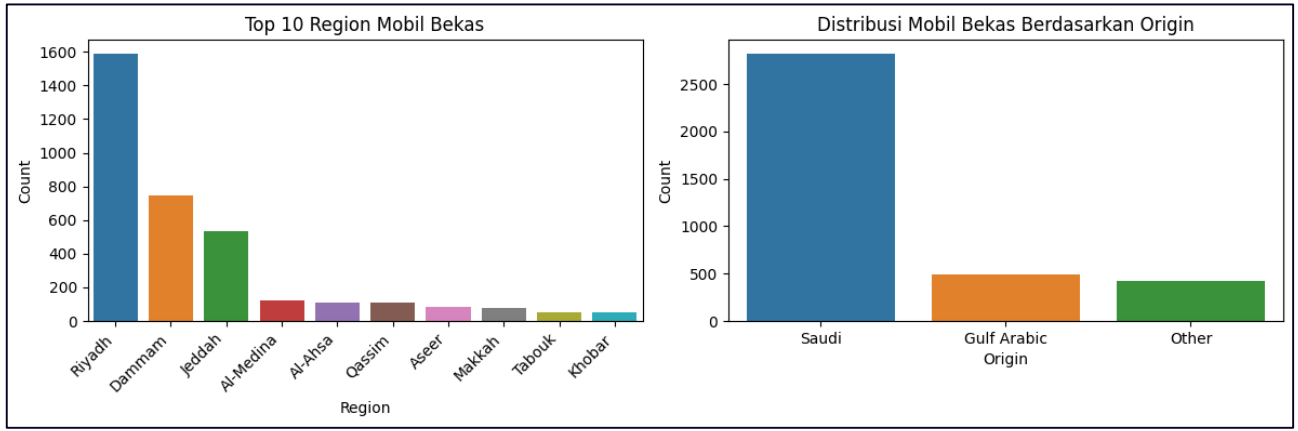


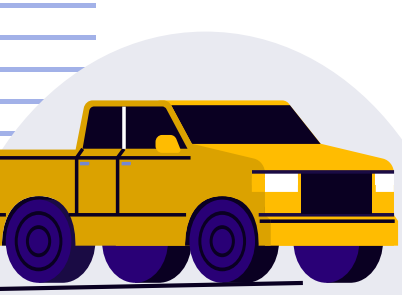


Exploratory Data Analysis (EDA)

2. Categorical Variabel

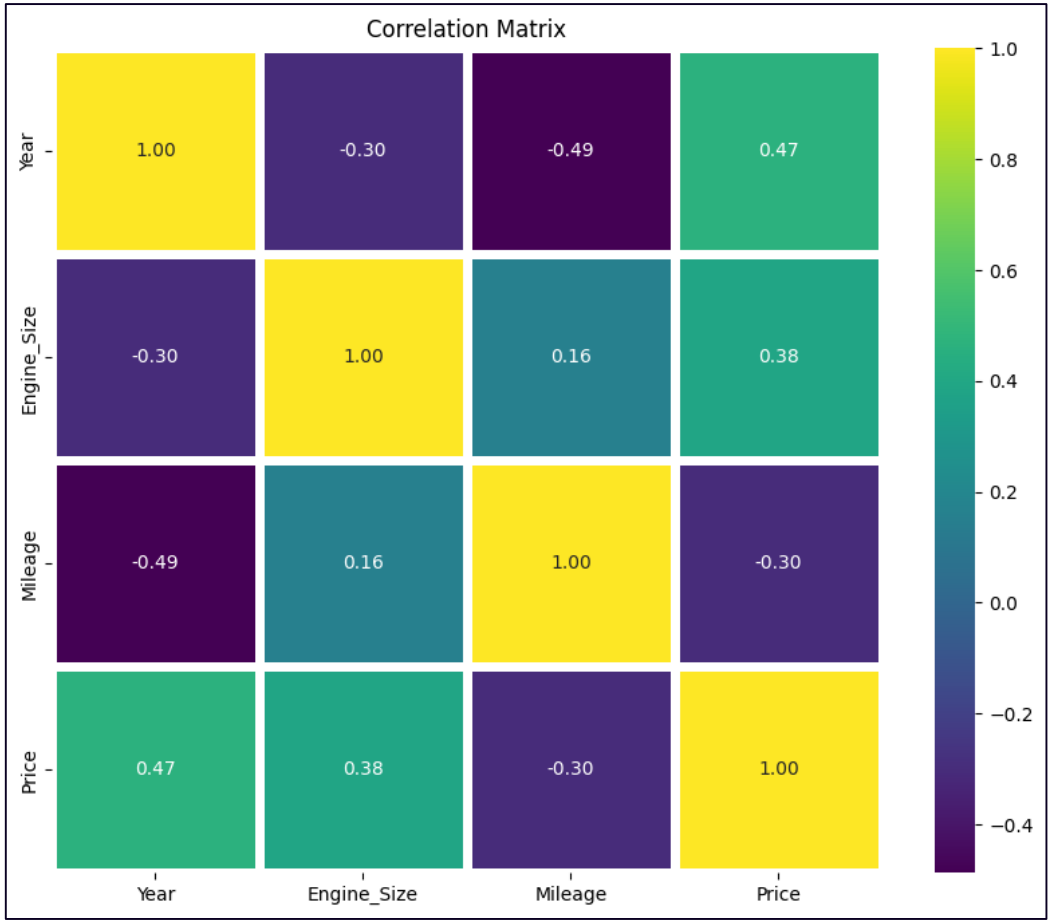


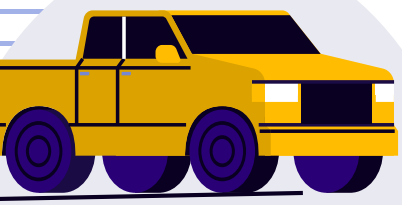




Exploratory Data Analysis (EDA)

3. Data Correlation

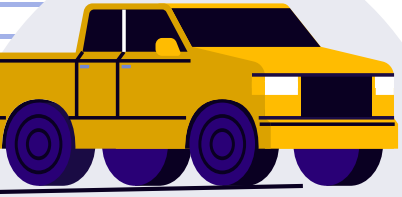




Feature Engineering

1. One Hot Encoding Method: Gear_Type, Origin, dan Options
2. Binary Encoding Method : Type, Region, dan Make

```
# Encode object features
transformer = ColumnTransformer([
    ('One Hot', OneHotEncoder(drop='first'), ['Gear_Type', 'Origin', 'Options']),
    ('Binary', ce.BinaryEncoder(), ['Type', 'Make','Region'])
], remainder='passthrough')
```



04.

Modeling and Evaluation

Model Selection

List Model:

1. Linear Regression
2. KNN Regressor
3. Decision Tree Regression
4. Ridge Regression
5. Lasso Regression
6. ElasticNet
7. Random Forest Regression
8. Xtreme Gradient Boosting Regression
9. LightGBM Regressor
10. CatBoostRegressor

Model Selection

	Model	RMSE	MAE	MAPE	R2
9	CatBoost	30615.804191	16112.003766	0.260762	0.821544
8	LGBM	33544.116956	17887.718416	0.276589	0.784668
7	XGBoost	34552.773941	18989.572481	0.308035	0.769885
6	RandomForest Regressor	37715.744390	20026.673022	0.321998	0.732233
2	Decision Tree	53140.095526	26781.190477	0.396052	0.425786
3	Ridge Regression	54903.099137	34027.590192	0.644500	0.429915
4	Lasso Regression	54905.594316	34045.311117	0.644984	0.429845
0	Linear Regression	54905.907278	34047.122213	0.645033	0.429836
5	Elastic Net	58306.366396	35880.937780	0.654986	0.360105
1	KNN	68108.954514	44035.234693	0.796988	0.123551

Berdasarkan nilai RMSE, MAE dan MAPE, **CatBoost** memiliki nilai paling rendah.

Hyperparameter Tuning

Hyper Parameter Tuner **: RandomizedSearchCV**

```
[143] # Hyperparameter tuning using RandomizedSearchCV
      param_dist = {
          'regressor__iterations': [500, 1000, 1500],
          'regressor__learning_rate': [0.01, 0.05, 0.1],
          'regressor__depth': [4, 6, 8],
          'regressor__l2_leaf_reg': [1, 3, 5]
      }

      # Wrap the CatBoostRegressor in a pipeline with transformers
      pipeline_catboost = Pipeline([
          ('transformer', transformer),
          ('regressor', CatBoostRegressor(verbose=0))
      ])

      random_search = RandomizedSearchCV(
          pipeline_catboost,
          param_distributions=param_dist,
          n_iter=10,
          cv=5,
          scoring='neg_root_mean_squared_error',
          verbose=2,
          random_state=42,
          n_jobs=-1
      )

      random_search.fit(X_train, y_train)
```

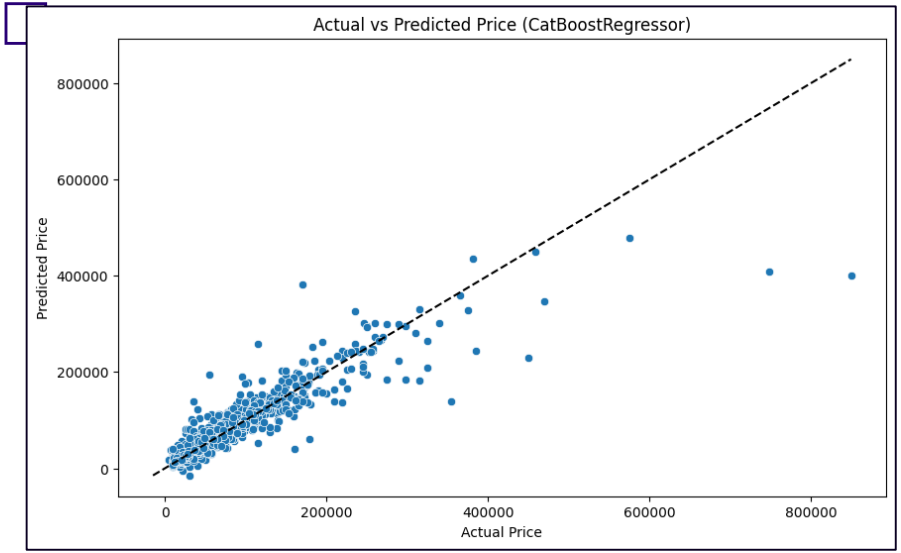
Best Hyperparameters:

1. 'regressor__learning_rate': 0.05,
2. 'regressor__l2_leaf_reg': 5,
3. 'regressor__iterations': 1000,
4. 'regressor__depth': 8

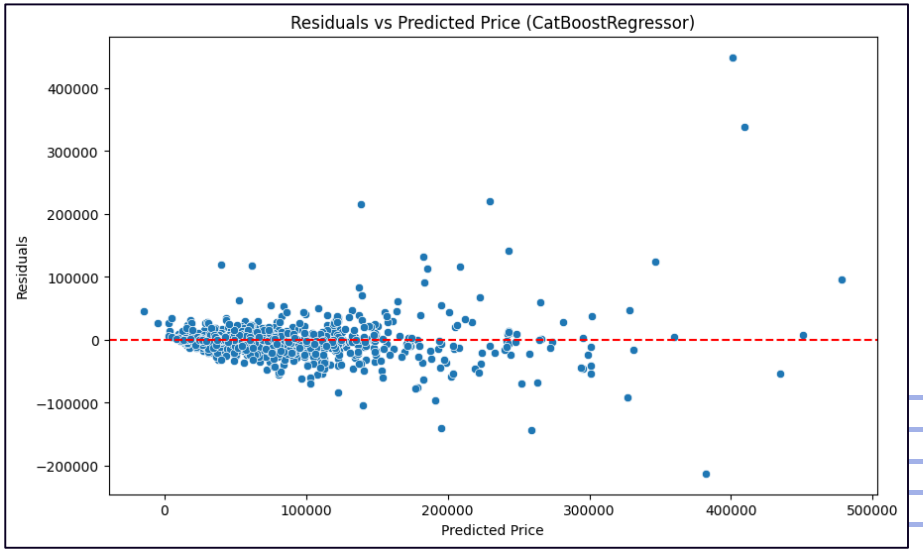
Performance Comparison

Model	RMSE	MAE	MAPE	R2
CatBoost	29929.203443	14676.161698	0.248398	0.822852
CatBoost (Tuned)	30261.867690	14180.460494	0.240108	0.818892

Actual vs Prediction Price (CatBoostRegressor)

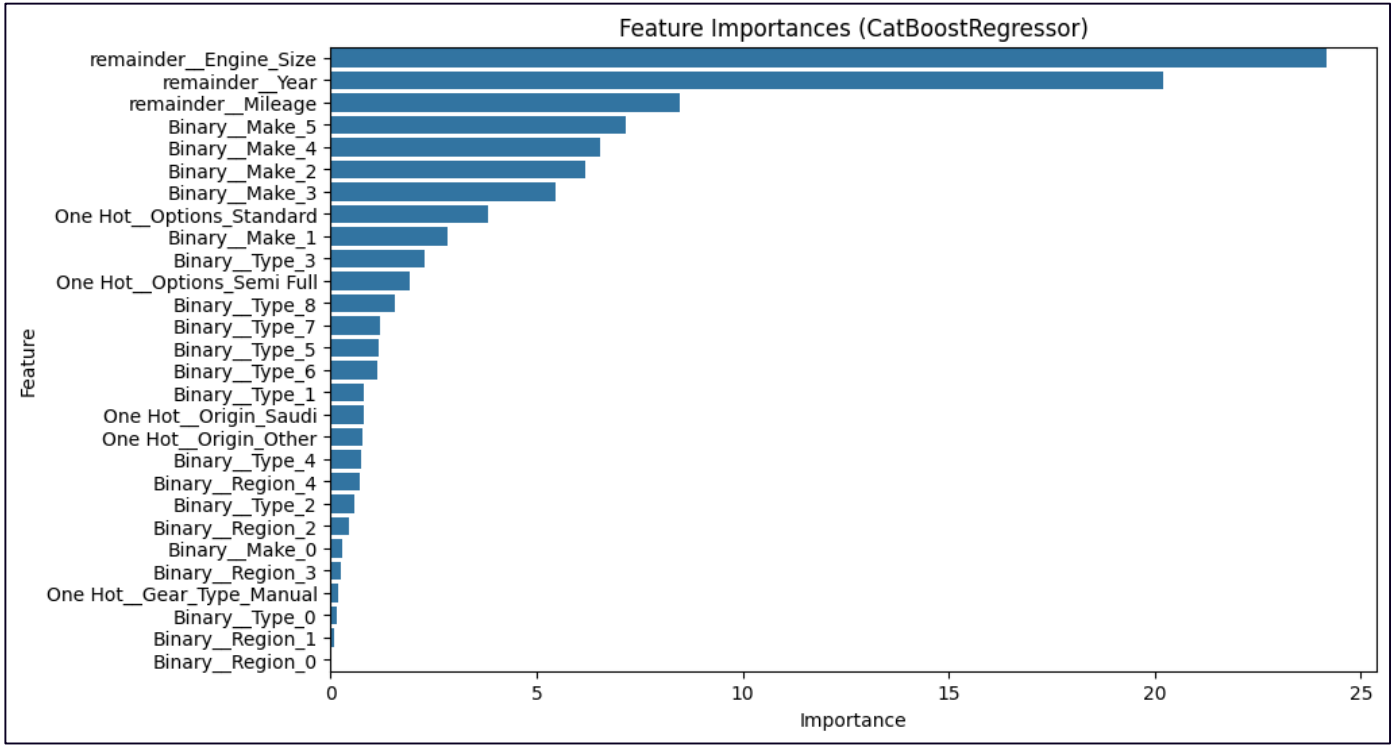


Pola penyebaran titik-titik data mendekati garis diagonal, menunjukkan bahwa prediksi harga secara umum mendekati harga sebenarnya. Semakin dekat titik-titik data dengan garis diagonal, semakin akurat prediksi model.



Berdasarkan grafik Residuals vs Predicted Price dapat diamati bahwa sebagian besar residual tersebar di sekitar garis nol, menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi harga mobil bekas.

Feature Importances



Fitur-fitur yang paling penting dalam memprediksi harga mobil bekas adalah **Year, Make, Engine Size,** dan **Mileage**.

05.

Conclusion and Recommendations

Conclusion

1. Model CatBoostRegressor memberikan performa terbaik dalam memprediksi harga mobil bekas di Arab Saudi, dengan RMSE terendah (29929.203) dan nilai R-squared yang tinggi (0.822) dibadingkan model-model lainnya.

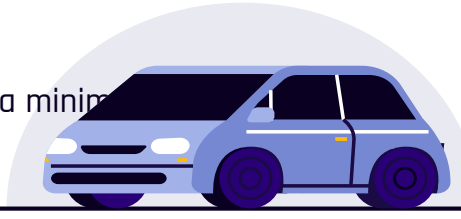
2. Hyperparameter tuning menggunakan RandomizedSearchCV tidak meningkatkan performa model CatBoostRegressor dengan RMSE yang lebih tinggi 30261.867 dan nilai R-squared yang lebih rendah 0.8189.

3. Nilai RMSE memiliki arti bahwa pada saat model digunakan untuk memprediksi harga mobil bekas, maka perkiraan harga rata- ratanya akan berbeda kurang lebih sebesar 29929 SAR dari harga yang seharusnya.

4. Fitur-fitur yang paling penting dalam memprediksi harga mobil bekas adalah **Year, Make, Engine Size**, dan **Mileage**.

5. Model ini dapat membantu bisnis dalam menentukan harga mobil bekas secara lebih kompetitif, meningkatkan pengambilan keputusan, dan meningkatkan pengalaman pelanggan.

6. Model ini memiliki keterbatasan dalam memprediksi harga mobil bekas di luar rentang harga minim (5000 SAR) dan maksimum (850000 SAR) dalam dataset.



Recommendations

1. Eksplorasi Fitur Lebih Lanjut:

- Fitur Kombinasi: Coba buat fitur baru dengan menggabungkan fitur yang ada. Misalnya, gabungkan "Make" dan "Model" untuk membuat fitur "Make_Model" yang lebih spesifik.

- Fitur Polinomial: Pertimbangkan untuk menambahkan fitur polinomial (misalnya, kuadrat atau kubik) dari fitur numerik seperti "Year", "Engine_Size", dan "Mileage" untuk menangkap hubungan non-linear.

2. Teknik Pemodelan Lanjutan:

- Neural Networks: Pertimbangkan untuk menggunakan jaringan saraf tiruan (neural networks) untuk pemodelan, terutama jika dataset memiliki banyak fitur dan hubungan non-linear yang kompleks.

Recommendations

3. Data Lebih Banyak:

- Data Tambahan: Jika memungkinkan, kumpulkan lebih banyak data untuk memperluas dataset dan meningkatkan kemampuan generalisasi model.

4. Validasi Model yang Lebih Kuat:

- Cross-Validation yang Lebih Ekstensif: Gunakan teknik cross-validation yang lebih ekstensif, seperti nested cross-validation, untuk mendapatkan estimasi performa model yang lebih andal.

- Metrik Evaluasi Tambahan: Pertimbangkan untuk menggunakan metrik evaluasi tambahan selain MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Squared Error), dan R^2 (R-squared), untuk mendapatkan gambaran yang lebih lengkap tentang performa model.

Thanks!