Saudi Arabia Used Cars - Regression

Iko Prasetyo Firdaus

Preview

1. **Business Problem Understanding**
2. **Data Understanding**
3. **Data Preprocessing**
4. **Modeling**
5. **Kesimpulan dan Rekomendasi**

Business Problem Understanding

Context and Problem Statement

Alat trasnportasi paling diminati

Saudi Negara penghasil minyak terbesar (cnbcindonesia.com)

Praktik bisnis yang digandrungi

Pencabutan larangan mengemudi untuk wanita

Penentuan harga agar tidak overprice atau underprice

Goals

Tool Machine Learning

Menentukan harga mobil bekas

Business Problem Understanding

Metric Evaluation

- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- R-Square
  - If Model are Linear

Source Dataset

kaggle

https://www.kaggle.com/datasets/turkibintalib/saudi-arabia-used-cars-dataset

Data Understanding

Feature Description

Features	Data Types	Description
Type	object	Tipe Mobil Bekas
Region	object	Wilayah Tempat Mobil Bekas Ditawarkan untuk Dijual
Make	object	Nama Perusahaan Pembuat Mobil
Gear_Type	object	Tipe Transmisi pada Mobil Bekas
Origin	object	Asal Mobil Bekas yang Dijual
Options	object	Pilihan untuk Mobil Bekas
Year	integer	Tahun Pembuatan Mobil
Engine_Size	float	Ukuran Mesin dari Mobil Bekas
Mileage	integer	Jarak Tempuh Mobil Bekas (Miles)
Negotiable	boolean	Negosiasi Harga, jika harganya 0 kemungkinan telah dilakukan negosiasi
Price	integer	Harga Mobil Bekas (SAR)

Shape

Baris	Kolom
5.624	11

Variable

Numerik
Year, Engine_Size, Mileage, Price

Kategorik
Type, Region, Make, Gear_Type, Origin, Options, Negotiable

Data Preprocessing

Missing Values

No missing values

Duplicates

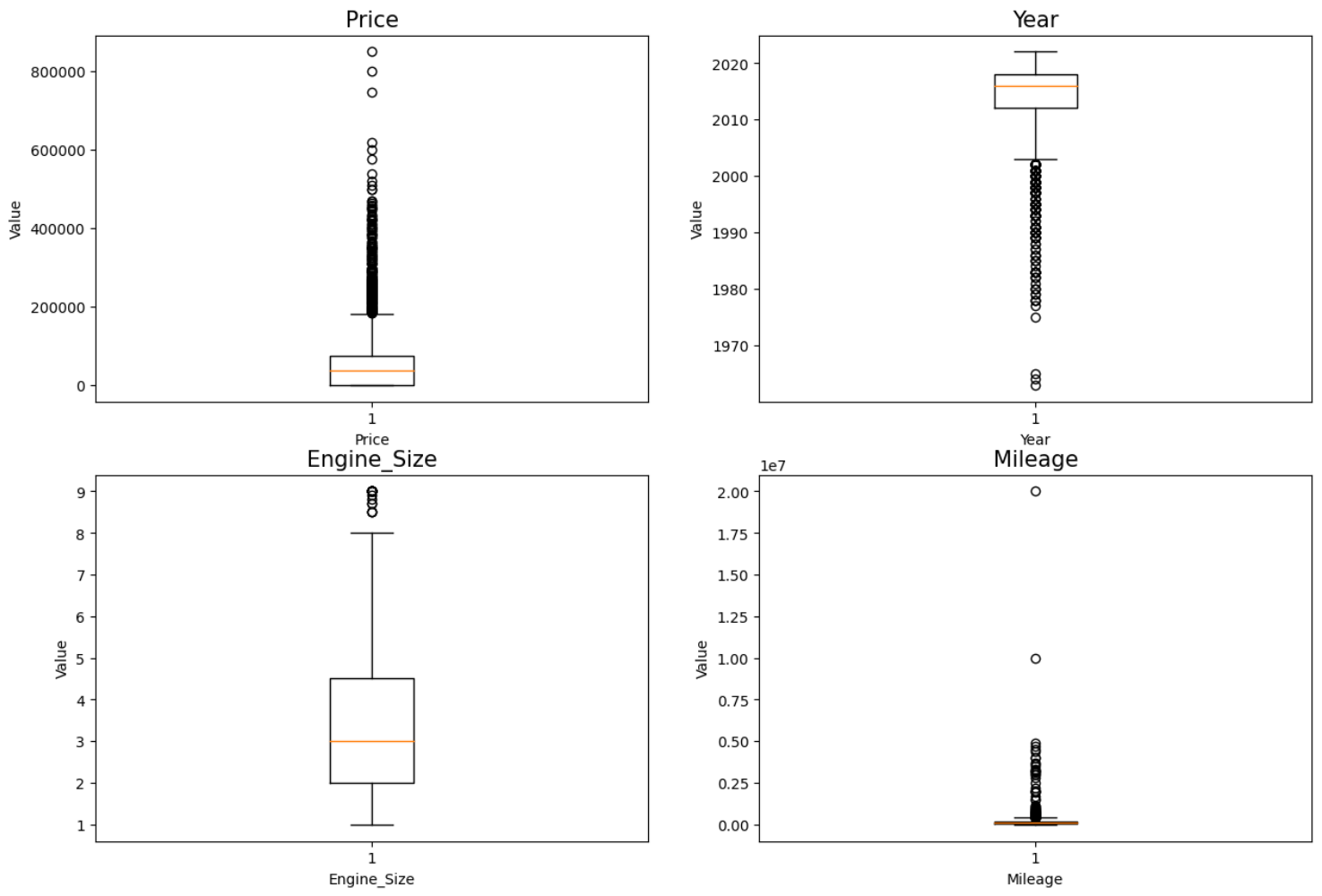
4 data duplicated

Feature Selection

drop = 'Negotiation', 'Origin'

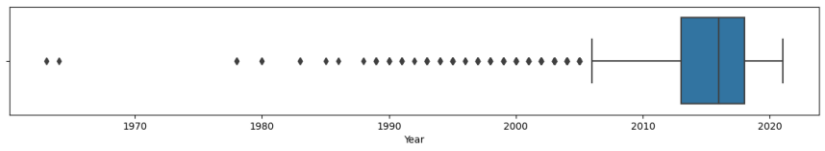
Dependent variable = 'Price'

Outliers (Boxplot Analization)



Data Preprocessing (Penanganan *Outlier*)

Year

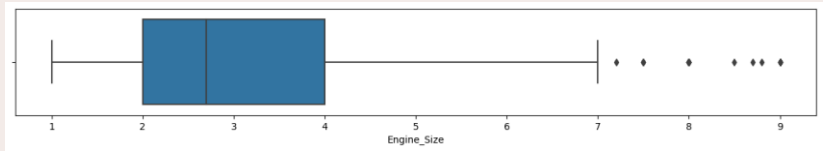


IQR : 5.0
Batas Bawah :2005.5
Batas Atas :2020.5

Adjusted

Batas Bawah :2000
Batas Atas :2021

Engine_Size

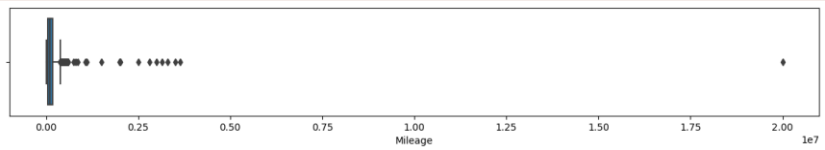


IQR : 2.0
Batas Bawah : -1.0
Batas Atas : 5.0

Adjusted

Batas Bawah : 1.0
Batas Atas : 7.0

Mileage

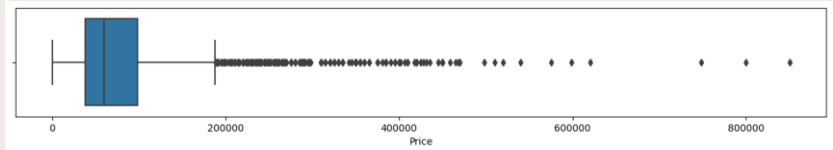


IQR : 132000.0
Batas Bawah :-149000.0
Batas Atas :247000.0

Adjusted

Batas Bawah : 100
Batas Atas : 328.064

Price



IQR : 60000.0
Batas Bawah :-52000.0
Batas Atas :128000.0

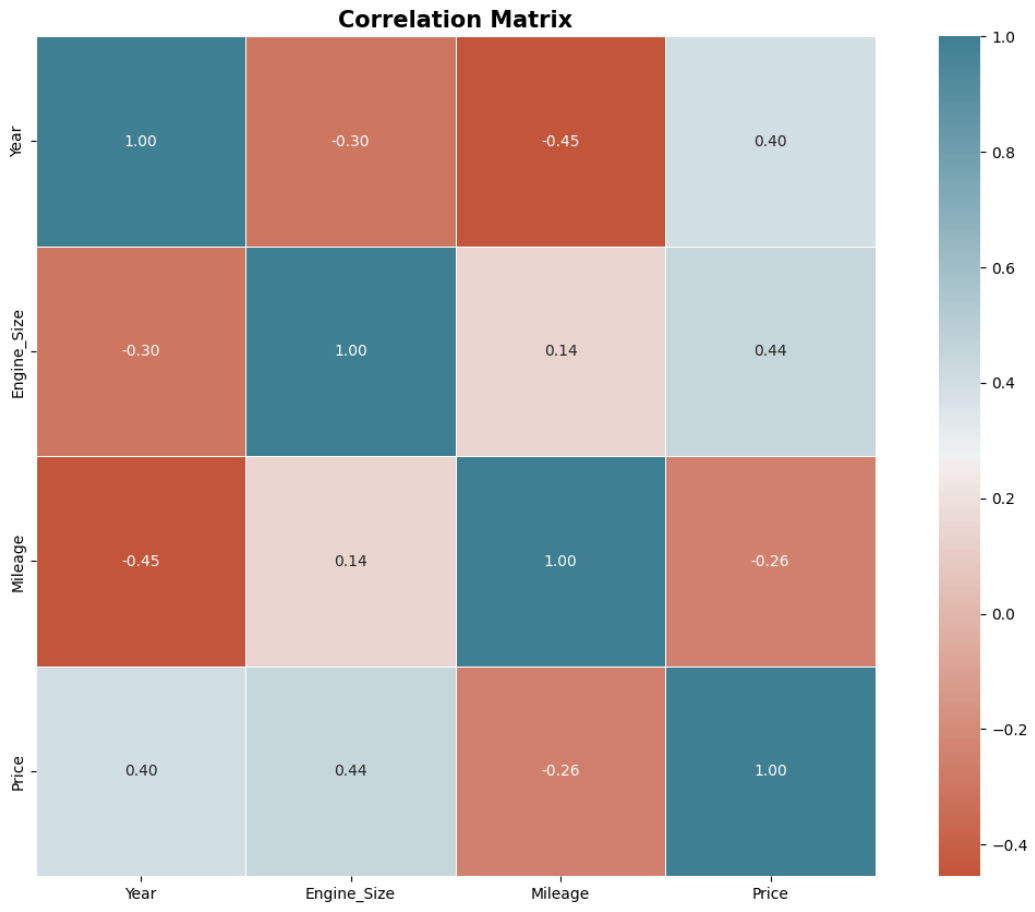
Adjusted

Batas Bawah : 4.000
Batas Atas : 450.000

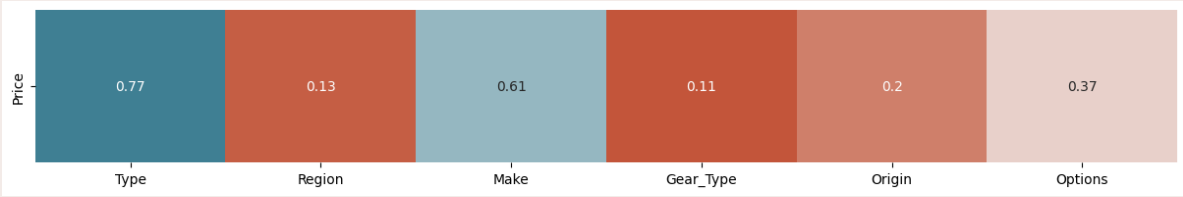
Data Preprocessing

Correlation Matrix

Numerical



Categorical



https://en.wikipedia.org/wiki/Provinces_of_Saudi_Arabia

Origin	Gulf Arabic	Other	Saudi
Region			
Abha	NaN	5.0	21.0
Al-Ahsa	13.0	24.0	59.0
Al-Baha	2.0	2.0	11.0
Al-Jouf	1.0	1.0	9.0
Al-Medina	7.0	12.0	87.0
Al-Namas	NaN	NaN	9.0
Arar	NaN	NaN	5.0
Aseer	4.0	9.0	59.0
Besha	NaN	1.0	2.0
Dammam	96.0	108.0	493.0

Modeling

Feature Engineering

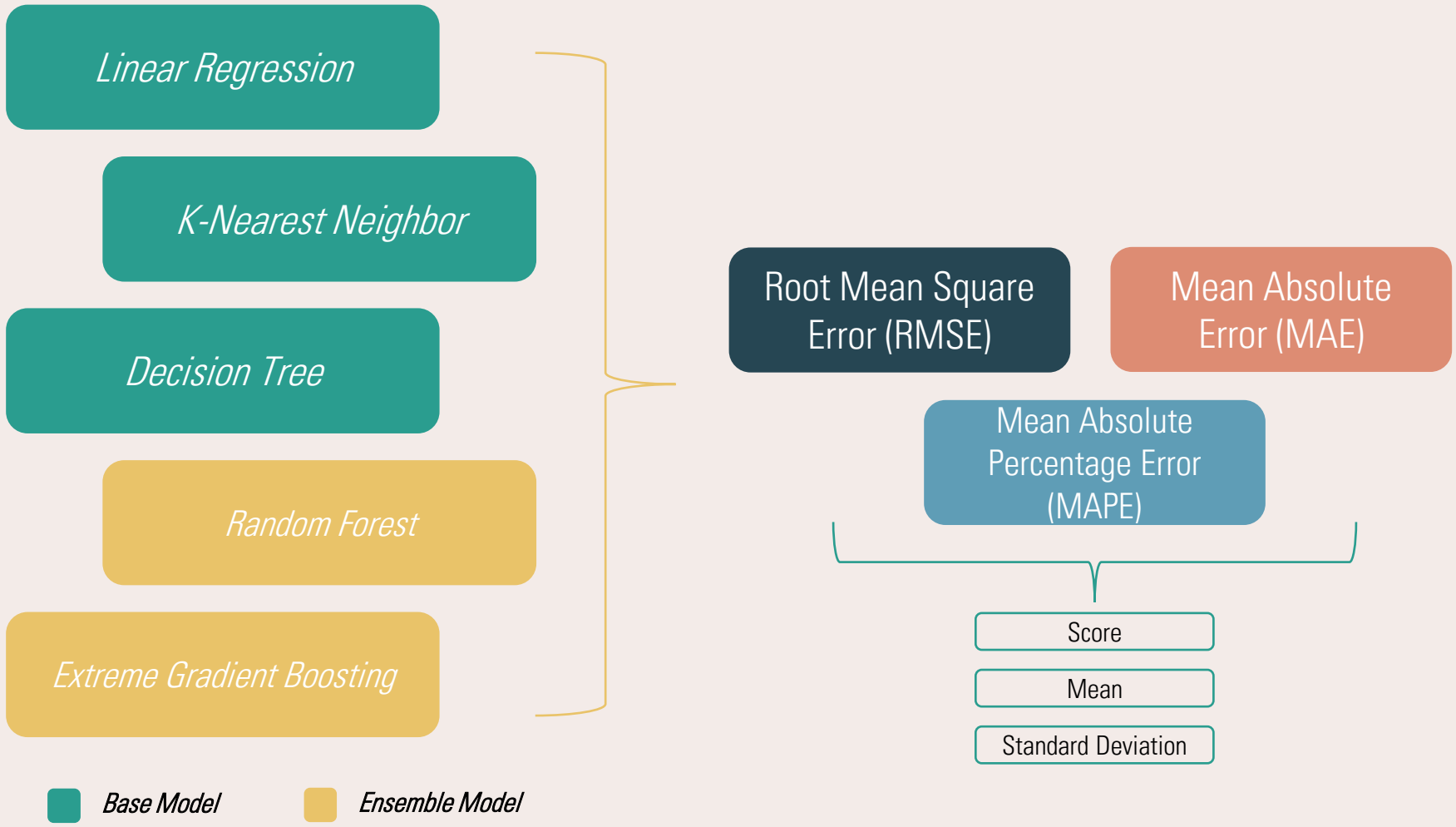
Encoding

Ordinal : 'Options'
OneHot : 'Gear_Type'
Binary : 'Make', 'Type', 'Region'

Data Split

Train 70 : Test 30

Algorythm



Modeling

Hasil test berdasarkan *Metric Evaluation*

	Model	mean_RMSE	std_RMSE	mean_MAE	std_mae	mean_MAPE	std_MAPE
0	Linear Regression	-40209.487144	3300.310750	-23707.257917	1573.701810	-0.330591	0.021244
1	KNN Regressor	-36162.149806	3678.665933	-21423.404825	1071.337848	-0.366653	0.027542
2	DecisionTree Regressor	-42228.522846	4857.356874	-22519.561988	1882.308074	-0.322669	0.021026
3	RandomForest Regressor	-29863.221545	3155.212263	-15993.895655	1295.146143	-0.225362	0.012010
4	XGBoost Regressor	-26847.315937	2360.421298	-14707.201019	736.651498	-0.202105	0.007541

Hasil test set pada *Benchmark* Model

	RMSE	MAE	MAPE
XGB	29640.909155	16738.998197	0.235723
RandomForest	30850.704176	16820.361607	0.238544

Modeling

Hyperparameter Tuning

https://xgboost.readthedocs.io/en/latest/parameter.html#general-parameters

Param_distribution

max_depth	= 1 – 10
learning_rate	= 0.1 – 0.99
n_estimators	= 100 – 200
subsample	= 0.2 – 0.9
gamma	= 1 – 10
colsample_bytree	= 0.1 – 0.9

Hyperparameter tuning

estimator	= XGBRegression
param_distribution	= param_distribution
n_iter	= 50
cv	= crossval
scoring	= ['neg_root_mean_squared_error', 'neg_mean_absolute_error', 'neg_mean_absolute_percentage_error'],
n_jobs	= -1
Refit	= 'neg_root_mean_squared_error',
random_state	= 42

Skor Parameter Terbaik

XGBoost	
Best_score	-27673.83
Best_params	
- model__subsample	0.4
- model__reg_alpha	1.29
- model__n_estimators	100
- model__max_depth	7
- model__learning_rate	0.06
- model__gamma	1
- model__colsample_bytree	0.7

Perbandingan model sebelum dan setelah tuning

	RMSE	MAE	MAPE
XGB	29640.909155	16738.998197	0.235723

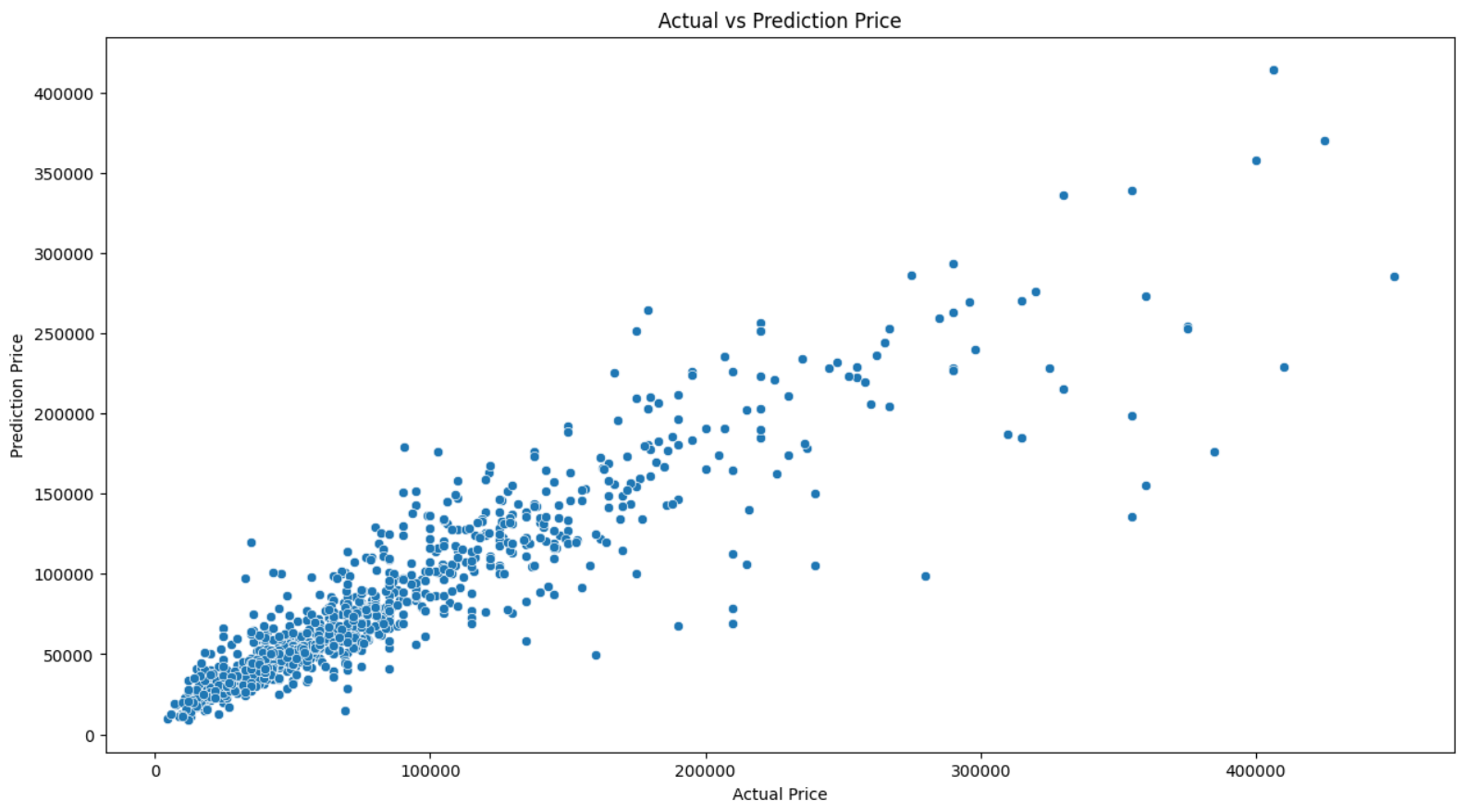
Before

	RMSE	MAE	MAPE
XGB	28357.391903	15703.403648	0.220568

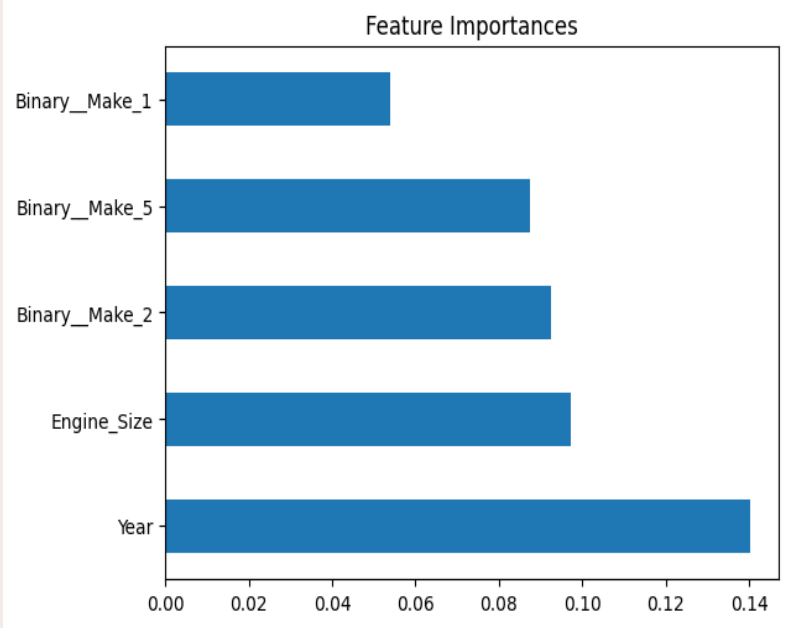
After

Modeling

Analisa Hasil Prediksi Model



Feature Importance



Kesimpulan

1. Dataset terlalu banyak nilai ‘0’ yg berkorelasi dengan fitur Negotiable. Dataset bersih hanya sebesar 60.84%.
2. Korelasi valid antara Price dengan Variabel Kategorikal. Fitur yang paling berpengaruh berdasarkan Feature Impoertance adalah Year.
3. Nilai MAPE dari model XGBoost adalah sebesar 22.05% dengan limitasi harga pada kisaran 4.000 – 450.000 SAR.
4. Perlu adanya beberapa fitur tambahan

Rekomendasi

1. Penambahan fitur-fitur yang relevan pada dataset dan perbaikan entry nilai negosiasi yang kemungkinan akan bisa dimanfaatkan apabila nilai akhirnya bukanlah ‘0’.
2. Melakukan User Testing untuk mendapatkan feedback, dan dari feedback tersebut bisa ditambahkan untuk fitur baru yang relevan seperti warna mobil.

Terima Kasih

Iko Prasetyo

prasetyonico@gmail.com