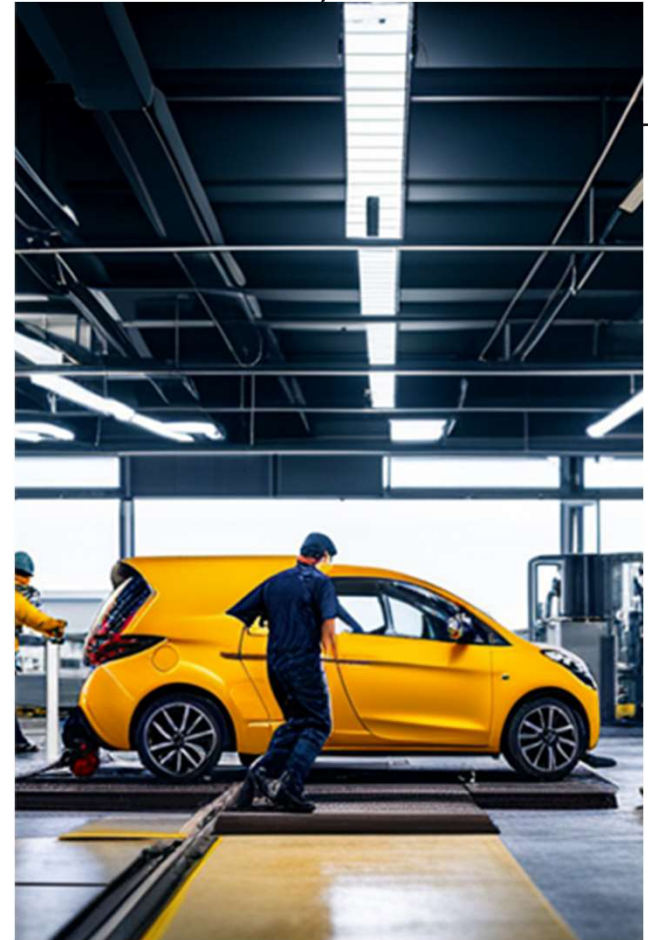


CAPSTONE PROJECT MODULE 3

SAUDI ARABIA USED CAR MACHINE LEARNING

RIZKI NUGRAHA - JCDS ONLINE LEARNING





OUTLINE

01 **BUSINESS
PROBLEM**

02 **DATA
UNDERSTANDING**

03 **DATA
PREPROCESSING**

04 **FEATURE
ENGINEERING**

05 **MODELLING**

06 **CONCLUSION &
RECOMENDATION**






BUSINESS PROBLEM

CONTEXT

- Pasar Mobil Bekas di Saudi Arabi telah mengalami pertumbuhan positif selama bertahun-tahun didukung oleh peningkatan populasi milenial di negara tersebut seiring dengan masuknya pemain baru ke pasar dengan growth stage 2500+ dealership.
- Bisnis mobil bekas memiliki persaingan yang ketat, keberhasilan dalam bisnis ini membutuhkan keterampilan dan pengetahuan yang mendalam mengenai mobil.
- Mobil bekas menjadi pilihan yang menarik karena harganya lebih terjangkau dibandingkan mobil baru. Namun, proses pembelian atau penjualan mobil bekas bisa menjadi tugas yang rumit dikarenakan terdapat banyak jenis mobil, terutama saat menentukan harga pasar yang terbaik.



PROBLEM STATEMENT

- Isu utama dalam industri pasar mobil bekas adalah menentukan harga mobil bekas untuk pembeli atau penjual agar harganya tidak terlalu tinggi untuk konsumen serta tidak terlalu rendah untuk penjual.
 - Dalam hal ini, diperlukan sebuah model yang dapat memberikan perkiraan harga mobil bekas berdasarkan data historis yang ada.
 - Model tersebut diharapkan memberikan kemudahan bagi pembeli dengan memberikan informasi tentang kualitas mobil, baik itu dalam kondisi yang sangat baik, baik, atau buruk. Di sisi lain, model ini juga dapat membantu para pelaku bisnis mobil bekas dalam menentukan harga jual yang kompetitif.
- 



BUSINESS PROBLEM



GOALS

Membuat sebuah model untuk dapat prediksi harga mobil bekas yang dapat diandalkan bagi pengguna yang ingin membeli atau menjual mobil berdasarkan rincian seperti merk mobil, tahun, jarak tempuh, dsb.

Sehingga jika seseorang ingin menjual mobilnya, kita dapat memberikan perkiraan harga berdasarkan tren pasar dan sesuai dengan spesifikasi mobil untuk membantu pengguna mengatasi kesulitan dalam menentukan harga yang kompetitif.

ANALYTIC APPROACH

Kita perlu melakukan analisis data untuk dapat menemukan pola dari fitur-fitur yang ada, yang membedakan satu fitur dengan yang lainnya.

Selanjutnya, kita akan membangun suatu model regresi yang akan membantu pengguna untuk mendapatkan sebuah alat prediksi harga mobil bekas yang kompetitif dan bertujuan untuk menghindari harga mobil bekas yang *overprice* atau *underprice*.





BUSINESS PROBLEM



EVALUATION METRIC



RMSE (Root of Mean Squared Error) - nilai rataan akar kuadrat dari error



MSE (Mean Squared Error) - rataan nilai absolut dari error



MAPE (Mean Absolute Percentage Error) - rataan persentase error yang dihasilkan oleh model regresi



R-Squared - digunakan untuk mengetahui seberapa baik model dapat merepresentasikan varians keseluruhan data





DATA UNDERSTANDING

Attribute Information

Attributes	Data Type, Length	Description
Type	Text	Brand Name of Used Car
Region	Text	The region in which the used car was offered for sale
Make	Text	The Company Name
Gear_Type	Text	Gear Type (AT/MT)
Origin	Text	Origin of Used Car (Saudi, Gulf, Other)
Option	Text	Full Options / Semi-Full / Standard
Year	Int	Year of Manufacturing
Engine_Size	Float	The engine size of used car
Mileage	Int	Mileage of Used Car
Negotiable	Bool	If True, the price is 0. This means the price is negotiable (Not Used)
Price	Int	Used Car Price (Riyal)



DATA UNDERSTANDING

DATASET

```
Jumlah baris dan kolom di dataset df adalah (5624, 11)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5624 entries, 0 to 5623
Data columns (total 11 columns):
#   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
dtypes: bool(1), float64(1), int64(3), object(6)
memory usage: 445.0+ KB
```

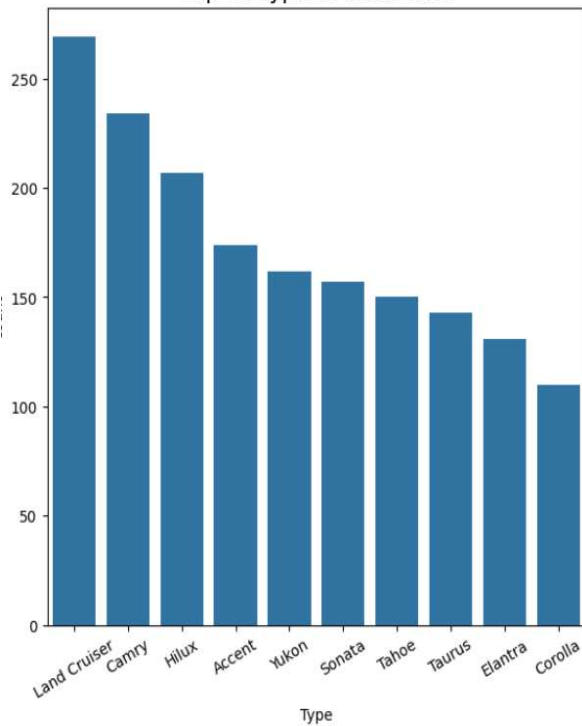


```
<class 'pandas.core.frame.DataFrame'>
Index: 3734 entries, 3513 to 5521
Data columns (total 10 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Type         3734 non-null   object
1   Region       3734 non-null   object
2   Make         3734 non-null   object
3   Gear_Type    3734 non-null   object
4   Origin       3734 non-null   object
5   Options      3734 non-null   object
6   Year         3734 non-null   int64
7   Engine_Size  3734 non-null   float64
8   Mileage      3734 non-null   int64
9   Price        3734 non-null   int64
dtypes: float64(1), int64(3), object(6)
memory usage: 320.9+ KB
```

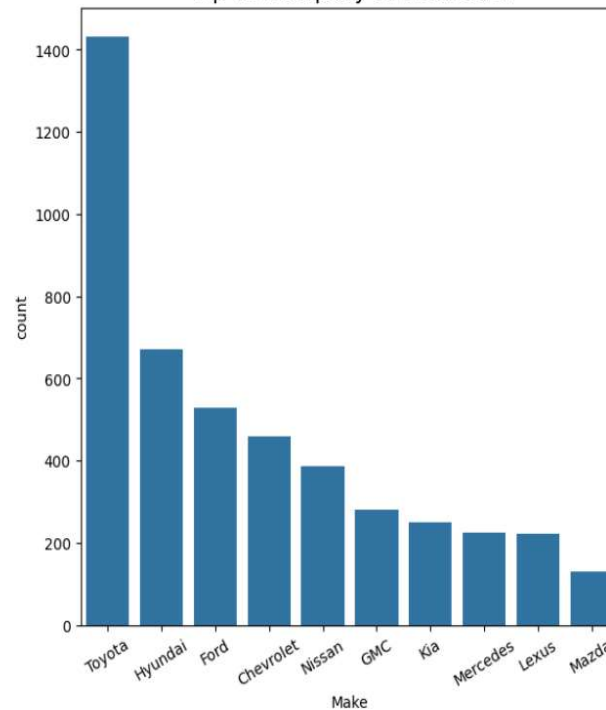
Data yang ada saat ini setelah kita lakukan cleaning data menjadi 3734 atau 66,55% dari data awal

DATA UNDERSTANDING

Top 10 Type of Used Cars



Top 10 Company of Used Cars

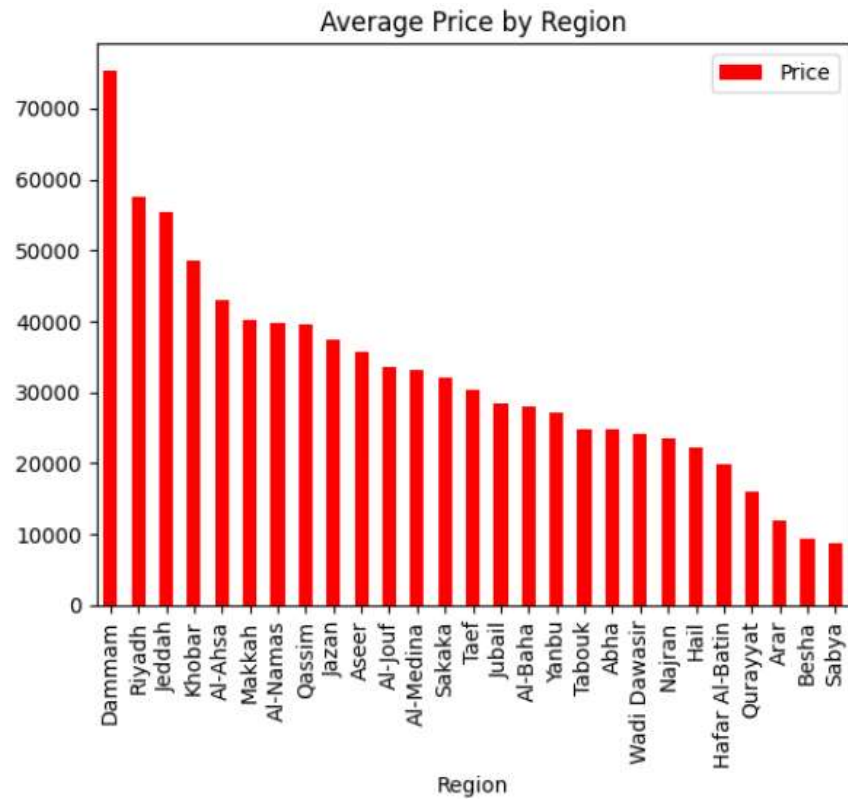


Insight:

Dari grafik di atas terlihat bahwa tipe mobil bekas yang tertinggi adalah berasal dari produsen mobil Toyota (Land Cruiser, Camry, Hilux, etc) yang mana juga merupakan produsen mobil bekas yang paling banyak dibandingkan dengan perusahaan lainnya. Kemudian, yang kedua tipe mobil bekas berasal dari produsen yang Hyundai (Accent, Sonata, dan Elantra) menempati urutan kedua tertinggi.

Jadi, berdasarkan grafik ini, kita dapat melihat adanya korelasi antara Categorical Variable seperti Merek mobil dan Tipe mobil.

DATA UNDERSTANDING



Insight:

Rata-rata harga mobil jika dilihat dari region terbanyak dari Dammam, Riyadh, Jeddah, dst.



DATA UNDERSTANDING

```
Type & Amount of Negotiable on Used Cars :
Negotiable
False      3828
True       1796
Name: count, dtype: int64
```

```
Type & Amount of Gear Type on Used Cars :
Gear_Type
Automatic   4875
Manual       749
Name: count, dtype: int64
```

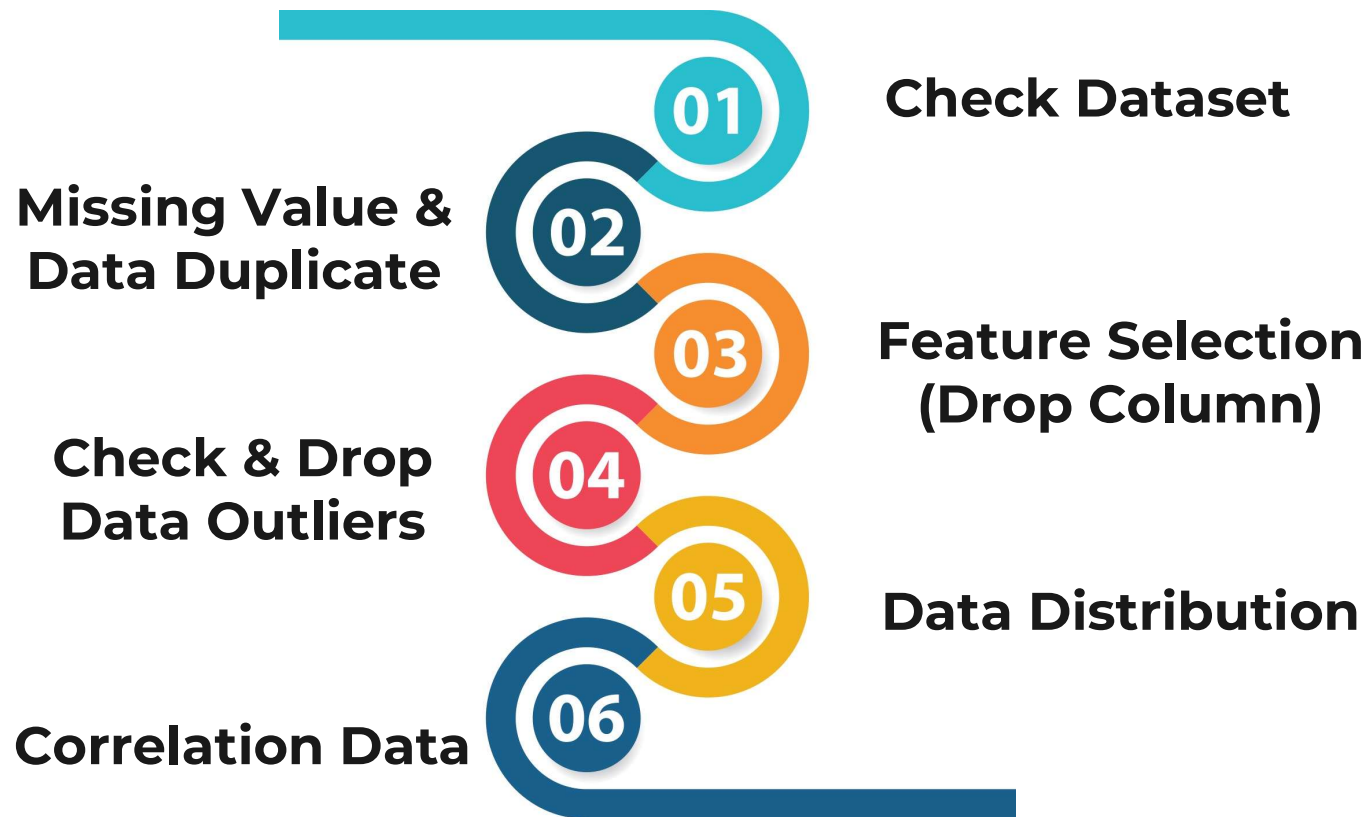
```
Type & Amount of Options on Used Cars :
Options
Full         2233
Standard     1822
Semi Full    1569
Name: count, dtype: int64
```

Insight:

- Kebanyakan untuk harga mobil yang ada sudah harga pasti dan tidak dapat dinegoisalkan.
- Mobil-mobil bekas yang ada didominasi oleh mobil bertransmisi Automatic
- Terdapat tiga jenis Opsi pada Mobil Bekas, yaitu Full (2233), Standar (1822), dan Semi Full (1569).



DATA PREPROCESSING





MISSING VALUE



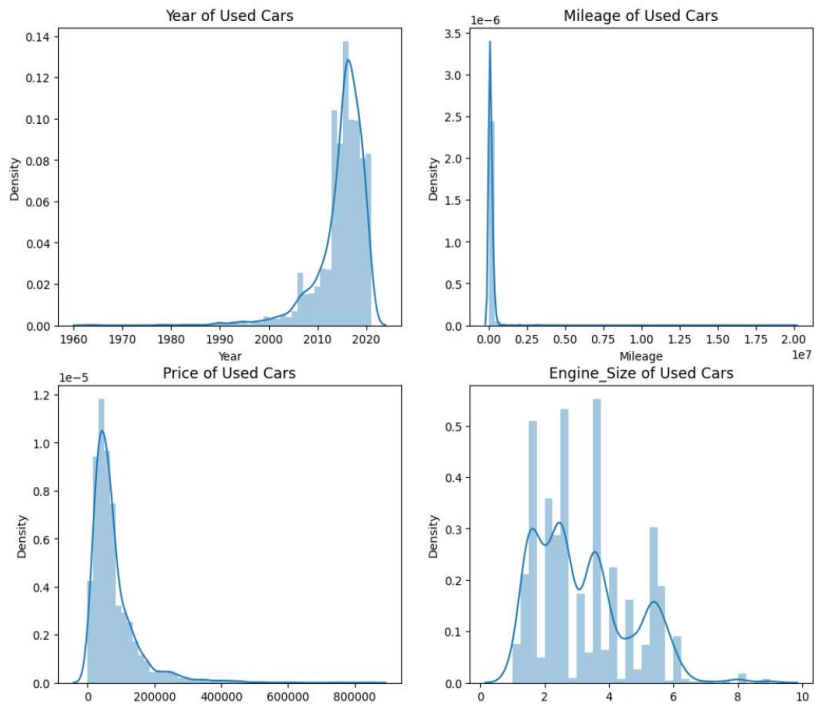
Insight:

Tidak terdapat *missing value* pada data.

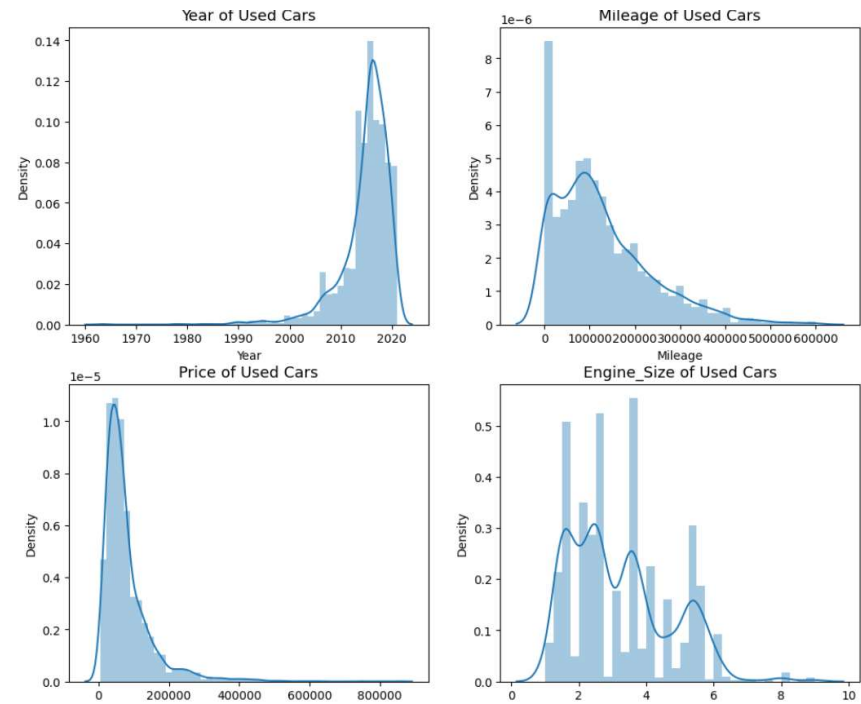




DATA OUTLIERS

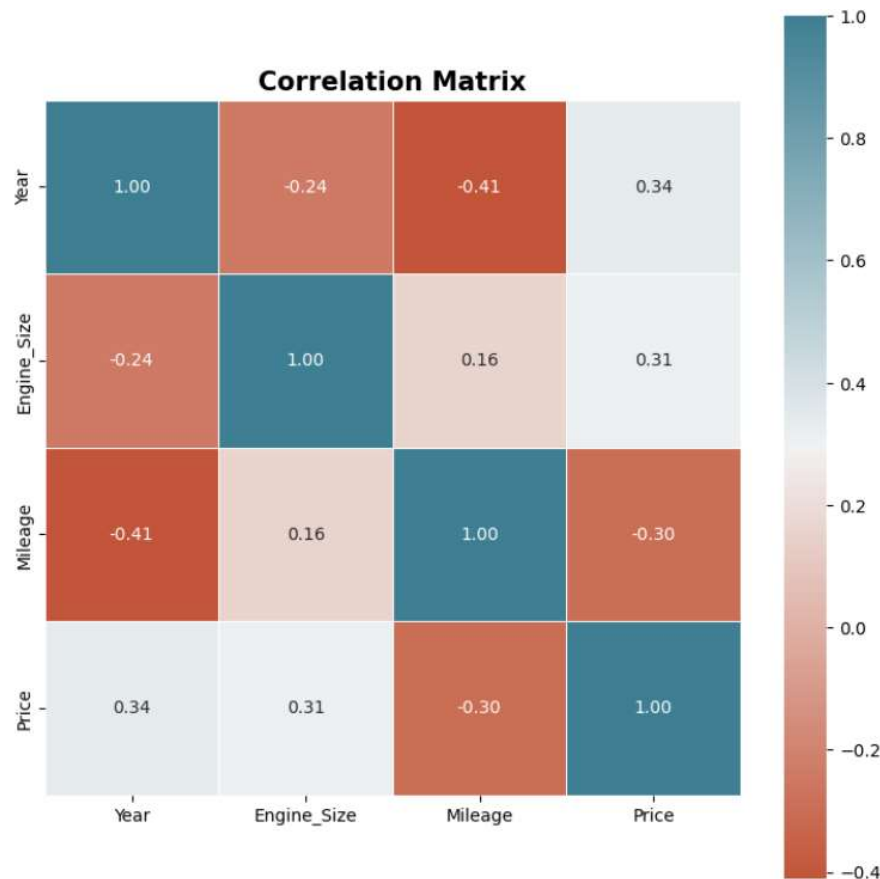


Before Cleaning



After Cleaning

DATA CORRELATION



Insight:

- Bisa kita lihat pada Correlation Matrix Heatmap di atas, terdapat high correlation antara Year dan Mileage dimana Mileage akan meningkat dari tahun ke tahun sesuai pemakaian kendaraan.
- Kita juga bisa lihat Price bergantung pada Year dan Engine Size.



FEATURE ENGINEERING



ENCODING

One-Hot Encoding: (variabel Gear_Type, Origin, dan Options)

Binary Encoding: variabel Type, Region, dan Make.

TRAIN & SPLITTING



Splitting data into training and test with propotion 70:30

MODELLING

Base Model:

1. Linear Regression
2. KNN Regressor
3. Decision Tree Regressor
4. Rigde Regression
5. Lasso Regression
6. Elastic Net

Ensemble Model:

1. Random Forest Regressor
 2. Ada Boost Regressor
 3. Gradient Boosting Regressor
 4. Xtreme Gradient Boosting Regessor
- 
- 



MODELLING

TEST RESULT

	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	Mean_MSLE	Std_MSLE	Mean_R2	Std_R2
0	Linear Regression	-48913.7	6240.4	-25865.1	2217.5	-0.4	0.0	-0.2	0.0	0.5	0.1
1	KNN Regressor	-45798.3	4442.5	-25198.2	1179.5	-0.4	0.0	-0.2	0.0	0.6	0.0
2	DecisionTree Regressor	-50591.9	5050.9	-23964.3	1860.6	-0.4	0.0	-0.2	0.0	0.5	0.1
3	Rigde Regression	-31318.9	4214.7	-15556.5	1296.2	-0.2	0.0	-0.1	0.0	0.8	0.0
4	Lasso Regression	-48946.3	6249.9	-25850.1	2214.5	-0.4	0.0	-0.2	0.0	0.5	0.1
5	Elastic Net	-48932.9	6245.2	-25858.4	2216.0	-0.4	0.0	-0.2	0.0	0.5	0.1
6	RandomForest Regressor	-35442.0	4753.1	-16757.8	1099.3	-0.2	0.0	-0.1	0.0	0.8	0.0
7	AdaBoost Regressor	-48486.7	4996.4	-27379.1	1583.7	-0.4	0.0	-0.2	0.0	0.5	0.1
8	Gradient Boosting Regressor	-37899.0	4271.6	-19032.9	1121.9	-0.3	0.0	-0.1	0.0	0.7	0.0
9	XGBoost Regressor	-32641.4	3576.2	-15792.8	502.3	-0.2	0.0	-0.1	0.0	0.8	0.0

INSIGHT

- Dalam pemilihan kandidat model yang akan digunakan kita dapat melihat pada nilai RMSE, MAE dan MAPE paling rendah, dan memiliki nilai R2 yang paling tinggi
- Dengan demikian Rigde Regression, RandomForest Regressor dan XGBoost Regressor akan dilakukan benchmark serta akan dilakukan prediksi menggunakan test set untuk kedua model tersebut



MODELLING

TEST RESULT


	RMSE	MAE	MAPE	R2
XGB	36270.7	18600.9	0.3	0.8
RandomForest	38506.8	19476.6	0.3	0.7
Ridge Regression	36871.3	19976.3	0.4	0.8



INSIGHT

Dari hasil tes di atas menunjukkan bahwa prediksi menggunakan **XGBoost** memiliki performa yang lebih baik dengan memiliki nilai yang lebih rendah dibanding RandomForest dan Ridge Regression.

Model yang akan digunakan adalah **XGBoost**






MODELLING

HYPERPARAMETER TUNING

	Before Hyperparameter Tuning	After Hyperparameter Tuning	Change in %
RMSE	36270.7	35639.0	1.7
MAE	18600.9	17467.6	6.1
MAPE	0.3	0.3	4.7
R2	0.8	0.8	-1.0

INSIGHT

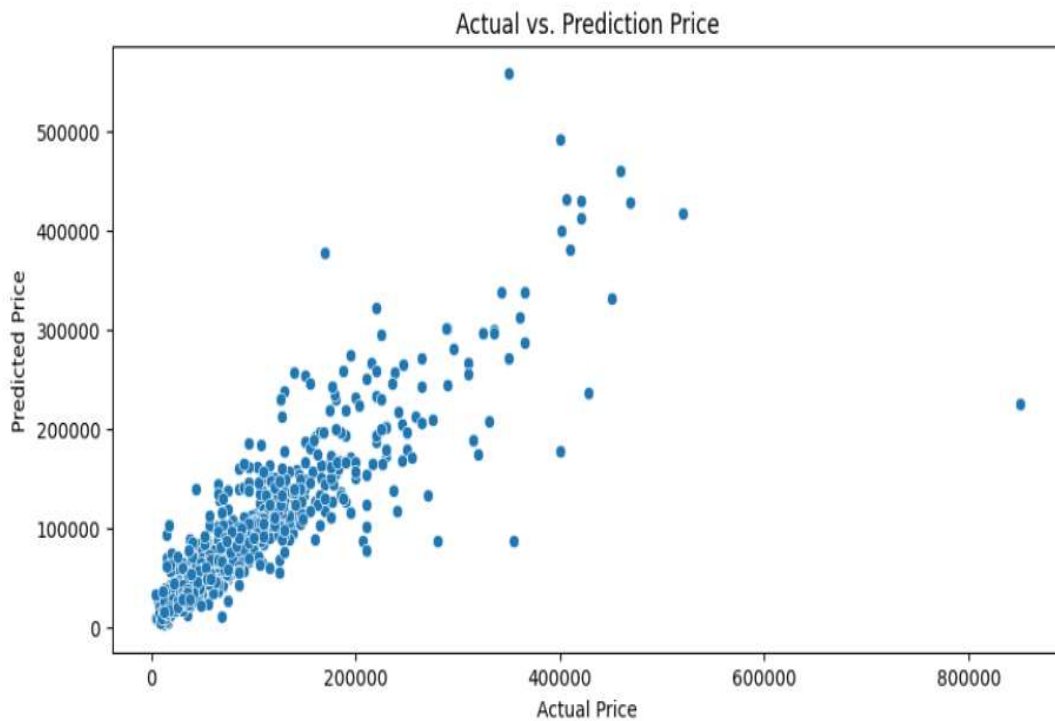
Dari tabel di atas kita bisa melihat, komparasi analisa sebelum dan sesudah dilakukan tuning, dimana nilai RMSE,MAE,MAPE mengalami perbaikan sebesar 1%-6%, namun nilai R2 naik 1%. Kedepannya hasil dari Hyperparameter Tuning dapat digunakan dalam model karena menghasilkan nilai lebih baik.



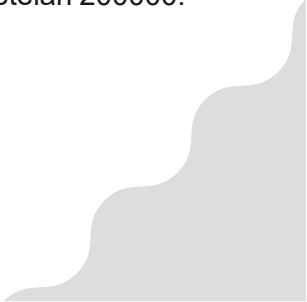


MODELLING

PREDICTION RESULTS (GRAPHIC)



INSIGHT

- Model yang dihasilkan tergolong cukup baik dan linear, dimana terlihat dari grafik di atas serta menghasilkan nilai small error metric values, (RMSE, MAE, MAPE) dan nilai R2 hampir mendekati 1 (0.8)
 - Beberapa nilai actual price yang rendah di atas jika dibandingkan dengan nilai predicted price mengalami error, hal ini disebabkan karena banyaknya data dengan nilai **Price** yang rendah.
 - Ketika Price more than 200000, distribusi plot mulai irregular. Kita dapat melihat bahwa kadang-kadang kita mempunyai harga yang diprediksi tinggi dan kadang-kadang rendah setelah 200000.
- 

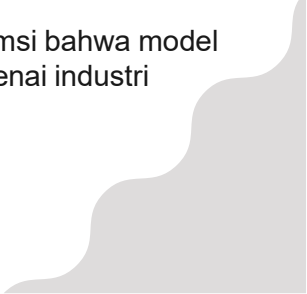



CONCLUSION & RECOMENDATION

Conclusion:

- Hasil dari model menunjukkan bahwa fitur-fitur yang paling signifikan pengaruhnya adalah Make, Year dan Engine Size.
- Performa model regresi dievaluasi menggunakan metrik RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) dan R-Squared . Setelah melalui proses hyperparameter tuning, model yang dihasilkan tergolong cukup baik dan linear (XGBoost) dimana menghasilkan nilai small error metric values,(RMSE, MAE, MAPE) dan nilai R2 hampir mendekati 1 (0.8).
- Nilai RMSE memiliki makna bahwa ketika model digunakan untuk memprediksi harga mobil bekas, perkiraan harga rata-ratanya dapat memiliki selisih sekitar 35.639 Riyal (RMSE) atau 17.467 (MAE) dari harga actual. Sedangkan nilai MAPE yang dihasilkan 0.3 yang menunjukkan error absolut pada Price yang diprediksi oleh model. Semakin kecil nilai MAPE berarti nilai taksiran semakin mendekati nilai sebenarnya.
- Sedangkan nilai R2 sebesar 0.8 berarti hubungan antara variabel dan variabel dependen dengan variabel independen sebesar 80%. Nilai R2 yang tinggi menunjukkan bahwa variabel independen mempunyai pengaruh yang besar terhadap variabel dependen.
- Ketika Price more than 200000, distribusi plot mulai irregular. Kita dapat melihat bahwa kadang-kadang kita mempunyai harga yang diprediksi tinggi dan kadang-kadang rendah setelah 200000.

Model yang diperoleh masih memiliki potensi untuk ditingkatkan melalui proses-proses tertentu. Namun, untuk saat ini, kami berasumsi bahwa model sudah mencapai hasil yang diharapkan. Selain itu, dalam proses pembuatan model, diperlukan pengetahuan yang mendalam mengenai industri mobil untuk dapat mengembangkan pengembangan model yang lebih baik lagi.





CONCLUSION & RECOMENDATION

Recommendation:

- Mengecek prediksi mana saja yang memiliki nilai error yang tinggi, kita dapat mengelompokkan error tersebut ke dalam grup overestimation dan underestimation. Lalu kita bisa mengecek hubungan antara error tersebut dengan tiap variabel independen. Pada akhirnya kita dapat mengetahui sebenarnya variabel mana saja dan aspek apa yang menyebabkan model menghasilkan error yang tinggi, sehingga kita bisa melakukan training ulang.
- Menambahkan fitur/variabel yang mengkategorikan jenis mobil (Sedan, SUV, dst / mobil klasik non klasik), warna mobil, dan lainnya. Dengan memasukkan fitur ini ke dalam model, kemungkinan besar akan meningkatkan akurasi prediksi harga mobil bekas.
- Jika tersedia tambahan data yang signifikan, dapat mencoba menggunakan model yang lebih kompleks.
- Analisis kolinearitas dengan menghitung nilai VIF pada setiap fitur. Diharapkan dapat diperoleh model yang lebih baik dengan menghilangkan fitur-fitur yang berkorelasi.





Thanks!