سيارة

Predict Used Car Prices in Saudi Arabia with Machine Learning on Syarah.com

Ghaisan Rabbani
13/01/2025

AGENDA

01 Business problem

02 Data understanding

03 Data preprocessing

04 Modeling

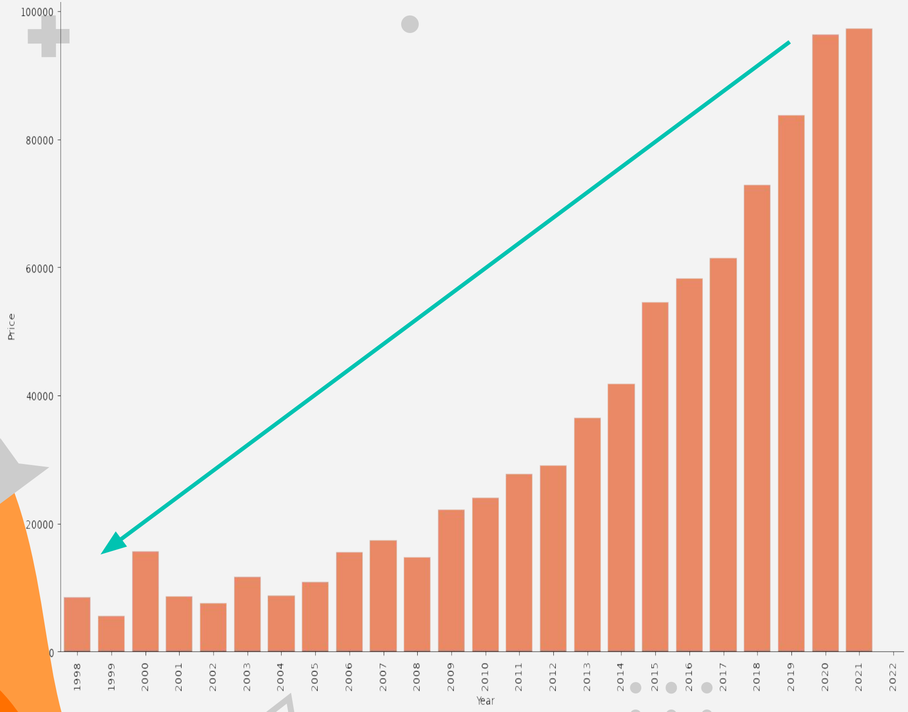
05 Conclusion

06 Recommendation

BUSINESS PROBLEM !

- Context Business
- Goals
- Success Criteria

Price depreciation is a significant factor, as car values naturally decrease over time



Price is also influenced by other features

Condition

Mileage

difficult to get the right price

Over Price

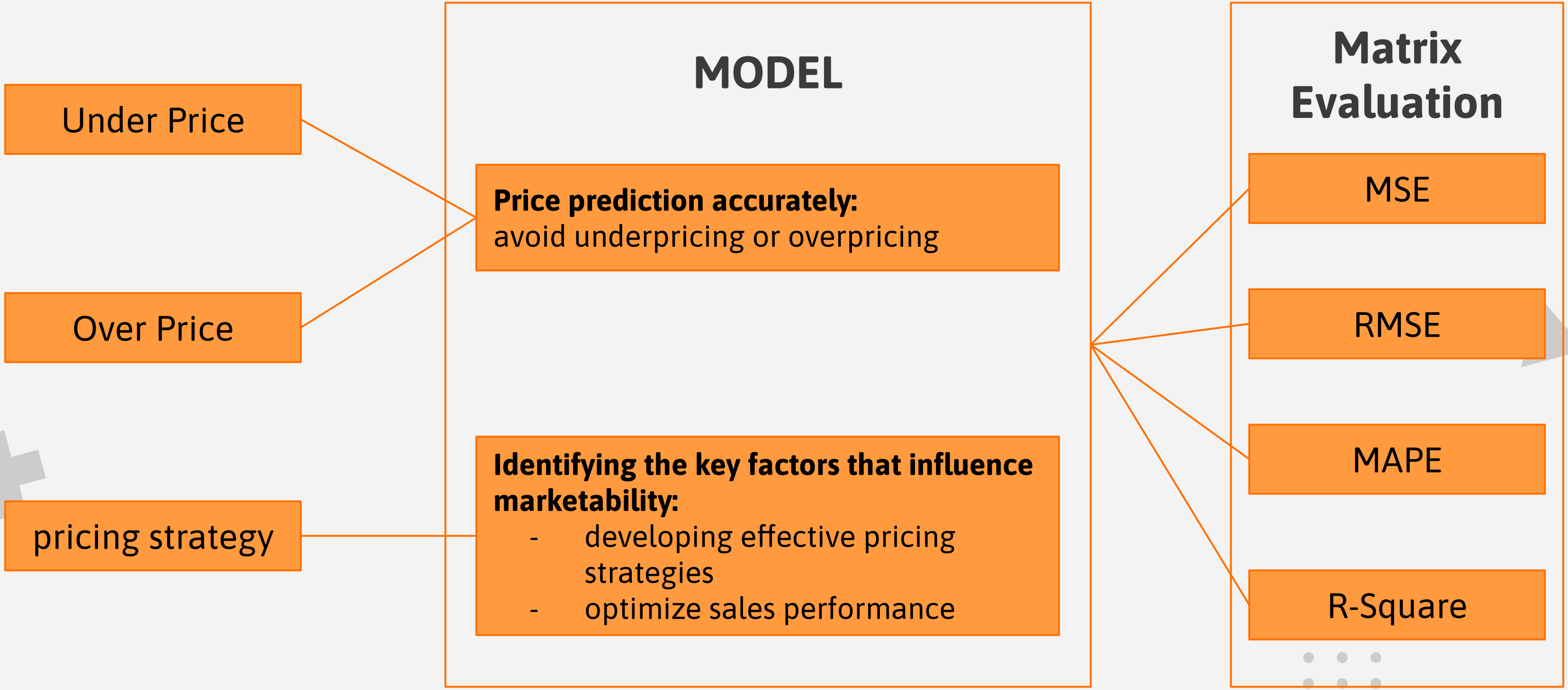
- slower sales
- increased storage costs
- further value loss

Under Price

- missed revenue opportunities

Solve problem with machine learning from predict price

Building a Model to solve the problem: price prediction and identification of influencing factors



Under Price

Over Price

pricing strategy

MODEL

Price prediction accurately:
avoid underpricing or overpricing

Identifying the key factors that influence marketability:

- developing effective pricing strategies
- optimize sales performance

Matrix Evaluation

MSE

RMSE

MAPE

R-Square

GOALS

System Prediction

Develop an Accurate and Competitive Price Prediction System

Strategy Optimization

Offer premium features to increase profits

Market Transparency

Providing clear information on factors influencing car prices

Growth Marketplace

accelerate transactions and support business growth.

Data Understanding !

Overview dataset used car in saudi arabia's

OVERVIEW DATA

Total Data:
5.624 Data

Description :
Saudi arabia's used car

Sumber :
syarah.com

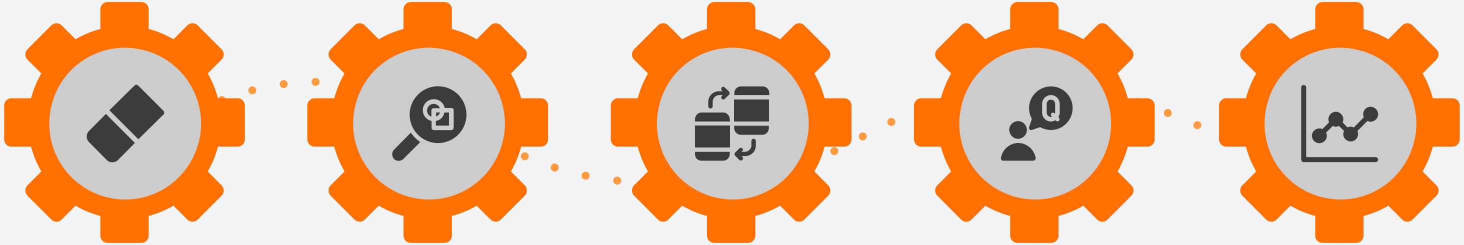
Feature	Description	Impact to Business
Type	Type of used car	Determining the types that are popular in market
Region	The region in which the used car was offered for sale	Understanding sales trends by location
Make	The company name	Find out the most popular car brands
Gear_Type	Gear type size of used car	Determining customer preference (auto or manual)
***Origin**	Origin of used car	Helps determine customers are interested in (local)
***Options**	Options of used car	Determining car values that can increase the price
Year	Manufacturing years	year of production affects for price and demand
Engine_Size	The engine size of used car	Determine buyer interest based on vehicle needs
Mileage	Mileage of used car	The kilometers traveled affect the selling price
Negotiable	True if the price is 0, that means it is negotiable	Demonstrate pricing flexibility
Price	Price used cars	important factors for buyers

Target

Data Preprocessing

Data cleaning and feature engineering before used in model

Data Cleaning



Feature Selection

The "Negotiable" column and data with "Price=0" were removed

Missing Values

all missing values will be removed

Duplicate Values

all Duplicated data will be removed

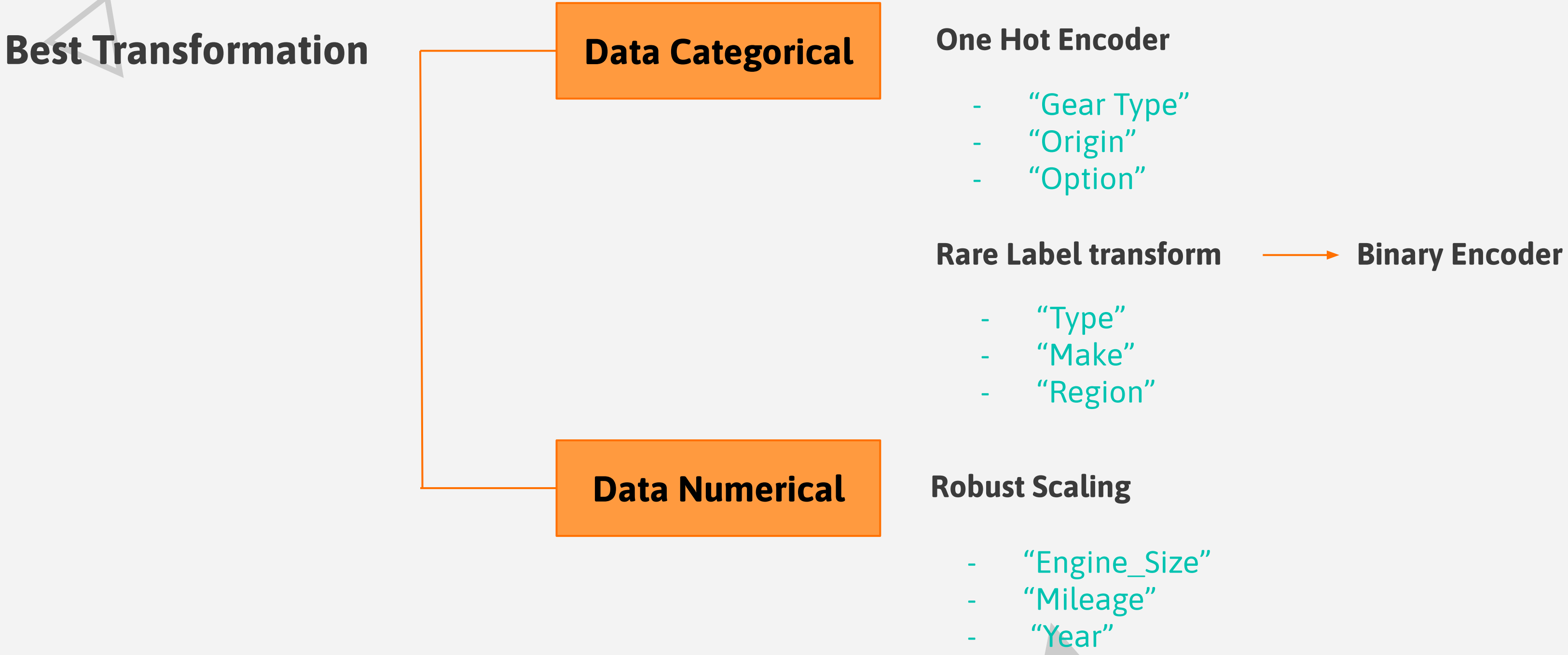
Spelling Error

"Origin" column needs to be improved. The "Unknown" value can be replaced with "Other"

Handling Outlier

- “Year” < 2000
- “Engine_Size” >= 8
- “Mileage” >= 600000
- “Price” < 5000

Feature Engineering



Best Transformation

Data Categorical

One Hot Encoder

- "Gear Type"
- "Origin"
- "Option"

Rare Label transform → Binary Encoder

- "Type"
- "Make"
- "Region"

Data Numerical

Robust Scaling

- "Engine_Size"
- "Mileage"
- "Year"

Modeling

Create the best model for predict used cars price

Model Benchmarking

Algorithm

- Decision Tree
- K- Nearest Neighbor
- Linear Regression
- Random Forest
- XGB
- Gradient Boosting
- ADA Boost

Transformation Target for Each Algorithm

very good to use for regression

Result Model

Name	Mean RMSE	STD RMSE	Mean MAE	STD MAE	Mean MAPE	STD MAPE	Mean R2	STD R2
XGBOOST	-34894.87	7458.09	-17121.38	1603.23	-0.22	0.02	0.77	0.06
Random Forest	-38706.96	8354.69	-17902.61	1521.11	-0.23	0.02	0.77	0.07
Gradient Boosting	-39508.94	7508.73	-19832.47	1639.88	-0.24	0.01	0.71	0.06
KNN	-40133.90	7564.80	-19736.78	1647.40	-0.28	0.02	0.70	0.06
Decision Tree	-48986.24	7739.486	-26114.97	1779.74	-0.33	0.01	0.56	0.07
Ada Boots	-50486.97	7505.82	-28563.18	2379.68	-0.38	0.03	0.53	0.05
Linear Regression	-56153.03	5765.47	-25652.30	1693.17	-0.35	0.02	0.41	0.10

Best Model

Tuning parameter

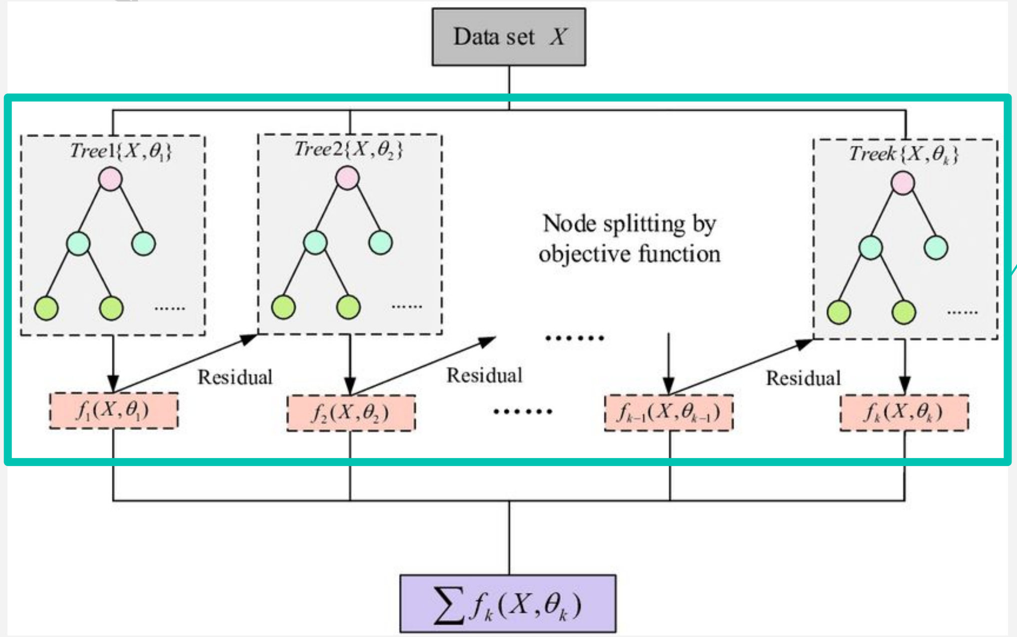
	RMSE	MAE	MAPE	R2
Before Tuning	28445.21	14624.80	0.195	0.831
After Tuning	27210.94	14244.15	0.188	0.845

model produces better results after the tuning process with parameters

- Gamma = 0
- Learning rate = 0.1
- Max depth = 5
- N estimator = 500
- Reg alpha = 0.1
- Col sample by tree = 0.5
- Subsample = 0.9

What is XGBOOST ??

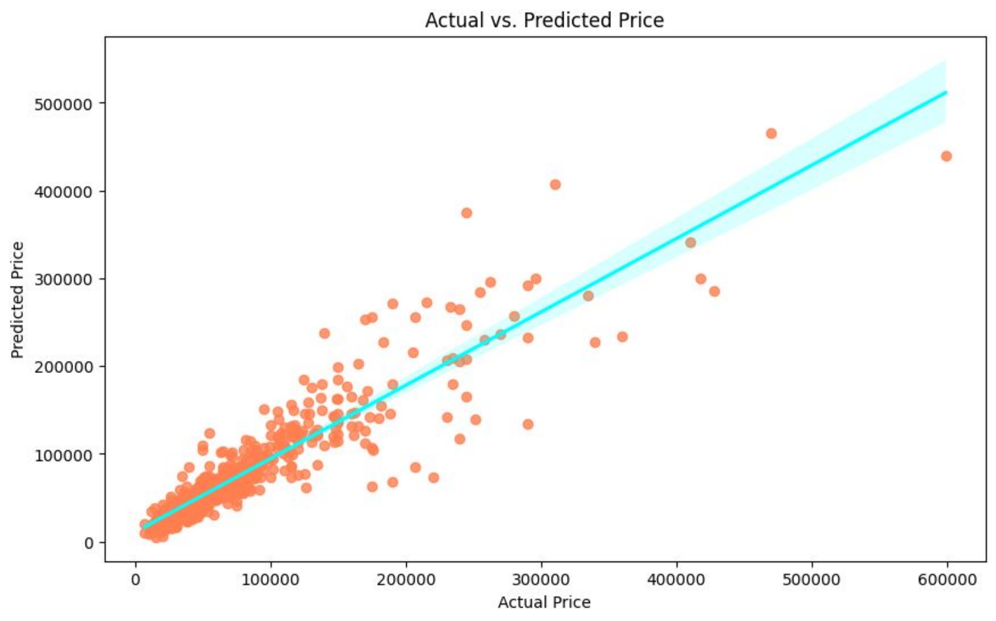
The XGBoost model is one of the techniques in machine learning designed to make highly **accurate predictions.**



Process XG BOOST

- XGBoost works by correcting the mistakes made by previous experts.
- Every step of the way, XGBoost strengthens these predictions so that the end result is highly accurate.

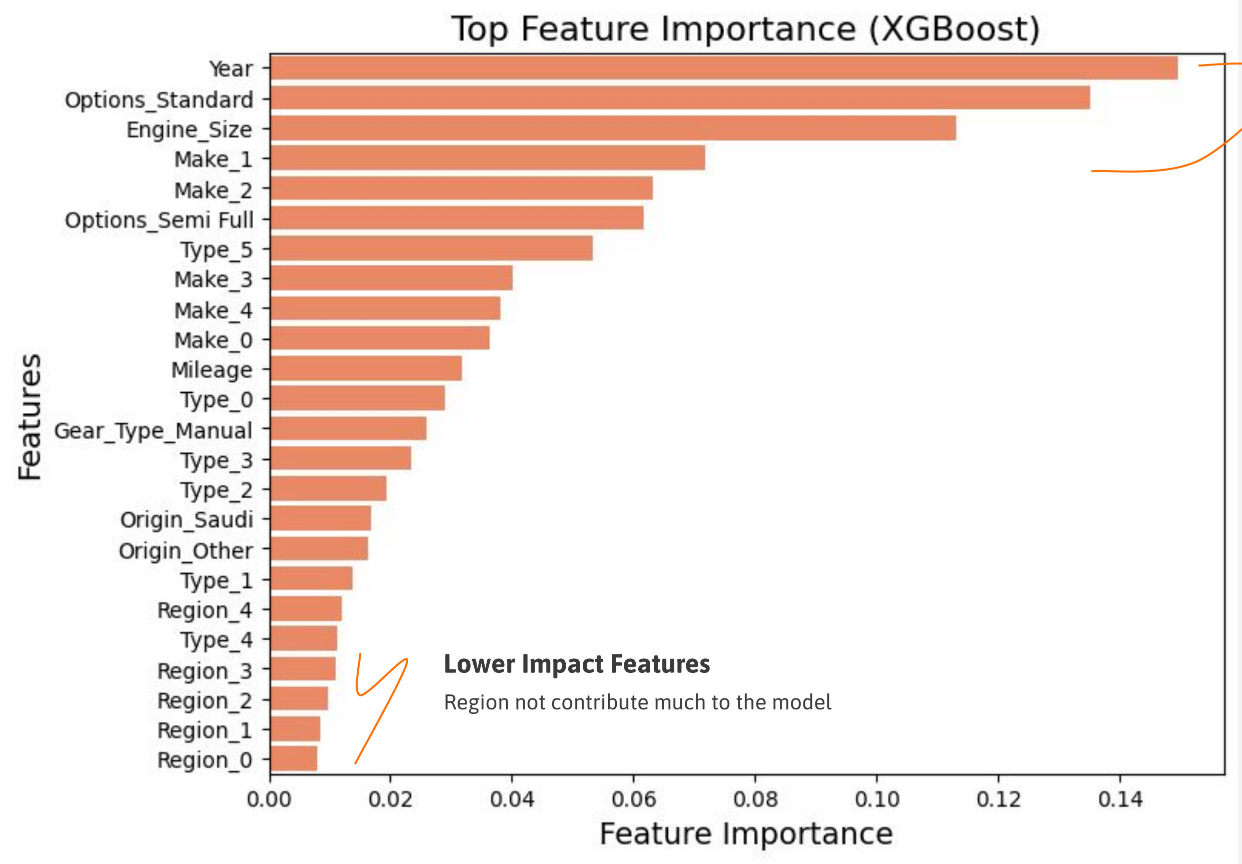
Assessing the Performance of Used Car Price Prediction Model



- **The plot demonstrates that the model performs well overall**, with predictions closely matching actual values for most cases.

- However, there is **room for improvement in handling higher actual prices** and reducing outliers.

the most influential feature in the model



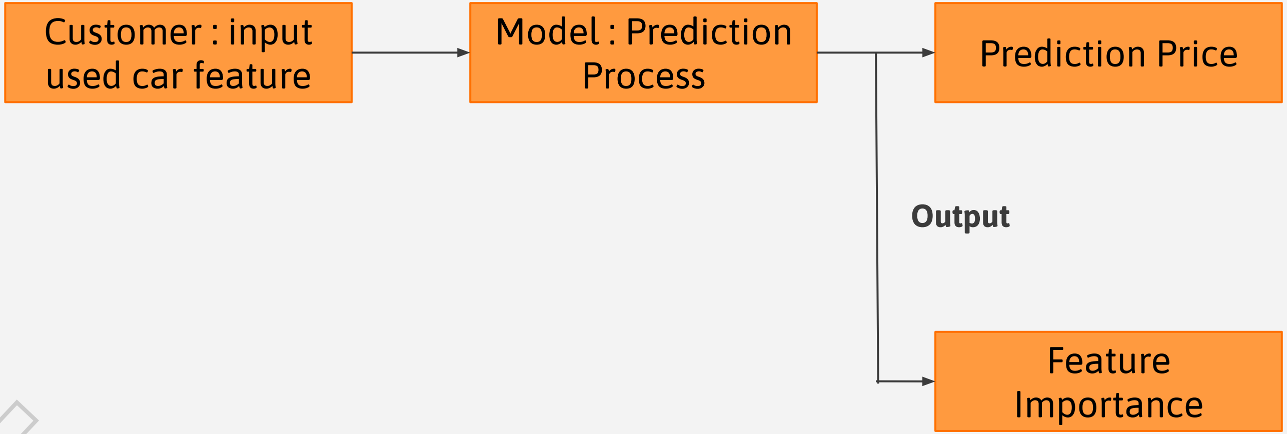
Lower Impact Features

Region not contribute much to the model

Top Features Drive Predictions

indicating that focusing on these features is crucial for the accuracy of the model.

How to use the model ?



strategies to increase profits :

Prediction Price

Free : generate price ranges to be able to estimate prices

Premium : display the model as accurately as possible

Feature Importance

Free : provides features that affect the price

Premium : provides recommended strategies for selling used cars

Conclusion and Recommendation

Conclusion for model and business and recommendation

Conclusion Model

1. Model produces the best results in new data

MAE Target ≤ 5%

R-Square ≥ 0.80

2. features that have the most influence on the predicted price

Standard Option

Year

Brand

Engine Size

3. Limitation Model

Price > SAR 5000

Mileage < 600000

Year > 2000

Engine Size = 8000 CC

max profit: the advantage that can be obtained from using machine learning

Premium Revenue Model for Paid Services

- Premium SAR 50
- Annual SAR 500

728 data

- 30 % premium member
- 10 % annual members

Development Machine learning

$13900.00 - \boxed{8000} =$ **SAR 5900**

Result Predict

total lost profits of all cheap car sellers (under predict) : **SAR 1125619.50**

Proportion total hard to sell cars (over predict) : **0.5**

Recommendation

Important

improve the model to produce better results and update the model regularly.

Middle

Optimize Premium Pricing: Regularly evaluate premium feature pricing to ensure competitiveness and profitability.

Low

Increase Premium Features : promotions and advertisements.

MERCI !

are there any questions?

ghaisanrabbani5@gmail.com
+6285156101050

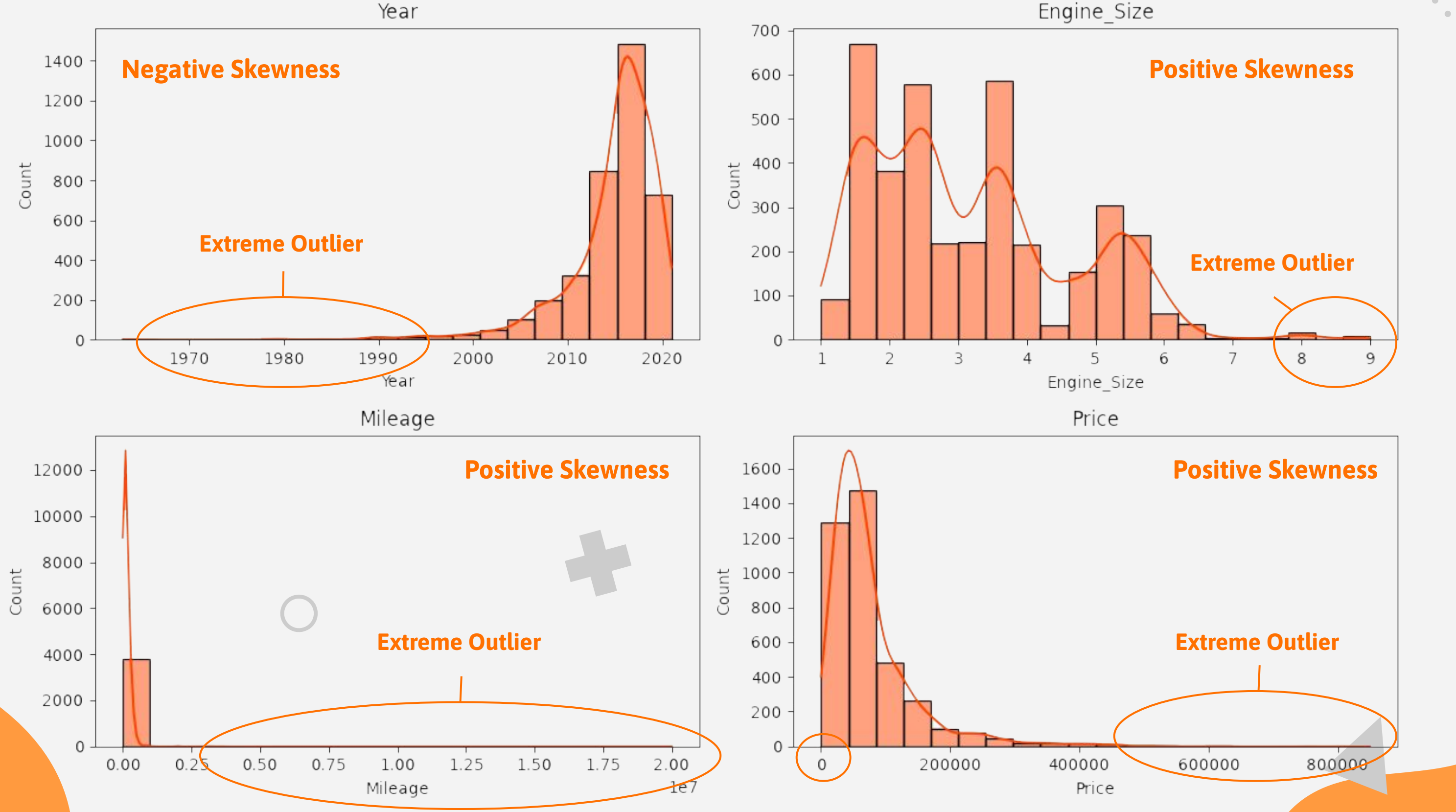
https://github.com/ghaisanr/Predict-used-cars-price-in-Saudi-Arabia-s

CRÉDITS: Ce modèle de présentation a été créé par **Slidesgo**, comprenant des icônes de **Flaticon**, des infographies et des images de **Freepik** et des illustrations de **Storyset**

APPENDIX

Data Understanding : EDA

Numerical Data Distribution Analysis Using Histograms: Pattern Identification and Outliers



Negative Skewness

Extreme Outlier

Positive Skewness

Extreme Outlier

Positive Skewness

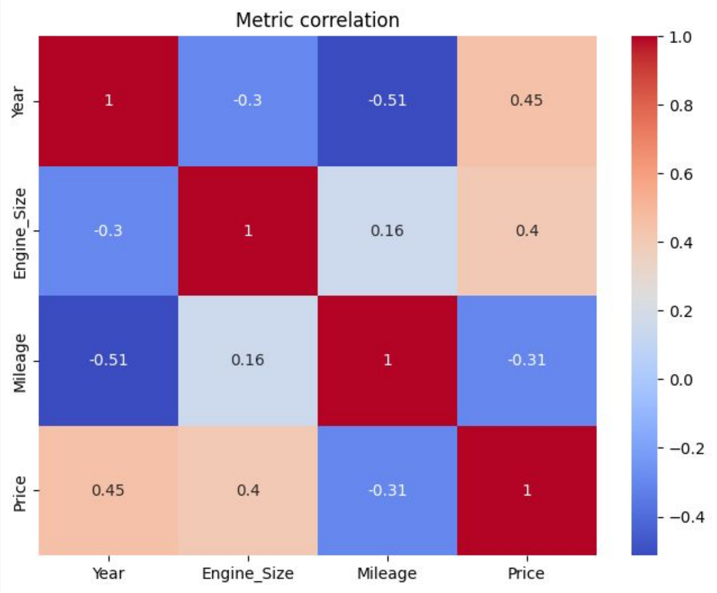
Extreme Outlier

Positive Skewness

Extreme Outlier

Data Understanding : EDA

Correlation Matrix Visualization: Understanding Relationships Between Variables in Numerical Data



	Year	Engine_Size	Mileage	Price
Year	1	-0.3	-0.51	0.45
Engine_Size	-0.3	1	0.16	0.4
Mileage	-0.51	0.16	1	-0.31
Price	0.45	0.4	-0.31	1

Based on the correlation matrix above, it can be seen:

1. **Positive correlation** between **”Year”** and **“Price”** of **0.45** This shows that the newer the year of the car, the higher the price tends to be.

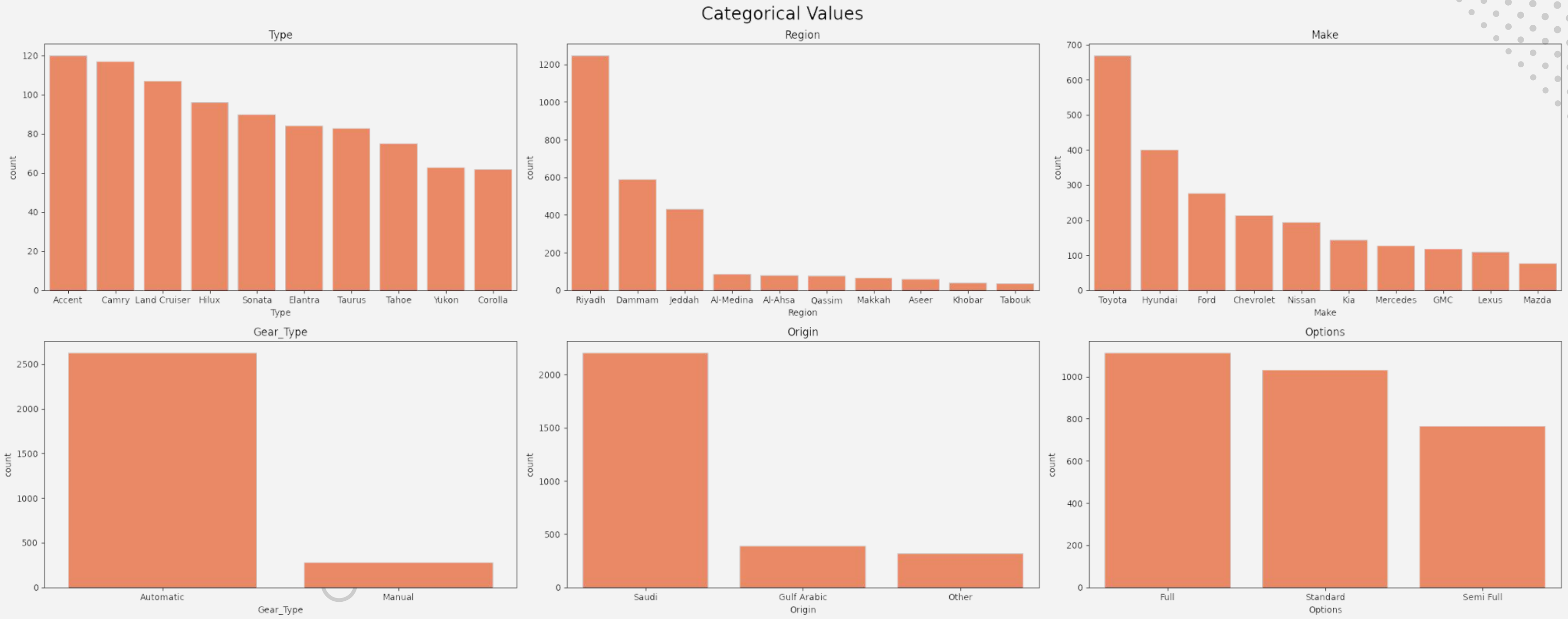
2. **Negative correlation** between **“Price”** and **“Mileage”** of **-0.31**. This shows that the greater the Mileage, the lower the price tends to be.

3. **Negative correlation** between **“Year”** and **“Mileage”** of **-0.51**. This shows that the newer the year of the car, the lower the mileage.

No one has a high correlation so there is no **multicollinearity problem**.

Data Understanding : EDA

Frequency Analysis of Categorical Data: Bar Plot for Top 10 Values



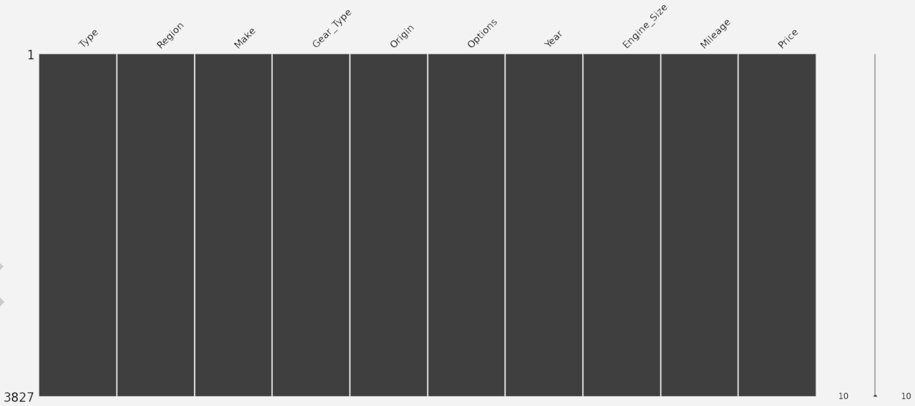
Data Preprocessing : Cleaning Data

Feature Selection

remove columns that are not needed for analysis : “Negotiable” → “price”= 0

because a price of 0 can result in machine learning results that are not appropriate

Missing Values



→ Not Have Missing Values

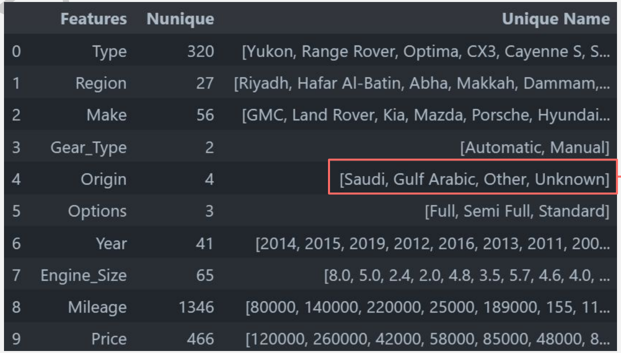
Duplicated Data

Total Duplicated data = 3 (Drop data) → Because efficiency in terms of time

Data Preprocessing : Cleaning Data

Spelling Error

Check spelling error with nunique code for each categorical column



	Features	Nunique	Unique Name
0	Type	320	[Yukon, Range Rover, Optima, CX3, Cayenne S, S...
1	Region	27	[Riyadh, Hafar Al-Batin, Abha, Makkah, Dammam,...
2	Make	56	[GMC, Land Rover, Kia, Mazda, Porsche, Hyundai...
3	Gear_Type	2	[Automatic, Manual]
4	Origin	4	[Saudi, Gulf Arabic, Other, Unknown]
5	Options	3	[Full, Semi Full, Standard]
6	Year	41	[2014, 2015, 2019, 2012, 2016, 2013, 2011, 200...
7	Engine_Size	65	[8.0, 5.0, 2.4, 2.0, 4.8, 3.5, 5.7, 4.6, 4.0, ...
8	Mileage	1346	[80000, 140000, 220000, 25000, 189000, 155, 11...
9	Price	466	[120000, 260000, 42000, 58000, 85000, 48000, 8...

has the same meaning and it would be a shame to delete it

Rows “Other” > “Unknown”

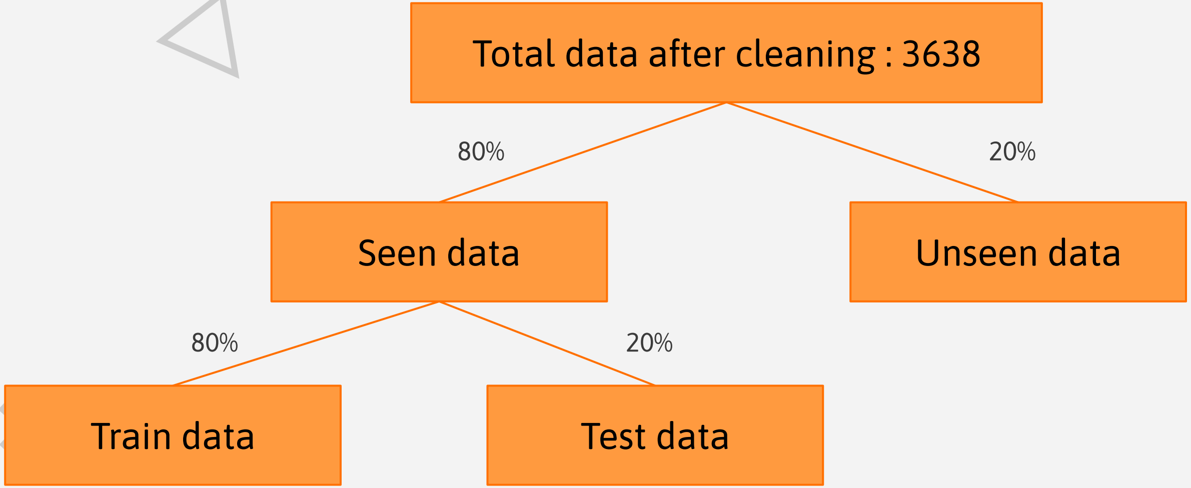
"Unknown" value can be replaced with "Other"

Handling Outlier

- Total drop “Year” < 2000 : 74
- Total drop “Engine_Size” >= 8 : 25
- Total drop “Mileage” >= 600000 : 21
- Total drop “Price” < 5000 : 66

Extreme Outlier = unreasonable value

Data Generation



Seen data

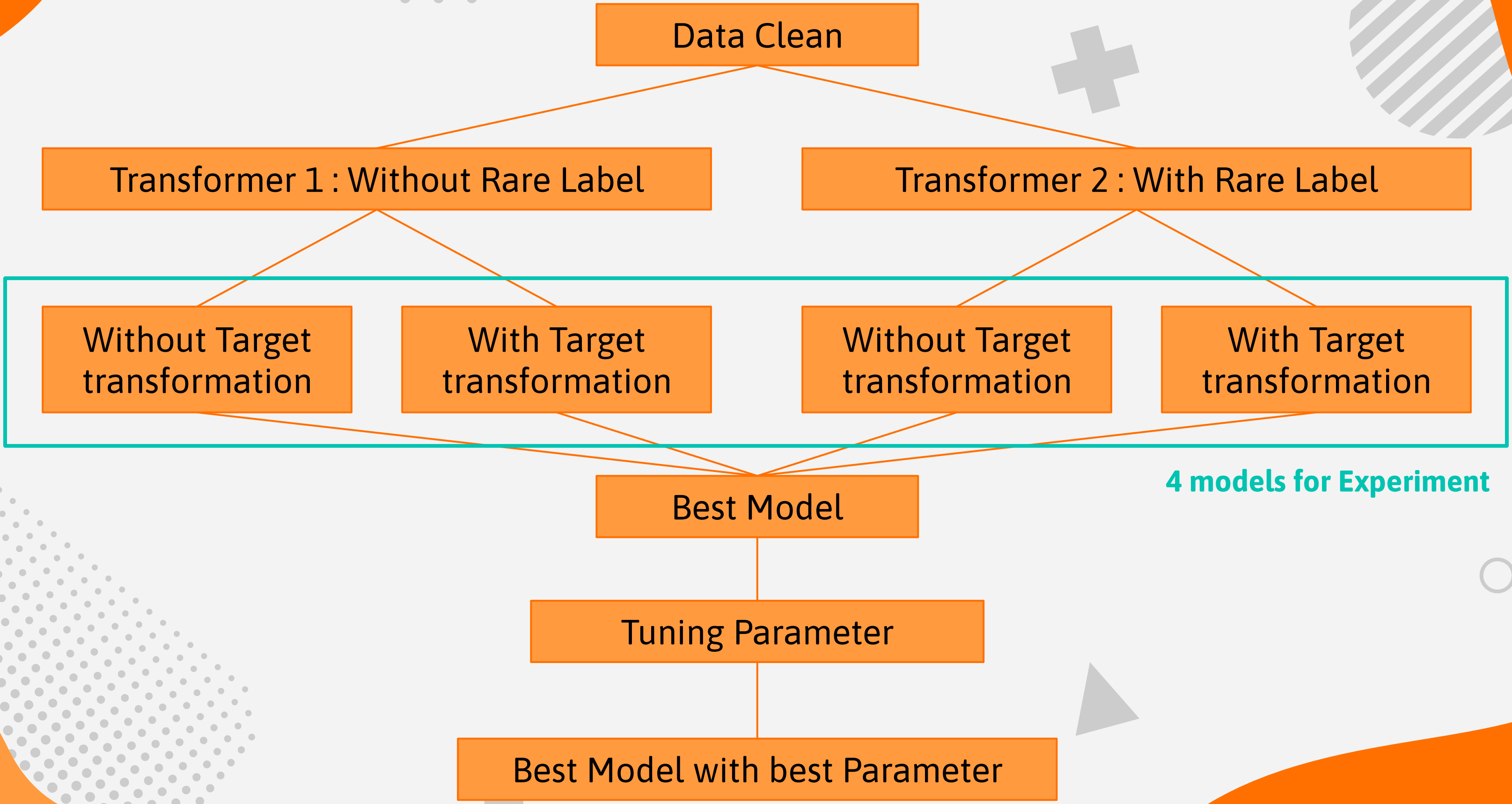
Develop machine learning model

- **Train data** : helps the model understand the trends and patterns
- **Test data** : evaluate the model's

Unseen data

- Business calculations
- Decision Making at Syarah.com

Modeling : Workflow experiment



Data Clean

Transformer 1 : Without Rare Label

Transformer 2 : With Rare Label

Without Target transformation

With Target transformation

Without Target transformation

With Target transformation

4 models for Experiment

Best Model

Tuning Parameter

Best Model with best Parameter

Modeling : Transformation Data Process

Transformer 1

Data Categorical

One Hot Encoder

- “Gear Type”
- “Origin”
- “Option”

Rare Label transform

- “Type”
- “Make”
- “Region”

Data Numerical

Robust Scaling

- “Engine_Size”
- “Mileage”
- “Year”

Transformer 2

Data Categorical

One Hot Encoder

- “Gear Type”
- “Origin”
- “Option”

Rare Label transform → **Binary Encoder**

- “Type”
- “Make”
- “Region”

Data Numerical

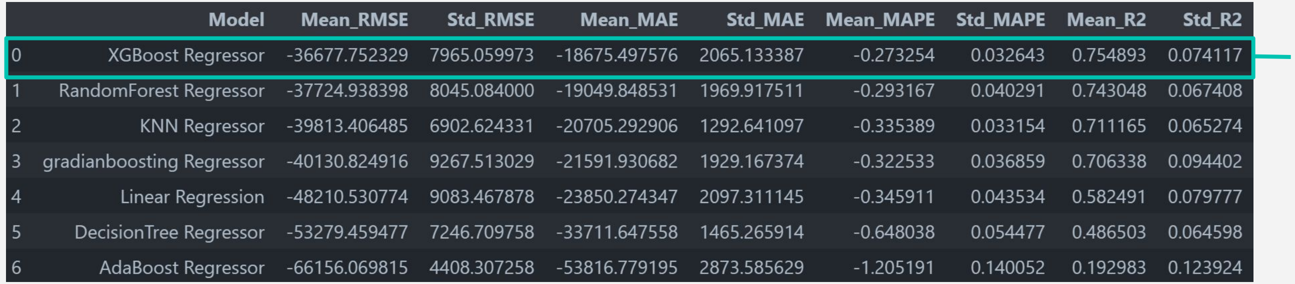
Robust Scaling

- “Engine_Size”
- “Mileage”
- “Year”

Modeling : Result the Model Experiment

Model 1

Transformer 1 : Without rare label handling
Model : without target transformation

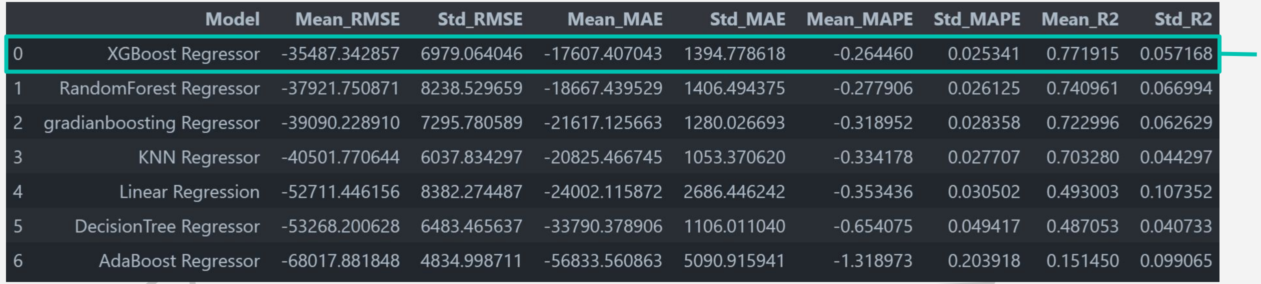


	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	Mean_R2	Std_R2
0	XGBoost Regressor	-36677.752329	7965.059973	-18675.497576	2065.133387	-0.273254	0.032643	0.754893	0.074117
1	RandomForest Regressor	-37724.938398	8045.084000	-19049.848531	1969.917511	-0.293167	0.040291	0.743048	0.067408
2	KNN Regressor	-39813.406485	6902.624331	-20705.292906	1292.641097	-0.335389	0.033154	0.711165	0.065274
3	gradianboosting Regressor	-40130.824916	9267.513029	-21591.930682	1929.167374	-0.322533	0.036859	0.706338	0.094402
4	Linear Regression	-48210.530774	9083.467878	-23850.274347	2097.311145	-0.345911	0.043534	0.582491	0.079777
5	DecisionTree Regressor	-53279.459477	7246.709758	-33711.647558	1465.265914	-0.648038	0.054477	0.486503	0.064598
6	AdaBoost Regressor	-66156.069815	4408.307258	-53816.779195	2873.585629	-1.205191	0.140052	0.192983	0.123924

Best Model

Model 2

Transformer 2 : With rare label handling
Model : without target transformation



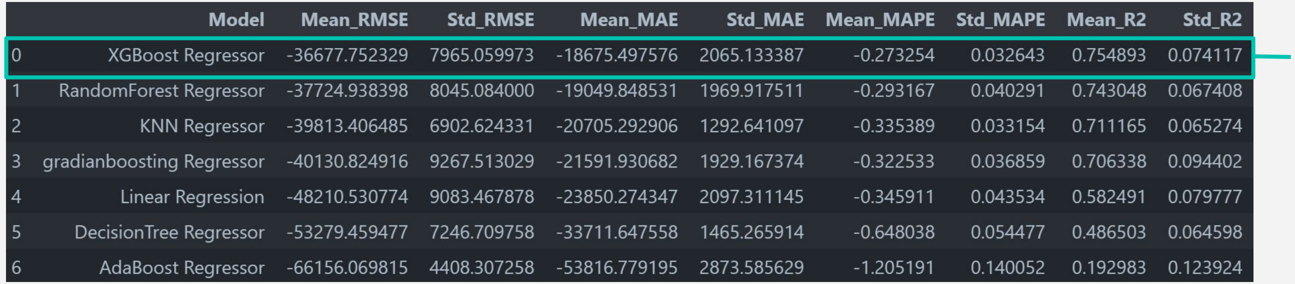
	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	Mean_R2	Std_R2
0	XGBoost Regressor	-35487.342857	6979.064046	-17607.407043	1394.778618	-0.264460	0.025341	0.771915	0.057168
1	RandomForest Regressor	-37921.750871	8238.529659	-18667.439529	1406.494375	-0.277906	0.026125	0.740961	0.066994
2	gradianboosting Regressor	-39090.228910	7295.780589	-21617.125663	1280.026693	-0.318952	0.028358	0.722996	0.062629
3	KNN Regressor	-40501.770644	6037.834297	-20825.466745	1053.370620	-0.334178	0.027707	0.703280	0.044297
4	Linear Regression	-52711.446156	8382.274487	-24002.115872	2686.446242	-0.353436	0.030502	0.493003	0.107352
5	DecisionTree Regressor	-53268.200628	6483.465637	-33790.378906	1106.011040	-0.654075	0.049417	0.487053	0.040733
6	AdaBoost Regressor	-68017.881848	4834.998711	-56833.560863	5090.915941	-1.318973	0.203918	0.151450	0.099065

Best Model

Modeling : Result the Model Experiment

Model 3

→ **Transformer 1** : Without rare label handling
Model : with target transformation

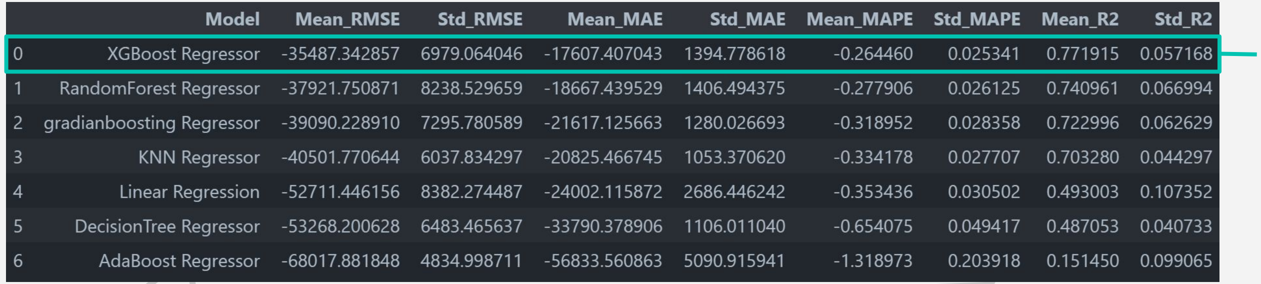


	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	Mean_R2	Std_R2
0	XGBoost Regressor	-36677.752329	7965.059973	-18675.497576	2065.133387	-0.273254	0.032643	0.754893	0.074117
1	RandomForest Regressor	-37724.938398	8045.084000	-19049.848531	1969.917511	-0.293167	0.040291	0.743048	0.067408
2	KNN Regressor	-39813.406485	6902.624331	-20705.292906	1292.641097	-0.335389	0.033154	0.711165	0.065274
3	gradianboosting Regressor	-40130.824916	9267.513029	-21591.930682	1929.167374	-0.322533	0.036859	0.706338	0.094402
4	Linear Regression	-48210.530774	9083.467878	-23850.274347	2097.311145	-0.345911	0.043534	0.582491	0.079777
5	DecisionTree Regressor	-53279.459477	7246.709758	-33711.647558	1465.265914	-0.648038	0.054477	0.486503	0.064598
6	AdaBoost Regressor	-66156.069815	4408.307258	-53816.779195	2873.585629	-1.205191	0.140052	0.192983	0.123924

Best Model

Model 4

→ **Transformer 2** : With rare label handling
Model : with target transformation



	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	Mean_R2	Std_R2
0	XGBoost Regressor	-35487.342857	6979.064046	-17607.407043	1394.778618	-0.264460	0.025341	0.771915	0.057168
1	RandomForest Regressor	-37921.750871	8238.529659	-18667.439529	1406.494375	-0.277906	0.026125	0.740961	0.066994
2	gradianboosting Regressor	-39090.228910	7295.780589	-21617.125663	1280.026693	-0.318952	0.028358	0.722996	0.062629
3	KNN Regressor	-40501.770644	6037.834297	-20825.466745	1053.370620	-0.334178	0.027707	0.703280	0.044297
4	Linear Regression	-52711.446156	8382.274487	-24002.115872	2686.446242	-0.353436	0.030502	0.493003	0.107352
5	DecisionTree Regressor	-53268.200628	6483.465637	-33790.378906	1106.011040	-0.654075	0.049417	0.487053	0.040733
6	AdaBoost Regressor	-68017.881848	4834.998711	-56833.560863	5090.915941	-1.318973	0.203918	0.151450	0.099065

Best Model

Model Performance Comparison: Model 4 Shows Best Results using XGBOOST

Model	Name	Mean RMSE	STD RMSE	Mean MAE	STD MAE	Mean MAPE	STD MAPE	Mean R2	STD R2
Model 4	XGBOOST	-34894.87	7458.09	-17121.38	1603.23	-0.22	0.02	0.77	0.06
Model 3	XGBOOST	-35000.67	7824.27	-17013.63	1463.80	-0.22	0.01	0.77	0.06
Model 2	XGBOOST	-35487.342	6979.06	-17607.40	1394.77	-0.26	0.02	0.77	0.05
Model 1	XGBOOST	-36677.75	7965.05	-18675.49	2065.133	-0.27	0.03	0.75	0.07

Best Model

- performs well overall, especially in terms of RMSE and R-squared.
- consistent performance with low standard deviations across metrics