Purwadhika
Digital Technology School

SAUDI ARABIA USED CAR PRICE PREDICTION

By **Amartiya Shalaisya Raihani** | JCDSOL - 014

OVERVIEW

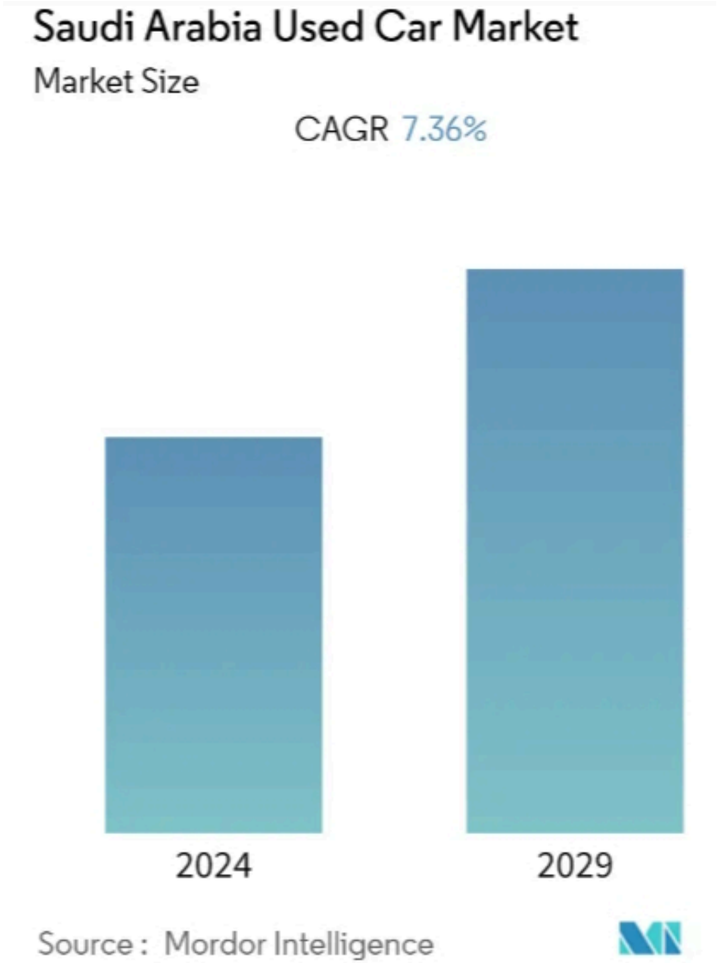
1. About Syarah.com & Business Problem
2. Data Understanding
3. Data Cleaning
4. Feature Selection, Feature Engineering
5. Analytics
6. Conclusion & Recommendation

About Syarah.com

Syarah.com, created by Syarah Company, is an online platform that facilitates the sale and purchase of guaranteed used cars.

The online platform helps used car sellers reach consumers more easily and quickly with competitive and free-setting selling prices.

Business Problem Statement



The Saudi Arabian used car market has experienced notable growth, with a 7.36% CAGR growth expected.

One of the biggest challenges for Syarah Company is **preventing overpricing and underpricing of used cars that price being determined by sellers**.

A price prediction model based on car specifications is needed, which Syarah.com can easily use to **provide price recommendations** to sellers. This means that after the seller inputs the specifications, a price recommendation will automatically appear.

GOALS

From these two goals below, if more transactions be occurred, it will achieve the main goal, which is increasing the company's profit.

Having a price recommendation feature, potentially attracting more sellers to sell their used cars on Syarah.com.

More people will buy used cars and transact on Syarah.com because of the well-fitted price offers according to the car specifications.

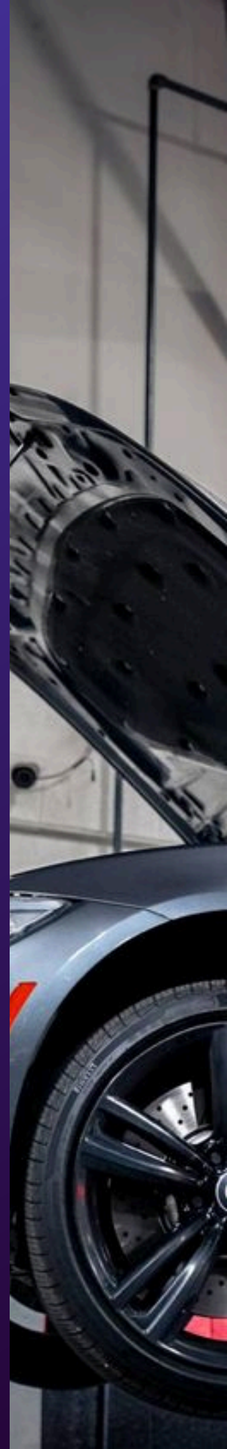
ANALYTIC APPROACH

Build a **regression model as price prediction tool** for newly listed used cars, which will be useful for sellers in determining the selling price of used cars.

METRICS EVALUATION

- Mean Absolute Percentage Error (MAPE): the average percentage error produced by the regression model
- Mean Absolute Error (MAE): the average absolute value of the error
- R-squared (R2): R-squared if the final selected model is a linear model. The R-squared value is used to determine how well the model can represent the overall variance of the data. The closer it is to 1, the better the model fits the observational data.

MAPE will be the primary reference for selecting the best model; the smaller the MAPE, the better the model.



DATA UNDERSTANDING

The dataset contains 5624 records of used cars collected from syarah.com.

- Type: Type of used car.
- Region: The region in which the used car was offered for sale.
- Make: The company name.
- Gear_Type: Gear type size of used car.
- Origin: Origin of used car.
- Options: Options of used car.
- Year: Manufacturing year.
- Engine_Size: The engine size of used car.
- Mileage: Mileage of used car
- Negotiable: True if the price is 0, that means it is negotiable.
- Price: Used car price.

DATA CLEANING

Check Missing Value

No Missing Value on Dataset.

Check & Handling Outliers

Only Extreme Outliers will be removed. Extreme outliers are only found in the 'Year', 'Mileage', and 'Price' columns.

Check Duplicate

Remove 3 duplicate data found.

Check & Handling Faulty Data

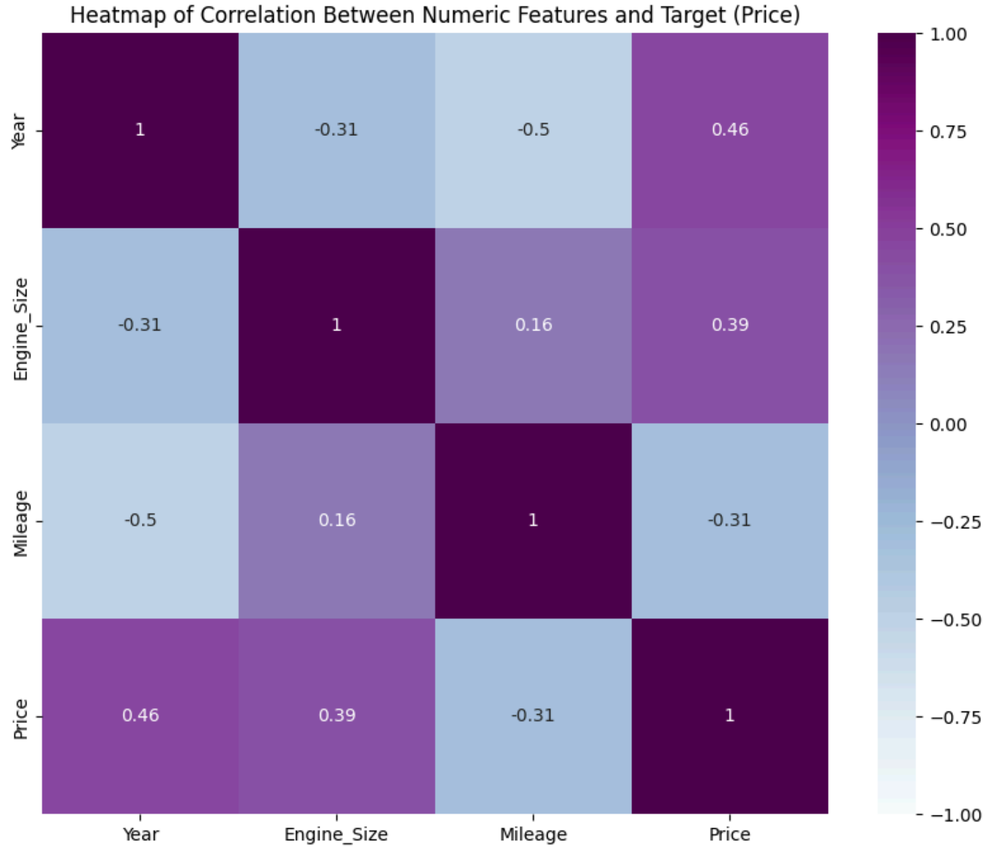
- Remove Data 'Price' = 0

- Remove Data Price' < 8000
The cheapest used car for sale in Riyadh is priced at 8000.

- Handling Faulty Data: 'Mileage'
The average mileage per year is assumed to be 25,750 km. Therefore, the maximum possible mileage of a car as of this year (2024) is 772,500 km.
Therefore, data with 'Mileage' > 772,500 will be dropped because it is not reliable as a reference for prediction.

FEATURE SELECTION

01 Numerical Features



'Year' has positive medium correlation with price; 'Engine_Size' positive low correlation with price; and 'Mileage' has negative low correlation with price.
All Numerical features will be used.

03 Categorical Features

'Negotiable' feature is not related to the actual price of the used car. Therefore, 'Negotiable' feature can be dropped.

Determining other categorical features will be assisted by ANOVA (for 'Type', 'Region', 'Make', 'Gear_Type', 'Origin', and Options').

All P-Value of other categorical features < 0.05 --> there is a price difference between various categories for each feature, then '**Type', 'Region', 'Make', 'Gear_Type', 'Origin', and 'Options' are features that will be used.**

Feature Engineering

01

Doesn't need to use an **imputer** because there are no missing values in the dataset.

02

Scaling is applied to equalize the scale of all numerical features using the Robust Scaler method (not sensitive to outliers).

03

Encoding is applied to transform categorical data into numerical format.
One Hot Encoder is used for categorical features with <= 5 categories, while Binary Encoder is used for categorical features with > 5 categories.

Analytics Flowchart

Model Benchmarking

- Base Model (KNN Regressor, Decision Tree Regressor, and Linear Regression)
- Voting & Stacking (Soft Voting, Stacking - KNN, Stacking - DT, Stacking - Linear Regression)
- Bagging (Linear Regression and Random Forest Regressor)
- Boosting (AdaBoost Regressor, Gradient Boosting Regressor, and XGBoost Regressor)

Hyperparameter Tunning

Hyperparameter Tunning will be applied on the 3 best models to determine the best parameters of each model.

Model Limitations

Check the absolute value of error and MAPE on each row of data. Data that has a MAPE > 50% is considered to have a high error, meaning that the model is not accurate when applied to that data.

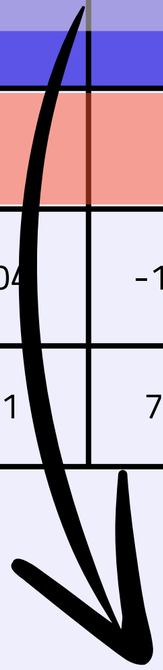
Check Residual & Feature Importance

Check residual plot to evaluate bias in the model. Check Feature Importance to determine the most important features that greatly affect the used car price prediction.

MODEL BENCHMARKING RESULT

Top 5 Models Based on the smallest MAPE (ignoring the negative sign).

Metrics	Random Forest Regressor	XGBoost Regressor	Stacking - KNN	Gradient Boosting Regressor	KNN
MAPE	-0.2790	-0.2796	-0.2949	-0.3259	-0.3423
MAE	-1.876499e+04	-1.899928e+04	-2.071254e+04	-2.202065e+04	-2.134954e+04
R2	7.429000e-01	7.454000e-01	7.150000e-01	7.120000e-01	7.159000e-01



Best 3 models that will be carried out hyperparameter tuning

HYPERPARAMETER TUNNING RESULT

MAPE Before Hyperparameter Tunning
- Random Forest Regressor (27.9%)
- XGBoost (27.96%)
- Stacking - KNN (29.5%)

MAPE After Hyperparameter Tunning
- Random Forest Regressor (27.9%)
- XGBoost (26%)
- Stacking - KNN (28.5%)

The FINAL MODEL selected based on the smallest MAPE value after Hyperparameter Tunning is --> XGBoost Regressor.

HYPERPARAMETER TUNNING RESULT

MAPE on Train Dataset After Tunning

- Random Forest Regressor (27.9%)
- XGBoost (26%)
- Stacking - KNN (28.5%)

MAPE on Test Dataset After Tunning

- Random Forest Regressor (25.8%)
- XGBoost (24.9%)
- Stacking - KNN (26.8%)

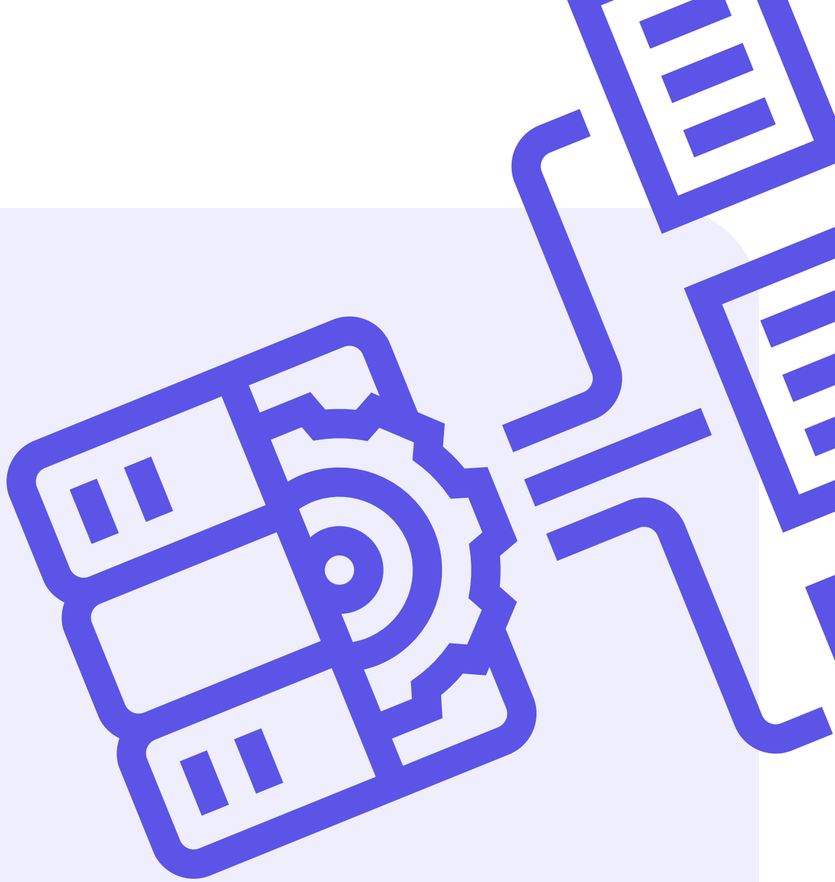
There is **no indication of overfitting or underfitting** in the final model (XGBoost Regressor) and the difference in MAPE values (between the models on train data and test data) is the most not significant. These indicate that XGBoost Regressor is the most suitable model for the case of used car price prediction.

MODEL LIMITATION

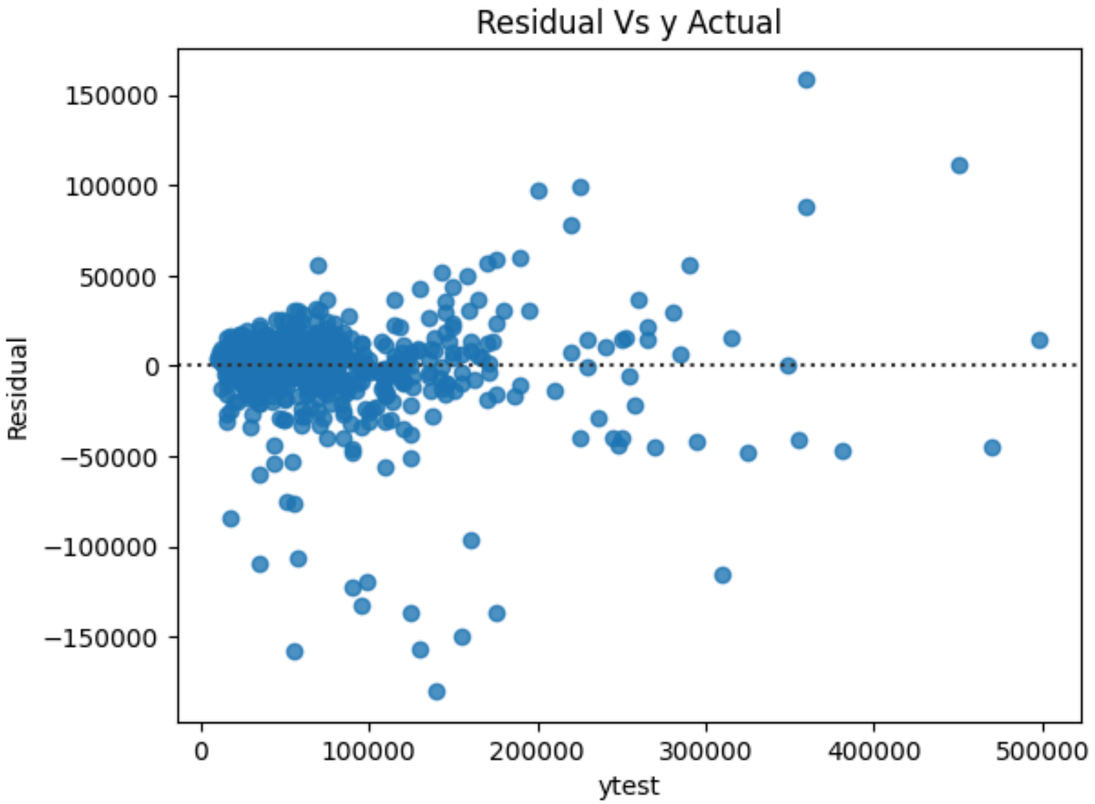
Model only provides accurate predictions for data used cars with actual **price range of 11.000 - 498,000 SR**; a **manufacturing year range of 1995 - 2021**; an **engine size range of 1 - 8.8 L**; and a **Mileage range of 100 - 570,000 KM**.

Also, model only provides accurate predictions for data used cars with the majority categorical characteristics:

- Used cars with **accent type**, **automatic gear type**, and **standard options.**
- Used cars in **Riyadh as their region for sale** and **Saudi as their origin** of car .
- Used cars from the **Toyota company.**

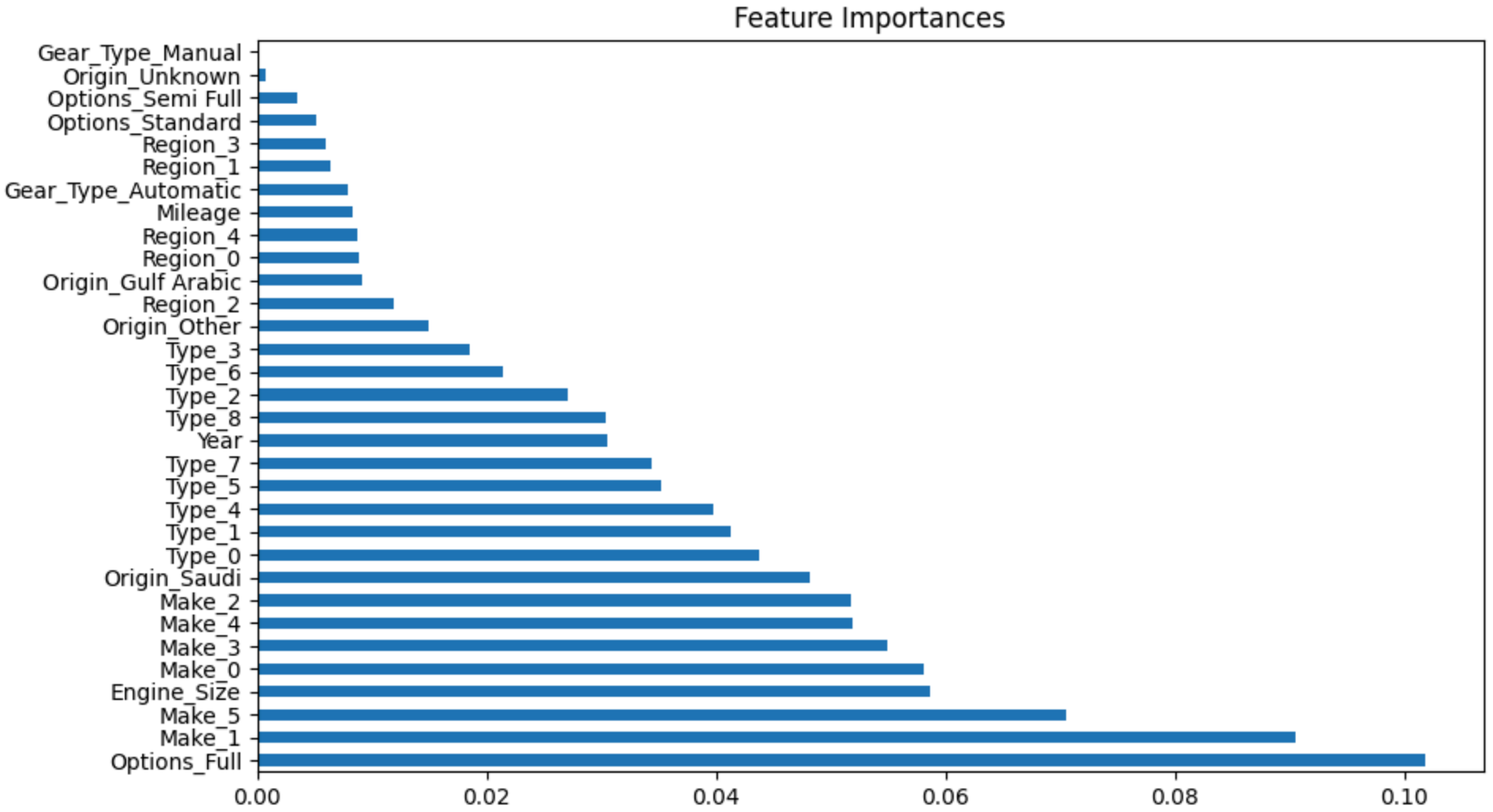


CHECK RESIDUAL



- Most points are close to the zero line, showing that the model's predictions are generally accurate.
- Points far from the zero line are outliers or inaccurate predictions, especially when the target value (Price) is higher or more variable (over 300,000 SR).
- The even spread of residuals around the zero line suggests that the model's predictions are not biased (residuals are not only above or below the line but on both sides).

FEATURE IMPORTANCE



- 'Options', 'Make', 'Engine-Size', 'Origin', and 'Type' are the most important features that greatly affect the used car price prediction.
- In contrast, feature 'Gear_Type' has a small affect on the prediction.

IMPACT OF MODEL IMPLEMENTATION

Comparing Used Car Price Predictions Using Two Different Methods (Conventional and using Model Regression).
Conventional method is predicting price based on average prices for cars of the same type and year.
This comparison involves checking the predicted prices for randomly selected car types and years using both methods.

	Type	Year	Predicted Conventional
225	Hiace	2016	65000.0

	Type	Year	Actual	Predicted_XGBRegressor
5177	Hiace	2016	70000	62909.558594
5586	Hiace	2016	60000	61799.503906

Cars of the same type and year can have different prices. Using a machine learning regression model for price prediction can lead to higher profits compared to conventional methods.

If Syarah.com sells a car for 70,000 SR:

- Profit with conventional prediction: 5,000 SR (70,000 - 65,000)
- Profit with regression prediction: 7,091 SR (70,000 - 62,909)

Price prediction using regression model can **increase profit up to 1.4%** for each car. While the per-car profit increase is small, applying this across all used cars can significantly boost overall profits.

CONCLUSION

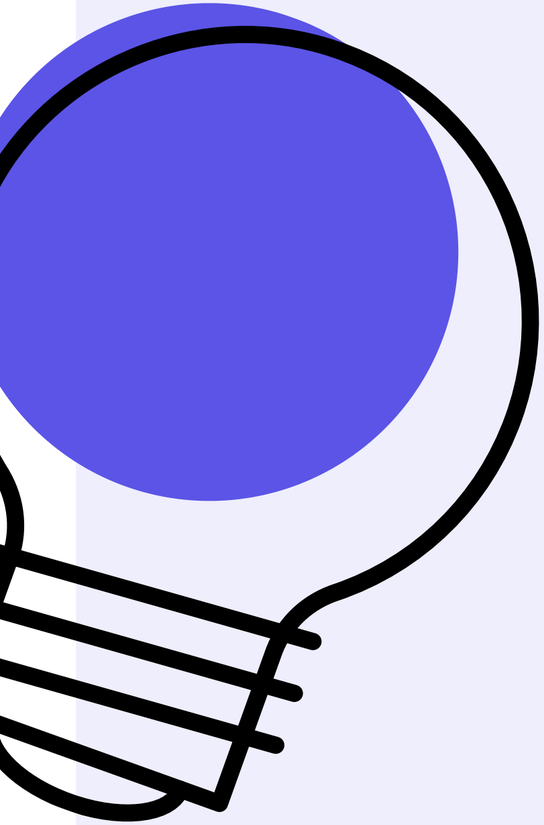
- **XGBoost Regressor** model is the best model with **MAPE 26%** for predicting used car sales prices with parameters: 'subsample': 0.6; 'n_estimators': 500; 'max_depth': 13; and 'learning_rate': 0.1.

- The features that most influence 'Price' are **'Options', 'Make', 'Engine-Size', 'Origin', and 'Type'**.

- Model only provides accurate predictions for data used cars with actual price range of 11.000 - 498,000 SR; a manufacturing year range of 1995 - 2021; an engine size range of 1 - 8.8 L; and a Mileage range of 100 - 570,000 KM.

- The impact of using this regression model if implemented by Syarah.com **could save time and cost** in conducting market research on competitive used car selling prices. Furthermore, Price prediction using regression model can **increase profit up to 1.4% for each car**.

RECOMMENDATION



- **Add the amount of data (especially for prices above 300,000 SR)** so that the model can train on more data and produce more accurate predictions with lower MAPE. The model for this prediction case was built using only 3688 clean rows of data (initially 5624 rows), so the data amount could be increased to around > 5000 clean rows.

- **Add features** that are likely to correlate with the target 'Price', such as the condition of the car's interior, exterior, and fuel type.

- **Update the data with more recent manufacturing years (> 2021)**, because the newest car manufacturing year used in this dataset is 2021.

THANK YOU

Purwadhika
Digital Technology School