MACHINE LEARNING PROJECT

USED CARS PRICE PREDICTION TO OPTIMIZE AUTOMOTIVE PLATFORM PRICING STRATEGY

BY: GRACE NATALIE CATHERINE

OUTLINE

1. BACKGROUND
2. DATA UNDERSTANDING
3. DATA PREPROCESSING
4. MACHINE LEARNING MODELLING
5. ANALYSIS
6. CONCLUSION & RECOMMENDATION

ABOUT COMPANY AND BUSINESS

Syarah is one of the largest **digital used-car platforms** in Saudi Arabia,

Sellers can advertise their used cars with specifications and set their own prices, while buyers can purchase available used cars directly through the platform.

Syarah also buys used cars directly, modifies them, and resells the used cars.

Company generates revenue from fee and sales margin on each transaction.

PROBLEM STATEMENT

Pricing is a critical factor in used-car transactions.

Pricing on Syarah is still fully determined by sellers, often leading to **mispriced cars that do not reflect true market values.**

incorrect prices cause slow sales, low buyer trust and ultimately **lower purchases from buyers and lower revenue for company.**

The key question that needed to be solved by Data Scientist:

“How can we build a machine-learning model to accurately predict used-car prices and maximize revenue?”

BACKGROUND

01 **Management Team**
Make business decisions based on the insights for pricing strategy to improve business performance

02 **Sellers (Car Owners/Dealers):**
Receive price recommendations to set competitive and fair prices.

03 **Buyers:**
Benefit from more transparent, market-aligned pricing.

GOALS

01

Build a machine learning model to predict used car selling prices based on various vehicle features, enabling the development of an optimal, realistic, and transparent pricing strategy.

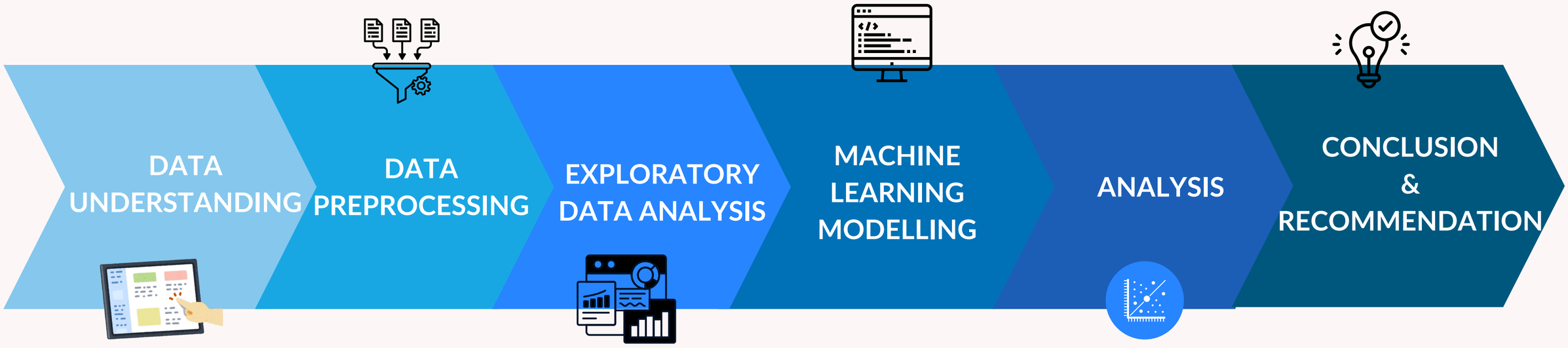
02

Identify the key factors influencing used-car prices to support stakeholders in defining actionable strategies.

BACKGROUND

ANALYTICAL APPROACH

PRICE is the TARGET → REGRESSION



BACKGROUND

KEY EVALUATION METRICS

MAPE

measures the average percentage difference between the predicted values and the actual values.

MAE

measures the average absolute difference between predicted values and actual values.

DATA UNDERSTANDING

TYPES OF DATA	
CATEGORICAL	**NUMERICAL**
Type: Type of used car. **Region: The area/region where the used car is offered for sale.** **Make: The brand of the vehicle.** **Gear_Type: The type of transmission in the used car.** **Origin: The origin of the used car.** **Options: The features/equipment included in the used car.** **Negotiable: True if the price is negotiable (price = 0), otherwise False.**	**Year: The manufacturing year of the used car** **Engine_Size: The engine capacity of the used car.** **Mileage: The mileage of the used car.** **Price: The price of the used car**

DATA PREPROCESSING

Data	Reason	Action
Duplicated values (4 rows)	Duplicated data can potentially affect data distribution and modelling	Drop values
Negotiable Variable	Only related to the negotiation process	Drop column
Origin (category: Unknown, 61 rows)	The origin of the car is unknown.	Drop values
Year (<1998)	Remove outliers and reduce skewness	Drop values
Mileage (<100 and >584744)	Remove outliers and reduce skewness	Drop values

DATA PREPROCESSING

Data	Reason	Action
Price = 0 / Zero values (1796 rows)	It’s not the real value of the price, it’s because the price is negotiable.	Drop values
Price (<5000 SAR and > 600.000 SAR)	Remove outliers and reduce skewness	Drop values
Make ('Škoda')	Transform/Replace 'Škoda' with Skoda	Replace method

DATA PREPROCESSING

Data	Action	Result
Make	Feature Engineering	New Feature ‘Make Tier’ for brand tier category
Type	Feature Engineering	New Feature ‘Body Type’ for body type of the cars category
All Columns	Rearrange columns	Columns rearranged
Multicolinearity	Check multicolinearity with Heatmap and VIF	No multicolinearity

DATA PREPROCESSING

Data	Action	Result
Features	Feature Selection	Use all features
Split Data	80% train, 20% test	Splitted data: train and test
Categorical Features	Preprocessing with Encoding (Binary, One Hot, Ordinal)	Encoded Categorical Features
Numerical Features	Preprocessing with Scalling	Scaled Numerical Features

MODELLING

1

MODEL SELECTION

2

MODEL TRAINING

3

MODEL EVALUATION (BENCHMARKING BASELINE MODELS) & HYPERPARAMETER TUNING

METRICS:
MAPE, MAE (select models with the lowest error values)

4

MODELS IMPROVEMENT

5

BEST MODEL SELECTION (BENCHMARKING BASELINE & TUNED MODELS)

MACHINE LEARNING MODELLING

REGRESSION MODELS

01 LINEAR REGRESSION

02 K-NEAREST NEIGHBORS (KNN)

03 DECISION TREE

04 SUPPORT VECTOR MACHINE (SVM)

05 RANDOM FOREST

06 ADAPTIVE BOOSTING (ADABOOST)

07 GRADIENT BOOSTING

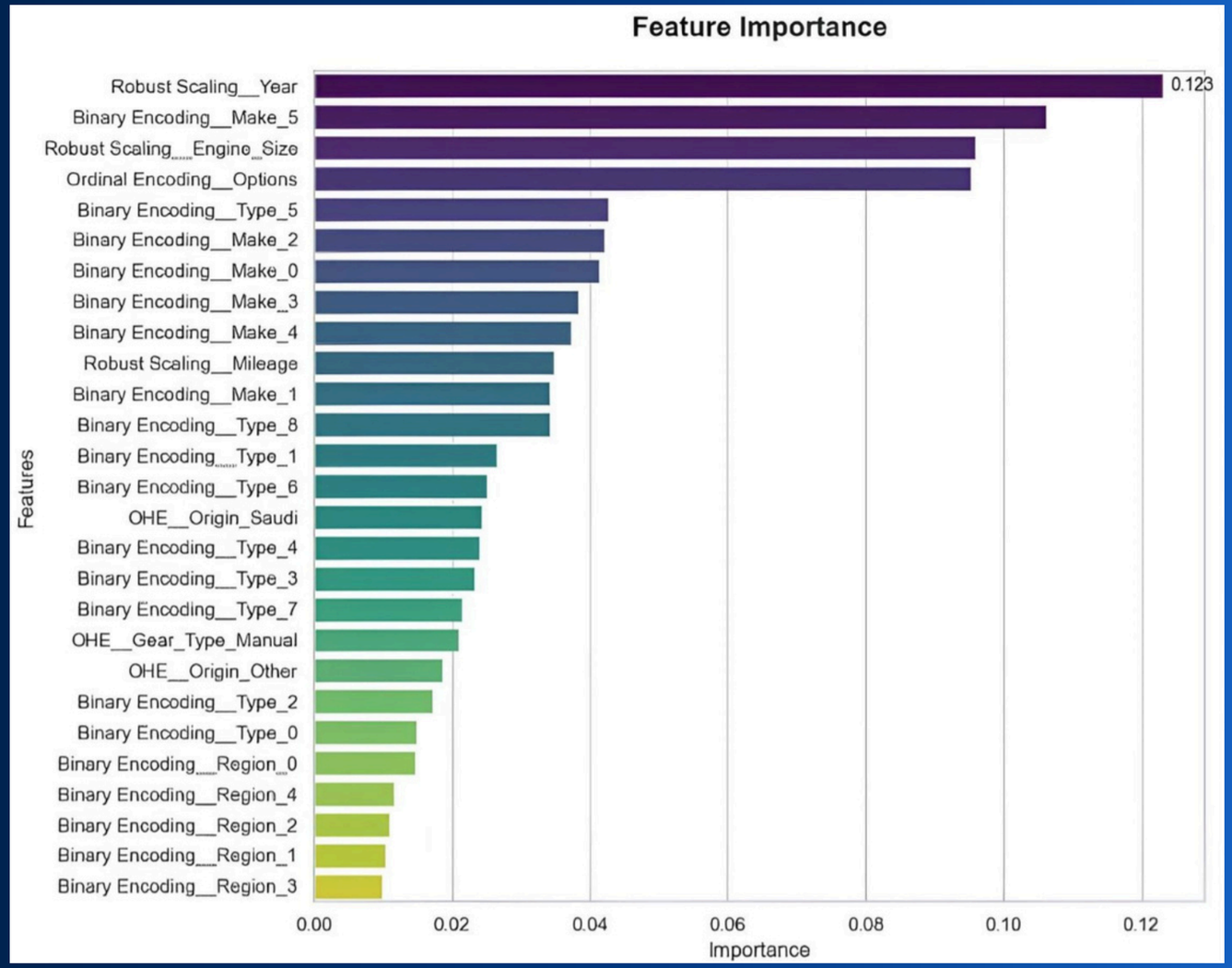
08 XTREME GRADIENT BOOSTING (XGBOOST)

MODEL PERFORMANCE BENCHMARKING

MODEL	TRAIN MAPE	TEST MAPE	MAPE DIFFERENCE	TRAIN MAE	TEST MAE
Xtreme Gradient Boosting (XGBoost) Tuned	0.241587 (24.16%)	0.222632 (22.26%)	0.018955 (1.9%)	15574.976367	13884.068359
Xtreme Gradient Boosting (XGBoost)	0.276100 (27.61%)	0.253760 (25.55%)	0.022340 (2.23%)	17657.511328	16103.273438
Random Forest Tuned	0.269926 (26.80%)	0.254632(25.46%)	0.015295 (1.5%)	17299.441002	15405.438794
Random Forest	0.289920 (28.99%)	0.280079 (26.96%)	0.009841 (0.9%)	18293.880563	16564.551367
Gradient Boosting	0.312122 (31.21%)	0.323192 (32.31%)	-0.011070 (1.1%)	20578.930818	19991.695363
Decision Tree	0.371351 (30.80%)	0.323942 (32.40%)	0.047409 (4.7%)	24019.705211	20806.913817
K-Nearest Neighbor	0.329573 (39.95%)	0.326450 (32.64%)	0.003123 (0.3%)	19320.977579	18471.855814
Linear Regression	0.618704 (61.87%)	0.652785 (65.27%)	-0.034081 (3.4%)	32378.147470	32623.073859
Support Vector Machine (SVM)	0.683208 (68.3%)	0.676769 (67.65%)	0.006439 (0.64%)	43174.392182	41544.791502
Adaptive Boosting	1.145118 (114%)	1.183831 (118%)	-0.038713 (3.8%)	50044.520511	50413.072835

MACHINE LEARNING MODELLING

FEATURE IMPORTANCE



The most significant feature is **Year**, followed by **Make** in second place and **Engine_Size** in third place

MODEL IMPROVEMENT

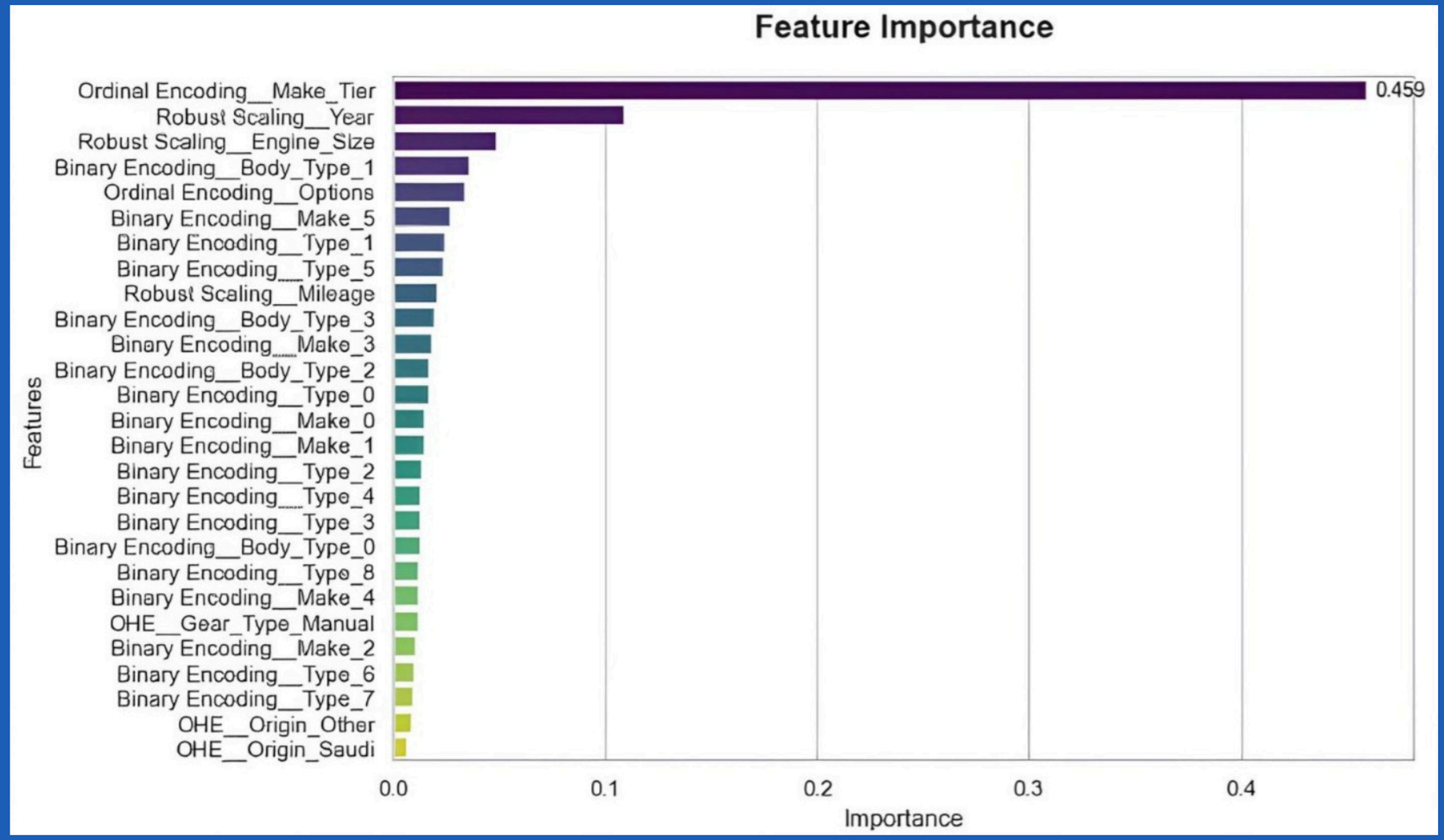
Model improvement is performed to address weak model performance/underfitting (MAPE > 20%) by:

- Feature Selection : increasing the number of used features
- Feature Engineering : adding new features Make_Tier and Body_Type.
- Encoding the new features using Ordinal and Binary Encoder

FINAL MODEL BENCHMARKING (MODEL IMPROVEMENT)

MODEL	TRAIN MAPE	TEST MAPE	MAPE DIFFERENCE	TRAIN MAE	TEST MAE
Xtreme Gradient Boosting (XGBoost) Tuned	0.197340 (19.73%)	0.190793 (19.07%)	0.006547 (0.65%)	13291.922656	11832.232422
Random Forest Tuned	0.218076 (21.8%)	0.211703 (21.17%)	0.006373 (0.63%)	14245.882060	13282.933266
Xtreme Gradient Boosting (XGBoost)	0.217629 (21.76%)	0.212081 (20.20%)	0.005548 (0.55%)	14296.873828	13076.559570
Random Forest	0.228379 (22.83%)	0.229336 (22.5%)	-0.000957 (0.09%)	14904.859706	14337.062343
Gradient Boosting	0.255710 (25.57%)	0.266989 (26.3%)	-0.011279 (1.1%)	17037.954794	16863.706842
Decision Tree	0.308845 (30.88%)	0.288679 (28.86%)	0.020166 (2%)	20767.050402	18812.142271
K-Nearest Neighbors (KNN)	0.311693 (31.16%)	0.308105(30.81%)	0.003588 (0.36%)	18165.936357	17946.821067
Linear Regression	0.560924 (59.9%)	0.551702 (55.4%)	0.009222 (0.5%)	28154.029024	27445.075883
Support Vector Machine (SVM)	0.683235 (68.3%)	0.676747 (67.7%)	0.006489 (0.65%)	43174.733655	41543.448759
Adaptive Boosting	0.868549 (87.4%)	0.917095 (91%)	-0.048546 (4.8%)	38307.402711	40083.293187

FEATURE IMPORTANCE



The most significant feature is **Make_Tier**

FINAL MODEL & PARAMETERS

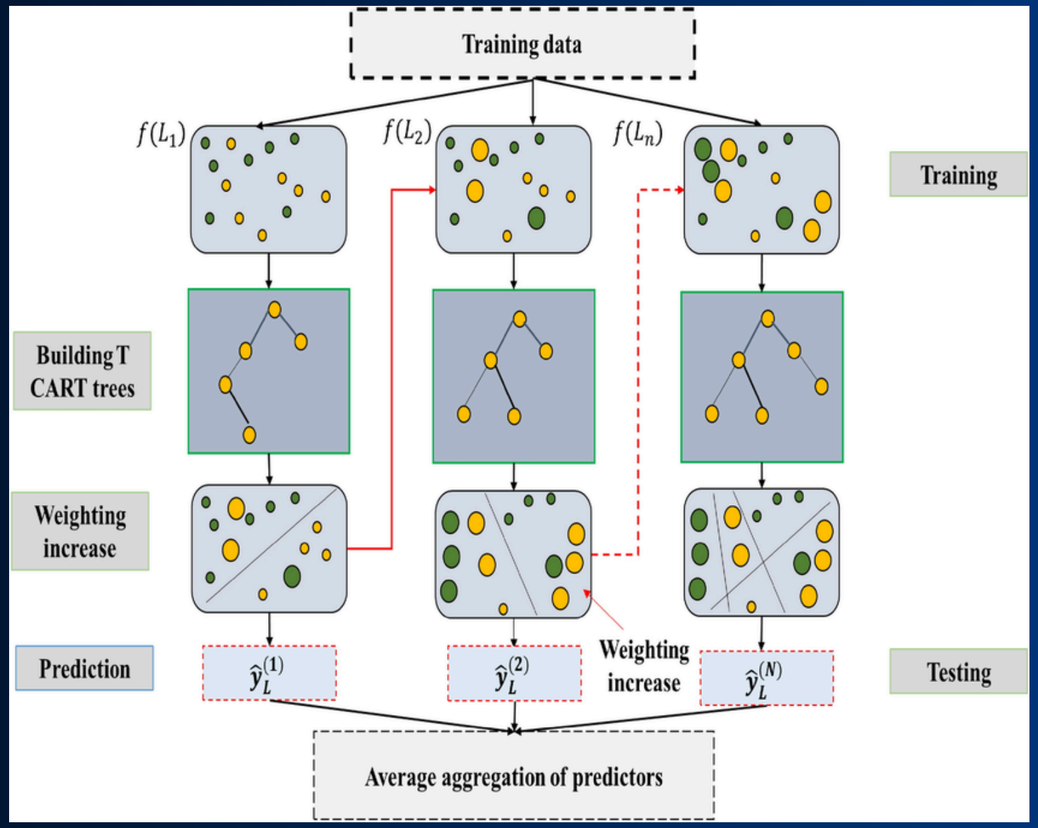
FINAL BEST MODEL :
XGBOOST TUNED

Metrics:
MAPE = 19%
(Good Forecasting,
Source: Klimberg et al., 2010)
MAE = 11832 SAR

BEST PARAMETERS

- learning_rate: 0.05
- gamma: 0
- max_depth: 7
- min_child_weight: 2
- n_estimators: 500
- reg_alpha: 0.5
- reg_lambda: 10
- subsample: 0.8
- colsample_bytree: 0.8

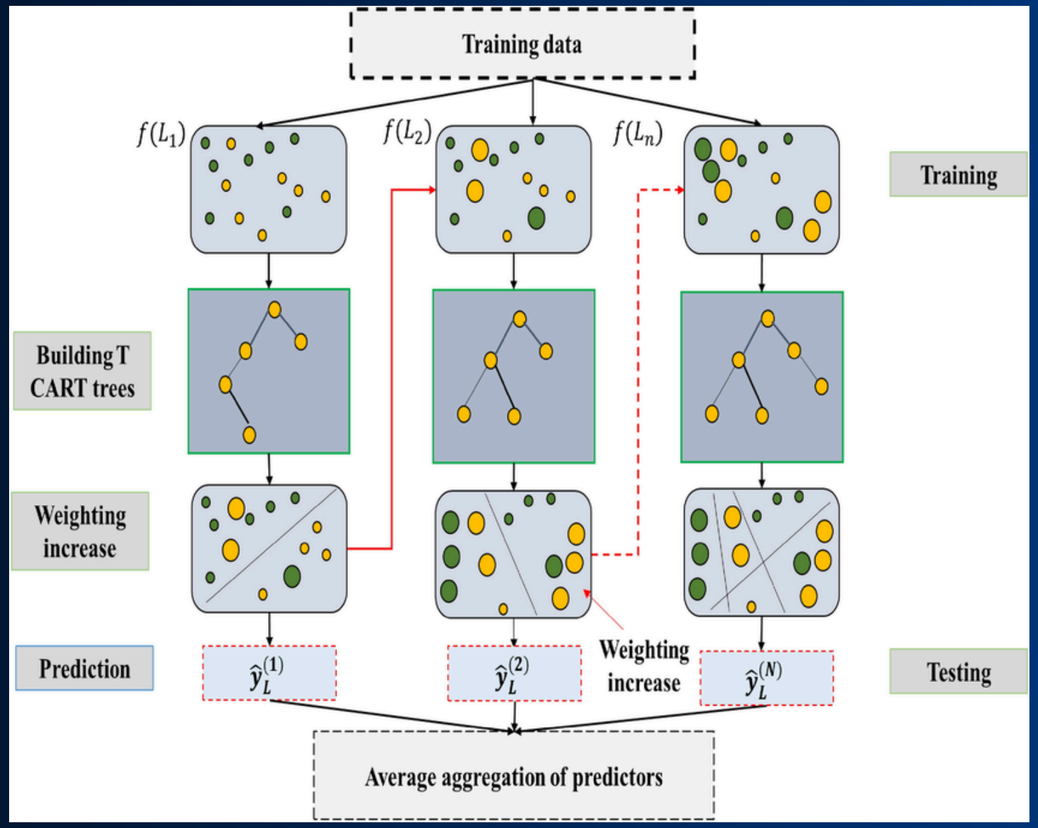
XGBOOST MODEL



XGBoost is a **fast**, **stable, and highly accurate gradient boosting algorithm**. It can handle large and complex datasets, captures non-linear patterns, and can be optimized**.**

Its efficiency and strong predictive performance make XGBoost one of the best algorithms for machine learning models and **widely used in industry**.

HOW XGBOOST ALGORITHM WORKS



XGBoost works by **building many trees sequentially**, where each new tree focuses on **correcting the errors made by the previous trees**.

It **gives more weight to samples with large error, continuously reducing the residuals**.

The final prediction is the combined output of all the trees, resulting in a strong and highly accurate model.

MODEL LIMITATIONS

- Production Year: 1998 – 2021
- Engine Size: 1.0 – 9.0
- Mileage: 100 – 584744 KM
- Price: 5,000 – 600,000 SAR
- Type: The type of cars sold in Saudi Arabia, according to dataset.
- Region: 27 regions in which the used car was offered for sale.
- Car Body Type/Shape: SUV, Sedan, Luxury, Hatchback, Pickup, Coupe/Sports, MPV, Van, Truck/Bus, and Others.
- Brands: 55 brands sold in Saudi Arabia, according to dataset.
- Brand Tier: Ultra Luxury, Premium Luxury, Midrange, and Entry Level.
- Transmission: Automatic and Manual.
- Feature Completeness: Full, Semi Full, Standard.
- Car Origin: Gulf Arabic, Saudi, and Others.

ACTUAL PRICE VS PREDICTED PRICE

Most points align closely with the reference line, showing good model performance, though there are actual prices that are predicted higher (overpredict) and vice versa (underpredict), which is likely caused by outliers, incomplete information, and noise in the data.

Overpredict risks mispricing above market value, while underpredict reduces margins.

Lower errors (19%) lead to more accurate pricing, faster sales, and support revenue growth.

COST BENEFIT ANALYSIS

WITHOUT MODEL

- Lost Platform Revenue from Underpredict (Potential Margin Loss): 5,768,559.70 SAR
- Lost Platform Revenue from Overpredict (Service Fee + Holding Cost): 8,891,197.76

TOTAL LOSS WITHOUT MODEL: 14,659,757.46 SAR

WITH MODEL

- Lost Platform Revenue From Underpredict (Potential Margin Loss): 2,031,262.24 SAR
- Lost Platform Revenue From Overpredict (Service Fee + Holding Cost): 2,067,104.49 SAR

TOTAL LOSS WITH MODEL: 4,098,366.74 SAR

ECONOMIC SAVINGS

TOTAL SAVINGS USING MODEL: 10,561,390.72 SAR

CONCLUSION

Machine Learning Model for Used Car Price Prediction

XGBoost is the best model with error (MAPE) ~19%, providing reliable used cars pricing guidance.

Important Factor

Brand Tier is the most influential feature, useful for marketing and pricing strategy.

Cost Benefit Analysis

Using the model potentially reduces losses by ~10.5 M SAR, the model provides significant financial benefits for the company

MODEL RECOMMENDATION

Feature Expansion, Bias & Range Improvement

Add more features such as **car condition, inspection results, service history, and color** to **reduce model bias, and expand data coverage to improve predictions** beyond the current model's limited range to **improve the models' capability.**

Data Size Increasement

Add more data from different periods to **improve the model's generalization capability.**

If the **dataset increase significantly, more complex models**, like recursive neural networks (RNN), **could be explored.**

Build Car Demand Prediction Model

Combining existing features with buyer-behavior features from the company's real data to **build a model to predict which cars are most in demand** among buyers for **inventory planning strategy**.

BUSINESS RECOMMENDATION

1. Dynamic Pricing
Adjust car prices based on Brand Tier, the most influential factor on resale value.

2. Feature Highlight
Emphasize the used car brands, brands tier (based on the most influential feature), and feature options to increase perceived value to buyers.

3. Explainable AI
Show key factors influencing price of used cars to improve transparency and build user trust.

4. Personalized Buyer Recommendations
Use price predictions to suggest used cars, offering the best value for money based on user preferences.

5. Integration, Deployment & Continuous Improvement
Integrate the XGBoost model into the platform to provide accurate price predictions for used cars, while implementing continuous monitoring and routine retraining with the latest market data

THANK YOU!