CAPSTONE PROJECT 3

SAUDI ARABIA USED CARS

Al Azhary Putera Satria
DTI DS 0106

Table of Contents

01. Business Understanding

02. Data Understanding

03. Data Preprocessing

04. Modelling, Evaluation, and Benchmarking

05. Hyperparameter Tuning and Explainable AI

06. Model Deployment on Cloud

07. Conclusion and Recomendation

01.
Business Understanding

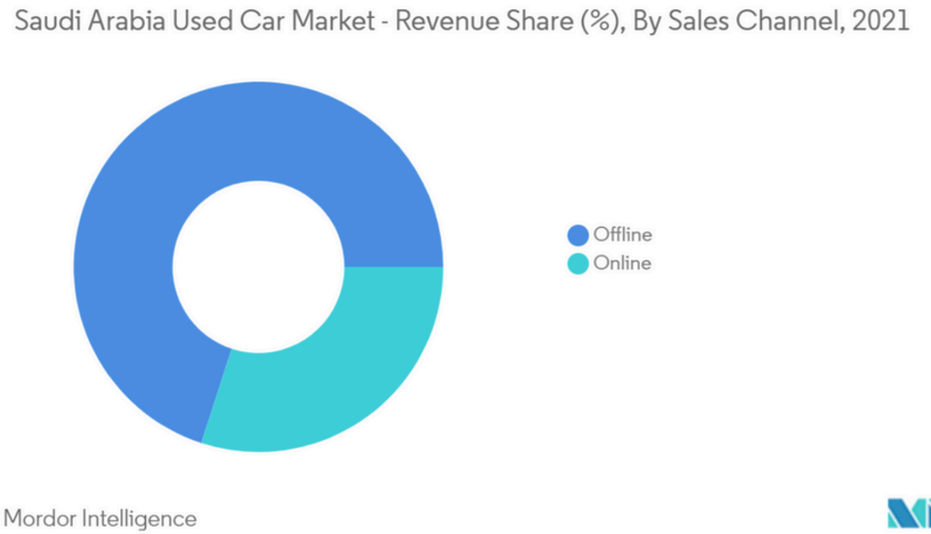
CAPSTONE PROJECT 3

CONTEXT

Saudi Arabia Used Car Market was valued at USD 4.91 billion in 2021 and is expected to surpass a net valuation of USD 8.69 billion by 2027 end, registering a solid CAGR growth of 7.36% over the forecast period.

Source:
https://www.mordorintelligence.com/industry-reports/saudi-arabia-used_cars_market_size

CONTEXT



As most of Revenue share from used cars market still lead by offline transaction. Growing trend presents a significant opportunity for Syarah.com, a leading online platform for buying and selling used cars in Saudi Arabia. By improving its services as online platform for used cars transaction, can lead to increasing sales and revenues

PROBLEM STATEMENT

Traditionally, car dealerships and individual sellers rely on experience and market research to estimate used car prices. This method is subjective and time-consuming. Machine learning can provide a more data-driven and objective approach for predicting used car prices in Saudi Arabia.

GOALS

The goal is to build a machine learning model that can predict the selling price of a used car in Saudi Arabia based on relevant features. Thus can help our customers to get the best price of their cars.

ANALYTIC APROACH

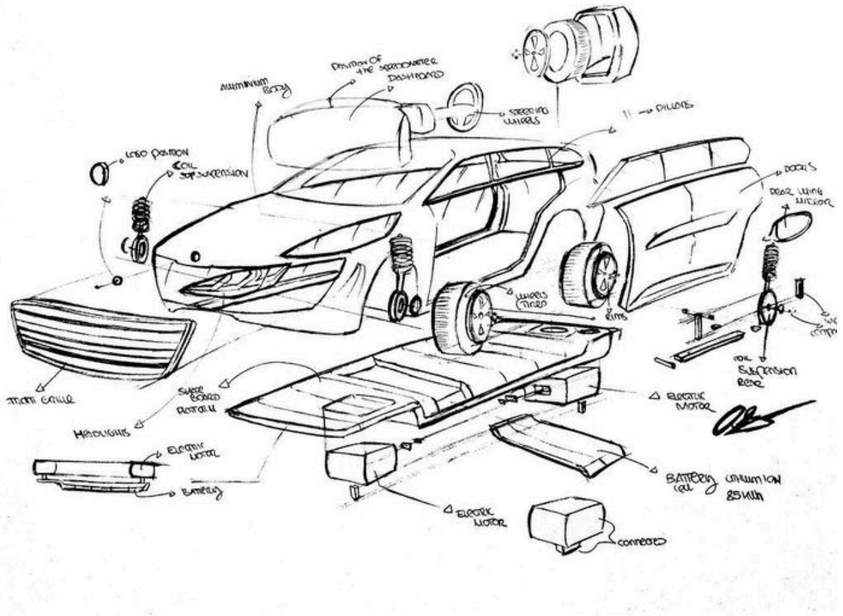
Based on car features and prices,we build a regression model that can predict prices of used cars.

EVALUATION METRICS

To determine the best models and performance of choosen machine learning model, our main evaluation metrics is Mean Absolute Percentage Error.

Check other metrics for comparison and business aproach :

- Root Mean Squared Error (RMSE)
- Mean Absolute Error
- Adjusted R-squared



Machine Learning as Solution

Calculations (Hypothetical Scenario):

- Assume Syarah.com charges a 3% commission on every successful transaction.
- Without machine learning, sellers might underprice their cars by an average of 10% compared to market value.
- Let's say a seller incorrectly prices their car at 10,000 SAR, when it's actually worth 11,000 SAR.
- Syarah.com earns a commission of 300 SAR (3% of SAR 10,000).

Machine Learning Impact:

- With accurate price prediction, the car gets listed at 11,000 SAR.
- Syarah.com earns a commission of 330 SAR (3% of 11,000 SAR), a 10% increase.
- This small percentage increase is multiplied across thousands of transactions, leading to a significant overall revenue boost.

Beyond commissions, improved user experience due to accurate pricing can lead to more platform usage and potentially higher revenue from other services like featured listings or value-added services.

02.

Data Understanding

CAPSTONE PROJECT 3

Attributes	Data Type, Length	Description
Type	Text	Brand name of car
Region	Text	The region in which the used car was offered for sale
Make	Text	Name of the car company
Gear_Type	Text	Automatic / Manual
Origin	Text	Country of importer (GulF / Saudi / Other)
Option	Text	Full Options / Semi-Full / Standard
Year	Int	Year of Manufacturing
Engine_Size	Float	The engine size of used car
Mileage	Int	The average distance that a vehicle can travel on (in km)
Negotiable	Bool	If True, the price is 0. This means the price is negotiable (not set)

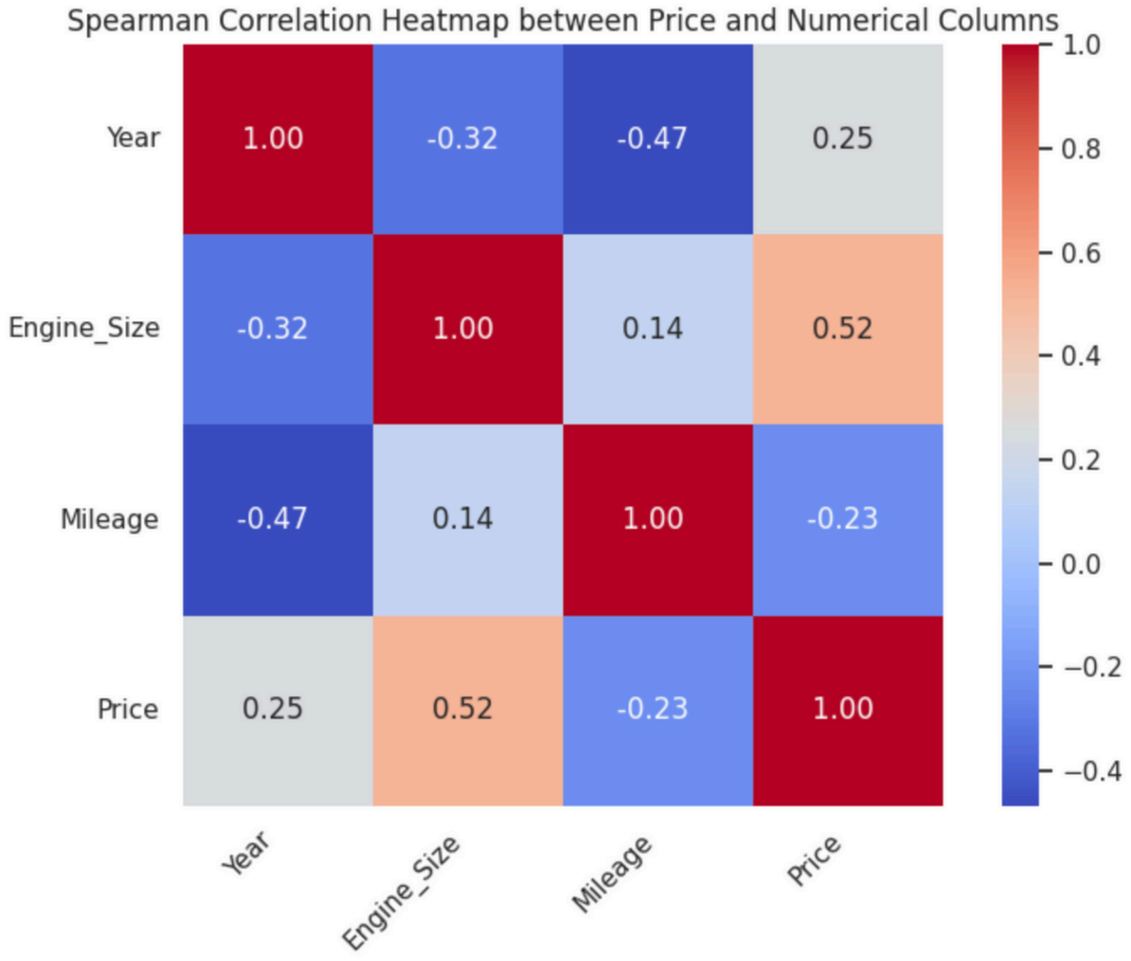
Data Understanding

Metric	Count	Mean	Std	Min	25%	50%	75%	Max
Year	5624	2014.101885	5.791606	1963	2012	2016	2018	2022
Engine_Size	5624	3.29543	1.515108	1	2	3	4.5	9
Mileage	5624	150923.375	382835.963	100	38000	103000	196000	20000000
Price	5624	53074.05814	70155.34061	0	0	36500	72932.5	850000

Data Understanding

Metric	Count	Unique	Top	Freq
Type	5624	347	Land Cruiser	269
Region	5624	27	Riyadh	2272
Make	5624	58	Toyota	1431
Gear_Type	5624	2	Automatic	4875
Origin	5624	4	Saudi	4188
Options	5624	3	Full	2233

Data Correlation



- Price and Year: There's a moderate positive correlation (0.25), suggesting that newer cars tend to be more expensive.
- Price and Engine Size: There's a moderate positive correlation (0.52), indicating that cars with larger engines are generally pricier.
- Price and Mileage: There's a weak negative correlation (-0.23), suggesting that cars with higher mileage tend to be slightly cheaper.

03.

Data Preprocessing

Data Preprocessing

Data Cleansing	Filter by Treshold	Feature Engineering	Log Transformation
Missing Value Duplicate Data Outliers	Filter data for Year, Mileage, Price, and Engine Size	Country Maker Car Age Mils per Year Condition	Transform Target Variable into logarithmic form

Data Cleansing

Check Type	Threshold	Action
Missing Value	0	None
Duplicate Data	3	Drop
Outlier (Price)	208	Contextual outliers

Data Filtering

Criteria	Min Value	Max Value
Year	2008	2024
Price	24000	1200000
Mileage	100	277000
Engine	1	6.6

Feature Engineering

Make → Mapping Function → **Country_make**

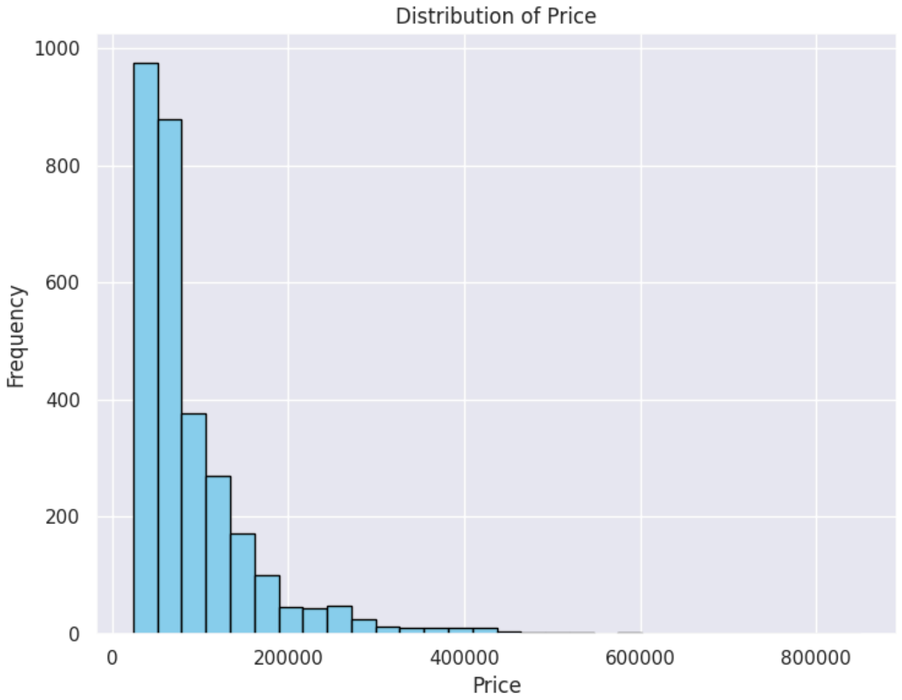
Condition

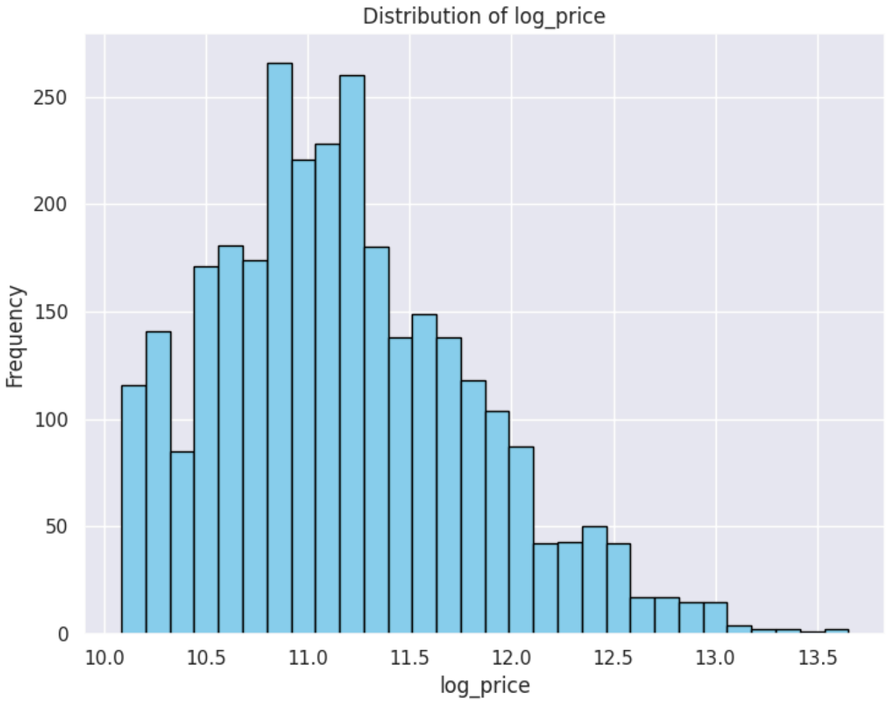
Mileage / car_age = Mileage_per_year →

Binning

- <10k Km/Year = Seldom
- 10k - 20k Km/Year = Normal
- 20k - 40k Km/Year = Over
- >40k Km/Year = Extreme

Log Transformation (Price)





04.

Modelling

CAPSTONE PROJECT 3

Modelling

Splitting Train and Test Data

Splitting data into training and test with propostion 70:30

Encoding and Scaling

One-Hot Encoding:

- Gear_Type
- Origin
- Options

Binary Encoding:

- Type
- Region
- Make

Ordinal Encoding:

- Condition

Scalling : RobustScaller

Modelling

- Linier Regression
- KKN Regressor
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost Regressor

Model Benchmarking

Training Model with CrossVal Result

	Model	RMSE	MAE	MAPE	Adjusted R-squared
0	LinearRegression	43039.746575	24217.065173	0.272717	0.669331
1	KNeighborsRegressor	39626.210088	20375.252557	0.230339	0.719703
2	DecisionTreeRegressor	50876.535312	24956.934687	0.276980	0.537950
3	RandomForestRegressor	35312.616638	17399.950313	0.189642	0.777406
4	XGBRegressor	33161.180364	17106.039732	0.189923	0.803703

Predict on Test Data Result

	Model	RMSE	MAE	MAPE	Adjusted R-squared
0	LinearRegression	41045.245856	22972.996054	0.270867	0.688424
1	KNeighborsRegressor	35529.925044	18267.696262	0.227015	0.766532
2	DecisionTreeRegressor	41601.560927	19961.795127	0.228727	0.679921
3	RandomForestRegressor	30078.861133	14931.903874	0.172260	0.832675
4	XGBRegressor	31150.017794	15279.285520	0.171527	0.820545

05.

Hyperparameter Tuning and Explainable AI

CAPSTONE PROJECT 3

Hyperparameter Tuning with Gridsearch

XGB Regressor

Hyperparameter	Value 1	Value 2	Value 3
model__n_estimators	100	200	300
model__learning_rate	0.01	0.1	0.2
model__max_depth	3	5	7
model__subsample	0.7	0.8	0.9
model__colsample_bytree	0.7	0.8	0.9

Random Forest

Hyperparameter	Value 1	Value 2	Value 3	Value 4
model__n_estimators	100	200	300	
model__max_depth	None	10	20	30
model__min_samples_split	2	5	10	
model__min_samples_leaf	1	2	4	
model__bootstrap	TRUE	FALSE		

Hyperparameter Tuning

MODEL	RMSE	MAE	MAPE	Adjusted R-squared
Random Forest Before	30078.861133	14931.903874	0.172260	0.832675
XGBoost Before	31150.017794	15279.285520	0.171527	0.820545
Random Forest After	30245.442032	15004.522949	0.172967	0.830816
XGBoost After	28496.559742	14185.527242	0.157845	0.849816

Extreme Gradient Boosting

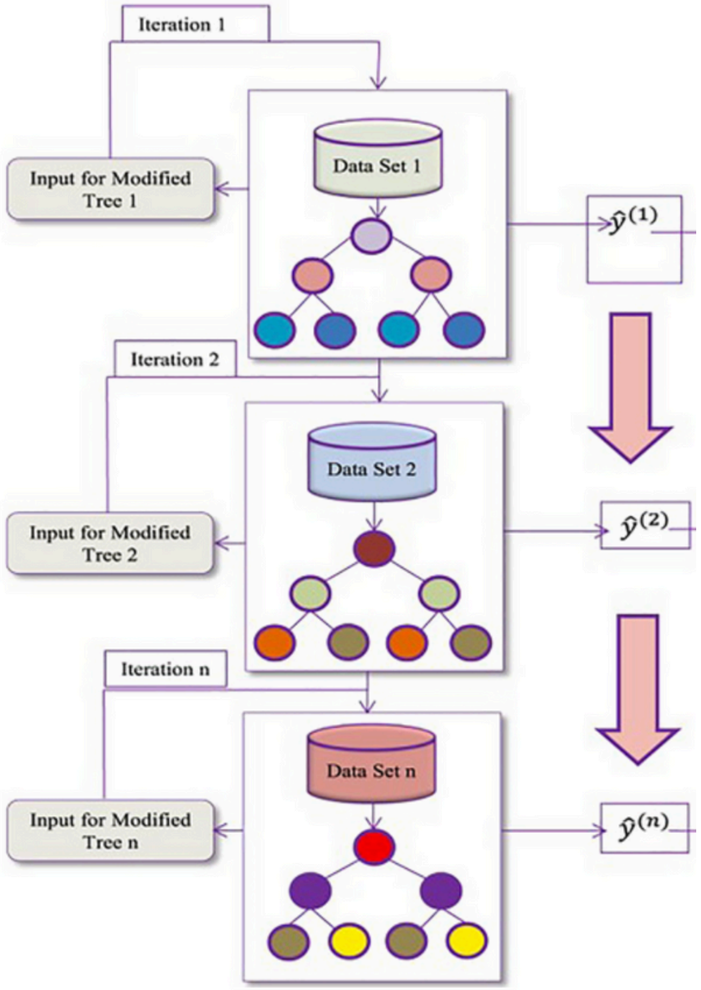
How this model works

At its core, **XGBoost** uses **decision trees** as the building blocks for its models. A decision tree is a flowchart-like structure where each **internal node represents a feature** (or attribute), each **branch represents a decision rule**, and each **leaf node represents an outcome** (label or value).

Ensemble method

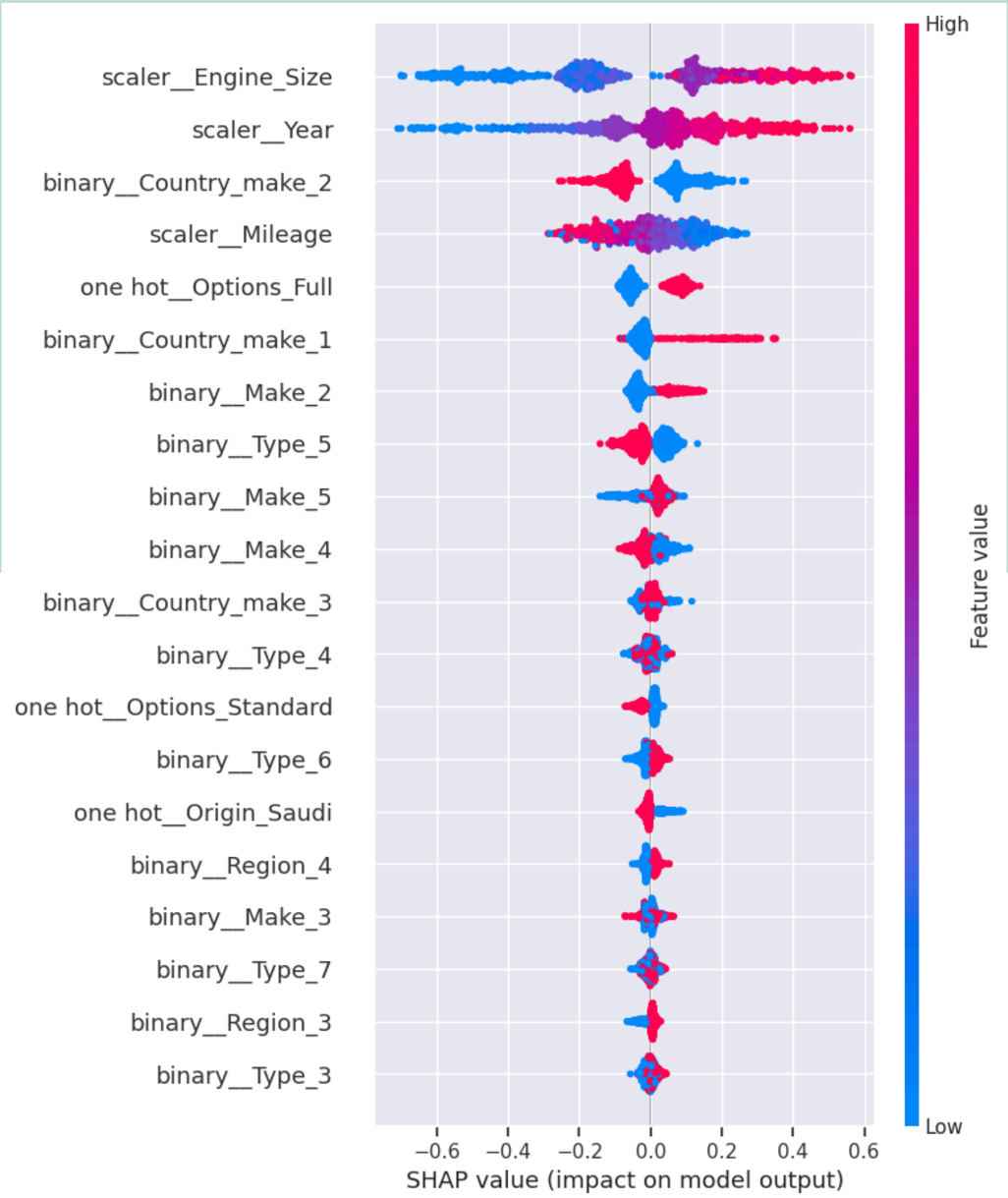
XGBoost is an **ensemble learning method**, meaning it builds **multiple decision trees** and **combines their outputs** to make more accurate predictions. Specifically, it uses boosting, a technique that builds trees sequentially, where **each new tree corrects the errors of the previous ones**.

Extreme Gradient Boosting

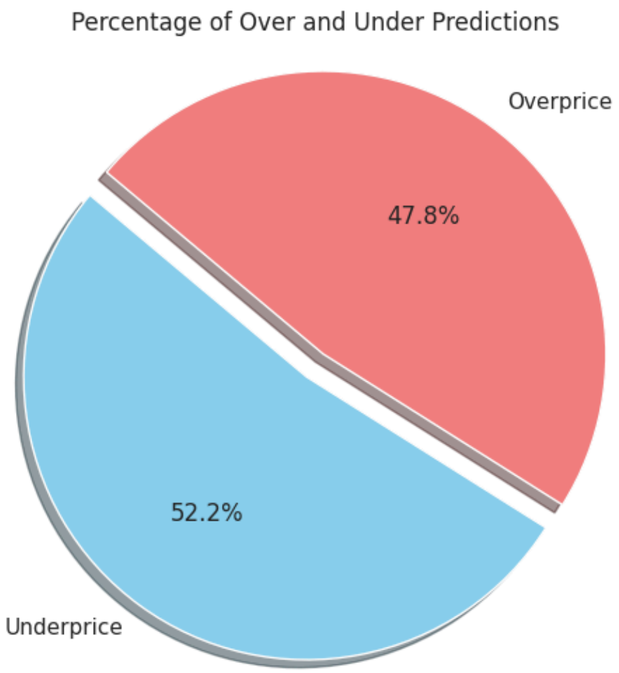


In essence, XGBoost gradually refines its predictions by combining the outputs of multiple trees in an additive manner.

Model Interpretation



Machine Learning as Solution



Calculations (Hypothetical Scenario):

- Lets assume for every 1000 Car sold at syarah.com, 522 Cars are underprice and 478 overprice. If the overprice compensate the underprice, there is stll 44 Cars that sold under its fair price based on our model.
- If we use mean average error as price difference (14.185 SAR), we can calculate the potential of profit because of underprice cars sold.
- If we assume that those 44 cars price is same as average Price at syarah.com = 90.014 SAR. Means that cars should be worth = 90.014+14.185 =104,199 SAR.
- Without Price Recommendation from Machine Learning Model :
  3% x 90.014 x 44 = 118.818,48 SAR
- With Price Recommendation from XGBoost Tuned Model :
  3% x 104,199 x 44 = 137.542,68 SAR
- By using machine learning model, we can generate 18.724 SAR per 1000 cars sold.

Beyond commissions, improved user experience due to accurate pricing can lead to more platform usage and potentially higher revenue from other services like featured listings or value-added services

06.

Model Deployment on Cloud

CAPSTONE PROJECT 3

07.

Conclusions and Recommendations

Conclusions

- XGBoost Regression after Hyperparameter Tuning model offers best prediction for used cars price. with mean absolute percentage error (MAPE) 15.78%.
- We can focus on feature Engine Size, Year, and Country Maker as those features are giving most impact to our prediction result. Certain Engine Size, Maker/Brand, and Manufacture Year are best choice for most buyer.
- Implementing ML models to predict used cars price may increase our revenue from shared commission by 18.724 SAR / 1000 Car sold.

Recommendations

- Analyze more dataset and including more features to get better insight and train better model.
- Analyzing trends and the needs of the majority of Arab society for certain features can help companies sell used vehicles that are indeed the people's favorites.
- Appliying machine learning to predict/give recommendation to used car seller, may lead to overall price increase. We can offer additional services like preselling inspection and after sales guaranty for the used cars sold from our platform.

Thank you