Maximizing Used Car Sales Value in Saudi Arabia through Price Prediction with Machine Learning

Chandra Driastama

Aug 30, 2024

Executive Summary

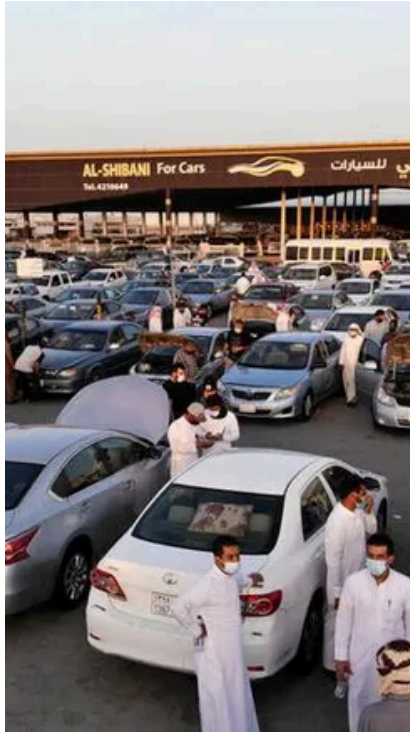
- The Importance of Accurate Pricing:
  - Mistakes in pricing can result in longer used car sales, increased storage costs, and decreased profitability.

- Opportunities for Optimization:
  - Although the current pricing is better than the market average, there is still a lot of room for further optimization.
  - Of the total 2,940 cars analyzed, 587 were sold below market value, resulting in a potential loss of around 20% of the actual value.

- Prediction Focus:
  - Predictive models highlight variables such as Year,Make, Engine Size, and Mileage of the car as key predictors of pricing.

Bussines Understanding

The used car market in Saudi Arabia is highly competitive, with many sellers and buyers transacting every day. Accurately pricing a used car is critical for both sellers, who want to maximize profits, and buyers, who are looking for fair value for their money. Mistakes in **pricing** can lead to prolonged sales cycles, inventory carrying costs, or lost revenue opportunities. In this market, understanding the factors that influence car prices can help dealers, individual sellers, and online car marketplaces set competitive prices, attract buyers, and accelerate sales. **Accurate pricing models** can also help buyers make more informed purchasing decisions.



What should we do now to maximize used car selling price predictions?

Goal & approach

Goal

- develop a predictive model that can accurately estimate the selling price of used cars in Saudi Arabia based on available data

Approach

1. **Analyze all data** to identify patterns of existing features and differences between one car and another.
2. **Build a regression model** that aims to help companies provide used car price prediction tools.
3. Selecting a regression model will be based on the **best metric evaluation** so that the most optimal final machine learning model can be determined.

Success measures

- Machine learning metric
  - MAE score: Evaluates how effectively our The model identifies the selling price of the car with an accuracy level approaching 0 SAR. It can be ascertained that the model works very accurately.

Agenda

1 Data Understanding

2 Data Preprocessing

3 Modelling

4 Summary and recommendations

Columns Describe

Kolom	Penjelasan
Type	Jenis mobil bekas.
Region	Wilayah tempat mobil bekas tersebut ditawarkan untuk dijual.
Make	Nama perusahaan atau merek mobil.
Gear_Type	Jenis atau ukuran transmisi mobil bekas.
Origin	Asal mobil bekas tersebut (misalnya, negara asal).
Options	Fitur atau opsi yang dimiliki oleh mobil bekas.
Year	Tahun pembuatan mobil.
Engine_Size	Ukuran mesin mobil bekas.
Mileage	Jarak tempuh yang telah dilalui oleh mobil bekas.
Negotiable	Menunjukkan apakah harga mobil dapat dinegosiasikan (True jika harga 0 berarti dapat dinegosiasikan).
Price	Harga mobil bekas.

Agenda

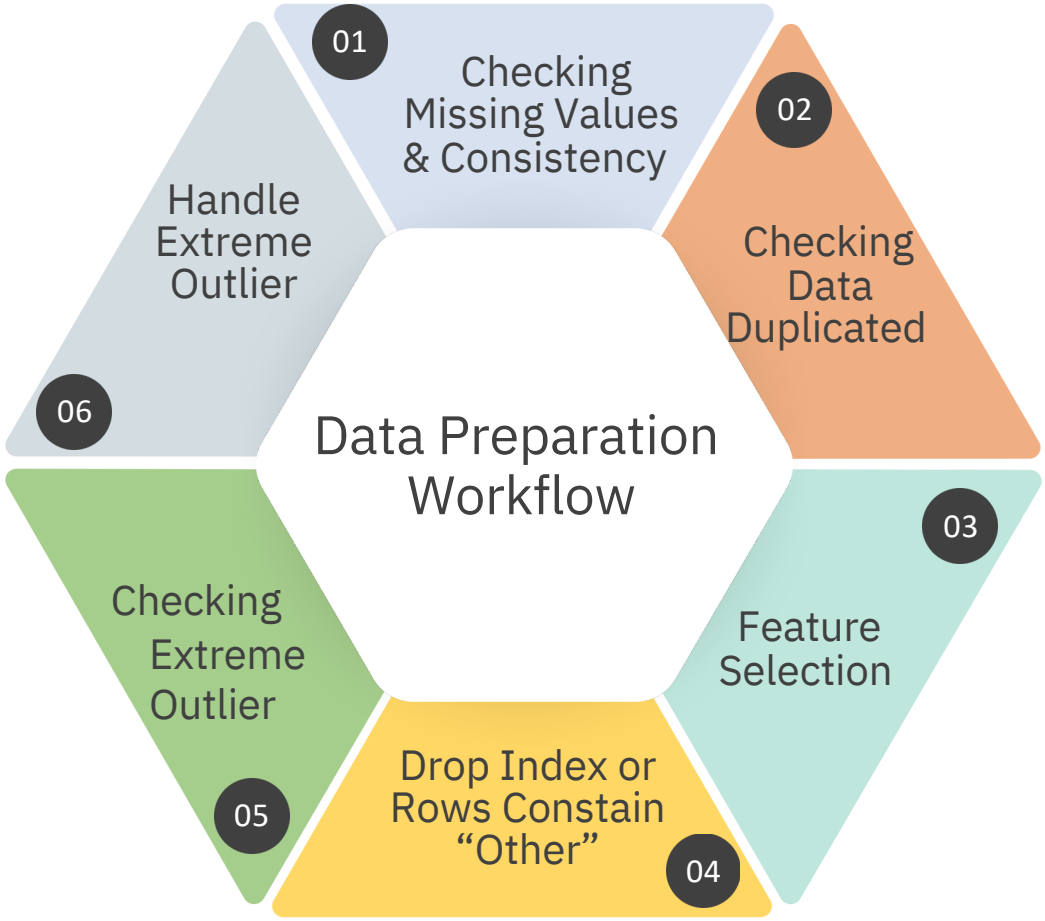
1 Data Understanding

2 Data Preprocessing

3 Modelling

4 Summary and recommendations

We make sure the data is ready for solution development



Agenda

1 Data Understanding

2 Data Preprocessing

3 Modelling

4 Summary and recommendations

Modeling in machine learning using regression

Encoding Feature

	Data Features	Data Types	Unique	Unique Sample
0	Type	object	268	[Victoria]
1	Region	object	27	[Qassim]
2	Make	object	49	[Foton]
3	Gear_Type	object	2	[Automatic]
4	Options	object	3	[Standard]

Key highlights

- **Binary Encoder** is used for categories with a large number of unique items.
- **One Hot Encoder** is used for categories with a relatively small number of unique items that can be counted on the fingers of one hand.

Modeling in machine learning using regression

Evaluating Model

	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE	Mean_RMSLE	Std_RMSLE	Mean_R2	Std_R2
0	Linear Regression	-24973.955675	1983.260269	-17051.258170	1042.652443	-0.281399	0.008396	0.347752	0.006862	0.577837	0.064320
1	KNN Regressor	-23575.441951	1592.805680	-16324.533552	1141.565903	-0.308872	0.015210	0.370393	0.008670	0.626353	0.032014
2	DecisionTree Regressor	-26864.128625	733.336762	-17128.116471	734.746055	-0.297857	0.006549	0.398849	0.008763	0.511134	0.060160
3	RandomForest Regressor	-19176.029126	1507.449091	-12428.387987	826.686037	-0.209419	0.002429	0.286794	0.005327	0.751521	0.035935
4	XGBoost Regressor	-17786.330782	869.877542	-11793.034285	592.152823	-0.201794	0.009766	0.277323	0.004127	0.786753	0.019170

	RMSE	MAE	MAPE	RMSLE	R2
XGB	16706.699798	11324.539724	0.218446	0.288821	0.792308
RandomForest	17267.507002	11678.579414	0.233236	0.297894	0.778131

Key highlights

- **Extreme Gradient Boost** is the best model in this table, showing superior performance in terms of accuracy and fit to the data. **Random Forest** is also a good choice, while models such as Linear Regression and KNN Regression show much lower performance and may be less effective for this dataset. Therefore, I chose Extreme Gradient Boost and Random Forest Regressor for my comparison.

Modeling in machine learning using regression

Compare Performance

- **Performance Comparison**

Berikut perbandingan hasil Datatest sebelum dan setelah di tuning

```
# Sebelum hyperparameter tuning
print('Skor sebelum di Tuning:' )
pd.DataFrame(score_before_tuning.loc['XGB']).T
```

✓ 0.0s

Skor sebelum di Tuning:

	RMSE	MAE	MAPE	RMSLE	R2
XGB	16706.699798	11324.539724	0.218446	0.288821	0.792308

```
# Setelah hyperparameter tuning
print('Skor setelah di Tuning:')
score_after_tuning
```

✓ 0.0s Open 'score_after_tuning' in Data Wrangler

Skor setelah di Tuning:

	RMSE	MAE	MAPE	RMSLE	R2
XGB	15861.869349	10855.717038	0.218178	0.276898	0.812782

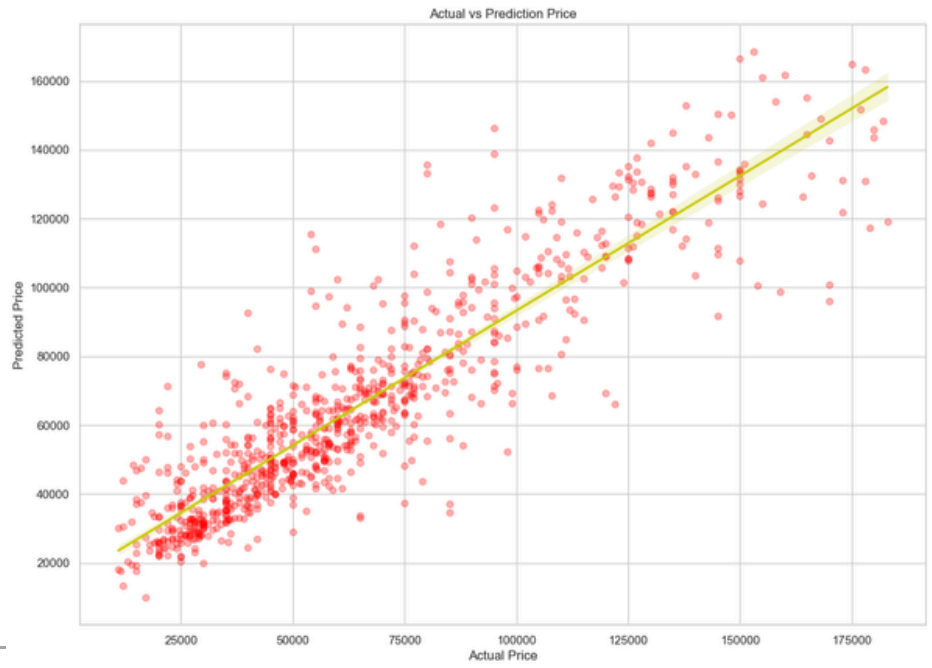
Key highlights

- After hyperparameter tuning, the performance of the XGBoost model experienced a **significant increase**, marked by a decrease in error values (RMSE, MAE, RMSLE) and an increase in R2 values. This shows that the model is more **accurate in predicting** used car prices after tuning.

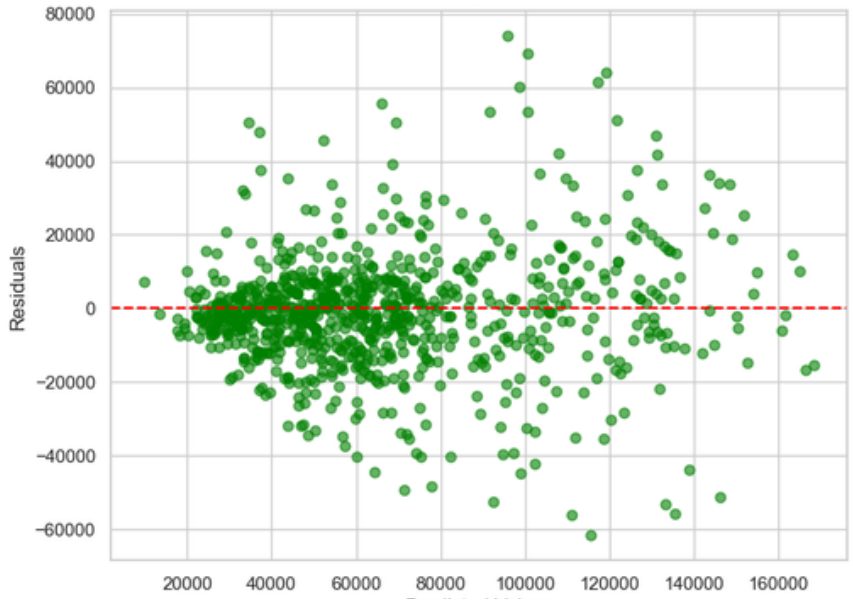
Actual vs Predicted Price Plot & Residual Plot

Residual Plot and **Actual vs Predicted Price Plot** are used to evaluate the performance of the used car **price prediction model**. Residual Plot shows the **difference between predicted** and actual values, with points randomly scattered around the horizontal line, indicating minimal bias but some outliers. Meanwhile, Actual vs Predicted Price Plot shows how close the model's predictions are to the actual prices; most points lying around the diagonal line indicate that the model has good accuracy, although there is variability that is not fully explained by the model.

Actual Price vs Predicted Price

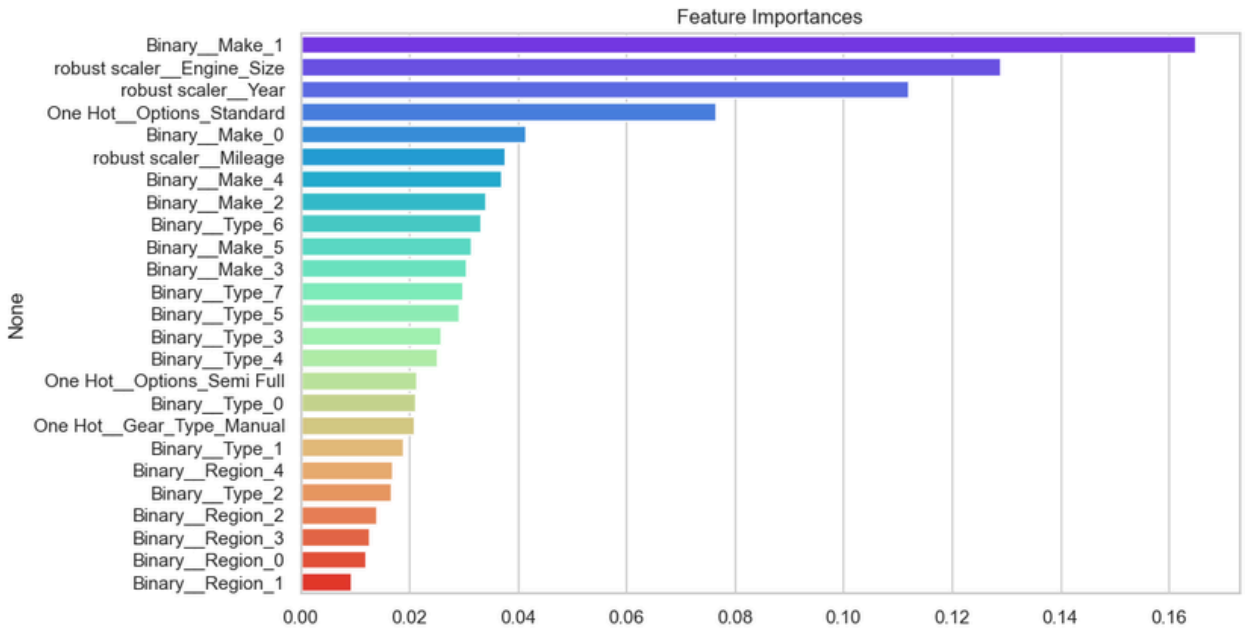


Residual Plot



Modeling in machine learning using regression

Feature Important

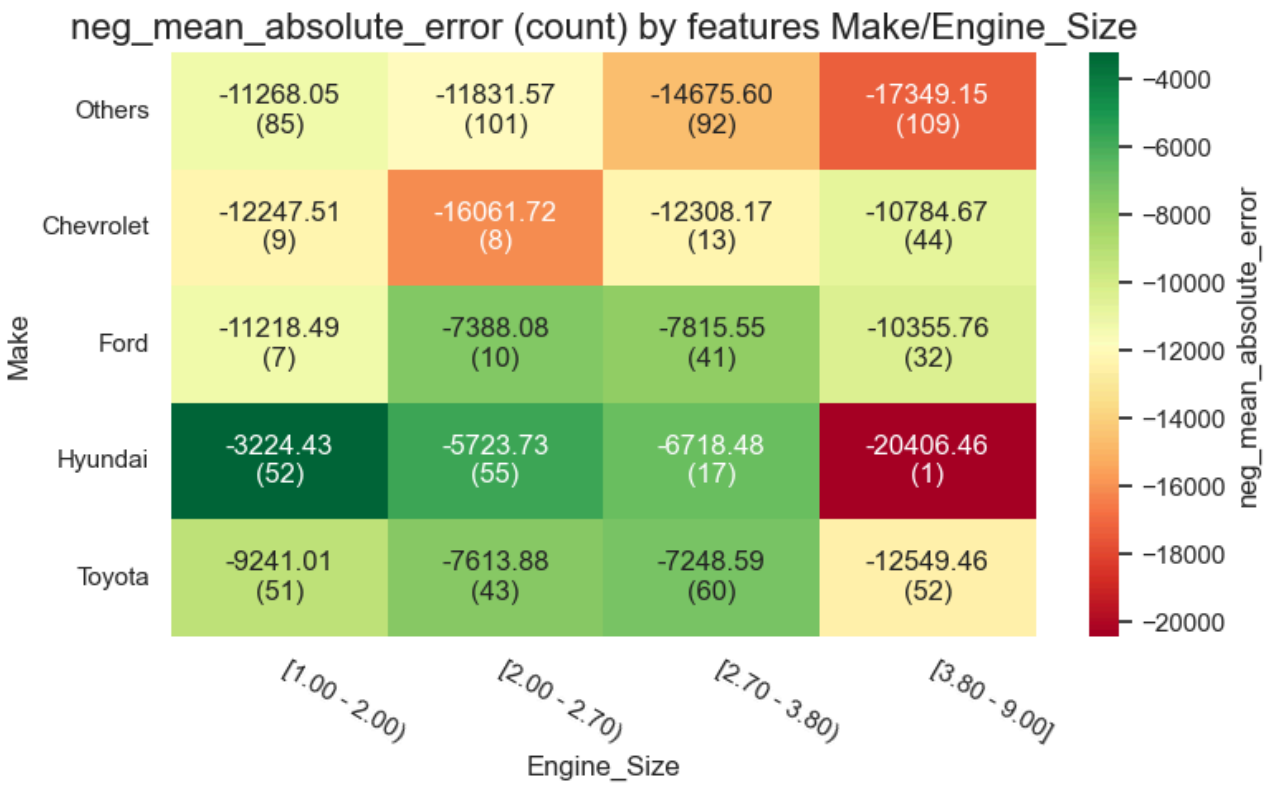


Key highlights

- The most important **features** that contribute to the price prediction are "**Make_1**", "**Engine Size**", and "**Year**". This shows that the car brand, engine size, and year of manufacture are the main factors to consider in determining the selling price of a used car.

Modeling in machine learning using regression

Segment Performance



Key highlights

- The model tends to be more **accurate in predicting prices for Hyundai and Toyota brands**, especially for smaller engine sizes (1.00-2.00). However, predictions become less accurate for larger engine brands, such as Chevrolet and “Others,” indicating a need for model improvement in this segment.

Agenda

1 Data Understanding

2 Data Preprocessing

3 Modelling

4 Summary and recommendations

Summary

Section	Findings
Objective	Utilizing predictive models to predict used car prices with a price limitation of 11,000 to 183,000 SAR. This model can be used as supporting information in determining selling or buying prices for companies or individuals who will conduct used car buying and selling transactions.
Feature Importance	The Make feature proved to be the most influential factor in determining the price of used cars in Saudi Arabia. In addition, engine size and year of manufacture also had a significant impact on the price.
Best Modeling	The selected model is XGBRegressor
Evaluation Model	Based on the evaluation, this model shows an average error of 10.855 SAR based on MAE or about 21% based on MAPE, indicating that the price prediction can deviate this far from the actual price.

Four strategic advice maximizing car sales use machine learning

High

Recommendation	Objective	Rationale
Optimization model	Improving the accuracy of used car pricing through optimal predictive models.	With better optimization models, companies can set prices that are more in line with market value, which can reduce time to sale and increase profit margins.
Feature Expansion	Expanding the use of new features in machine learning models to improve the accuracy of price predictions.	Adding features such as vehicle condition, regional pricing trends, and car repair history can provide a richer information model to make predictions that are more accurate and relevant to market conditions.
expansion of integration features with sales systems	Integrate predictive models with sales systems more broadly to support automated pricing decision making.	With tighter integration, the sales system can automatically adjust prices based on the latest predictions, ensuring prices always follow market dynamics without the need for significant manual intervention.
monitoring and updating	Monitor model performance periodically and make updates to the model to maintain prediction accuracy as market trends change.	The used car market can change rapidly, and without regular monitoring and updating, models can become less effective. Monitoring allows companies to respond quickly to changes and maintain a competitive edge.

Low

Thank You!

chandra879012