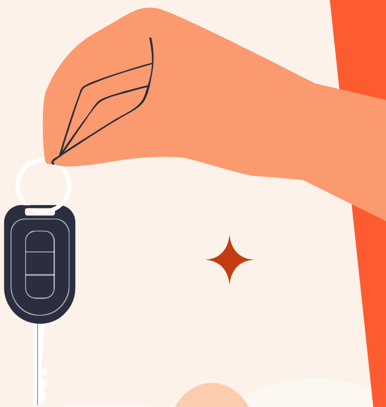
Used Car Price Prediction

Fadhiil Dzaki Mulyana



01

Business Understanding

Background

The automotive industry is one of the largest in the world, with the used car market playing a significant role. Understanding the factors that influence the pricing of used cars is crucial for various stakeholders, including buyers, sellers, and dealerships. Factors such as mileage, brand, model, count of previous owners, condition, ext. can all impact the price of a used car. Analyzing these factors through regression analysis can provide valuable insights into pricing trends and help stakeholders make informed decisions.

Goal

create a robust regression model that accurately predicts the prices of used cars.

problem

we dont know yet the best method that accurately estimates the price of used cars based on various attributes.

Use full for:

- Dealer
- Car Entushiast
- People who wants to buy a car

Objektive

- Data Visualization.
- End-to-End Regression Project.
- Hyperparameter tuning.

02

Used Cars Prices in UK Dataset

About Dataset

Used Car Prices in UK Dataset is a comprehensive collection of automotive information extracted from the popular automotive marketplace website, autotrader.co.uk. This dataset comprises 3,685 data points, each representing a unique vehicle listing, and includes thirteen distinct features providing valuable insights into the world of automobiles.

```
duplicated values: 826
```

[Automobile Dataset (kaggle.com)]

```
missing values: 128.79%
Previous Owners    1409
Engine               45
Doors                25
Seats                35
Emission Class       87
Service history    3145
```

14 COLUMNS X 3685 ROWS

Unnamed: 0	-
Title	Brand & Model
Price	Sale price (pounds.)
Mileage	Travelled distance (miles)
Registration year	officially registered year
Previous owners	Count of previous owners
Fuel type	Type of fuel used
Body type	General vehicle's shape
Engine	Engine's displacement
Gearbox	Transmission type
Doors	Count of doors
Seats	Seating capacity
Emission class	Standard emission (Euro)
Service history	Service completion

Important Columns	Unsused Columns
• Registration year • Mileage • Price	• Unnamed: 0 • Service history

03

Data Preprocessing

Preprocessing Step

01

Data Cleaning

02

Data Transformation

03

Feature Engineering

04

Feature Selection

05

Scaling

Data Cleaning

Unused Features

- Drop service history and unnamed: 0 columns.

```
columns: 14
```

```
columns: 12
```

Duplicated Value

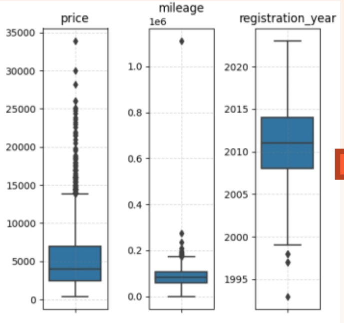
- Drop duplicate

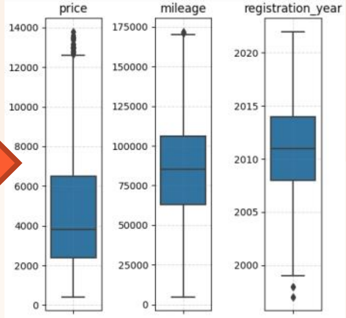
```
duplicated values: 826
3685 baris
```

```
duplicated values: 0
tersisa 2859 baris
```

Outliers

- Delete with IQR Method






```
tersisa 2612 baris
```

Missing Value

dropna

KNN Imput

```
missing values: 39.14%
previous_owners    34.38
engine              0.59
doors               0.87
seats               1.22
emission_class      2.06
```

```
missing values: 34.16%
previous_owners    34.16
tersisa 2793 baris
```

```
missing values: 0.0%
tersisa 2793 baris
```

Feature Engineering & Data Transformation

New Features:

Brand continent:
Extracted from Title feature.

- Europe
- Asia
- North America

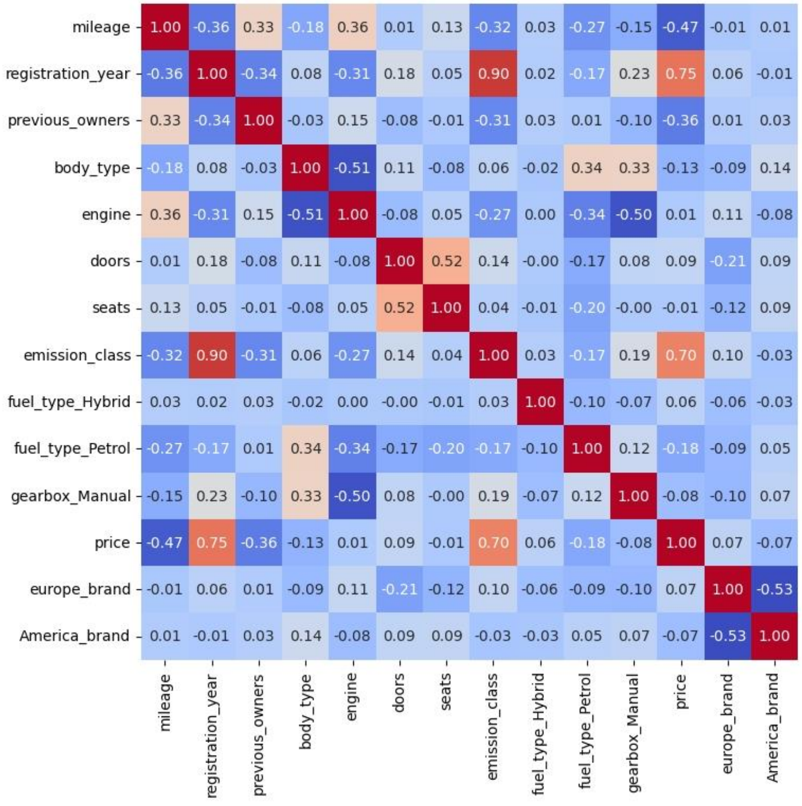
Regroup Fuel

Before	After
Petrol	Petrol
Diesel	Diesel
Electric	Electric
Petrol Hybrid	Hybrid
Petrol Plug-in Hybrid	
Diesel hybrid	

Convert Dtype

title	object	object
price	int64	float64
mileage	int64	float64
registration_year	int64	int64
previous_owners	int64	int64
fuel_type	object	object
body_type	object	object
engine	object	float64
gearbox	object	object
doors	float64	int32
seats	float64	int32
emission_class	object	object

Feature Engineering & Data Transformation



Feature Selection

Registration year and emission class seems to have high VIF score, indicates multicollinearity.

Registration year has more correlation with the target than emission class.

Drop emisission class.

Categorical Encode

Labeling: 2 categories

One-hot: >2 categories

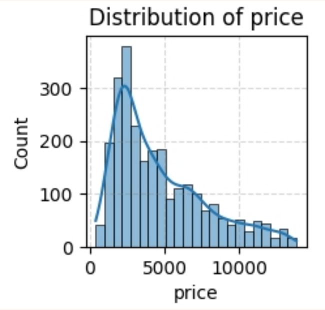
Count : >10 categories

```
mileage              float64
registration_year    float64
previous_owners      int64
body_type            float64
engine               float64
doors                float64
seats                float64
emission_class       float64
fuel_type_Hybrid     float64
fuel_type_Petrol     float64
gearbox_Manual       float64
price                float64
europe_brand         uint8
America_brand        uint8
```

03

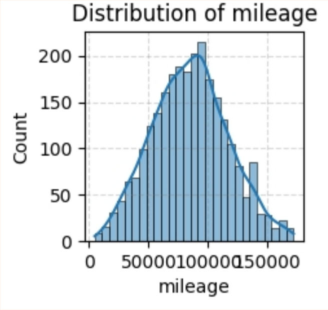
EDA

Continuous Features Distribution



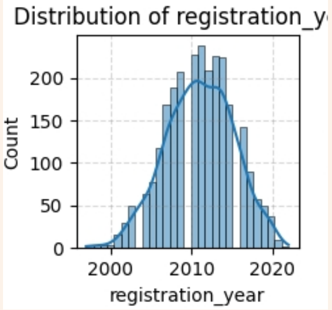
Price

The majority of vehicles price are low, but there are very expensive vehicles in the market. There could be special demand for rare vehicles, which could drive their prices higher than average.



Mileage

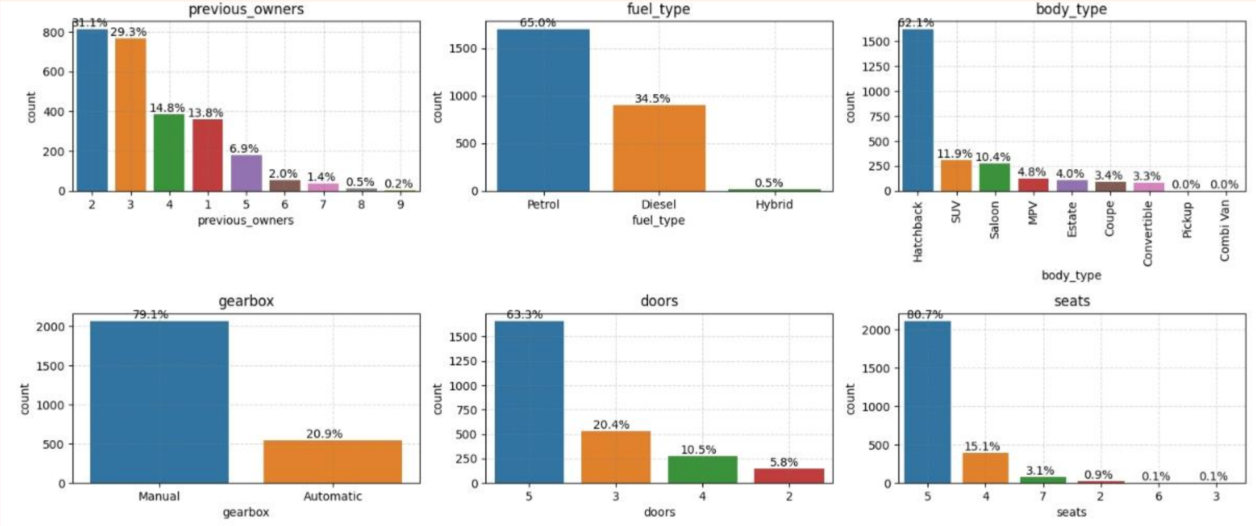
Most vehicles in the dataset have mileage values that are representative of typical usage patterns for vehicles of their type and age.

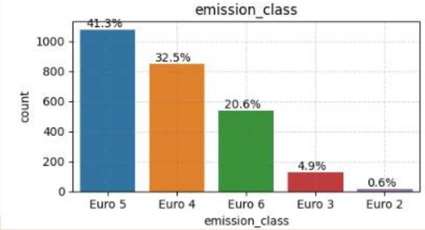


Reg. Year

most vehicles were registered evenly across different years, without a significant bias towards any particular time period.

Categorical Features

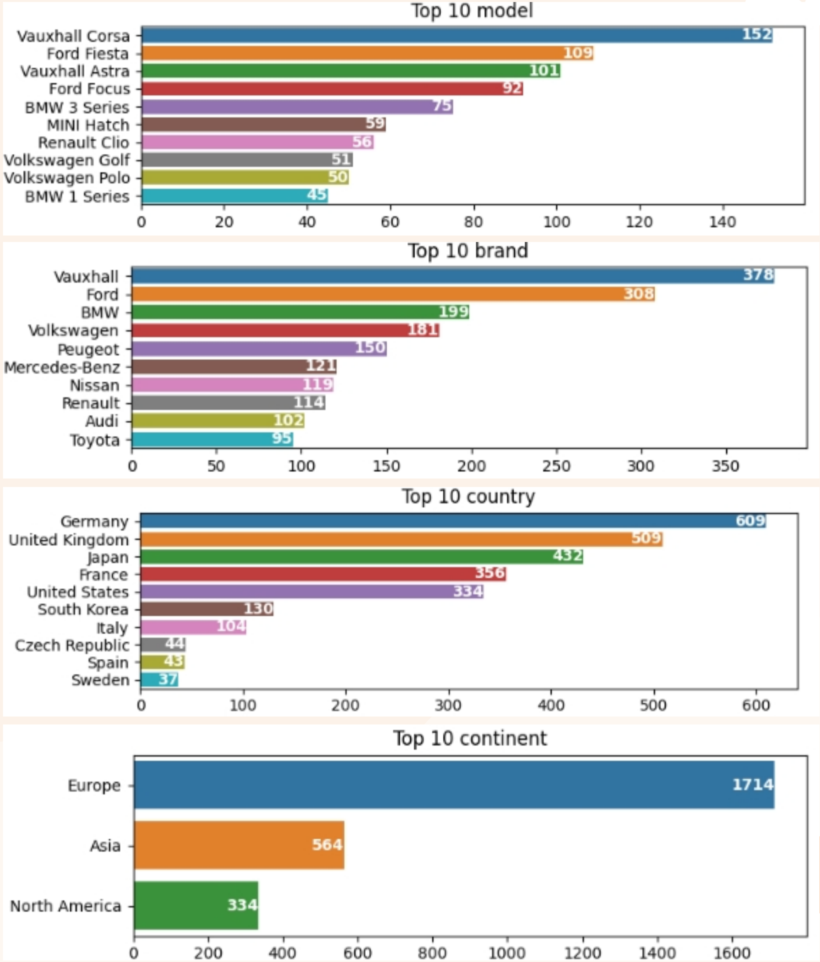




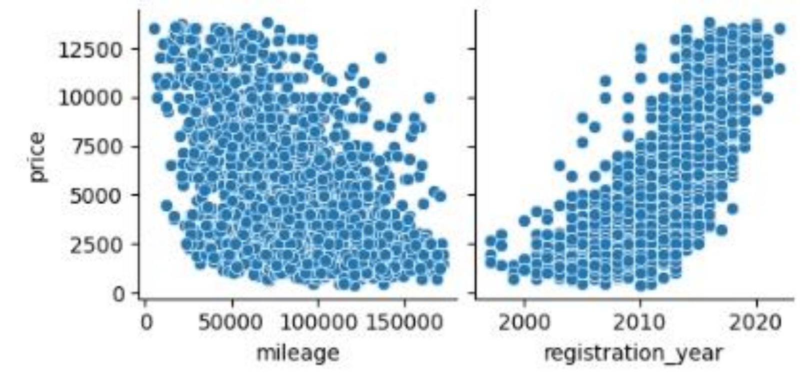
- Majority of features have imbalance distribution.
- Certain categories have very few instances like Hybrid fuel type, other than hatchback, 5 seater, and/or 5 doors.

Title Extraction Features

- The majority of vehicles are coming from Vauxhall with Corsa model.
- The majority of brands are European, as 6 of the top 10 brand countries are in Europe.



Price VS Continuous Features



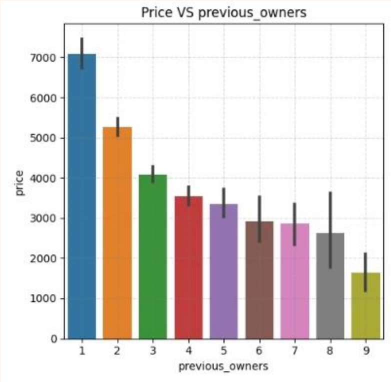
VS mileage

The further the distance travelled, the lower the price will be.

VS Reg. Year

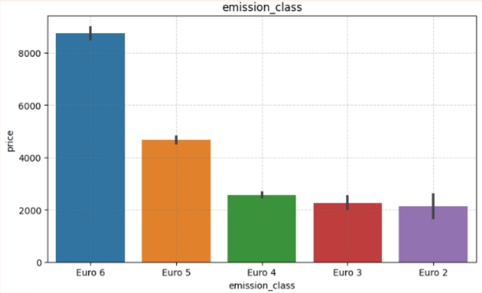
the newer the registration year, the higher the price will be.

Price VS Categorical Features



VS Previous Owner

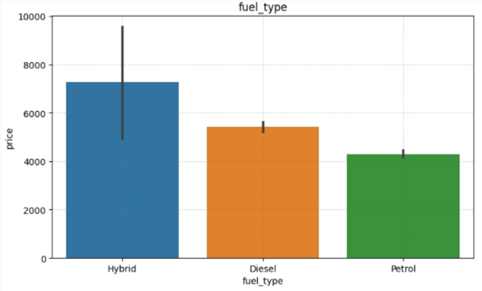
The more owners there are, the lower the price will be.



VS Emission Class

The higher the emission standard of a vehicle, the higher the price.

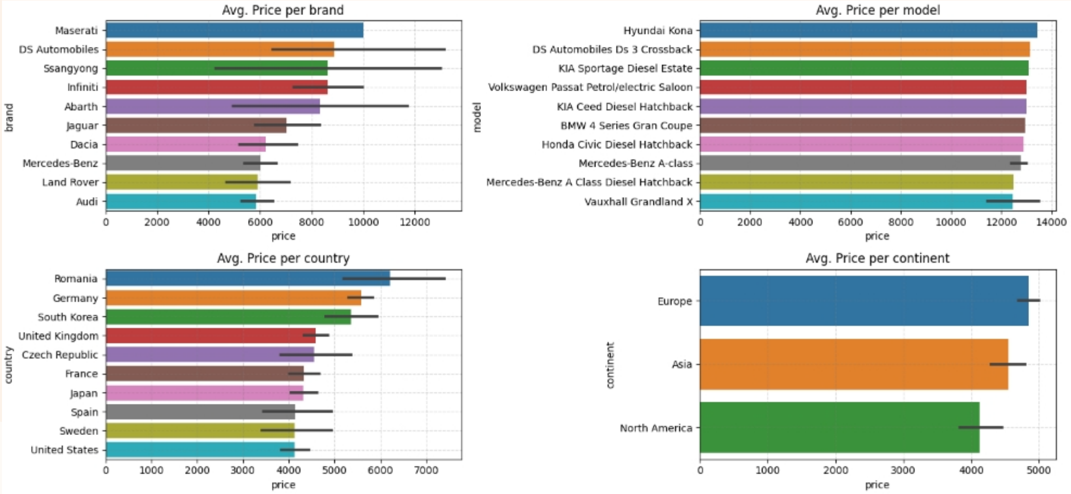
This may be due to the strong correlation between model year and emissions class.



VS Fuel Type

Ignoring Hybrid fuel, diesel vehicle has higher price than petrol.

Title Extracted columns



- The mosh expensive car in this dataset is Hyundai Kona, but Hyundai isn’t in top10 average brand price list. This may because Hyundai isn’t quite much and the price of other model is low.

- Europe has the highest price average and 7 of europe brands are in the top 10 brands country average prices.

As you can see, model column has less outliers than the other title extaracted columns, but 469 unique value are there (too much).

Brand or country has less unique value than model, but there are too many outliers.

Brand continent is the more stable features, because it has only 3 unique, the outliers not as much as country or brand, and can represent title very well.

Take a look at these plots

These are plots separated by transmission type.

Distribution between manual and automatic price is the same,

Most of the plots shows that automatic transmission has higher price than manual.

05

Machine Learning

Model Building

Model	R2		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Linear	0.739213	0.739583	1224.402536	1237.239468	2.455937e+06	2.415864e+06
Ridge	0.739212	0.739580	1224.382666	1237.202174	2.455939e+06	2.415894e+06
Lasso	0.739211	0.739553	1224.313579	1237.237101	2.455952e+06	2.416145e+06
ElasticNet	0.661870	0.663136	1379.822427	1381.488873	3.184306e+06	3.125055e+06
Decision Tree	0.999875	0.596637	1.385831	1381.013384	1.174466e+03	3.741965e+06
Random Forest	0.972173	0.786783	369.142032	1074.890739	2.620548e+05	1.977993e+06
Gradient Boosting	0.856403	0.813298	854.548768	1008.101459	1.352308e+06	1.732021e+06
Neural Network	-0.448340	-0.422354	2954.646243	2889.601282	1.363959e+07	1.319506e+07

Choosing The Best Model

- The Random Forest model has the highest test R2 (0.786783) and relatively low test MAE and test MSE compared to other models.
- Decision Tree model has exceptionally high R2 on the training set (0.999875), but it doesn't generalize well on the test set.
- Linear, Ridge, Lasso and ElasticNet have a stable score on test and train data for all metrics, but the score is very low.
- Neural Network seems to perform the worst based on all metrics.

Gradient Boosting	0.856403	0.813298	854.548768	1008.101459	1.352308e+06	1.732021e+06

Best Model

The **Gradient Boosting** model does indeed have strong performance, especially in terms of test R2 (0.813298). It also has relatively low test MAE and test MSE compared to many other models.

The **Gradient Boosting** model captures about 81.33% of the variability of the unseen data in the target variable, which is considered quite good.

Hyperparameter Tuning

Model	R2		MAE		MSE	
	Train	Test	Train	Test	Train	Test
Gradient Boosting	0.856403	0.813298	854.548768	1008.101459	1.352308e+06	1.732021e+06
Gradient Boosting (Tuned)	0.889526	0.821414	762.330661	977.320034	1.040374e+06	1.656728e+06

- Overall, hyperparameter tuning seems to have enhanced the Gradient Boosting model's performance.
- The R2 score increased for both the training and test sets, indicating better model fit.
- The MAE and MSE decreased, which suggests improved accuracy and precision in predictions.

06

Conclusion & Recommendation

Conclusion

- Important Features: registration_year and mileage are the most important features to determine prices due to imbalance data in other features.
- Vehicles with lower mileage, newer year, fewer owners, European brands, and/or automatic transmission are more likely to have a higher price.
- Best model: Gradient Boosting Regressor with Hyperparameters. Accuracy increased, MAE and MSE decreased.

Note:
The available data may be incomplete or may not cover all factors that influence the price of used cars.

Recomendation

- Utilize the insights gained from the prediction model to adjust pricing strategies for used cars.
- Segment customers based on their preferences and predicted buying behavior using the insights from the prediction model.
- Establish a feedback loop to continuously monitor the performance of the prediction model and incorporate new data or insights to improve its accuracy over time.
- choose Vehicles with lower mileage, newer year, fewer owners, European brands, and/or automatic transmission to get more stable price.

ThakYou

linkedin.com/in/fadhiildzaki

https://github.com/FadhiilDzaki/used_car_price_prediction.git