What Drives Used Car Value?

Predicting Used Car Prices in 2022

Robert Davies
General Assembly DSI21

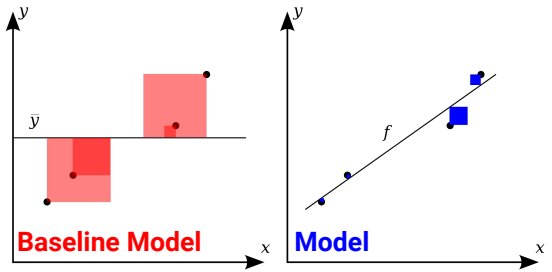
Goals

- The **primary** goal was to generate a predictive model for used car price in the UK in 2022, which is also interpretable.
- The **secondary** goal was to use the resulting model to test the hypothesis:

  When all other car attributes are equal, a Dacia branded car is cheaper than a Volvo branded car.

Success Criteria

- The **primary** metric used for model evaluation was **R^2 score**.
  - This can be interpreted as the fraction of target (car price) variance explained by the model which was not explained by the baseline (average) model.
  - An R^2 score of 1 is perfect prediction, whilst an R^2 score of 0 equates to a model equivalent to the average.
  - R^2 score is a unitless metric.



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- The **secondary** metric of **RMSE** (Root Mean Square Error) was used to further describe model performance.
  - This is defined as the square root of mean squared errors.
  - As error values are squared, larger errors are penalised more than smaller errors.
  - The units of RMSE are in £.

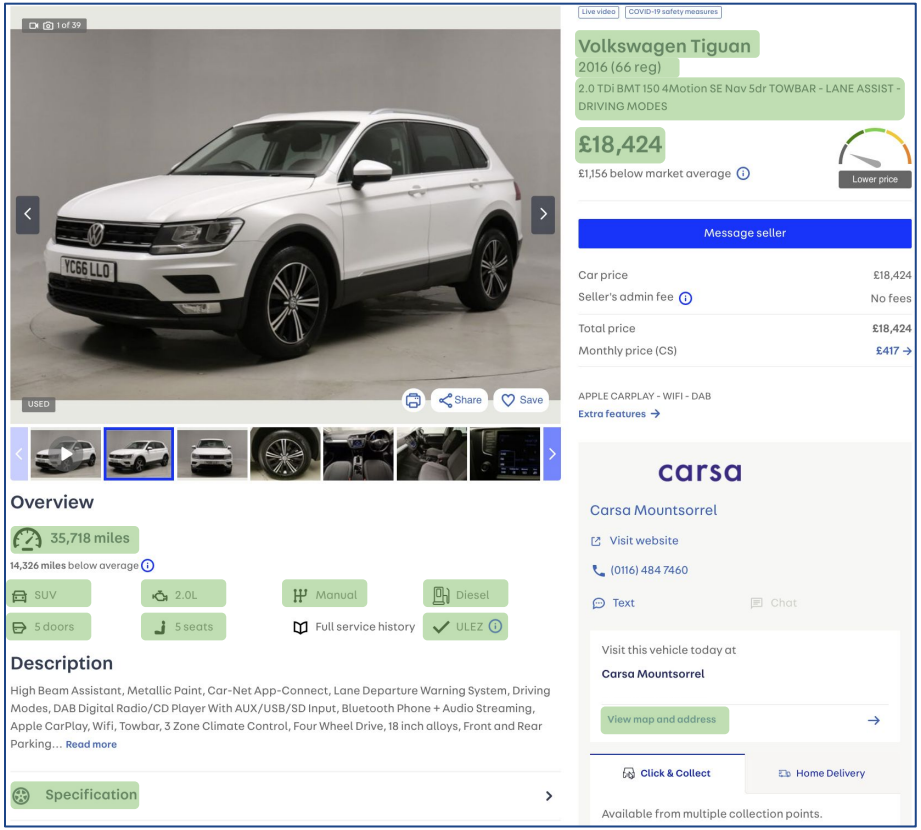
The Data - Where It Comes From

- 400,247 New and Used cars were scraped initially.
  - **369,730** Used cars at input to modelling after data cleaning
- Scraping problems encountered:
  - Limited search results
  - Cloudflare website security
- For each car, effort was made to scrape the data below. Ultimately, not all of it was used in modelling.

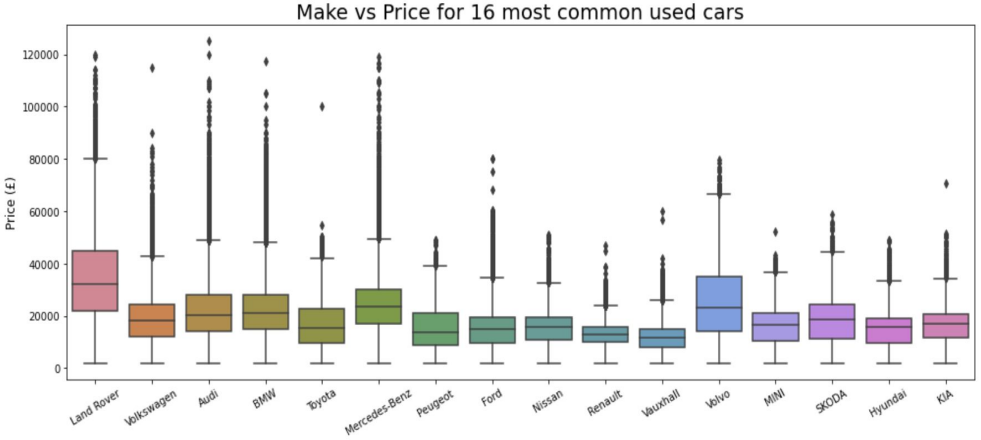
Engine Size, Drivetrain, MPG, Listing ID, Fuel Type, ULEZ, Price, Brand, BHP, Dealer City, Mileage, Transmission, Seats, Year Made, Body, Owners, Dealer Rating, Doors, Listing URL

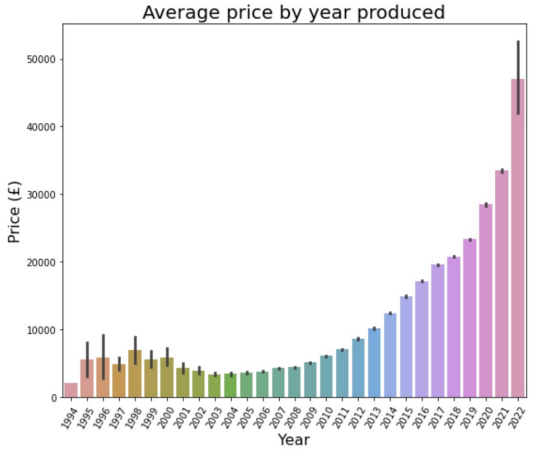
Used in Modelling | **Not Used in Modelling**

AutoTrader.co.uk

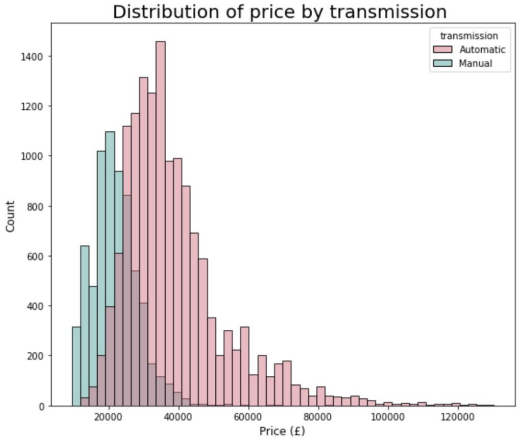


The Data - Brief Overview





- Land Rover, Mercedes-Benz and Volvo stand out to be the most expensive amongst the most common brands.
- It is evident which brands are targeting the high-end market and which are not.
- The relationship between year produced and price is not linear.
- Categorical predictors such as transmission may be useful in predicting car price



Selecting A Model

- Nine different models were tested on a random 50,000 car subset of the dataset.
- These models have been categorised into linear methods and tree based methods.

Linear Methods

Log-transform price variable

↓

80:20 train-test split

↓

Standardise

↓

Modelling with gridsearch and 5 fold cross validation

↓

Convert output predictions from log-price to price

↓

Evaluate success metrics on train and test data

Tree Based Methods

80:20 train-test split

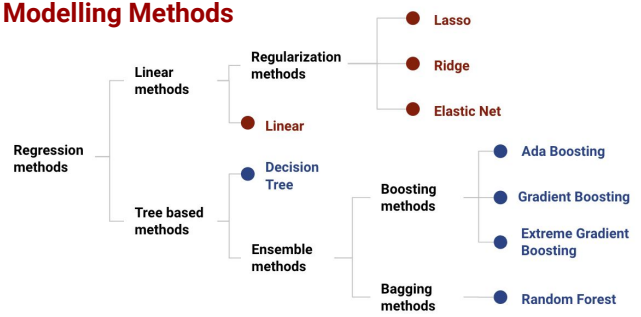
↓

Modelling with gridsearch and 5 fold cross validation

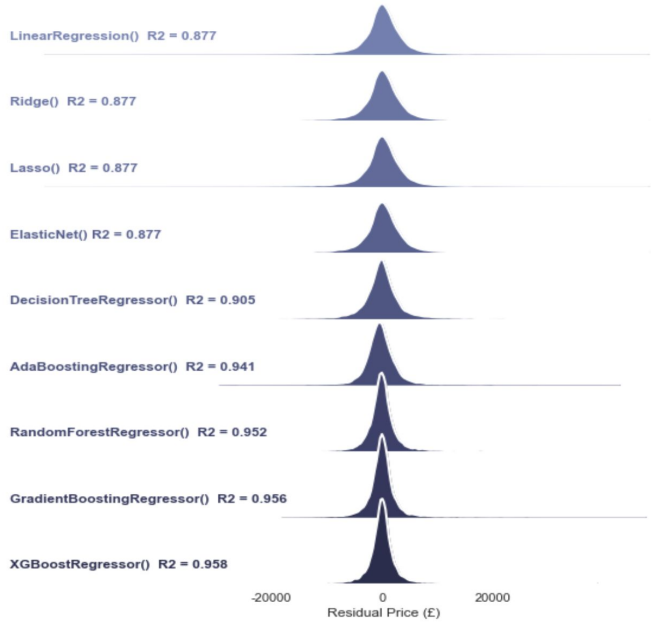
↓

Evaluate success metrics on train and test data

Modelling Methods

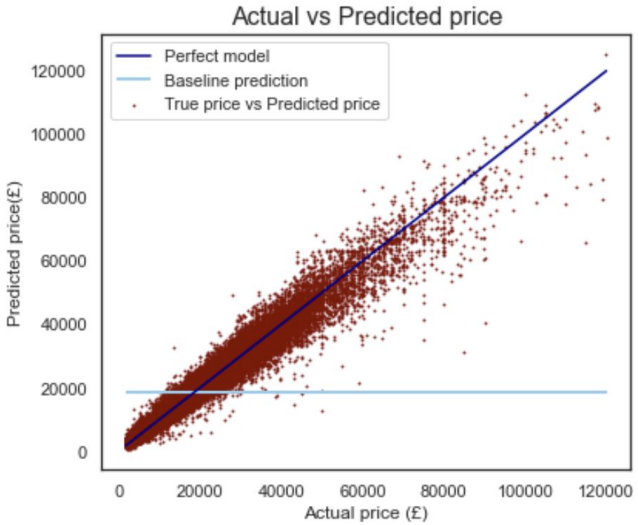


Price Residuals For Each Model Tested

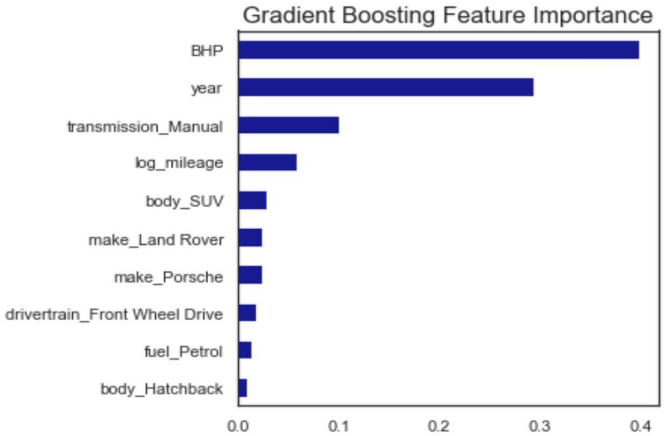


Gradient Boosting Model

- Gradient boosting regression is a which starts by simply predicting the average price of a car for every car and computing the price residuals.
- The model then uses a stepwise method to improve the result by iteratively modelling and reducing the residuals.
- After each step, new residuals are computed and the process repeats itself.
- The model is considered optimised when adding further steps fails to reduce the residuals.

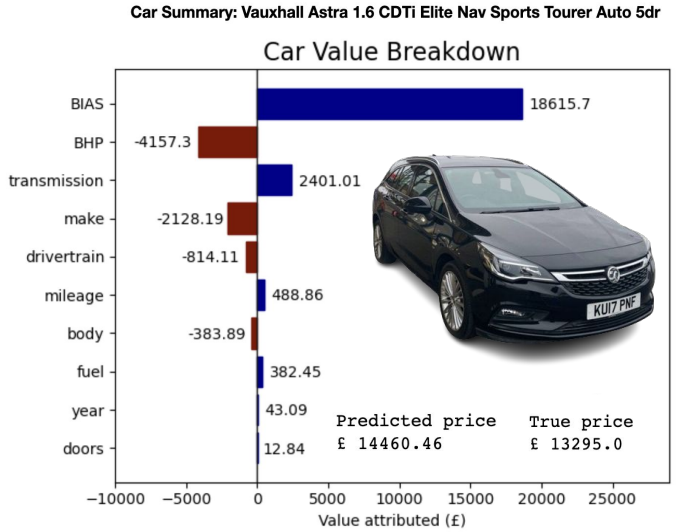


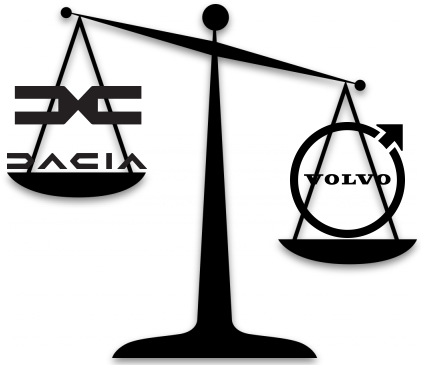
Metric	R^2 Score	RMSE (£)
Test data	0.954	2496.72
Train data	0.955	2553.55

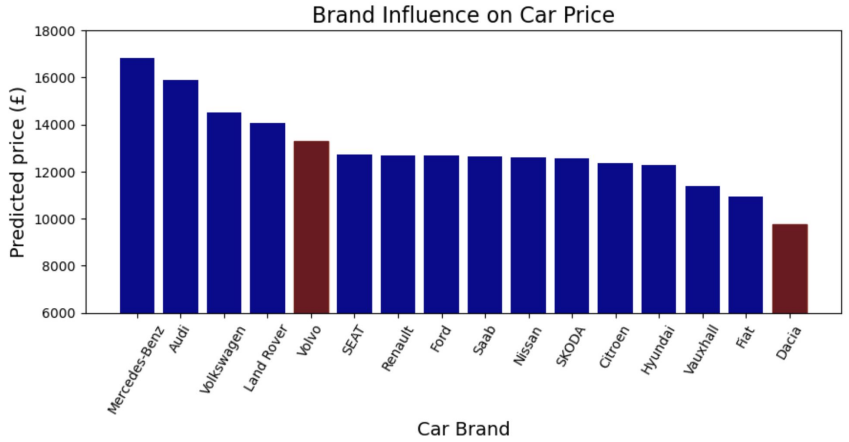


Under The Hood

- One minor problem with Gradient Boosting Methods is that they aren't easy to interpret in detail.
- The Explain it Like I'm Five (ELI5) package from Sklearn has been used to obtain an itemised value breakdown for any car
- This has been made possible through the computation of permutation importance rather than feature importance, or Gini importance from the model.
- The original hypothesis has been answered - A Dacia is indeed cheaper than Volvo







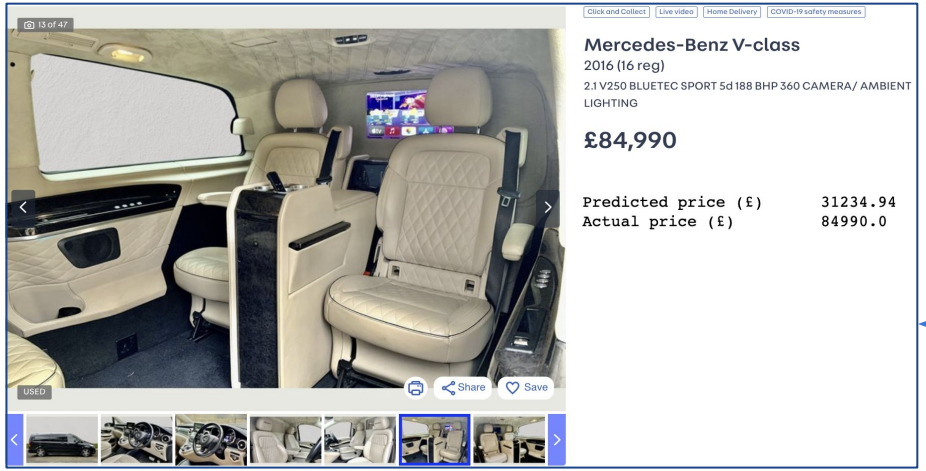
Risks And Limitations

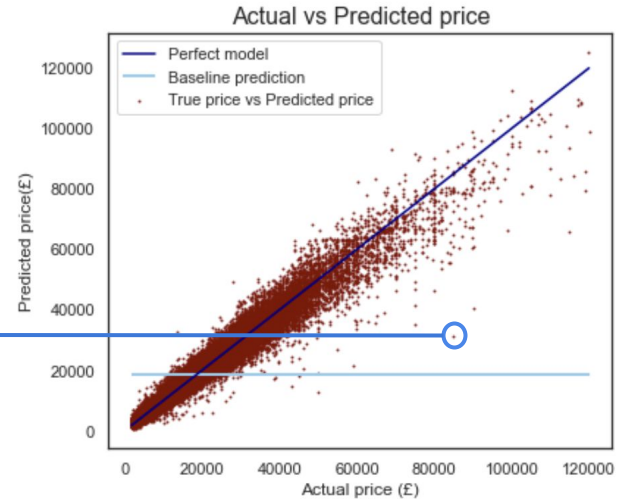
- The data may go stale quickly if the UK market changes.
- Outlier analysis has allowed me to sift out some erroneous rows caused by poor data entry. Non-remarkable data entry errors would harm the model but are difficult to identify.
- ‘Car spec’ is not captured in the dataset. Besides possibly the BHP variable, any non-standard upgrades are not captured.

Suspicious SKODA

Erroneous SKODA (?) entry	Description
name	skoda 911
name_subtitle	911 996 GT3 3.6 Coupe Manual Petrol 5dr
make	SKODA
price	119990.0
year	2004
body	Hatchback
mileage	6400.0
engine	1.9
BHP	105.0

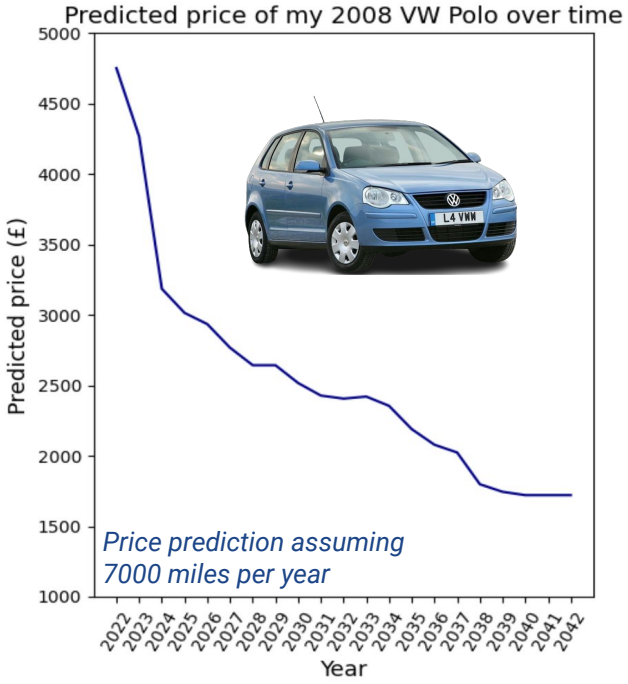
Largest Residual





Conclusions, Impact and Further Work

- All project goals have been met! With a 0.954 R^2 score it can be said that the used car model can be used to accurately predict used car price.
- Using permutation analysis with ELI5 has made it possible to explain why the Gradient Boosting Regressor does what it does.
- The model has been manipulated to rank common car brands in terms of influence on price and this could be extended to the other variables.
- With the model, we can predict the price of used cars in future.
- This model may be improved in future by implementing deep learning modelling methods.
- The model may be improved through enrichment of the predictor variables. It seems intuitive that car boot volume would impact the price of a car.
- It would be interesting to benchmark the model price predictions against AutoTrader's own value gauge. This isn't available on every listing and would require a new web scrape.

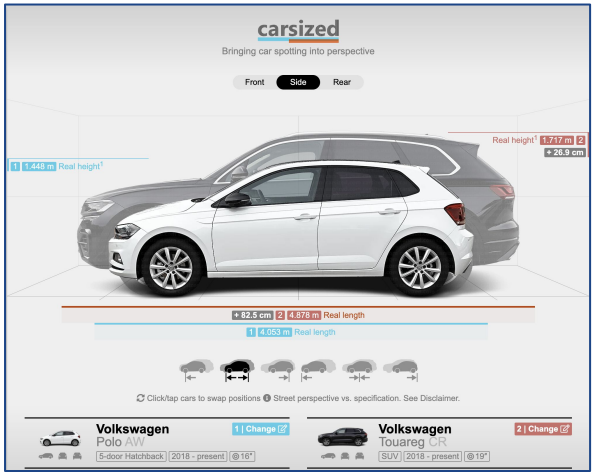


AutoTrader already have a feature which comments on the value for money!

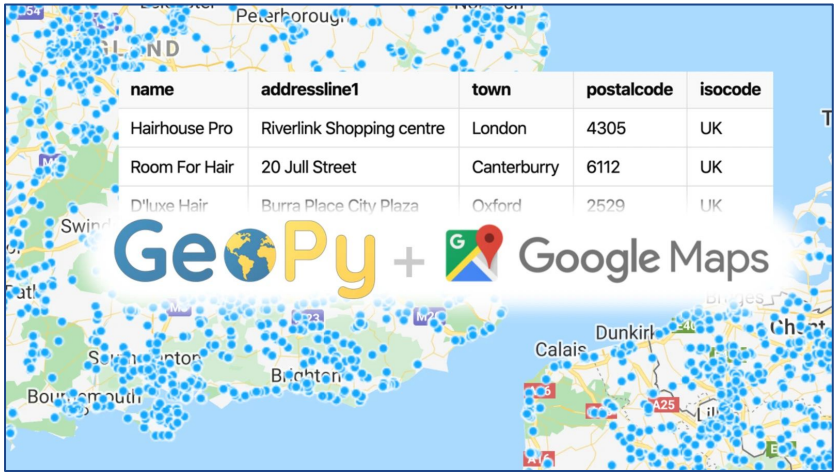
Further Work - Enriching Predictors

- Some of it has been done already. Additional sources besides AutoTrader have been used to enrich the predictors. This sequence is awful but reflects how it was done.

Carsized.com



GeoPy with Google Maps API



Fuel Type | Drivetrain | Price | Brand | Year Made
Dealer County | Mileage | Transmission | Doors | BHP | Cargo Volume | Body

Original | **Additional**

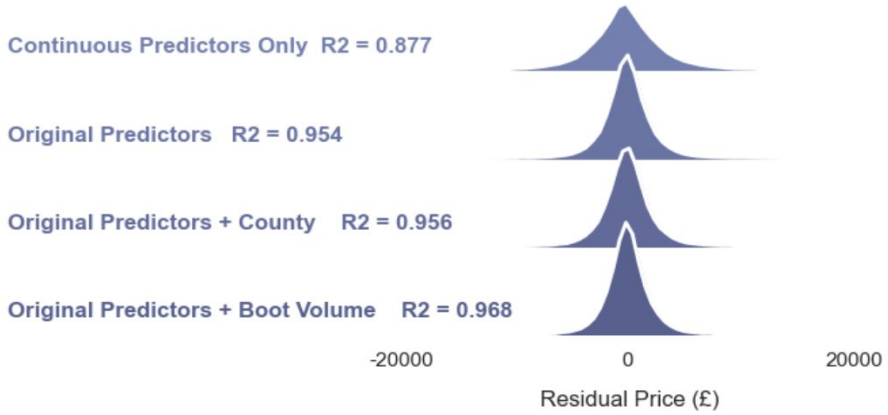
- **319,633** Complete Used Cars at input to further work
- 50,097 didn't have dealer county or cargo volume data

Further Work

Fuel Type | Drivetrain | Price | Brand | Year Made
Dealer County | Mileage | Transmission | Doors | BHP | Cargo Volume | Body

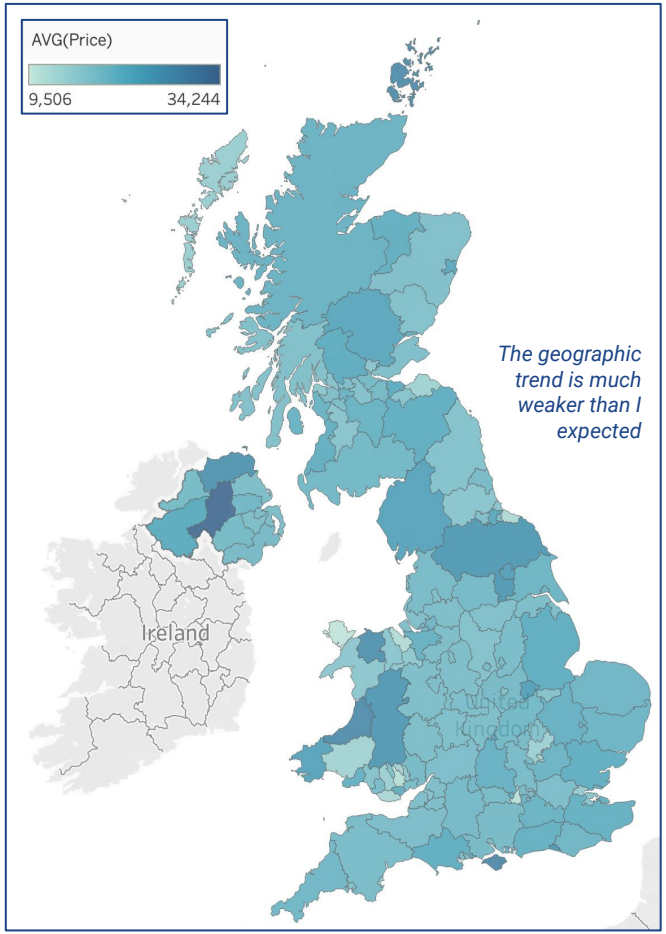
Original Continuous | Original Categorical | Additional

Price Residuals



Metric	R^2 Score		RMSE (£)	
	Original	Upgraded	Original	Upgraded
Test data	0.954	0.968	2477.30	2077.61
Train data	0.958	0.968	2356.69	2069.74

UK Counties Average Used Car Price Choropleth



The geographic trend is much weaker than I expected

Questions?


LGW 809G
GPF 146G
HMP 729G


TROTTERS
INDEPENDENT TRADING Co
BWC 94F


BMT 216A


01


53
53
OFP857

Car Value Breakdown with ELI5 - Example 1

Alternative Slide

y (score **35023.935**) top features

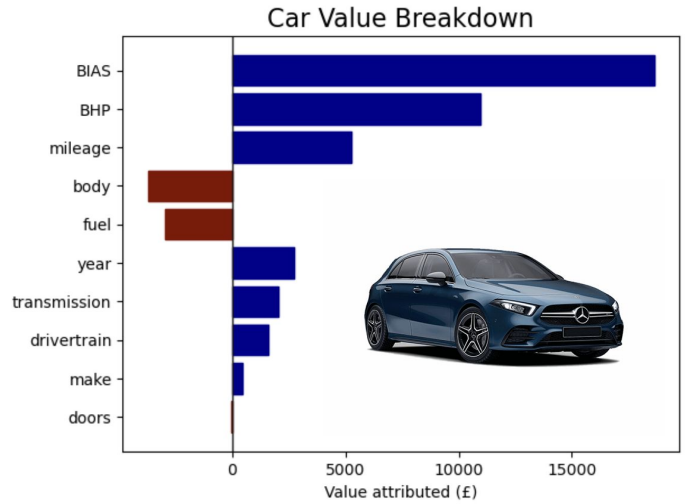
Contribution[?]	Feature	Value
+18627.367	<BIAS>	1.000
+10981.325	BHP	301.775
+5272.278	log_mileage	9.507
+2718.439	year	2019.000
+2213.293	make_Mercedes-Benz	1.000
+2037.989	transmission_Manual	0.000
+1472.747	drivertrain_Front Wheel Drive	0.000
	… 30 more positive …	
	… 16 more negative …	
-661.444	make_Porsche	0.000
-837.760	fuel_Diesel	0.000
-1214.711	body_Hatchback	1.000
-1220.384	make_Land Rover	0.000
-2073.431	fuel_Petrol	1.000
-2345.195	body_SUV	0.000

Predictor	Attribute	Price Contribution (£)	Cumulative Car Price (£)
BIAS	0	18627.367022	18627.367022
BHP	301.775148	10981.324640	29608.691662
mileage	13447.0	5272.278382	34880.970044
body	Hatchback	-3720.365950	31639.152757
fuel	Petrol	-2944.782282	28694.370475
year	2019	2718.438685	31412.809160
transmission	Automatic	2037.988511	33450.797671
drivertrain	Four Wheel Drive	1590.931155	35041.728826
make	Mercedes-Benz	478.548663	35359.518707
doors	5dr	-17.793867	35023.934959

Car Summary: Mercedes-Benz A Class A35 4Matic Premium Plus 5dr Auto 2.0

Predicted price	True price
£ 35023.93	£ 38398.0

https://www.autotrader.co.uk/car-details/202110028032322



Car Value Breakdown with ELI5 - Example 1

Alternative Slide

With permutation importance from ELI5 it is possible to quantify how each used car predictor impacted the final price prediction. The top row, labelled BIAS is the average car price.

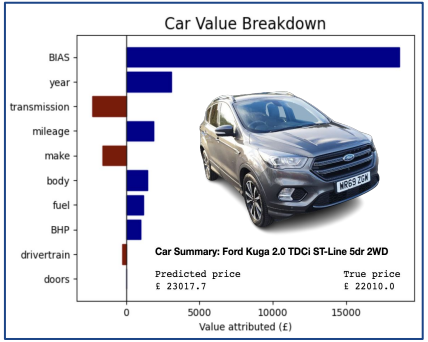
A function has been created to collate contributions from previously dummified categorical variables with accumulated price impacts displayed in a Pandas DataFrame.

Finally the car value breakdown is displayed in a horizontal bar chart with all ten predictor price contributions on the y-axis.

y (score **23017.701**) top features

Contribution[?]	Feature	Value
+18627.367	<BIAS>	1.000
+3100.195	year	2019.000
+1856.395	log_mileage	9.465
+1052.183	fuel_Petrol	0.000
+1037.665	body_Hatchback	0.000
+993.617	BHP	147.929
+403.330	body_SUV	1.000
	… 27 more positive …	
	… 17 more negative …	
-167.954	make_Mercedes-Benz	0.000
-198.167	make_Audi	0.000
-257.477	make_Land Rover	0.000
-285.174	drivertrain_Front Wheel Drive	1.000
-910.598	make_Ford	1.000
-2321.886	transmission_Manual	1.000

Predictor	Attribute	Price Contribution (£)	Cumulative Car Price (£)
BIAS	0	18627.367022	18627.367022
year	2019	3100.195227	21727.562248
transmission	Manual	-2321.885676	19405.676572
mileage	12898.0	1856.395476	19665.501516
make	Ford	-1596.570532	17809.106040
body	SUV	1469.615395	21135.116911
fuel	Diesel	1152.982247	22288.099158
BHP	147.928994	993.616896	23281.716054
drivertrain	Front Wheel Drive	-283.486230	22998.229824
doors	5dr	19.470730	23017.700554



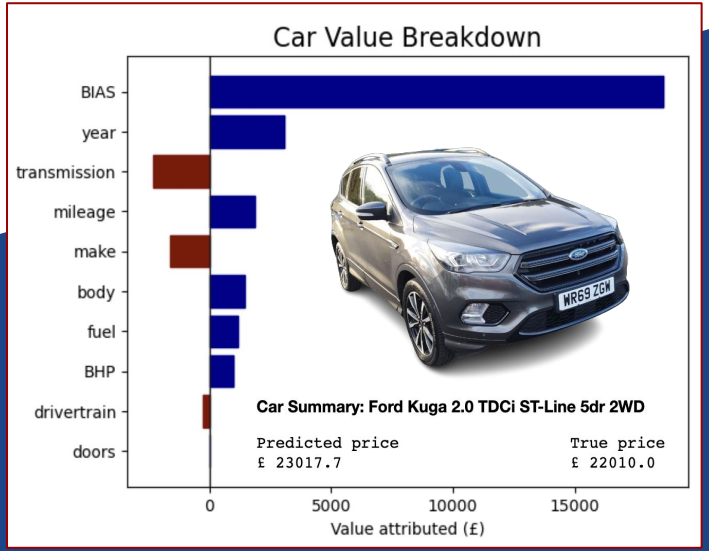
Car Value Breakdown with ELI5 - Example 2

y (score **23017.701**) top features

Contribution[?]	Feature	Value
+18627.367	<BIAS>	1.000
+3100.195	year	2019.000
+1856.395	log_mileage	9.465
+1052.183	fuel_Petrol	0.000
+1037.665	body_Hatchback	0.000
+993.617	BHP	147.929
+403.330	body_SUV	1.000
	… 27 more positive …	
	… 17 more negative …	
-167.954	make_Mercedes-Benz	0.000
-198.167	make_Audi	0.000
-257.477	make_Land Rover	0.000
-285.174	drivertrain_Front Wheel Drive	1.000
-910.598	make_Ford	1.000
-2321.886	transmission_Manual	1.000

Alternative Slide

Predictor	Attribute	Price Contribution (£)	Cumulative Car Price (£)
BIAS	0	18627.367022	18627.367022
year	2019	3100.195227	21727.562248
transmission	Manual	-2321.885676	19405.676572
mileage	12898.0	1856.395476	19665.501516
make	Ford	-1596.570532	17809.106040
body	SUV	1469.615395	21135.116911
fuel	Diesel	1152.982247	22288.099158
BHP	147.928994	993.616896	23281.716054
drivertrain	Front Wheel Drive	-283.486230	22998.229824
doors	5dr	19.470730	23017.700554

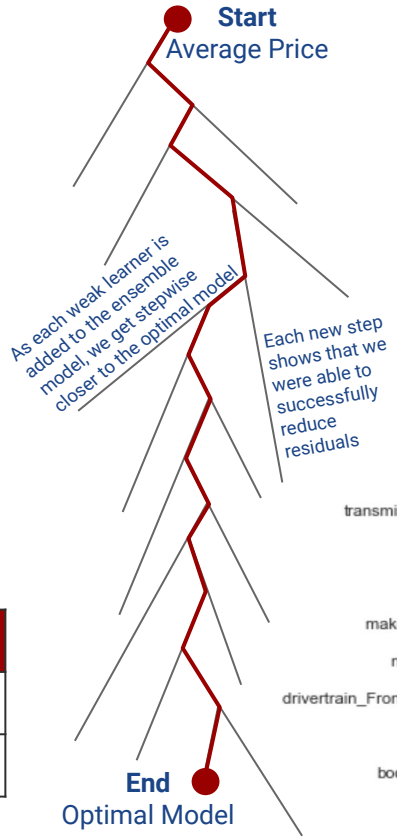


Gradient Boosting Model

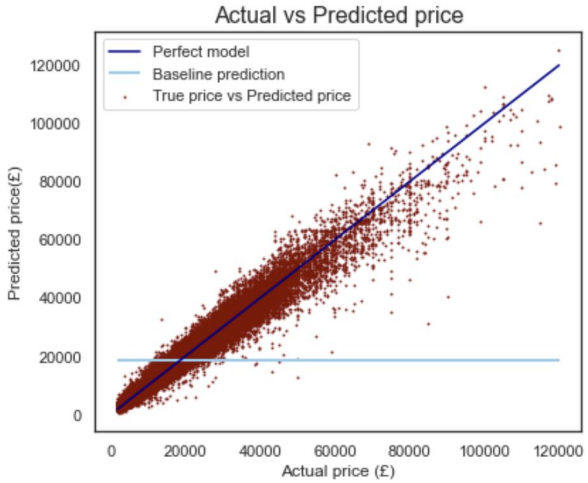
- Gradient boosting regression is an ensemble method which starts by predicting the average price of a car for every car and computing the price residual.
- It then constructs a sequence of predictors which are designed to minimise the these residuals.
- Each subsequent predictor is weighted by a small fraction to obtain a 'weak learner' before adding it to the ensemble. In this way, we take many small steps towards the optimal model.
- The model is considered optimised when adding further weak learners fail to reduce the residuals.

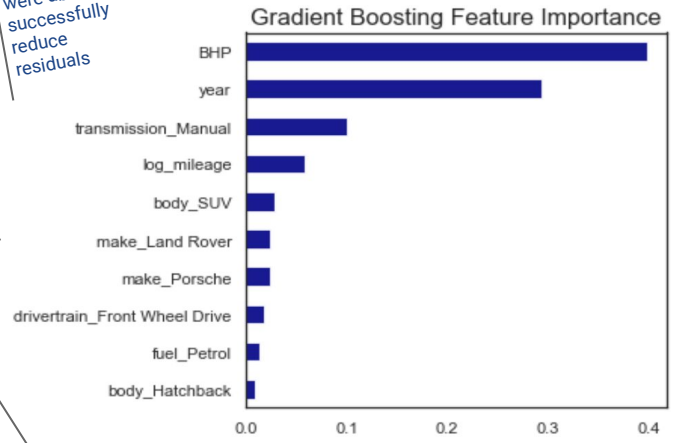
Metric	R^2 Score	RMSE (£)
Test data	0.954	2496.72
Train data	0.955	2553.55

Simplified model schematic



Alternative Explanation





AutoTrader Search Results

Alternative Explanation

Another outlier example

Alternative Slide