UK Used Car Price Prediction

Yang Zheng

December 5th, 2022

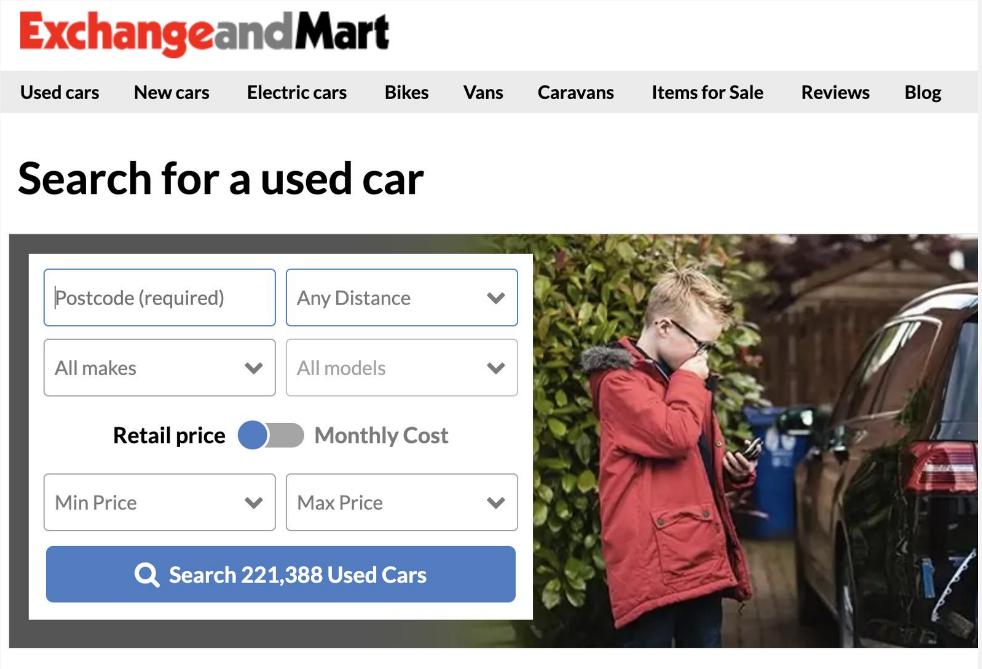
Data Science Initiative

DATA SCIENCE INITIATIVE
BROWN UNIVERSITY

Github Repository: https://github.com/yangzheng-brown/project_data1030

1 - Recap

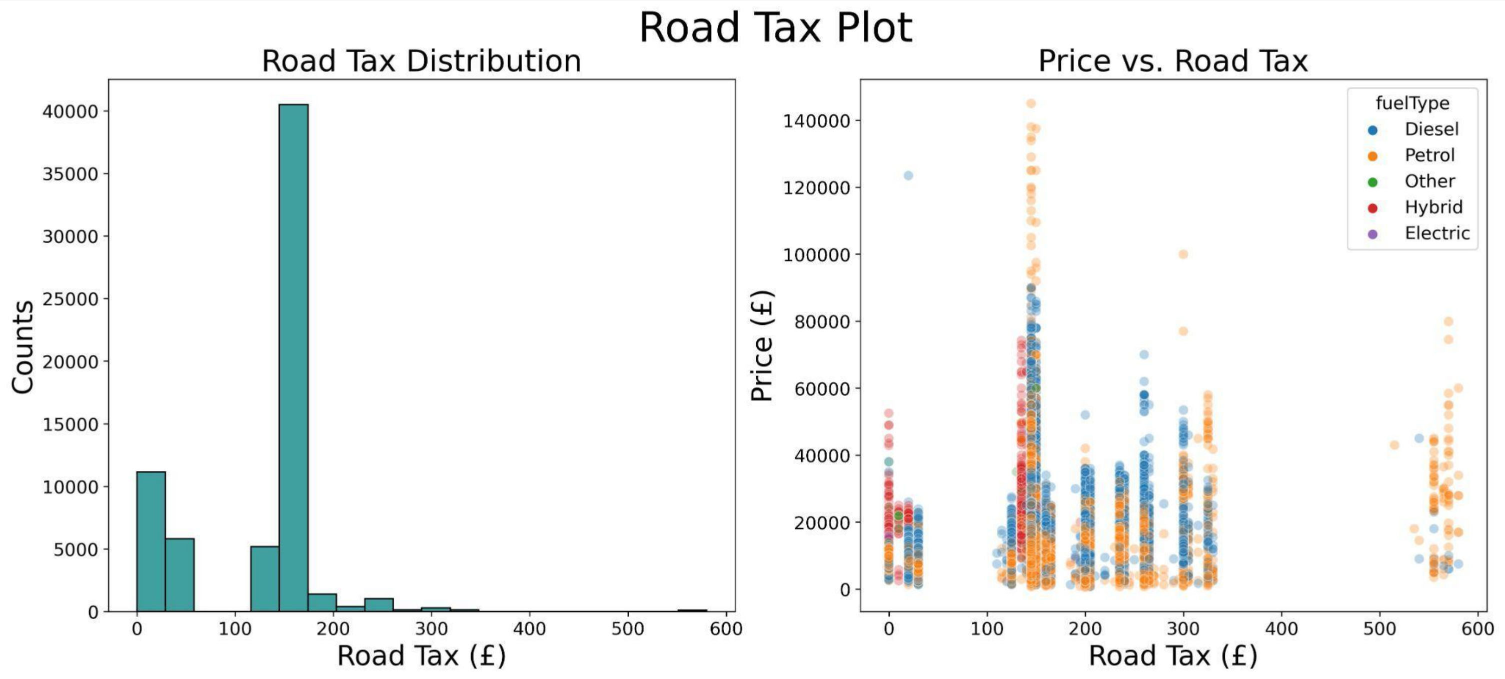
❖ Background



- To predict the price of used cars
- Dataset from Exchange & Mart

1 - Recap

- EDA

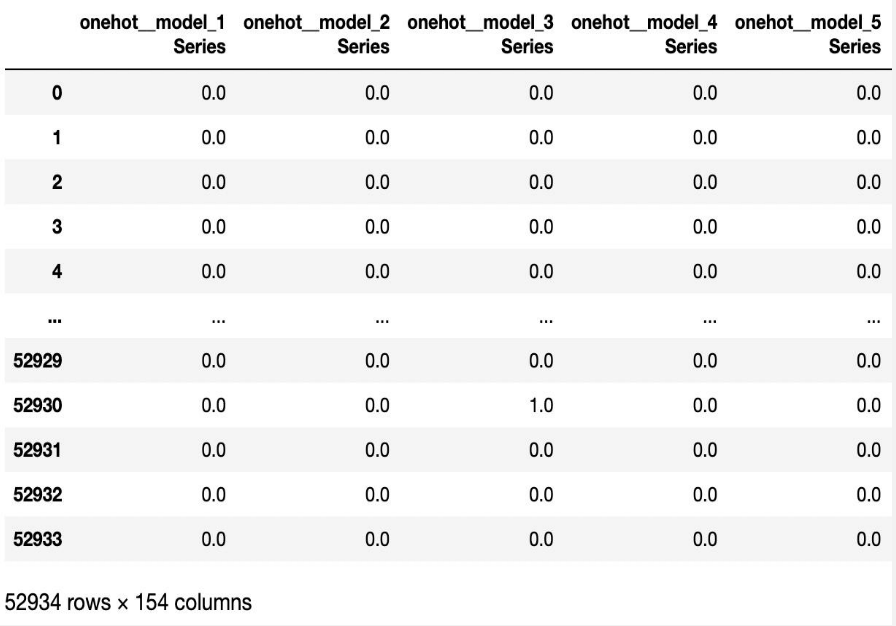


Road tax is based on fuel type and CO2 emissions,
so road tax has no strong linear correlation with price.

1 - Recap

❖ Preprocessing

- Method
    - OneHotEncoder for str
    - MinmaxScaler: engineSize
        - Bounded
    - StandardScaler: Rest of features
        - Tailed distribution
- No missing value
- Columns before prep: 9
  Columns after prep: 154



	onehot__model_1 Series	onehot__model_2 Series	onehot__model_3 Series	onehot__model_4 Series	onehot__model_5 Series
0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...
52929	0.0	0.0	0.0	0.0	0.0
52930	0.0	0.0	1.0	0.0	0.0
52931	0.0	0.0	0.0	0.0	0.0
52932	0.0	0.0	0.0	0.0	0.0
52933	0.0	0.0	0.0	0.0	0.0

52934 rows × 154 columns

2 - Cross Validation

- Basic hyperparameter tuning:
  - 5 random_state
  - 80%-10%-10% splitting
  - Preprocessing
  - Loop through all combinations of hyperparameter combos
  - Print out best model and best test score of each state

2 - Cross Validation

- ML Algorithms:
  - Linear: **Lasso, Ridge**
    - Parameter tuned: Alpha
  - Non-linear: **Random Forest, K-nearest neighbors, XGBoost**
    - Random Forest: max_depth, max_features
    - K-nearest neighbors: n_neighbors, weights
    - XGBoost: max_depth

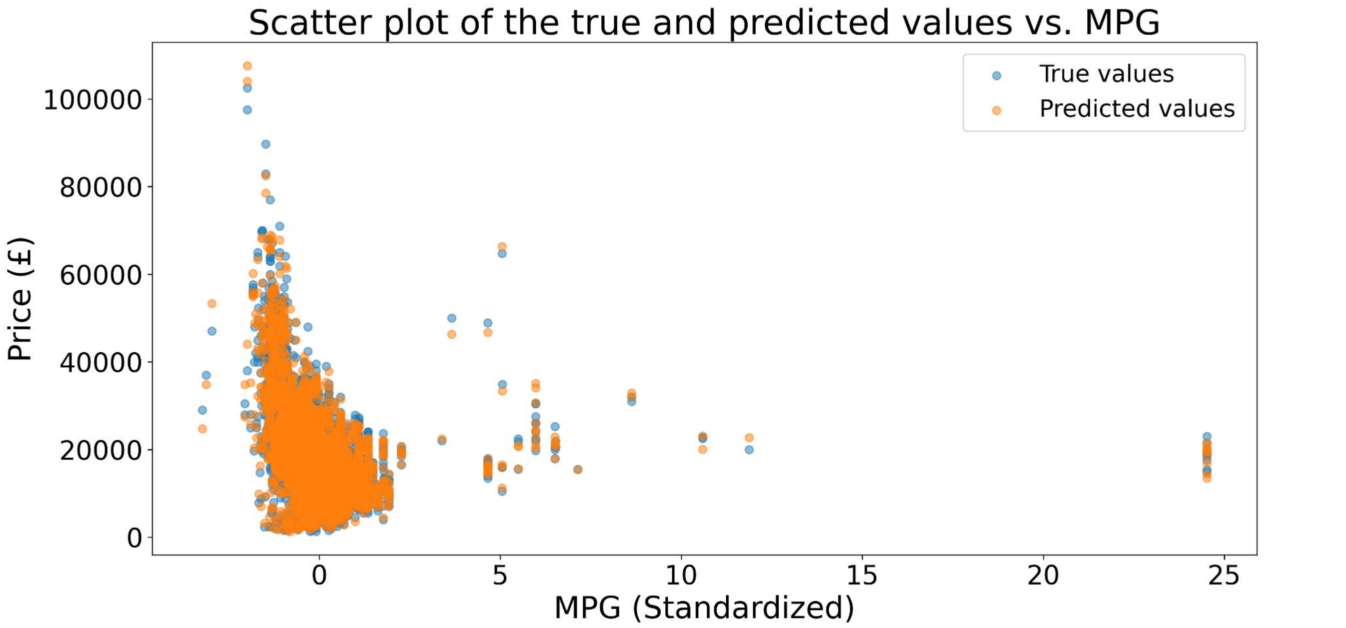
3 - Results

❖ Performance

	ML models	Mean RMSE (£)	Standard deviation (std) (£)	Baseline RMSE (£)
1	Lasso	3182.45	114.12	9340.22
2	Ridge	3189.63	109.52	
3	Random forest	1793.22	157.23	
4	KNeighbors	2404.34	183.34	
5	XGBoost	**1717.54**	167.92	

3 - Results

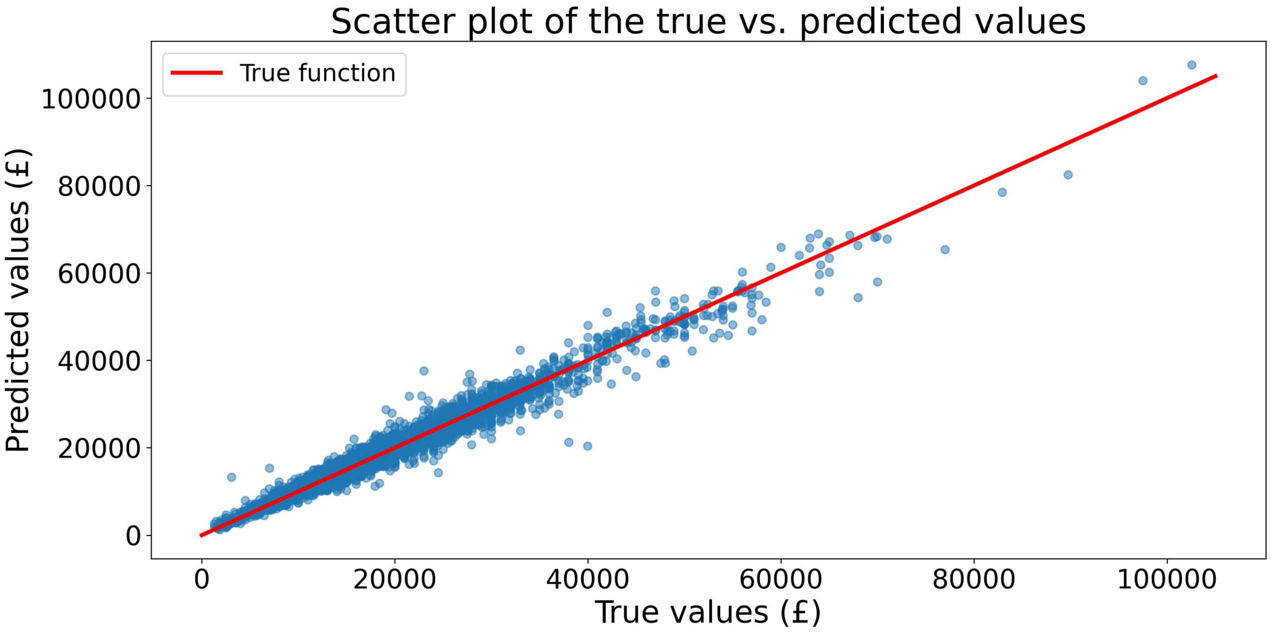
- Scatter plot of the true vs predicted values



Most predicted values are close to the true values within £2,000 tolerance.

3 - Results

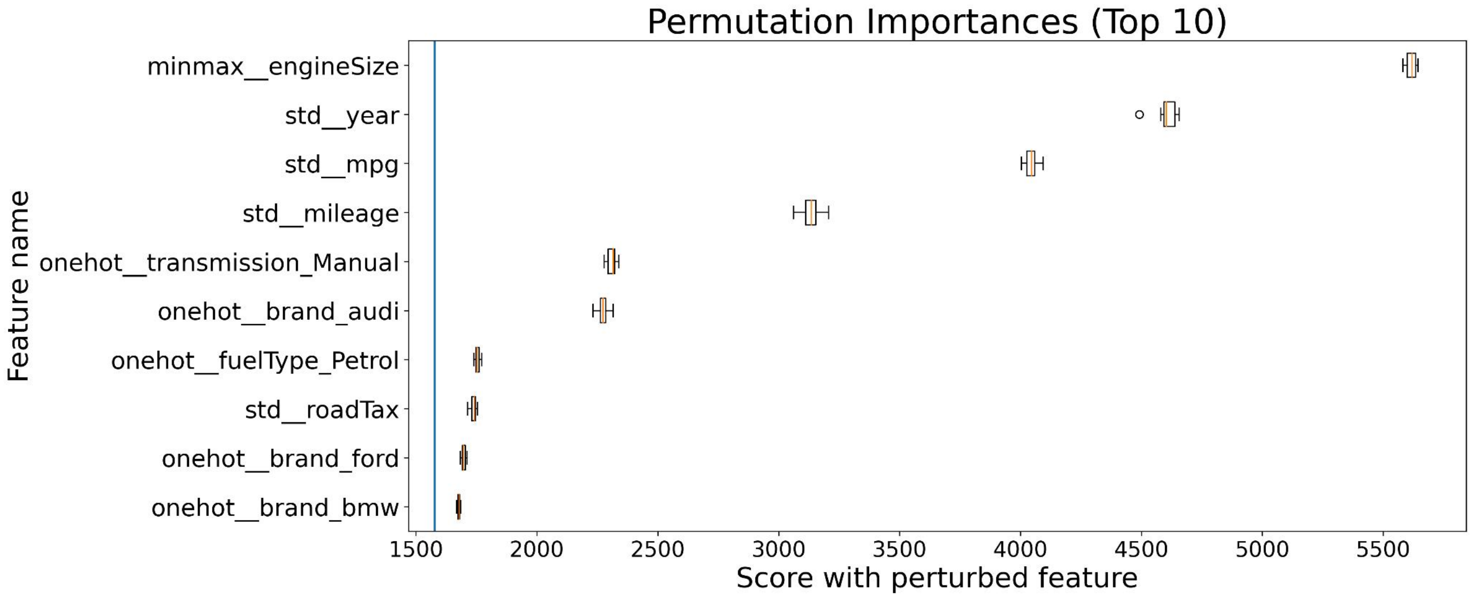
- Scatter plot of the true vs predicted values


Scatter plot of the true vs. predicted values

The most predicted values are along the true function line.

3 - Results

❖ Global feature importance

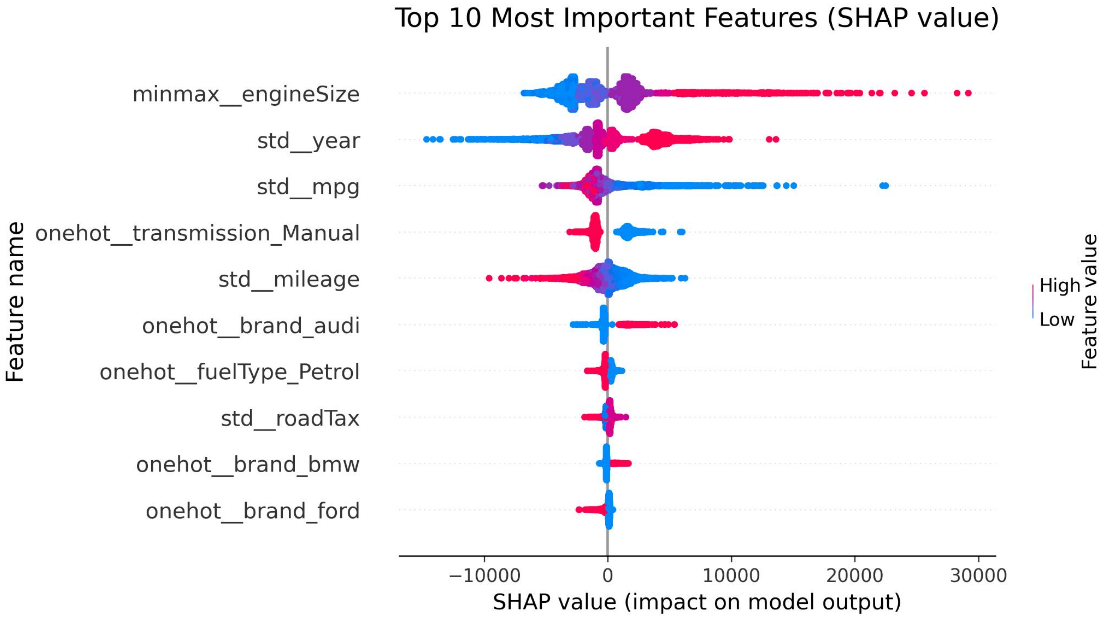


Engine size, year and MPG are the most important feature in permutation method.

DATA SCIENCE INITIATIVE
BROWN UNIVERSITY

3 - Results

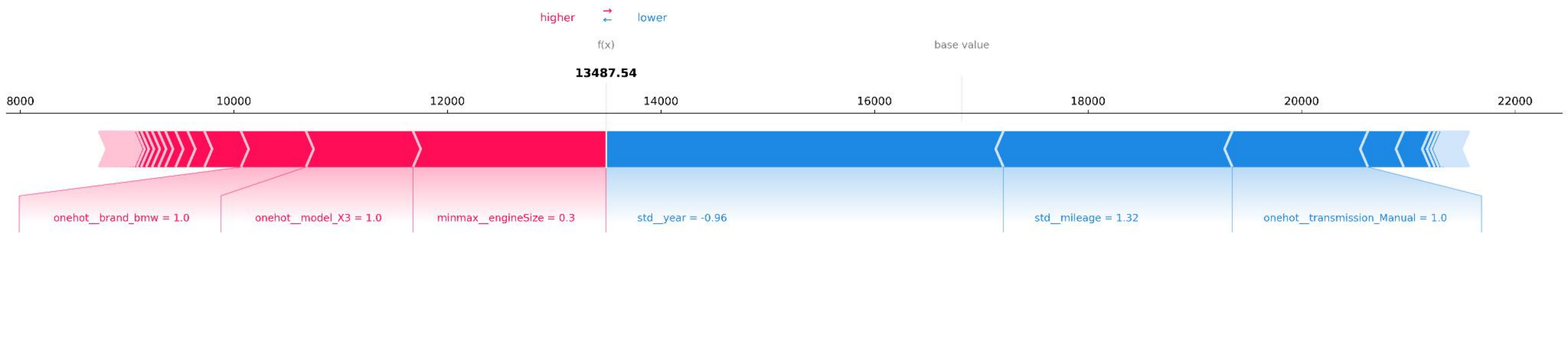
❖ Global feature importance



Engine size, year and MPG are the most important feature in SHAP method.

3 - Results

❖ Local feature importance: Index 0



- The positive factor: engine size, model_X3, brand_bmw
- The negative factor: year, mileage, transmission_manual

DATA SCIENCE INITIATIVE
BROWN UNIVERSITY

4 - Outlooks

- Improve the model
  - Try more algorithms
  - Tune more hyperparameters
  - Feature engineering
  - Add more features
  - A better computer

Thank you

DATA SCIENCE INITIATIVE
BROWN UNIVERSITY