Car price prediction

For the car price prediction challenge, 6 different models were trained. From these **2 were chosen to be discussed in this report**. One is an XGB Regressor model (XGB)[[1]](#footnote-1) based on trees, and the other is a Linear Regression model (LR)[[2]](#footnote-2). These models were selected based on technical factors like measurements and generalization, as well as on business factors like relevance and suitability to address the questions.

# Most important features

Feature importances were determined using gain, coefficients, and SHAP values.

## XGB

The top 7 feature importances calculated based on gain are:

| **feature** | **importance** |
| --- | --- |
| model\_key | 0.292450 |
| feature\_7 | 0.136117 |
| age\_in\_months\_when\_sold | 0.118553 |
| feature\_8 | 0.076528 |
| engine\_power | 0.059833 |
| mileage | 0.049596 |
| model\_initial | 0.041771 |

The features calculated based on SHAP values are:

A blue and white bar graph

Description automatically generated

Both methods agree that the most important feature is model\_key. Besides that, age\_in\_months\_when\_sold, mileage, feature\_8, and engine\_power are also among the top 7 for both methods.

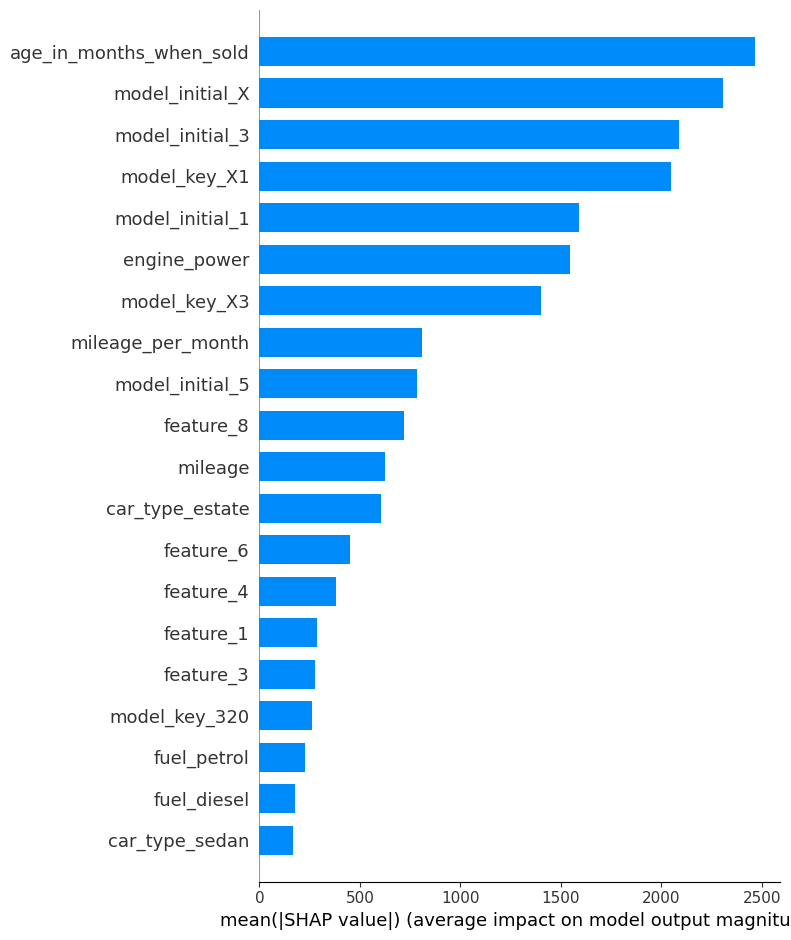
## LR

The largest regression coefficients, in absolute value, were:

| **feature** | **importance** |
| --- | --- |
| model\_key\_i8 | 41117.183922 |
| model\_key\_i3 | -23978.791365 |
| model\_initial\_i | 17138.392556 |
| model\_key\_M4 | 16772.597677 |
| model\_key\_650 | -16461.724894 |
| model\_key\_X1 | -12736.289962 |
| model\_key\_X6 M | 12098.244366 |

These are all indicator variables coming from model\_key or model\_initial (car model series). The values can be easily compared amongst themselves because they are on the same scale (0 or 1), but not to other features in different scales, like car age or mileage.

To measure how much impact each feature had on the model output, we can use SHAP:



This shows that the feature with the most impact was age\_in\_months\_when\_sold; and engine power was also among the top 7 features.

NOTE: none of the coefficients were individually significant, but together they were (the p-value for the F-test was 1e-16) meaning that the LR model is better than calculating an average price.

# Change in estimated value over time

This section gives examples of how price and estimated price change over time for different feature values.

## Change by car features 1 to 8

### Descriptive (no model)

The following insights were found after exploring the data.

* Except for feature\_7, **cars with True in the other car features have in average higher prices than cars without them**. This suggests that these features will be important for predictive models.
* There seems to be a peak average price in Aug 2018. This was caused by a **sale of a highly priced car precisely in August**.
* **Hybrid and electrical cars are more expensive on average**.

A graph showing the growth of the company

Description automatically generated with medium confidence

A graph showing the value of a graph

Description automatically generated

### XGB

The model captures the data well and produces very similar insights. Note how it seems to overestimate extreme values (Aug).

A graph with red and blue lines

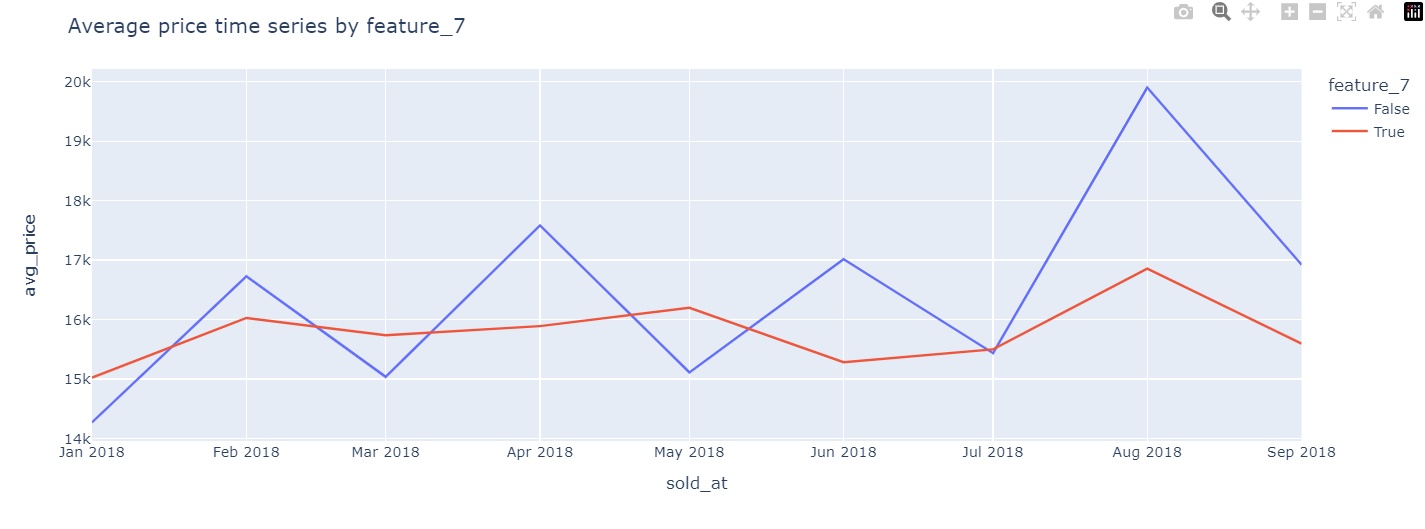
Description automatically generated

A graph showing the value of a number of days

Description automatically generated with medium confidence

### LR

The model captures the data well and produces very similar insights. Note how it seems to underestimate extreme values (Aug).

 A graph showing the value of a graph

Description automatically generated with medium confidence

## Seasonal patterns for other categorical variables

### Descriptive (no model)

* **Electrical cars average prices were stable** from winter to summer and were not sold in autumn.
* **Diesel and petrol cars had similar average prices**, although petrol cars had a drop in average prices starting summer 2018.
* **The most expensive car type is, on average, SUV,** although coupe was the most expensive at the start of the year and then dropped below SUV, also starting in summer.
* **Coupé and convertible cars were, on average, more expensive in winter than in summer**.
* **Vans were more expensive, on average, in spring, summer, and autumn, than in winter.**
* Subcompact had generally the lowest average prices.
* Paint color does not seem to generally determine or be associated with the average price, except for **color green, which consistently had prices much lower than other colors**. Maybe it is not very popular.
* **Orange and white cars were sold for more, on average, during summer** than during winter and spring.
* **Red cars were the opposite,** with lower average prices during summer than during winter and spring.

### XGB

The model captures the data well and produces very similar insights.

In addition, t-tests for the difference of means were performed to find if there were significant differences between the average estimated prices during summer and winter for different categorical features values. Below are the cases where the tests found significant differences (alpha=0.05).

|  |  |  |  |
| --- | --- | --- | --- |
| **feature** | **feature\_value** | **t\_stat** | **p\_val** |
| car\_type | sedan | 4.321558 | 0.000018 |
| car\_type | coupe | 3.560600 | 0.000936 |
| car\_type | estate | 2.881445 | 0.004071 |
| car\_type | van | -2.601545 | 0.018620 |
| paint\_color | white | -2.311843 | 0.021562 |

### LR

The model captures the data well and produces very similar insights.

In addition, t-tests for the difference of means were performed to find if there were significant differences between the average estimated prices during summer and winter for different categorical features values. Below are the cases where the tests found significant differences (alpha=0.05).

|  |  |  |  |
| --- | --- | --- | --- |
| **feature** | **feature\_value** | **t\_stat** | **p\_val** |
| car\_type | sedan | 4.295961 | 0.000021 |
| car\_type | coupe | 2.675371 | 0.010595 |
| paint\_color | white | -2.559119 | 0.011056 |
| car\_type | estate | 1.999919 | 0.045869 |

# What car to buy to minimize loss after 1 year

To answer this question the models were used to predict the price of cars today and 1 year after. It was decided to **set today to March 2024** (time of presenting these results), **but also to a month after the last auction, as data today is already more than 5 years old**.

As the results show, **it is preferable to choose the linear model to answer this question**, because it better captures the general fact that as time passes cars’ prices decrease.

In summary, **models of the series X were found to be good candidates** for buying. Using the LR model, **a diesel fuel BWM X5 with an average mileage per month of 30.6** was found to be the best option to buy as of March 2024.

NOTE: When changing the date of buying and selling, the features age\_in\_months\_when\_sold and mileage were changed accordingly. Mileage was updated using the mileage\_per\_month calculated based on the mileage and age from the original data set. Other time dependent variables like month\_sold\_at and season\_sold\_at were only updated for the XGB model, where they showed some importance.

## XGB

### Today set to March 2024

The car, in the original dataset, with the lowest loss was:

maker\_key BMW

model\_key M3

mileage 29925

engine\_power 309

registration\_date 2012-09-01 00:00:00

fuel petrol

paint\_color silver

car\_type coupe

feature\_1 True

feature\_2 True

feature\_3 False

feature\_4 False

feature\_5 True

feature\_6 True

feature\_7 True

feature\_8 True

price 47000

sold\_at 2018-04-01 00:00:00

age\_in\_months\_when\_sold 67

month\_sold\_at 4

season\_sold\_at spring

model\_initial M

mileage\_per\_month 446.641791

The calculated loss based on predictions was -184.783203. It is expected that the real loss won’t be negative, but relatively low.

Between the candidates with low calculated loss values (less than 2,000) these many cars by model\_key were found:

X4 13

X6 3

M3 3

640 Gran Coupé 3

X5 M 2

The **predicted prices as well as the SHAP values** for the predictions are shown in the plots below.

Prediction today:

A graph with numbers and text

Description automatically generated

Prediction 1 year later:

A graph with numbers and a graph

Description automatically generated with medium confidence

### “Today” set to a month after last auction

The car, in the original dataset, with the lowest loss was:

maker\_key BMW

model\_key X3

mileage 46409

engine\_power 135

registration\_date 2014-04-01 00:00:00

fuel diesel

paint\_color silver

car\_type suv

feature\_1 True

feature\_2 True

feature\_3 False

feature\_4 False

feature\_5 False

feature\_6 False

feature\_7 False

feature\_8 True

price 22800

sold\_at 2018-05-01 00:00:00

age\_in\_months\_when\_sold 49

month\_sold\_at 5

season\_sold\_at spring

model\_initial X

mileage\_per\_month 947.122449

The calculated loss based on predictions was -15594.402344. This result **does not make sense in the context of the question**.

Between the candidates with low calculated loss values (less than 2,000) these many cars by model\_key were found:

X3 34

X4 27

420 23

520 23

420 Gran Coupé 19

The **predicted prices as well as the SHAP values** for the predictions are shown in the plots below.

Prediction today:

A graph with numbers and red rectangles

Description automatically generated

Prediction 1 year later:

A graph with numbers and a bar chart

Description automatically generated with medium confidence

## LR

### Today set to March 2024

The car, in the original dataset, with the lowest loss was:

maker\_key BMW

model\_key X5

mileage 612

engine\_power 183

registration\_date 2016-10-01

fuel diesel

paint\_color black

car\_type suv

feature\_1 True

feature\_2 True

feature\_3 False

feature\_4 False

feature\_5 False

feature\_6 False

feature\_7 True

feature\_8 True

price 49100

sold\_at 2018-06-01

age\_in\_months\_when\_sold 20

month\_sold\_at 6

season\_sold\_at summer

model\_initial X

mileage\_per\_month 30.6

The calculated loss based on predictions was 1508.460034.

Between the candidates with low calculated loss values (less than 2,000) these many cars by model\_key were found:

X5 114

X4 35

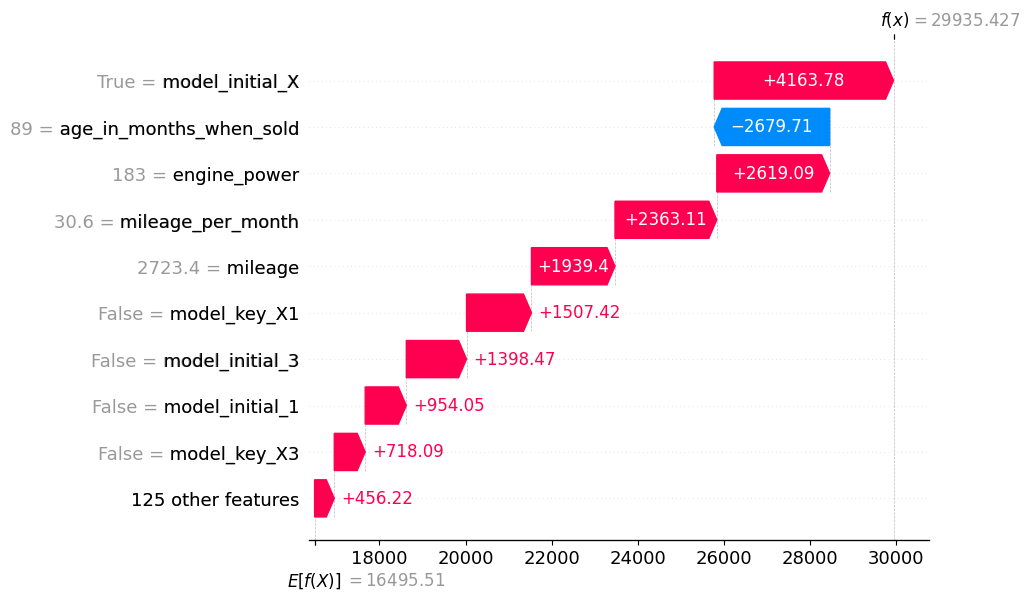
X6 30

740 15

X5 M 15

The **predicted prices as well as the SHAP values** for the predictions are shown in the plots below.

Prediction today:



Prediction 1 year later:

A graph with red squares and numbers

Description automatically generated

### “Today” set to a month after last auction

The car, in the original dataset, with the lowest loss was:

maker\_key BMW

model\_key 420 Gran Coupé

mileage 476

engine\_power 120

registration\_date 2014-05-01

fuel diesel

paint\_color blue

car\_type hatchback

feature\_1 True

feature\_2 True

feature\_3 False

feature\_4 False

feature\_5 False

feature\_6 True

feature\_7 True

feature\_8 True

price 30300

sold\_at 2018-08-01

age\_in\_months\_when\_sold 51

month\_sold\_at 8

season\_sold\_at summer

model\_initial 4

mileage\_per\_month 9.333333

The calculated loss based on predictions was 1504.906043.

Between the candidates with low calculated loss values (less than 2,000) these many cars by model\_key were found:

X5 182

X3 106

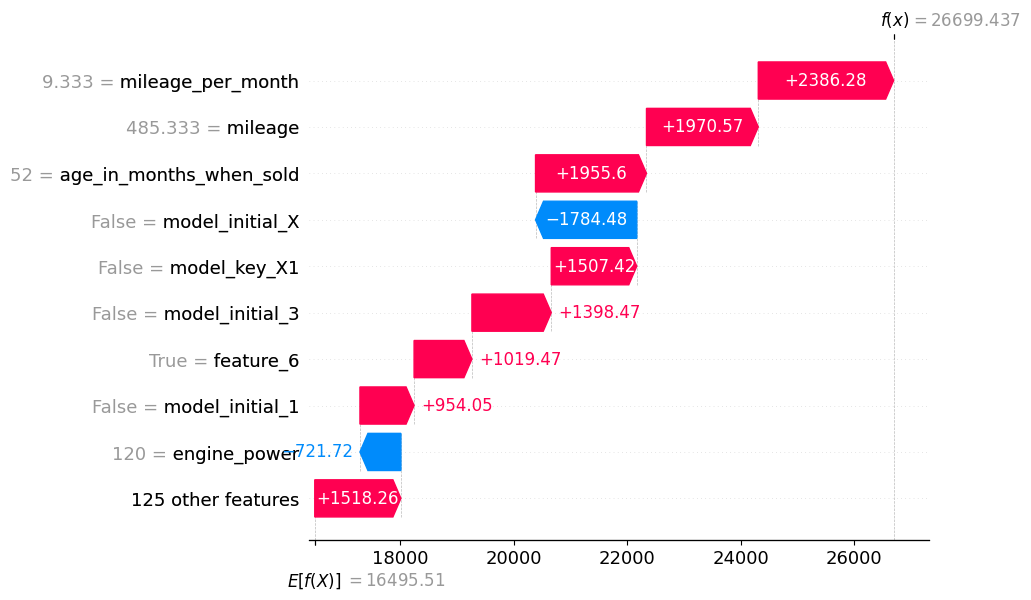
530 59

520 47

X6 40

The **predicted prices as well as the SHAP values** for the predictions are shown in the plots below.

Prediction today:



Prediction 1 year later:

A graph with numbers and symbols

Description automatically generated

# Out of sample metrics for the model

The models were initially trained on 80% of the data and tested on 20% of the data. The data split was random.

## XGB

The train metrics for the model were:

Train MSE: 7223111.16

Train RMSE: 2687.58

Train MAE: 1524.56

Train R2: 0.92

The test metrics for the model were:

Test MSE: 8705488.27

Test RMSE: 2950.51

Test MAE: 1914.02

Test R2: 0.87

NOTE: other models showed more signs of overfitting e.g. XGB without tuning had Test R2: 0.87 (Test MAE: 1919) and Train R2: 0.99 (Test MAE: 532).

## LR

The train metrics for the model were:

Train MSE: 22040928.92

Train RMSE: 4694.78

Train MAE: 2355.72

Train R2: 0.75

The test metrics for the model were:

Test MSE: 11259962.94

Test RMSE: 3355.59

Test MAE: 2273.41

Test R2: 0.83

# Additional insights

## Data quality

There are no null values in the data.

There are no infinite values in the data.

There are no duplicates.

From the number of unique features values we can observe that:

- **There is only one maker (BWM).** So, this feature will not give information to the models.

- There are 199 different registration dates.

- feature\_1 to feature\_8 are binary variables.

- The auction happened on 9 different dates.

Looking at the description of numerical features, hints that there are possibly erroneous observations. For example:

- **a car with -64 miles**,

- **a car with 0 (I assume hp) engine power**,

- **and a car that cost 100 (I assume USD).**

Therefore, data needs some further cleaning.

There is a 640 Gran Coupé with negative mileage which is not possible. Also, there are other 18 cars of the same model key, so this row will be removed.

This is a 13-year-old car (159 months) with more than a million miles. Although strange, it is not impossible that it has driven this many miles (about 210 on average daily). So, this observation is not recommended to be removed.

**There is a wrong observation with 0 engine power for an X1 which is a SUV**. This is impossible. Since there are more than 200 other X1 this observation can be removed.

**Very likely the engine power of 25 (hp) for two i3 is wrong.** These cars should have 75 (hp) engine power, so this is probably a typo. Since there are very few other i3 cars, data imputation might be a better alternative here to dropping the records. According to most values, a good candidate value for imputation is 75.

Regarding models 316 and 318, since there are more than 200 hundred other observations with the same model, and these cars typically have at least 75 of engine power, these records can also be deleted.

**There were 62 cars sold at less than 1,000 which is very unusual**. Looking at a description of the numerical features of these cars we find:

- The newest car sold at this price was less than 3 years old. This seems unusual.

- The oldest car was 24 years old.

- Minimum and max mileage seem sensible.

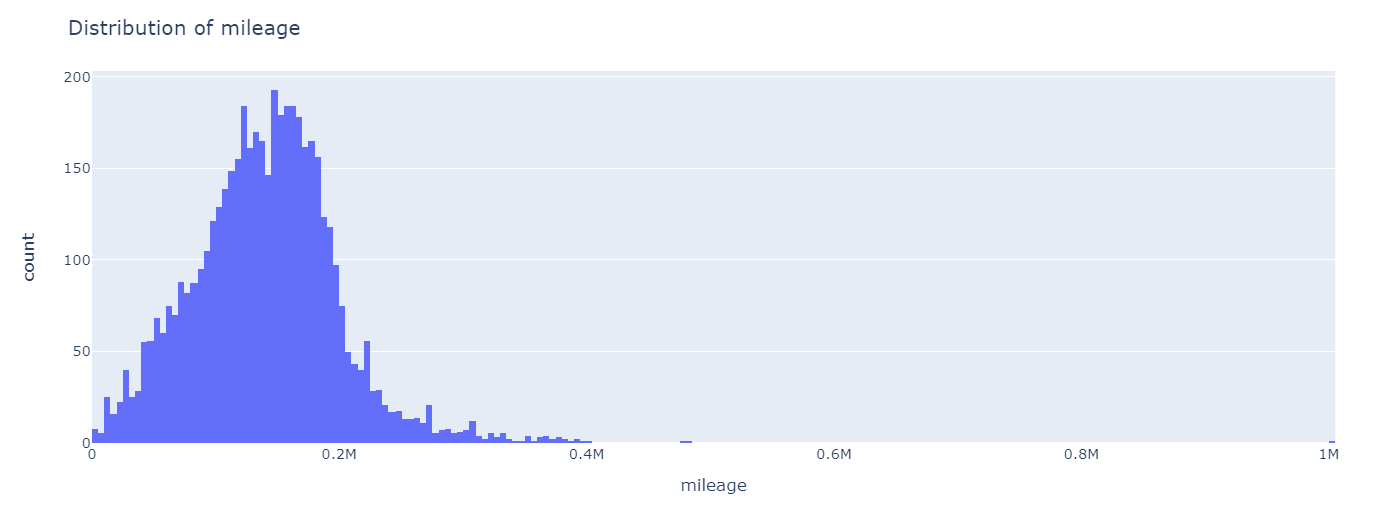
**Prices below 800 seem more unusual** and might correspond to cars with important damages. Since cars with engine damage were removed before, these other cars with important damages could be removed as well and maybe priced with another strategy.

## Data distribution (after cleaning)

The features still have very high values after removing the data that probably had errors.

Examples are:

- **1M miles drove**. This can be rare but possible for old cars.



- **A few cars cost more than 100k**, which is also possible depending on the car.

A graph with a number of lines

Description automatically generated with medium confidence

- There are **some cars with engine power above 400**, which can happen for sports cars.

A graph with blue and white bars

Description automatically generated

- There are **cars as old as 22 years**, which is also possible.

A graph of a graph

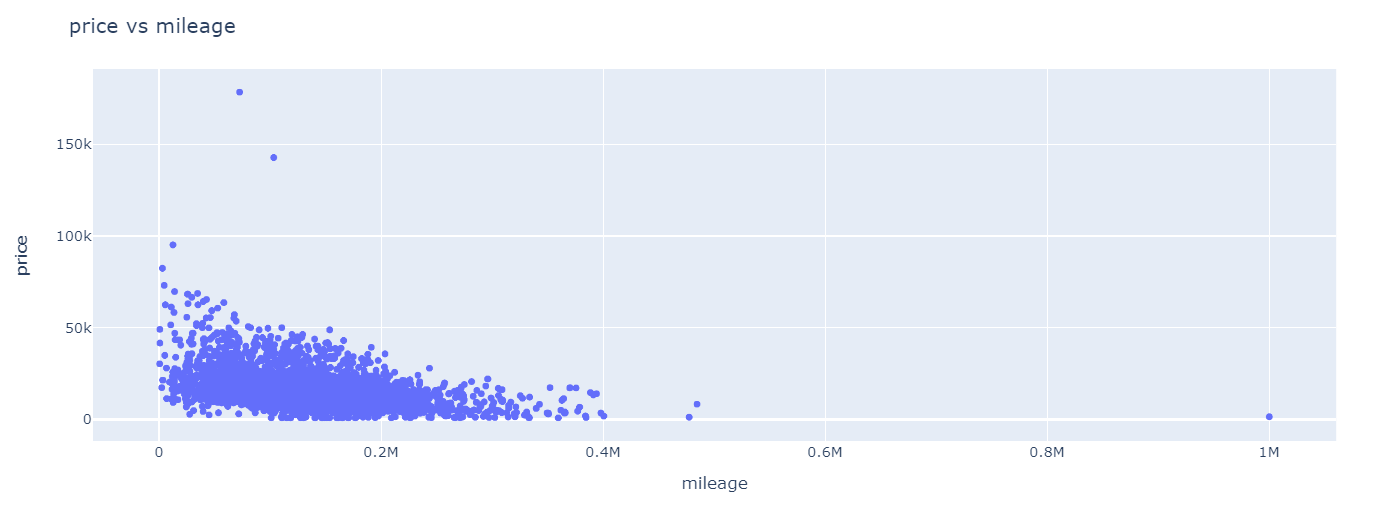
Description automatically generated with medium confidence

These values do not suggest that further data cleaning is needed.

## Numerical features vs. price

Based on the scatter plots, the following observations can be made:

- As expected, prices tend to decrease with mileage.



- As expected, prices tend to increase with engine power.

A graph showing a number of numbers

Description automatically generated with medium confidence

- As expected, prices tend to decrease with age.

A graph showing a number of data

Description automatically generated with medium confidence

## Improvements

In the context of the problem of finding the best car to buy, it can be a good idea to **retrain a model after removing cars with prices over 50k**, since recommendations are for cars with prices (as recorded in the data) between 20k and 50k. This can help models achieve lower MAE.

Also, if predictions were to be made as of March 2024, **more recent data is needed**.

1. Comes from notebook 04\_xgb\_trees\_model\_tuned.ipynb in the project’s code. [↑](#footnote-ref-1)
2. Comes from notebook 08\_linear\_model.ipynb in the project’s code. [↑](#footnote-ref-2)