Car value prediction

Ltaief Mohamed

5/14/2021

Contents

1 Introduction

The automotive industry is one of the largest industries by revenue . It is regarded as one of the most competitive and innovative in the world. That being the case, a good understanding of the market is a necessary task for business and consumers and a price forecasting model can be of values for both parties.

2 Project objective

We will be creating a car price predictor using machine learning methods based on data from one of the largest Europe's car market. Our report will showcase two methods we selected to execute our task: Using a matrix factorization and regression trees.

2.0.1 RMSE:

To describe the behavior of our price outcome, our approach is to define the loss function. A function that quantifies the deviation of the observed outcome from the prediction (residual). In our case we used the root square error(RMSE) known as the standard deviation of the residuals. It has the same units as the measured and calculated data. Smaller values indicate a better performance of our system.

$$RMSE = \sqrt{(\sum_{i=1}^{n}(X\ observation, i\ - X\ model, i\ )^2)}$$

3 Exploratory data analysis

3.1 Features and processing

3.1.1 The dataset

We are using a ***Kaggle*** dataset of 46376 cars collected between 2011 and 2021 and scraped from the autoscout24 ***website*** an online marketplace for purchase and sale of different type of vehicles.

Our data lists some of the car features:
mileage, make, model, fuel, gear, Offer type, Horse power (hp), year and the price.

```
## 'data.frame':    46405 obs. of  9 variables:
##  $ mileage  : int  235000 92800 149300 96200 156000 147000 91894 127500 115000 104 ...
##  $ make     : Factor w/ 77 levels "9ff","Abarth",..: 11 75 64 62 57 72 62 55 47 30 ...
##  $ model    : Factor w/ 842 levels "","107","108",..: 35 402 331 521 34 137 684 839 28 758 ...
##  $ fuel     : Factor w/ 11 levels "-/- (Fuel)","CNG",..: 3 8 8 8 8 6 3 8 8 3 ...
##  $ gear     : Factor w/ 4 levels "","Automatic",..: 3 3 3 3 3 2 3 3 2 3 ...
##  $ offerType: Factor w/ 5 levels "Demonstration",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ price    : int  6800 6877 6900 6950 6950 6950 6970 6972 6980 6990 ...
##  $ hp       : int  116 122 160 110 156 99 131 116 150 86 ...
##  $ year     : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
```

3.1.2 Data wrangling:

Before moving on to algorithm implementation, there is some data wrangling to take into account. In order to simplify the computation process we are going to convert some features units: * The year column will

be transformed in age in years. * Since we are working with car data in german market we can normalize the mileage metric accordingly. As reported by the Odyssee-mure ***website*** the average distance traveled in germany since 2011 is around 14000 Km/year. Given this information we can define a new metric "mileage ratio" that describes whether the car has been used more than the average as follows:

$$mileageratio = mileage/14000 - age$$

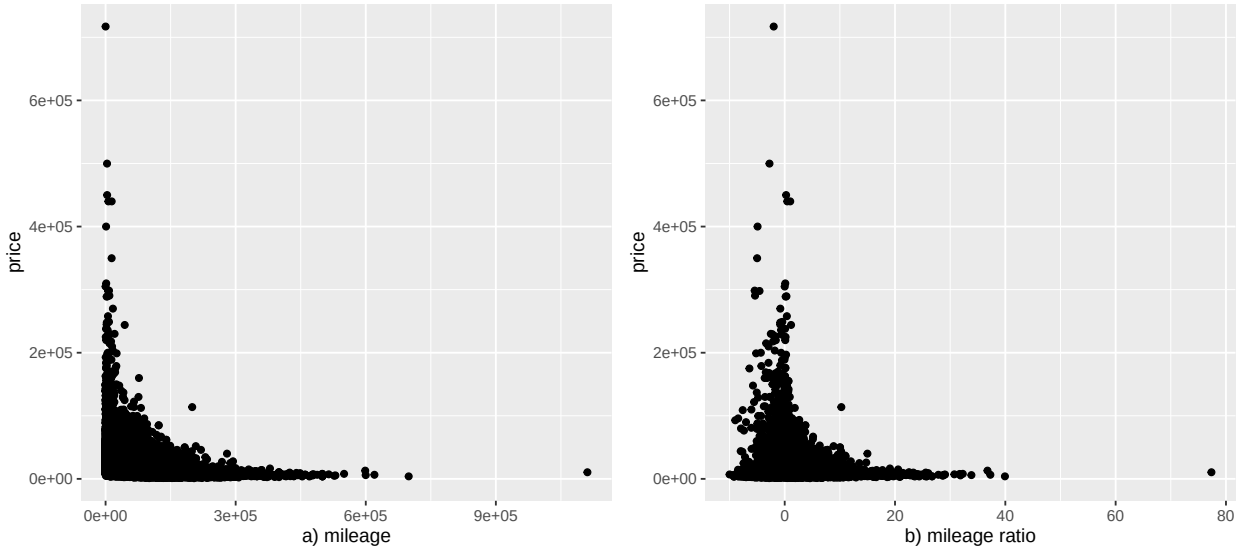
This gives us a qualitative description of the mileage metric.

We deliberately removed the far outliers according to Tukey definition. That is to say consider the prices between the 75th percentile (top whisker of the boxplot) plus 3 * the interquantile range (IQR) and the 25th percentile (bottom whisker of the boxplot) minus 3 * IQR. An outlier is anything outside this range.

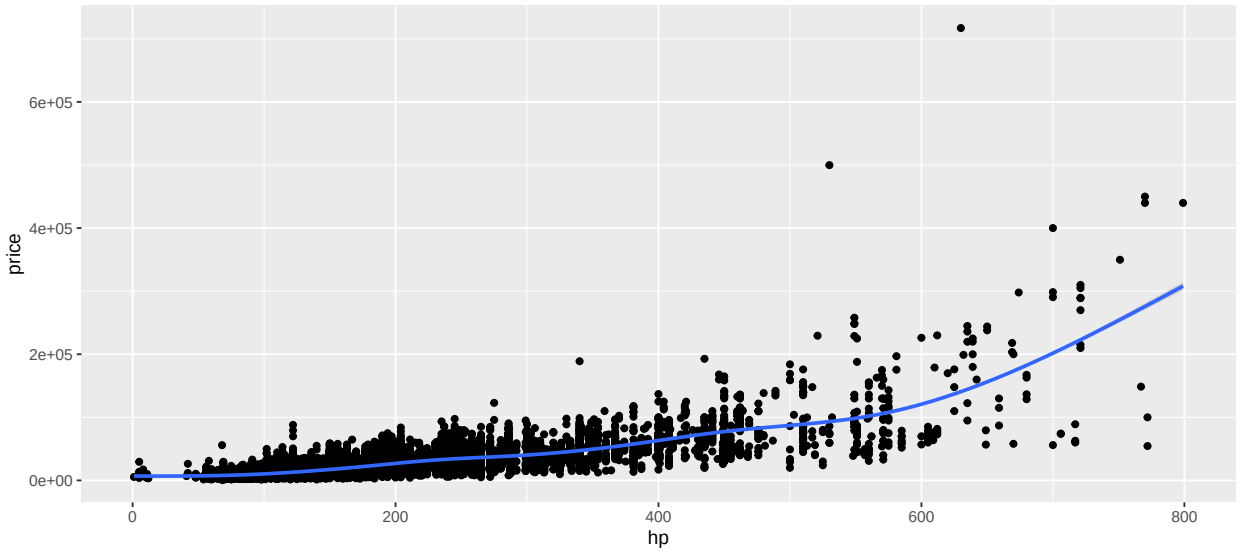
3.1.3 Data exploration:

After removing all missing data and partitioning our data into two datasets for testing purposes lets take a look at the distribution of our price according to the different features.

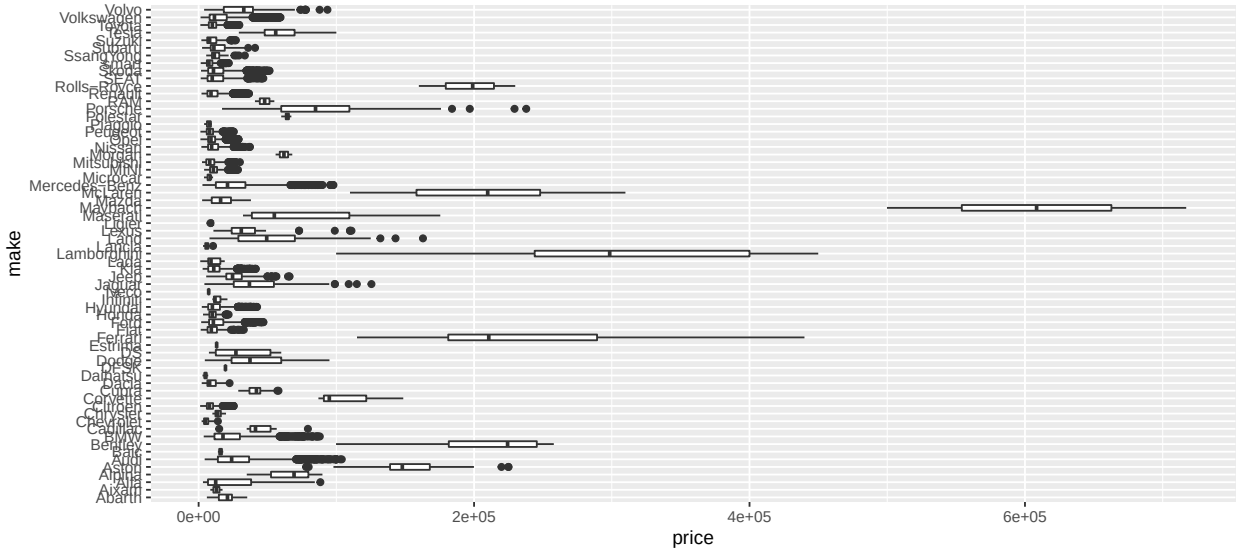
3.1.3.1 Mileage : As we can see here the prices are relatively low for the cars with higher mileage (figure 1.a). The cars that haven't been used much have more value than the more driven ones as the cost of maintenance and servicing gets higher with wear and tear. If ye look at the cars that have a mileage ratio close to 0 (a car that it normally driven according to the german market) the prices seem to be higher. (figure 1.b)



3.1.3.2 Horse power: People love a powerful vehicle. When it comes to horse power, price increase dramatically for cars with a better performance on the road. It is one of the main features that attracts buyers attention.



3.1.3.3 Make: When it comes to car prices, some clearly exceed others by far. In fact some are considered as luxuary items. The Maybach for example has a minimum price of 500 thousand euros. On the other hand, a Citroen has a minimum value of 1100 euros.



4 Model selection:

4.1 Matrix factorization:

In the course of our project we adopted a matrix factorization method to construct the first version of our algorithm. We assume the price y is the same to all entries with the difference explained by random variations (bias). Thereby the goal is to minimize the residual ϵ for each observation k with b the biases total.

$$\epsilon\ k\ = y\ k\ - b\ k$$

Given that the average of all rating as a value of μ minimize the residual ϵ we will start by identifying the first element of our formula $\hat{y}\ k$. The idea is to work with the average price $y\ k$ and gradually add the

different biases caused by the main features. The default model performance (without considering the bias) is characterized by the following RMSE:

```
## [1] 13894.51
```

Based on these observations we are building our model characterized by:
* car model
* make
* horse power
* mileage ratio

Our method is analytically described by the formula:

$$Y\ i,u,t\ = \hat{\mu} + \hat{b}\ m\ + \hat{b}\ h\ + \hat{b}\ r\ + \hat{b}\ f\ + \epsilon$$

where
** $\hat{b}\ o$ is a make effect
** $\hat{b}\ h$ is a horse power effect ** $\hat{b}\ r$ is a mileage ratio effect ** $\hat{b}\ f$ is a fuel effect penalized with the independent error ϵ

```
## [1] 10045.58
```

```
## [1] 7803.656
```

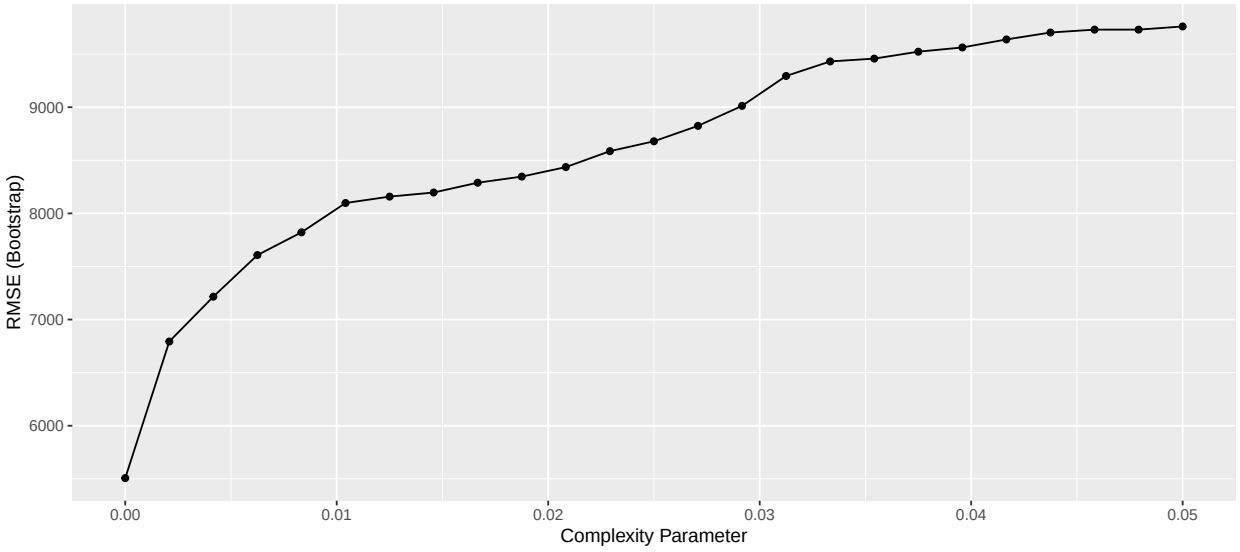
```
## [1] 4724.169
```

```
## [1] 4711.335
```

4.2 Regression tree:

As a second version of our algorithm, we decided to pick the regression tree method. The idea is to build a decision tree, at the end of each node, obtain a predictor.

By such a method our data is partitioned recursively into partitions for a number of times J in non overlapping regions R. Each selected partition x is split to create new partitions, characterized by the predictor j and the value s. Finally we obtain a number of predictors that we will act upon. We can select the complexity parameter by performing a cross validation that minimizes the root mean square error.



5 Results:

In this section we are going to quantify the performance of our model by presenting the root mean square errors obtained by each method.

method	RMSE
Average	13894.509
Matrix factorization	4711.335
Regression tree	2895.562

As illustrated in the results table, our regression tree has outperformed the matrix factorization model.

6 Conclusions:

In order to have an idea on the car price behavior in one of the largest marketplaces, we performed two different machine learning algorithms. We did perform a prediction that has an RMSE of 2800 euros. The regression tree model used in this report had the upper hand in computation facility and coding effort. However with the European Green Deal banning the Gas cars by 2035, the prices studied in this report may vary in the near future. The most important of future work is to expand the knowledge acquired during the execution of this project working on a new challenge .

7 References:

https://grouplens.org/datasets/movielens/

Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning, Data mining, inference and prediction. second edition.

https://leanpub.com/datasciencebook

https://www.carbibles.com/

https://www.odyssee-mure.eu/

https://en.wikipedia.org/