The influence of car attributes on price

Martin Kent Kraus
Danciulescu Dragos-Gabriel

January 2022

1 Abstract

This project concerns a "problem", how cars attributes affect their price. Thirteen properties of cars were used to determine their relation to prices of 42 thousand cars from Czechia and Slovakia. To this end, linear regression and the Random trees algorithm were used in the Weka environment. The initial assumption would be that there is a strong correlation between increases in desirable attributes and price, as these increases reflect the costs involved in building a car. This is treated as a regression problem. Therefore, we apply 2 algorithms and compare their results. We reached the conclusion that Random Trees is the better algorithm for evaluating the dataset. Our initial assumption proved correct, as price had a r^2 of 0.84 with the other attributes.

2 Application domain and research problem

Cars are an expensive, yet important commodity. In 2019, the average EU car after taxes cost 30 485 euros [1]. Moreover, more people in the European union have cars than do not, with 569 cars for 1000 people [2]. And demand for new cars is still high. In 2019, over 14 million cars were sold in western Europe [3]. Households in the EU 11.5 % of their budgets on transportation [4]. Cars make up a significant portion of that percentage, as can be seen in Belgium . All this goes to show that cars are an important market in the EU.

During the Corona crisis, car sales stagnated. This was due to lowered demand during the first year of the pandemic, and in 2021 compounded by a lack of microchips. As a result, many car manufacturers scaled down production or closed temporarily [5][6]. With lower supply, prices of used cars increased - up by 40 percent in the US. [7][8]

Because of the size and importance of used cars within the automobile market today, it would be useful to see how the prices of these cars are determined. What attributes of the car play a role?

3 Previous attempts

The relationship between the attributes of a car and its price is a relatively well researched. For example, one study with a sample size of 63 thousand German cars found that there was a price gap between cars that were registered in January and those registered in December of the previous year. [9]

Machine learning, specifically Random Forest classifiers have also been applied to the problem.For example, one paper found the combination of "brand, powerPS, kilometer, sellingTime, VehicleAge" provided the best results, with other attributes like fuel type being discarded. [10] A paper in the TEM journal found that applying just one machine learning algorithm at a time yielded generally poor results, but applying a combination increased accuracy in tests significantly. [11]

4 Data and Pre-processing.

Our data set contained 47 thousand cars. These are taken from a variety of Czech and Slovak car-dealership websites, and collected by XTRA software, which allowed us access to the data set. Each car has 13 attributes. These are: make, country, body type, price, registered, speedometer, fuel type, gearbox, seats, doors, engine size and colour. The values were standardized, so that for example the Czech and Slovak words for a certain colour (and their synonyms/misspellings/shortenings) show up as one colour in English. Prices are shown in Euros.

Originally, the data set also had a model of the car, but this had to be removed, as our environment couldn't decide whether it was a string or an integer. Each car also had a unique link to where it is being sold. Our dataset sadly contained a significant amount of outliers (like a car having 2020201908 doors), caused by the algorithm involved in getting the cars data from the websites. The most extreme of these were removed. Still, some cars may have incorrect values, as we did not inspect many non outlier cars.

Moreover, some websites do not have all the relevant information about their cars, leading to many cars missing the number of doors or seats. In that case, the attribute was assigned a "0". There were too many of these cases to remove. Since lacking these attributes was caused by the site from which the cars were downloaded, and since most car dealerships have a range of cars for different customers, we do not think we significantly affected the results of our analysis. Cars with 0 values for engine size,engine power and register year were removed from the data set as no real car actually has those values. Removing them improved the performance of the algorithms. In the end, our data set contained 42 138 cars.

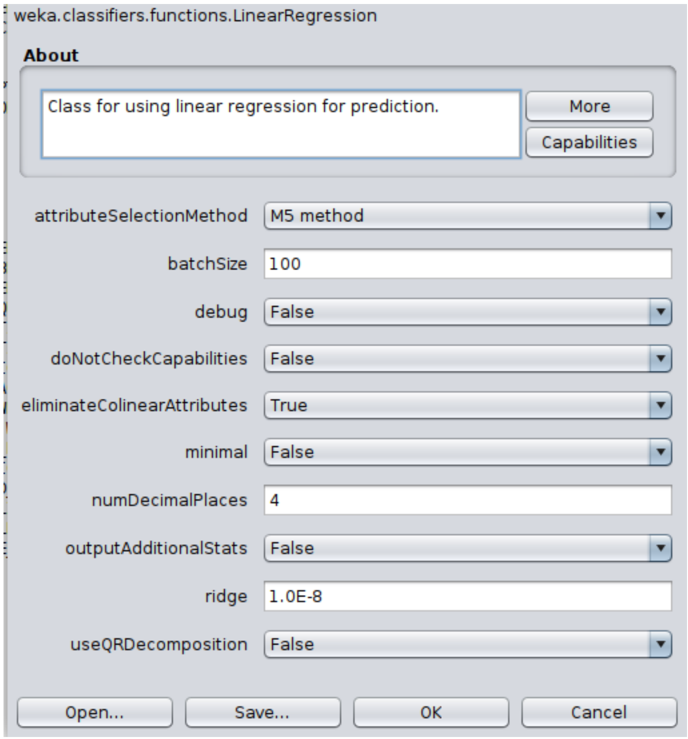
5 Approach

Our purpose is to see how much price is impacted by the other attributes of a car, which is approached as a typical regression problem. A high correlation is expected, as price should be set by the features of a car.
The Weka environment was used in order to apply 2 different regression algorithms and their results compared. Also, the algorithms were run multiple times, removing non-significant attributes on each run in order to identify which ones impact price the most. The algorithms we will compare are Linear Regression and Random Trees. For a detailed explanation of the code Weka implements refer to [12] for linear regression and [13] for Random Tree.

Linear regression is the process of fitting a line through data, in order to visualise the relationship between attributes.The response variable is on the Y axis, and X representing the other attributes. An increase in x by one would represent increases by one in, for example, the speedometer and the engine power. Qualitative variables, which can't be meaningfully increased or decreased, are converted into "dummy variables", where they impact the starting offset of the line. This algorithm was chosen because of its ease of interpretation, and because it provides a simple equation to predict future car prices.

Random trees is an algorithm that constructs a tree which considers K randomly chosen attributes at each node. It performs no pruning. It is a supervised classifier, an ensemble learning algorithm that generates many individual learners. The algorithm works as follows: the classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes".In our case of regression, the classifier response is the average of the responses over all trees in the forest. Refer to article [14] in sources for more details. We are using a K-value of 13, representing the number of attributes in our dataset, as we found that offers the best results.

The settings of Weka's linear regression can be observed in figure 1. It uses the M5 method for attribute selection and model tree generation. The M5 method minimizes standard deviation in its nodes, and prunes itself in order to avoid over fitting. Weka also eliminates co-linear attributes, which are attributes that are very similar to each other and can negatively impact the results of the algorithm. We are also using the percentage split method in order to avoid over fitting, using 70% for training and 30% for testing in both algorithms.We've also tried splitting $80 - 20$ and $90 - 10$, but got the best results when using $70 - 30$. Linear regression can also accept nominal attributes in the final equation, although it cannot predict them. If nominal attribute is equal to any of the ones in the equation, then the value is a one, otherwise, the value is a zero



weka.classifiers.functions.LinearRegression

About

Class for using linear regression for prediction.

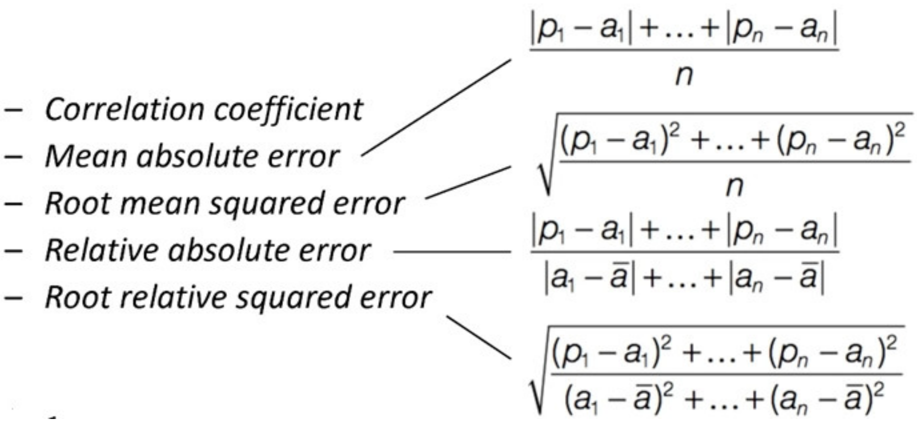
attributeSelectionMethod: M5 method
batchSize: 100
debug: False
doNotCheckCapabilities: False
eliminateColinearAttributes: True
minimal: False
numDecimalPlaces: 4
outputAdditionalStats: False
ridge: 1.0E-8
useQRDecomposition: False

Figure 1: LinReg Settings

Weka offers 5 statistical measures which we will use for evaluating the algorithms. We will mainly focus on the correlation coefficient, because it determines the strength of a correlation and the root mean square error, because it shows inaccuracies in the prediction models. These measures are defined in figure 2.

The correlation coefficient, is a measure of any linear trend between two variables. It's value ranges between -1 and 1. We expect it to be bigger than 0, as we will check how price is impacted by the other attributes, and all the attributes should affect price.The smaller the value of the coefficient the worse the data can be visualized by a single linear relationship.

The root mean squared error is a metric that tells us the average distance between the predicted values from the model and the actual values in the data set. The lower it is, the better a given model is able to "fit" a data set.



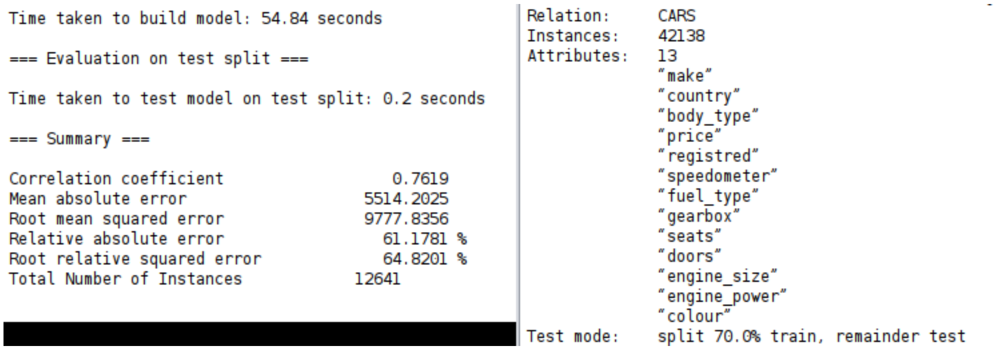
– Correlation coefficient
– Mean absolute error
– Root mean squared error
– Relative absolute error
– Root relative squared error
$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n}$
$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$
$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \bar{a}| + \ldots + |a_n - \bar{a}|}$
$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \ldots + (a_n - \bar{a})^2}}$

Figure 2: Measures

6 Results

Linear Regression

At first, the algorithm used all 13 attributes. The result is a moderately high value for the r^2, being 0.58, although the values of errors are extremely high, with the root mean square error and relative absolute error both being over 60%.The output can be observed in figure 3.

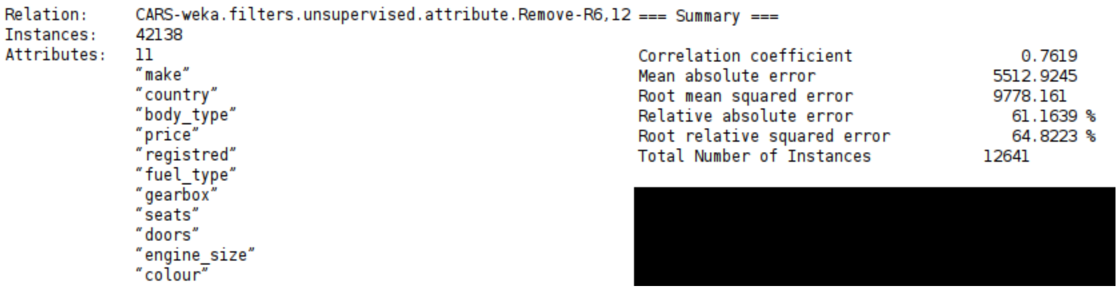


```
Time taken to build model: 54.84 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.2 seconds

=== Summary ===

Correlation coefficient                  0.7619
Mean absolute error                   5514.2025
Root mean squared error               9777.8356
Relative absolute error                 61.1781 %
Root relative squared error             64.8201 %
Total Number of Instances            12641
```

```
Relation:     CARS
Instances:    42138
Attributes:   13
              "make"
              "country"
              "body_type"
              "price"
              "registred"
              "speedometer"
              "fuel_type"
              "gearbox"
              "seats"
              "doors"
              "engine_size"
              "engine_power"
              "colour"
Test mode:    split 70.0% train, remainder test
```

Figure 3: LinRegResults1

The equation in total has over a 100 lines, and is not therefore fully included in the report. However, speedometer and engine power are not included in the equation. Figure 4 shows that correlation is not changed with their non-inclusion, and that error rates are only minimally affected,by a negligble amount. Therefore, we can conclude that, so far, the attributes speedometer and engine power do not significantly influence the price.



```
Relation:     CARS-weka.filters.unsupervised.attribute.Remove-R6,12
Instances:    42138
Attributes:   11
              ”make”
              ”country”
              ”body_type”
              ”price”
              ”registred”
              ”fuel_type”
              ”gearbox”
              ”seats”
              ”doors”
              ”engine_size”
              ”colour”
```

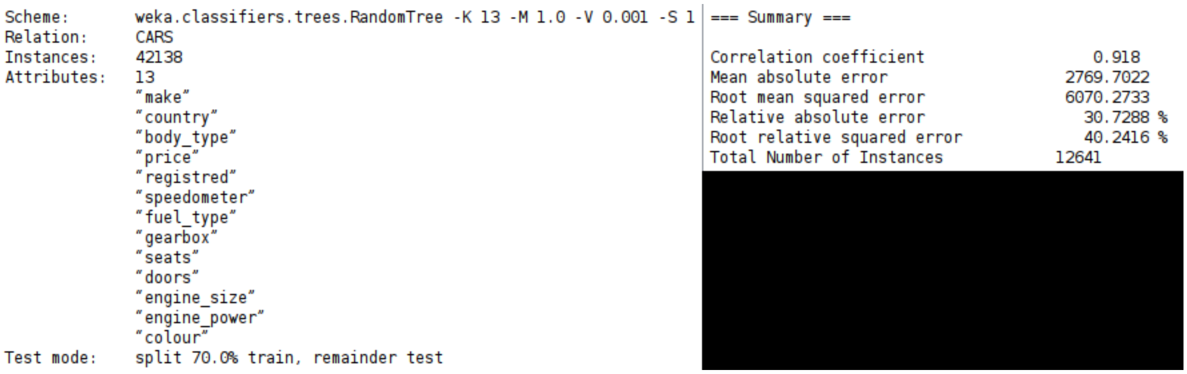
```
=== Summary ===

Correlation coefficient                  0.7619
Mean absolute error                   5512.9245
Root mean squared error               9778.161
Relative absolute error                 61.1639 %
Root relative squared error             64.8223 %
Total Number of Instances            12641
```

[redacted]

Figure 4: LinRegResults2

Linear regression provided an interesting result, showing that price is heavily impacted by the other attributes, but the errors were also quite big, meaning that there are many data points that were far away from the predicted line.

RandomTree Next, we run the RandomTree algorithm on all 13 attributes. The results are much better, with a high correlation of 0.918 or an r^2 of 0.84, and errors almost half as small as the ones in linear regression. They can be observed in figure 5.



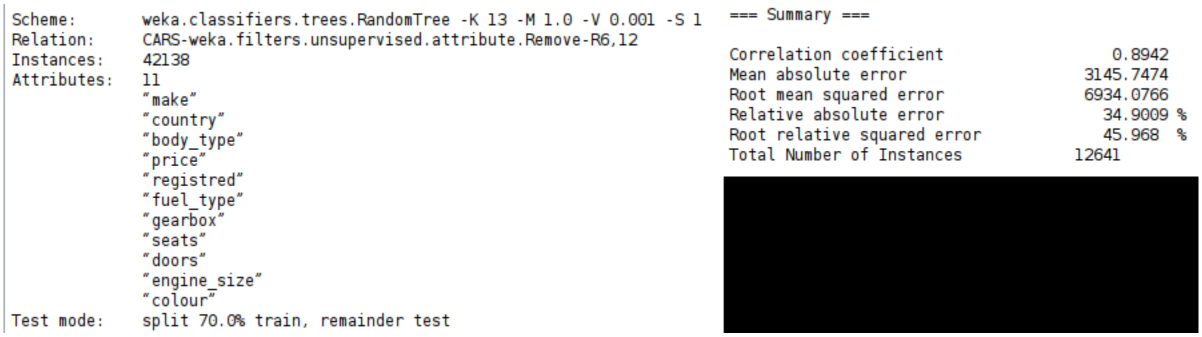
```
Scheme:       weka.classifiers.trees.RandomTree -K 13 -M 1.0 -V 0.001 -S 1
Relation:     CARS
Instances:    42138
Attributes:   13
              ”make”
              ”country”
              ”body_type”
              ”price”
              ”registred”
              ”speedometer”
              ”fuel_type”
              ”gearbox”
              ”seats”
              ”doors”
              ”engine_size”
              ”engine_power”
              ”colour”
Test mode:    split 70.0% train, remainder test
```

```
=== Summary ===

Correlation coefficient                  0.918
Mean absolute error                   2769.7022
Root mean squared error               6070.2733
Relative absolute error                 30.7288 %
Root relative squared error             40.2416 %
Total Number of Instances            12641
```

[redacted]

Figure 5: RandomTree1

Now, we will again remove the engine power and speedometer attributes to see if they indeed have no influence of the price. It can be seen in figure 6 that this algorithm reveals that the attributes actually influence the price, but by a small margin. It is an interesting result as the 2 attributes at first look like they should influence the price a lot. Engines, reflected in engine power, are expensive to make. Speedometer show how much a car has been driven, with cars that have been driven a long time usually being cheaper , due to being more worn down.Actually, they influence the price by a very small margin compared to the other attributes in the dataset.



```
Scheme:       weka.classifiers.trees.RandomTree -K 13 -M 1.0 -V 0.001 -S 1
Relation:     CARS-weka.filters.unsupervised.attribute.Remove-R6,12
Instances:    42138
Attributes:   11
              "make"
              "country"
              "body_type"
              "price"
              "registred"
              "fuel_type"
              "gearbox"
              "seats"
              "doors"
              "engine_size"
              "colour"
Test mode:    split 70.0% train, remainder test
```

```
=== Summary ===

Correlation coefficient                  0.8942
Mean absolute error                   3145.7474
Root mean squared error               6934.0766
Relative absolute error                 34.9009 %
Root relative squared error             45.968  %
Total Number of Instances            12641
```

[redacted]

Figure 6: RandomTree2

In conclusion, we can now say that the RandomTree algorithm performs much better on our dataset than linear regression and that price is heavily impacted by the other attributes of a car. Thus, our initial assumption was correct.

7 Suggestions for improvement of the results

The results could be improved with a cleaner dataset, where all of the cars matched their counterparts on the websites. This not being the case interfered with our model. Also, the removal of attributes with values of 0, such as doors and seats which are not possible in real life would most likely improve the results. It would also be useful to repeat the comparison with a dataset of cars from the same websites next year. It could be that the relationships found were influenced by the current economic situation and the lack of supply of new cars. Long-term observations would help confirm the results.

8 Interpretation of results

The results show that the price of a car is indeed related to it's attributes. This makes sense, as aspects such as the make of a car or it's engine size determine price in real life scenarios. However, why did engine power and speedometer not show much impact on the price? It could be that their information is already contained within other attributes. Engine size also provides information about how good the engine is, and the year in which a car was built provides similar information to how many kilometers it has driven.

9 Sources

In the order of appearance:

[1]
https://www.statista.com/statistics/425095/eu-car-sales-average-prices-in-by-country/
[2]
https://www.acea.auto/figure/motorisation-rates-in-the-eu-by-country-and-vehicle-type/
[3]
https://www.forbes.com/sites/neilwinton/2021/12/07/solid-european-car-sales-expected-after-c
[4]
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Household_budget_survey_-
[5]
https://edition.cnn.com/2021/09/03/business/gm-plant-closings-chip-shortage/index.html
[6]
https://www.volkswagenag.com/en/news/2020/03/coronavirus_pandemic.html
[7]
https://www.vox.com/the-goods/21507739/coronavirus-car-market-used-expensive
[8]
https://www.businessinsider.com/why-are-used-cars-so-expensive-now-shortages-pandemic-rental
[9]
https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2714
[10]
https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf
[11]
https://www.ceeol.com/search/article-detail?id=746689
[12]
https://weka.sourceforge.io/doc.stable/weka/classifiers/functions/LinearRegression.html
[13]
https://weka.sourceforge.io/doc.stable/weka/classifiers/trees/RandomTree.html
[14]
http://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf