Business analytics
Analysis of car advertisement data

Alina Ivanova, Moritz Hangen, Simon Gosch

2024/03/24

Contents

List of Figures

List of Tables

1 Introduction

[1]

2 Theoretical background

3 Methodology

3.1 Analyzation method

As the goal is to make a data driven decision on how much a Volkswagen (VW) Golf arriving at the dealership should be advertised for, linear regression is a fitting method to get a model for the price determination. After getting the car's specifications as inputs it calculates the price based on actual data of past advertisements.

3.2 Data processing

3.2.1 Preprocessing

First step of preprocessing the data is the reduction to only data points with value "VW Golf" for the car model attribute. Therefore, the Comma-separated values (CSV) file containing the raw data is imported into Microsoft Excel (Excel) and a filter to the car model column is applied. As some cells are blank for certain attributes, another filter removing the affected rows is set for every column.

3.2.2 Preparations for Orange

After the previous steps the data set is still not ready to be processed in Orange. The problem is that certain cells not only contain the actual numeric value but also the unit. For example the column of the car's average miles per gallon (MPG) always has the text "mpg" after the value, causing Orange to not recognize it as numeric value. To fix this issue the first approach is replacing the unit with blank text using the "Find and replace" functionality in Excel. But right after the text is replaced, Excel is converting some values to dates even if the cell format is set number. An example is the value 25.4 which is converted to "25. Apr".

The second approach to remove the units is opening the CSV file with a text editor and use its “Find and replace” functionality. This is working as the texts, which have to be removed, only appear in two other places, where they are manually added back. So both the “mpg” from the average MPG column and the “L” from the engine size column were removed like this.

3.2.3 Processing in Orange

Correlations with the price attribute

To understand how each attribute affects the car’s price, the correlations are calculated using the Pearson correlation coefficient. The result is a value between -1 and 1 for each attribute. The closer an attribute’s value is to 0, the less influence it has on the price.

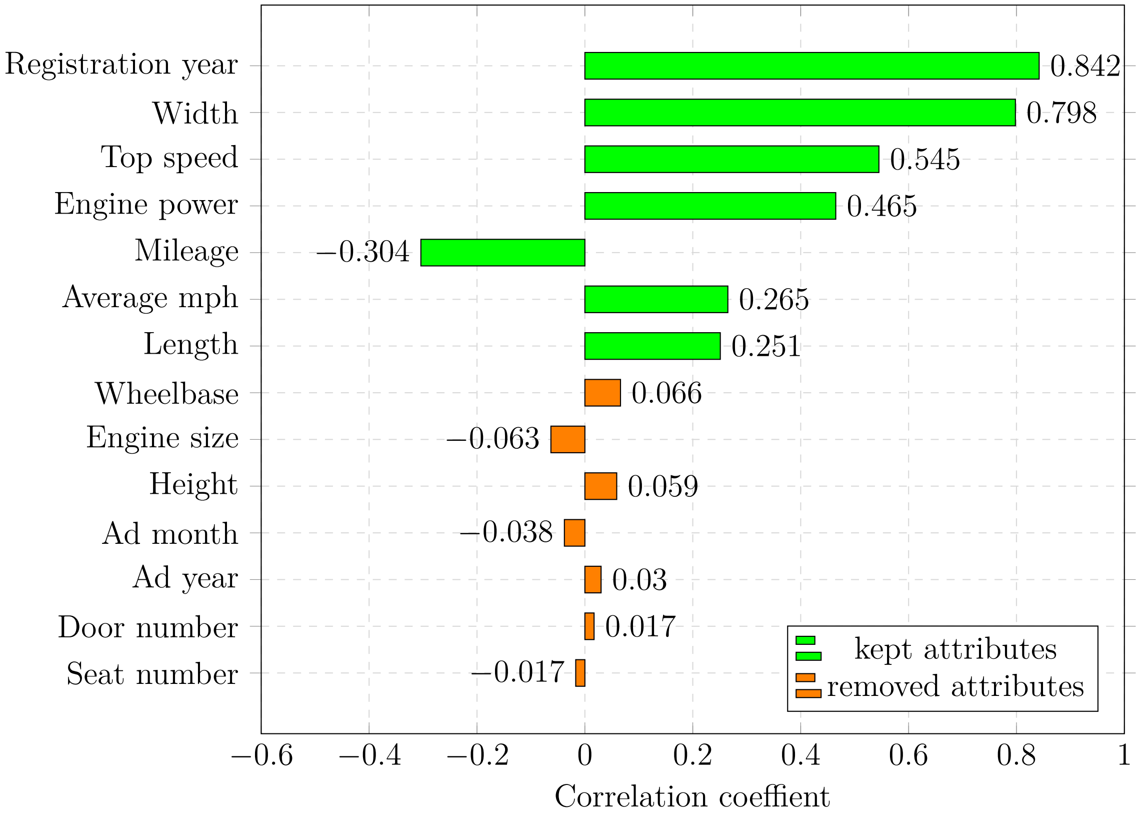

Figure 3.1: Correlation with the price attribute

Figure 3.1 shows the results which include both positive and negative correlation. Positive correlation means that if the value of attribute 1 increases, the value of attribute 2 does as well. A logical example is the engine power, as it makes sense that a more powerful car sells for a higher price. On the other hand negative correlation describes the opposite. A higher mileage results in a lower price as it indicates higher wear.

In the figure the attributes are listed from the highest to the least influence on the price. As the ones marked in color orange do not really have an affect, they are removed from the data set and only those highlighted in green are kept.

Removing attributes

Dividing data set

Calculating Linear Regression

3.3 Linear Regression Model

3.3.1 Results

3.3.2 Visualization

4 Findings and discussion

4.1 Model usage

4.1.1 Assessing expectations

Registration year	**Mileage**	Horsepower	Width (mm)	Length (mm)	Average mpg	Top speed (mph)	**Predicted price**
2019	**60000**	135	2027	4284	49	116	**17596**
2019	**130000**	135	2027	4284	49	116	**15267**

Table 4.1: Assessing expectations

4.1.2 Usage of example data

Registration year	Mileage	Horsepower	Width (mm)	Length (mm)	Average mpg	Top speed (mph)	**Predicted price (£)**
2014	180000	110	1799	4204	45	110	**5601**
2016	150000	120	2027	4255	48	112	**12130**
2018	80000	130	2027	4255	50	115	**16136**
2015	190000	115	1799	4204	44	108	**5819**

Table 4.2: Assessing model for business use cases

4.2 Findings

4.2.1 In line with presumptions

Mileage ↔ Price

Engine size ↔ Price

4.2.2 Outliers

Engine size ↔ Price

Width ↔ Price

Year ↔ Price

TODO include example of higher influence of mileage the older the car is

5 Conclusion

5.1 Further questions

Lack of sales data regarding advertisement

While the given dataset included cohesive data regarding the advertisement price of cars in the used vehicle market, there is no indicator given, whether the car was actually sold at the price that it has been advertised for.
Nevertheless, while you can expect a few dealerships to over- / undershoot their prices, considering the scale of the dataset for the overall market you can expect that effect to even out. However, evaluating whether there is a trend that used cars are systematically under- / overpriced in advertisements is only possible if you compare the given advertisement data set to data containing real sales.

Inter-model comparison of findings

Assess value depreciation

One possible further research topic is the comparison of a car's new price to its future advertisement prices. For each model in the used-car advertisement dataset, there is a corresponding data point that contains, among others, the manufacturer's suggested retail price (MSRP).
Given that information, the following questions could be analyzed:

- Which model retains the most value compared to its MSRP?
- Given five years of use, which car's price decreased the most?
- Is there a correlation between value depreciation and the manufacturer of the car?

This information can be useful to customers considering the purchase of a new car in order to assess its potential resale value in the future.

5.2 Résumé

Bibliography

[1] *Industrie-Roboter - Hirata Engineering Europe GmbH.* URL: `https://www.hirata.de/de/produkte/scara-roboter` (visited on 02/02/2024).