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

IndustrieRoboterHirataEngineering

2 Theoretical background

3 Analysis

3.1 Analyzation method

The goal is to make a data driven decision on how much a Volkswagen (VW) Golf arriving at the dealership should be advertised for based on the car's specifications, assuming there is some kind of dependence. To check whether this dependence is given, the correlations between the specifications and the price are evaluated. After that, to get a model which predicts the price based on the other attributes, linear regression is used. Its simplicity when it comes to understanding and calculating the model as well as interpreting the results make it fitting for this use case.

3.2 Data processing

3.2.1 Preprocessing in Excel

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 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 CSV file is opened in a text editor and its "Find and replace" functionality is used to replace the unit with blank text. 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.

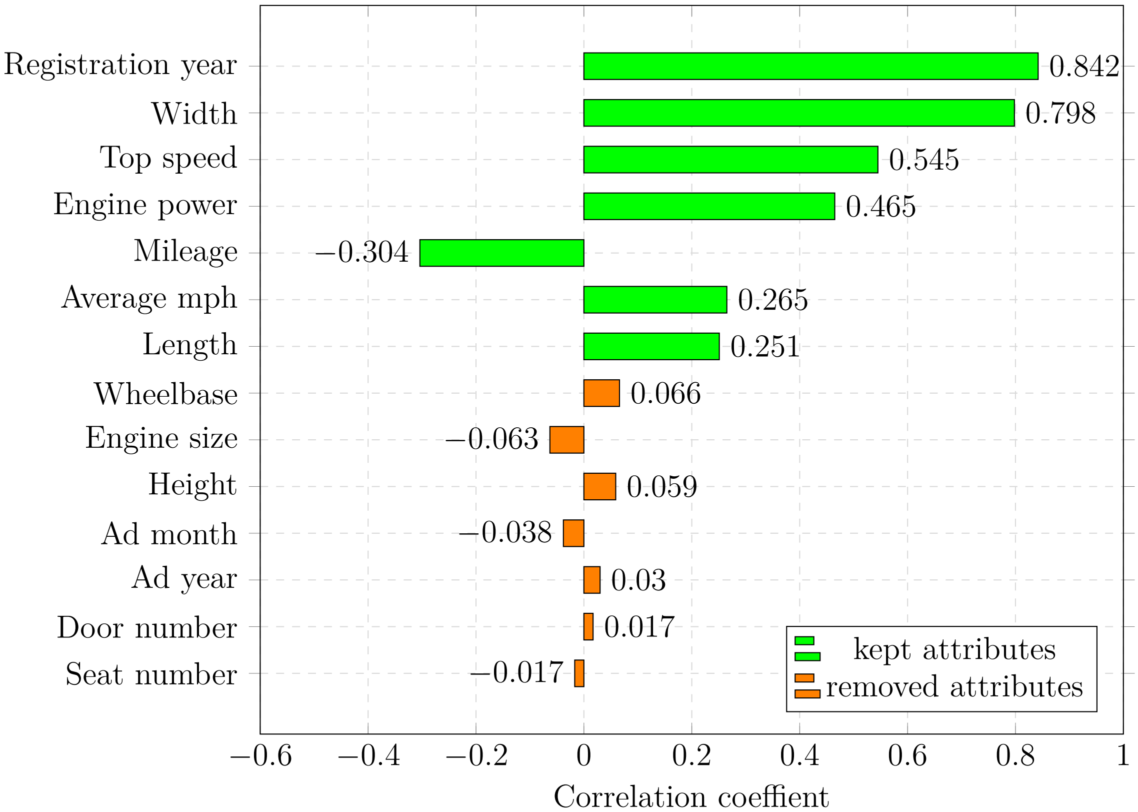

Figure 3.1: Correlation with the price attribute

Figure 3.1 shows a horizontal bar chart visualizing the correlation coefficient for each attribute. As a bar chart lists the different values of the coefficients side-to-side and in an easy-to-read way, it is perfect for the given purpose of focusing on comparison between the attributes. The horizontal layout allows good readability for the long labels and is more fitting to put the actual values next to the bars as there are no space limitations along the y-axis.

Having the results of the correlation analysis allows to answer the first research question: “What factors influence Volkswagen Golf’s price the most?”. Figure 3.1 provides the answer as the attributes are listed from highest to lowest influence on the price. Those highlighted in the color green have an eminent impact, while the rest doesn’t. Therefore, those highlighted in the color orange are not taken into account for the further data analysis.

Removing attributes & Setting target

In the next step not only the attributes with low correlation, but also the non-numeric ones are removed. This is the case as they can’t be used in the calculation of the linear regression model without further preprocessing. These additional preprocessing steps are not performed as the resulting linear regression model already is very good and, therefore it is not worth the extra effort. Additionally, the price is set as target value for the linear regression calculation.

Dividing data set & Calculating linear regression

As a final step the data set is split into training and test data and the linear regression model is calculated based on the training data. Insights about the training split as well as quality and specifications of the model are discussed in the next section of the report.

3.3 Linear regression model

3.3.1 Specifications

For the calculation of the linear regression model a training split of 60% training and 40% test data is used. When testing the model the results are very satisfactory with a mean squared error of 3750250.922 and r^2 of 0.912. Especially the r^2 being very close to 1 means a high linear relationship between the model and the target attribute, which means that the model predicts the price accurate.

3.3.2 Visualization

The reason for visualizing the linear regression model is that it is always easier to analyze and interpret visual results rather than only working with numeric values. Used for that matter is a scatter plot shown in Figure 3.2.

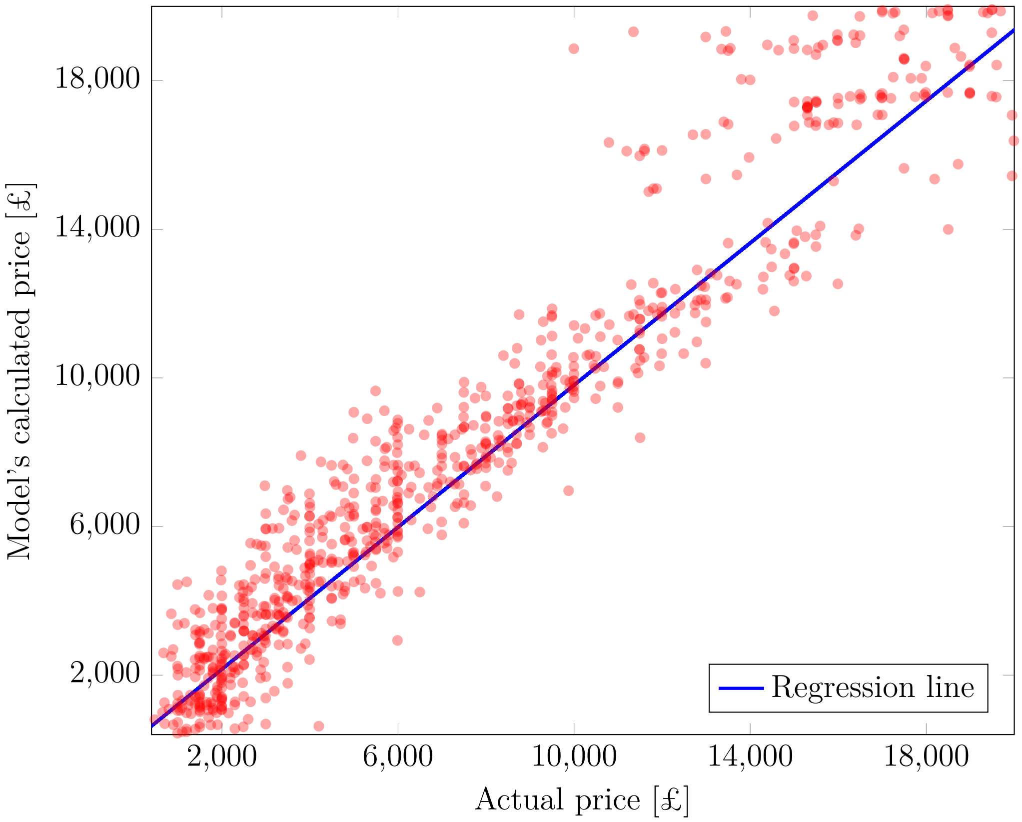

Figure 3.2: Plotted Linear Regression Model

Scatter plots are optimal to show the relationship between two variables, in this case the actual price from the test data and the price predicted by the linear regression model. The closer the data points are to the regression line, the better the model performs. By knowing this, it is easy to analyze the quality of the model even without any prior knowledge. So when looking at Figure 3.2, one can see that the data points portray the regression line quite precisely. This reinforces the prior statement, that the linear regression model is very good.

4 Findings and discussion

4.1 Model usage

After examining the theoretical background of the regression model, it can be applied to example data to investigate its behavior.

4.1.1 Assessing expectations

Intuitive assumptions draw you to the conclusion that the vehicle's mileage has an inverse relationship to the predicted advertisement price. To evaluate if the model also follows this behavior, it was applied to manually created data differentiating only by mileage.

Registration year	**Mileage (mi)**	Horse-power	Width (mm)	Length (mm)	Average mpg	Top speed (mph)	**Predicted price (£)**
2017	**60000**	135	2027	4284	49	116	**16157**
2017	**130000**	135	2027	4284	49	116	**13828**

Table 4.1: Impact of mileage on recently registered cars

As evident in Table 4.1, a higher mileage in fact reduces the predicted price. However, for an over 2-fold increase in miles, the depreciation is with 14.4 % not as high as initially expected.

While comparing two otherwise identical cars, it should be noted that the other values still measurably contribute to the result.
In particular the registration year, which, as shown before, strongly correlates with the target variable, has an effect on the limited influence of the mileage here. Given the data set's sampling of data up to 2017, both of the VW Golfs in Table 4.1 have been first registered very recently, thus the base price is higher. If the same example is applied to cars registered in 2010, which therefore have been running for seven years, the influence of the mileage grows.

Registration year	**Mileage (mi)**	Horse-power	Width (mm)	Length (mm)	Average mpg	Top speed (mph)	**Predicted price (£)**
2010	**60000**	135	2027	4284	49	116	**11122**
2010	**130000**	135	2027	4284	49	116	**8793**

Table 4.2: Influence of mileage on older cars

As shown in Table 4.2, the gap between the two otherwise identical vehicles has widened to 20.9 %, an increase by 45.3 %. Potential reasons for the strong correlation between year and target value will be investigated further in section 4.2.2. Nevertheless, for that small subset of data, the efficacy of the model is evident.

4.1.2 Usage of example data

However, this small example is not cohesive enough to demonstrate the ability to solve the overall business problem. To apply the model to a day-to-day use case as it regularly appears in a dealership, it was used to predict the advertisement price of four automobiles for demonstration purposes.

Registration year	Mileage (mi)	Horse-power	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 4.3: Assessing model for business use cases

Table 4.3 illustrates that the model is able to set a competitive advertisement price for a VW Golf. As its performance with $R^2 = 0.927$ is very good, it can be seen as a reliable source of information for the dealer. Thus, he does not have to rely solely on human estimate but can instead decide based on a data driven prediction.

4.2 Findings

In section 4.1 the application of the model to exemplary vehicles has been demonstrated. Here, some trends that could have already been anticipated after the correlation analysis, have emerged more clearly.
These trends that have been found regarding the model and the underlying data can be divided into two groups, the predicted and the unexpected.

4.2.1 Consistent with expectations

In this first part, the findings which can be considered logical or self-evident will be discussed.

Mileage $\leftrightarrow$ Price

Mileage as a parameter can be seen as an indicator for wear of the vehicle. Apart from that, maintenance parts are closer to their end of life and need to be replaced sooner, which costs customers money and time.

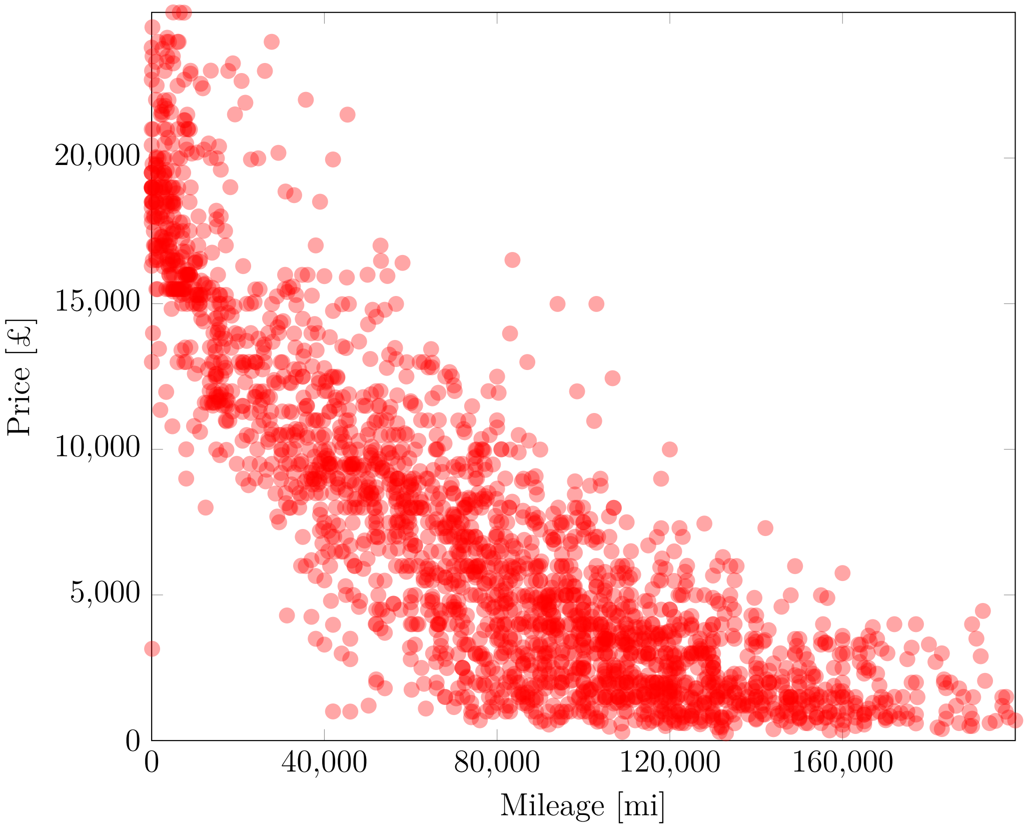

Figure 4.1: Effect of mileage on price

Looking at the data in Figure 4.1, it can be seen that there is a steep decline of the vehicle's value in the beginning up to ≈ 50000 miles, with the effect of mileage on the price notably decreasing from ≈ 100000 mi onwards. A potential reason for this could be that once vehicles reach certain thresholds, most of the common maintenance has already been executed, so a high distance travelled becomes an indicator of the car's reliability, countering the diminishing effect on the price to a certain extent.

This also means that the relationship between the two markers, resembling the arm of a parabola, is not clearly following a linear trend, therefore making it difficult to fully explain with the linear Pearson correlation only. Nevertheless, its negative value of -0.304 is in line with predictions, albeit not as strong.

Horsepower ↔ Price

Among the essential specifications of a vehicle is its power, measured in horsepower, significantly influencing price with a correlation coefficient of 0.465. It affects the maximum acceleration, the top speed, the fuel consumption and other key indicators of a car's capabilities. The improved performance is a reason why customers are willing to spend extra for a more powerful Golf.

4.2.2 Outliers

Nevertheless, there are also some outliers which ought to be investigated in more detail.

Engine size ↔ Price

The engine displacement, informally also referred to as engine size, describes the volume of air and fuel inside an engine's pistons [1] and is measured in liters in the data set. By definition, it is also related to its power output, in particular in older vehicles.

Given that, the intuitive prediction is that it will positively correlate with the final price. However, the analysis has shown it is not significant, with the coefficient of -0.063 approaching 0.

Exploring the correlation of the engine size with other parameters, the lack of effect on the final price can be explained by looking at three key indicators:

- **Top speed:** With 0.552, there is a significant effect of the engine size on the top speed, which in turn increases the predicted value of the Golf.
- **Registration year:** There is a measurable trend that for newer cars, the engines become smaller, notable by a coefficient of -0.295. Due to the strong effect of the registration year on the predicted price, the higher engine sizes negatively influence the target value, compensating the aforementioned effect of top speed.
- **Horsepower:** In newer cars, horsepower is not as limited by engine size as in the past [2]. Looking at the correlation value of -0.076, this effect is also resembled in VW Golfs. Thus, one of the deciding factors determining the end price is not

related to the engine size.

To conclude, the two measurable effects even each other out and for the horsepower, where correlation might be present, there is not any, unexpectedly rendering the engine size negligible for our analysis.

Width ↔ Price

At the first glance, a very strong tie between the width of a Golf and its advertisement price is not clear. From a potential customer's perspective, the width of a car can be a deciding factor, for example when it comes to narrow parking spots and small neighborhood roads. Nonetheless, there is a strong positive correlation with the target value, implying that the dimensions of the car could play a deciding factor in the purchase process.

Investigating the matter based on the data, the reason for the connection between width and price becomes evident: The width is very positively correlated with the registration year, which in turn is very strongly correlated with the Golf's advertisement price. This can be attributed to the fact that VW Golfs get wider every model generation, as for instance a VW Golf V is 1759 mm, while the current 8th generation's width is already 1789 mm [3].

Overall, there is a clear trend observable that in line with the registration year, the width increases, resulting in the unforeseen correlation of 0.715.

Year ↔ Price

As expected, there is a relationship in between the registration year of the car and the final price. However, its strength with 0.842 is unprecedented and can not solely be clarified by the analyzed data. The main reason for this is the absence of the exact model generation (Golf V, Golf VI, Golf VII...), as well as its respective base price, in the data. Nonetheless, it can be assumed that year of registration is approximately equal to the manufacturing date and therefore also to the model.

As prices have risen between the model generations, for instance a 16.3 % price increase happened between comparable models of Golf V and Golf VII [4], it can be expected that

this trend is reflected in advertisement prices as well. Additionally, newer cars include more optional features which can drive up the price if they are present. If this affects the price in the given case cannot be checked using the available data, as there is no notice of a car's selected options.
To add to that, there might also be differences in a car's maintenance cost and reliability depending on the exact model, manifesting in price differences in pre-owned vehicles. Newer models can also contain more innovations, whose absence can make older vehicles disproportionately less attractive, with features such as air conditioning, heated seats, navigation and more being considered standard nowadays.

Overall, trends and innovations that originate from the cost and behavior of new vehicles also manifest in the data set. By indirectly specifying the model, the registration year is thus a strong indicator of a Golf's future advertisement price.

5 Conclusion

5.1 Limitations of the analysis

5.1.1 Recency of the data

The training data has been collected between 2016 and 2017, thus even the most recent data is now over 6 years old. Given recent events such as the steep increase of inflation and the COVID-19 pandemic, predictions by the model might not accurately reflect current trends.

5.1.2 Lack of sales data regarding advertisement

Additionally, while the given dataset includes conclusive 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 one may anticipate a few dealerships to over or underestimate their prices, considering the scale of the dataset, that effect levels out for the overall market. However, evaluating whether there is a trend that pre-owned cars are systematically under- / overpriced in commercials is only possible if you compare the given data set to real sales information.

5.2 Further research topics

Given the scale of the available data, more potential research questions may also be examined.

5.2.1 Inter-model comparison of findings

The analysis is currently only valid for the small subset of the data including the VW Golf. To elevate generalizability to the whole used vehicle market and to assess potential disparities as well as similarities among car models, expanding the scope to the entirety of available data is recommended.

5.2.2 Assess value depreciation

After evaluating inter-model differences, a possible further research topic is the comparison of each car's new price to its future advertisement prices. For each model in the advertisement dataset, there is a corresponding data point in the "basic information" table that contains, among others, the manufacturer's suggested retail price (MSRP).
With 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 for customers considering the purchase of a new car in order to assess its potential resale value in the future.

5.3 Résumé

The conducted analysis provides a data-driven approach to the business problem of a competitive advertisement price of a given VW Golf. After preparing the raw data, key influences have been separated using Pearson-correlation and used to train a linear regression model.

Considering the parameters

- Registration year
- Mileage (mi)

- Horsepower
- Width (mm)
- Length (mm)
- Average mpg
- Top speed (mph)

it accurately predicts a suitable advertisement price with $R^2 = 0.927$, solving the business problem thereby. The end results mostly align with logical assumptions, for instance that newer cars sell for higher prices, yet also reveal surprising behaviors such as the significant effect of the Golf's width. To summarize, they provide further insight by revealing each factor's contribution to the vehicle's advertisement price and investigating their cause.

Given those results, further potential research areas are examined and left open for future analysis.

Bibliography

[1] *Engine Displacement*. In: *Wikipedia*. Feb. 28, 2024. URL: `https://en.wikipedia.org/w/index.php?title=Engine_displacement&oldid=1210756038` (visited on 03/27/2024).

[2] *What Is Engine Displacement and Does It Matter?* Capital One Auto Navigator. URL: `https://www.capitalone.com/cars/learn/finding-the-right-car/what-is-engine-displacement-and-does-it-matter/2055` (visited on 03/27/2024).

[3] *Volkswagen Golf | Technical Specs, Fuel Consumption, Dimensions*. URL: `https://www.auto-data.net/en/volkswagen-golf-model-896` (visited on 03/29/2024).

[4] *Duel: VW Golf V 1.6 FSi vs VW Golf VII 1.0 TSi*. ZePerfs. URL: `https://zeperfs.com/en/duel2995-5920.htm` (visited on 03/29/2024).