

Predicting the resale value of used cars

Created using R Markdown

Simone Andreetto
01635069

Felix Winterleitner
01612776



Figure 1: This is a teaser

ABSTRACT

This is Assignment 3 in Business Intelligence @ TU Wien in the winter term of 2020.

KEYWORDS

Business Intelligence, R, Rmd, Spark

1 INTRODUCTION

For this project, we used a dataset found on Kaggle.com provided by user Aditya [1]. It contains a collection of different used car listings obtained by searching through online marketplaces using a web scraper. The dataset is split into different files, one per car brand. The brands for which data is available are:

- Audi
- BMW
- Ford
- Hyundai
- Mercedes
- Skoda
- Toyota
- Vauxhall (= Opel in Great Britain)
- VW

Additionally, the data set contains files with premade subsets of above mentioned car brands, for example *coclass.csv*, which contains only listings for the Mercedes model C Class. We chose to only utilize the unfiltered datasets.

Apart from the car brand, there are a number of other attributes available for each data entry.

- car model

- year of first registration
- transmission type
- mileage
- fuel type
- tax
- miles per gallon of fuel
- engine size

as well as the target variable price.

2 BUSINESS UNDERSTANDING

a. Scenario

A group of entrepreneurs in the used car business want to counteract the ongoing trend of people selling their cars to other private individuals directly without involving commercial reseller, which has become very easy given the availability of online market places for used goods. The idea is the following: Customers are offered a new web-based platform where they can enter the most important key fact about the car they would like to sell. The platform immediately returns a first estimate of the price the platform owners would pay for the car. This estimate should be based on a model created from the used car listing data available.

b. Business Objectives

The business objectives is in short:

What is the expected value of a used car based on the given data entries?

Answering this question helps the platform in multiple ways.

- make it more convenient for customers to sell their used car but getting an accurate first estimate right after entering the data
- increase revenues by missing out on fewer chances to buy used cars (more cars resold via the platform instead of directly to other buyers)
- speed up final evaluation of car value by offering a good starting point
- base offers made to customers on true market values

c. Business Success Criteria

The following criteria need to be met by the prediction:

- The estimate should lead to a conversion rate of more than 30%, meaning that at least 30% of the users that enter their car data on the website actually proceed to sell their car on the platform.
- The estimations should never lead to an effective loss for the company. Therefore, estimations that are too high need to be avoided.

d. Data Mining Goals

In order to fulfill the business objective of determining an accurate price estimate, a regression problem needs to be solved. The input data consists of the 9 attributes mentioned above.

e. Data Mining Success Criteria

Regarding the result of the estimation, one important success criterion is:

- The estimate needs to be within a range of the actual price +/- 10% for 95% of the estimations made.

This is important because estimates that are further off the actual price may lead to:

- People aborting the process when the estimate is much lower than their expectation
- People entering the negotiations with far inflated expectation, effectively reducing the changes of the platform owners to score a good deal

Citation example.[1] Plot example. **The maturation of the song over time is shown in Figure 2.**

Small plot example.

3 DATA UNDERSTANDING

In the following section, a data description report containing data types, statistical properties, data quality aspects as well as a visual exploration of data properties is presented.

a. Data Types

The attributes in the data set have the data types shown in **Table 1**.

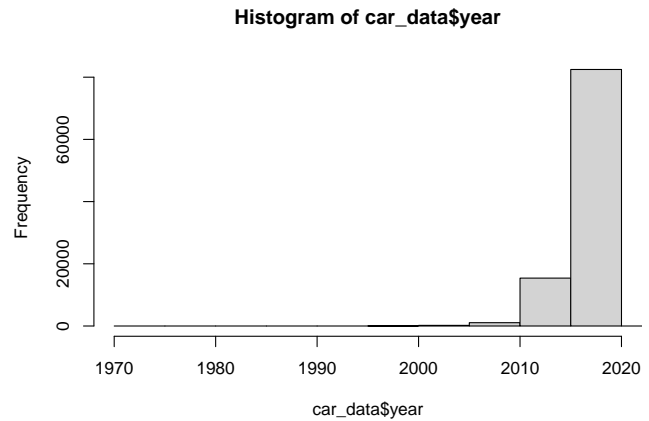


Figure 2: This is how great Tribute gets over time

Table 1: Data Types of the source data

Attribute	Type
model	String, nominal
year	Integer, interval
transmission	String, nominal
mileage	Integer, ratio
fuelType	String, nominal
tax	Integer, ratio
mpg	Float, ratio
engineSize	Float, ratio
price	Integer, ratio

b. Statistical Properties

```
## [1] 39176
```

```
## # A tibble: 1 x 10
```

```
## brand model year price transmission mileage fuelType ta
```

```
## <chr> <chr> <dbl> <dbl> <chr> <dbl> <chr> <dbl> <
```

```
## 1 Ford Fiesta 2060 6495 Automatic 54807 Petrol 205
```

Data Description Report presenting a. Data types, b. Statistical properties c. Data Quality aspects d. Visual Exploration of data properties and hypotheses

d. Visual Exploration

In the following figures, boxplots illustrate the ranges of the numeric (ratio) variables.

There are a few things we can learn from these initial diagrams. For example, it is interesting to see car listings contained in the data set are mostly for rather new cars, with a mileage median of less than 25000 miles. Taking a look at **Figure 5**, this suspicion is confirmed. The vast majority of cars in the dataset is indeed less than five years old.

Table example.

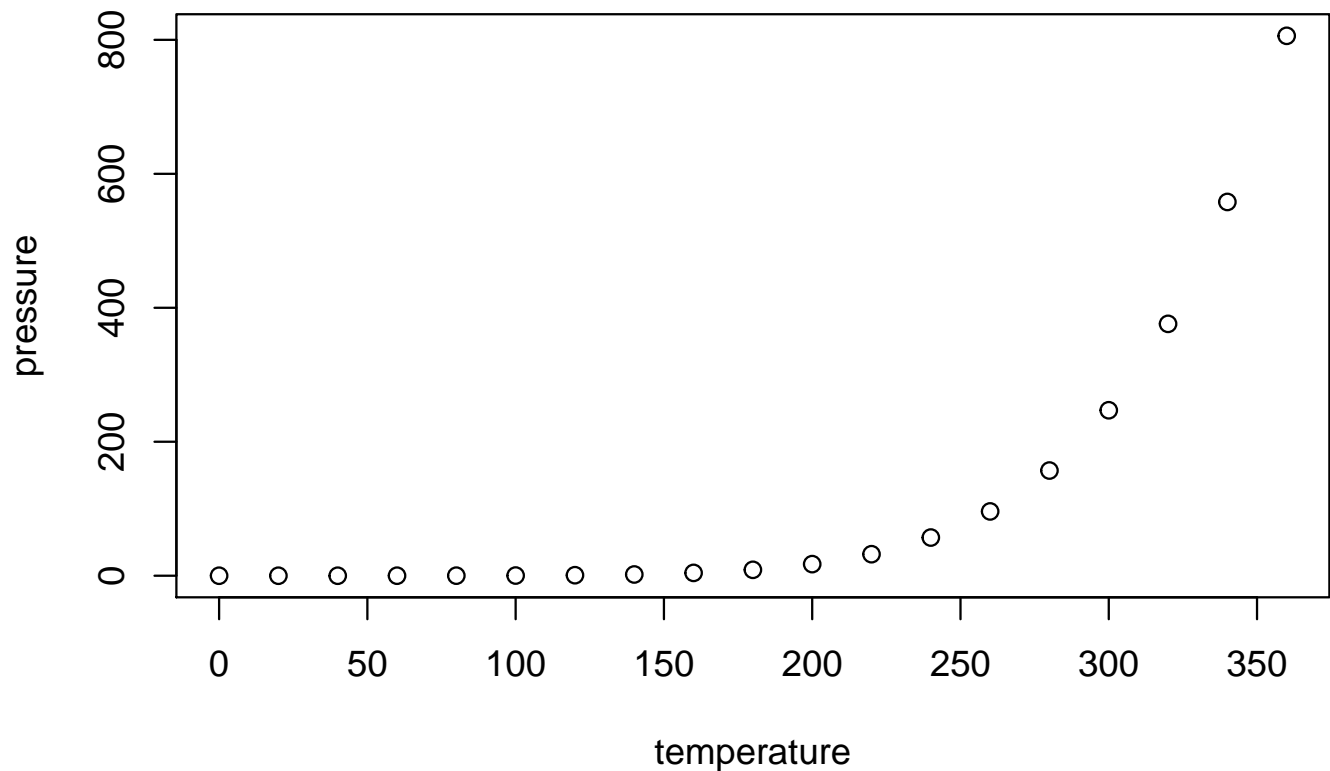


Figure 3: This is a two-column plot of how great Tribute gets over time

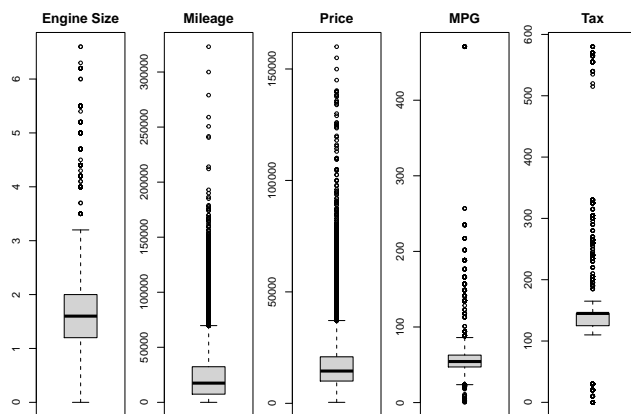


Figure 4: Boxplots on the distribution of the numeric attributes

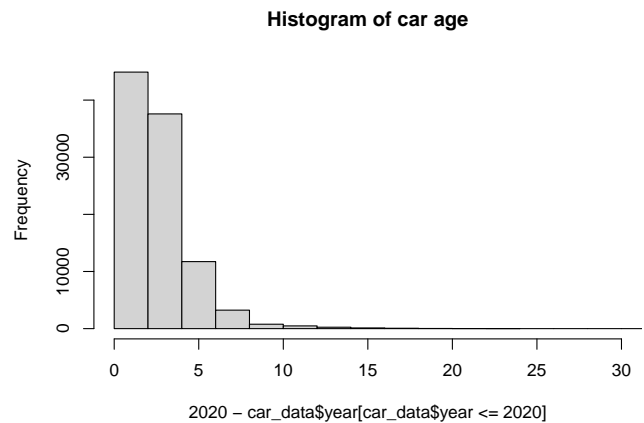


Figure 5: Histogram of car age.

4 DATA PREPARATION REPORT

Insert content here.

- a. Analyze options and potential for derived attributes (note: if the potential is considered low, these obviously do not necessarily have to be applied for your analysis, but options should be documented)

- b. Analyze options for additional external data sources, attributes that might be useful to better address the business objectives or data mining goals (Note: this description may be hypothetical, i.e. you are not necessarily required to actually obtain and integrate the external data for the analysis)

e.g. Car Horse Power.

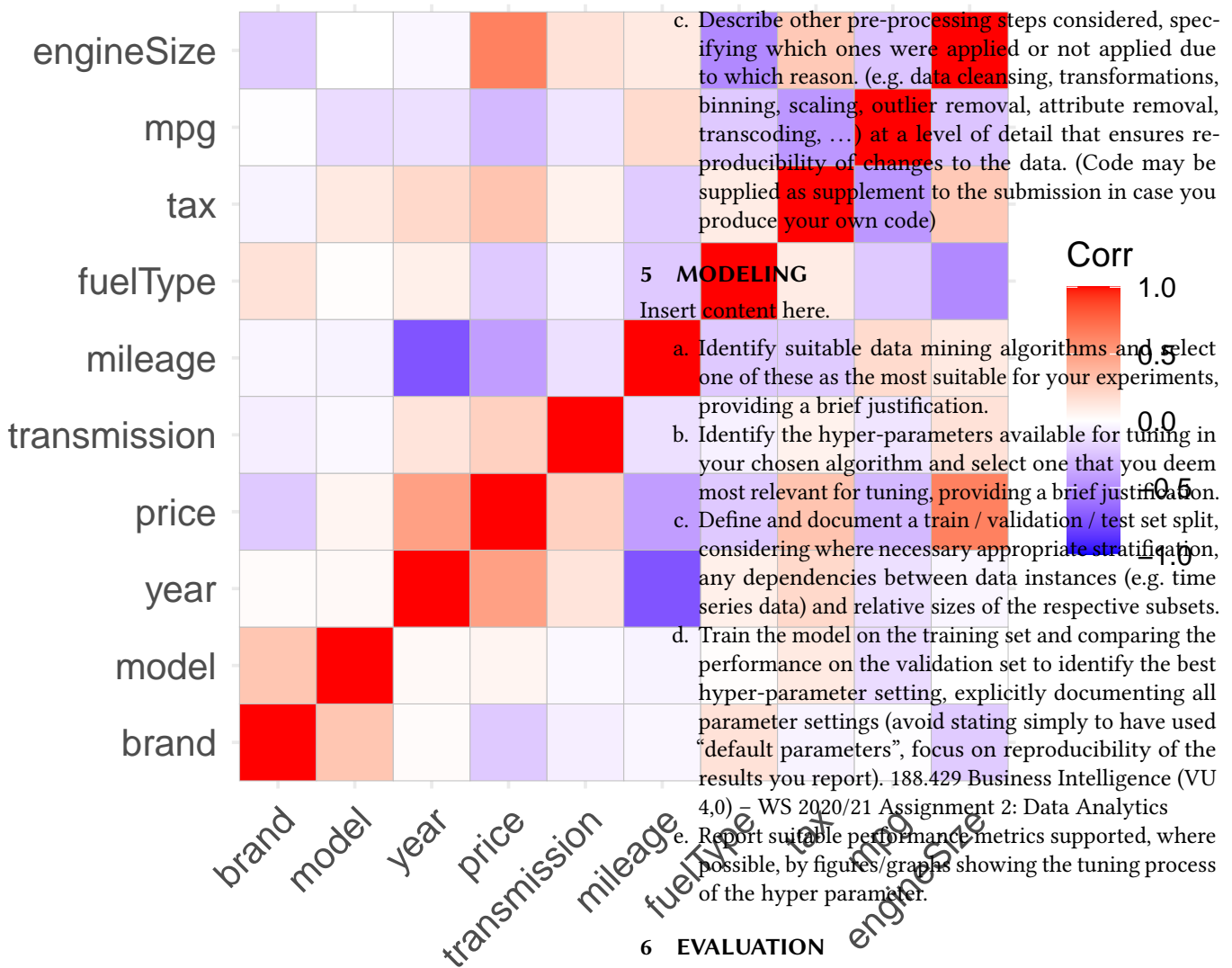


Figure 6: Pairs of all numeric attributes

Table 2: Example Table Showcasing capabilities

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5.0	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.9	3.1	1.5	0.1

of the data space, ... (Note your goal is not necessarily to obtain a better result than what has been reported in the state of the art, this is not a grading criterion! On the other hand, if the performance of your classifiers is massively below the state of the art (or even below a random baseline or trivial acceptor / rejecter) you may want to investigate the reason...)

- e. Compare the performance obtained with the Data Mining success criteria defined in the business understanding phase.

7 DEPLOYMENT

Insert content here.

- a. Compare the performance obtained with respect to the needs for addressing the business success criteria and provide recommendations for deployment (fully automatic, hybrid solutions, deploying only for a part of the data space, ...) as well as recommendations for subsequent analysis.
- b. Consider and briefly document potential ethical aspects as well as impact assessment / risks identified in deployment
- c. Document aspects to be monitored during deployment, specifying triggers that should lead to intervention.
- d. Briefly re-visit reproducibility aspects reflecting on aspects well documented and those that might pose a risk in terms of reproducibility based solely on the information provided in this report

8 SUMMARY OF FINDINGS

Insert content here.

- a. Briefly summarize your overall findings and lessons learned
- b. (optional) Provide feedback on this exercise in general: which parts were useful / less useful; which other kind of experiment would have been interesting, ... (this section is, obviously, optional and will not be considered for grading. You may also decide to provide that kind of feedback anonymously via the feedback mechanism in TISS – in any case we would appreciate learning about it to adjust the exercises for next year following a major re-structuring this year based on feedback obtained.)

REFERENCES

- [1] Aditya. 2020. 100,000 UK Used Car Data set. <https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes?select=vw.csv>