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 *cclass.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 facts 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 by 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 +/- 15% 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

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.**

Table 1: Data Types of the source data

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

b. Statistical Properties

```
options(width = 60)
summary(car_data)
```

```
##     brand              model                 year
## Length:99187      Length:99187      Min.   :1970
## Class :character  Class :character  1st Qu.:2016
## Mode  :character  Mode  :character  Median :2017
##                                     Mean   :2017
##                                     3rd Qu.:2019
##                                     Max.   :2060
##  transmission         mileage         fuelType
## Length:99187      Min.   :     1   Length:99187
## Class :character  1st Qu.:  7425   Class :character
## Mode  :character  Median : 17460   Mode  :character
##                   Mean   : 23059
##                   3rd Qu.: 32339
##                   Max.   :323000
##       tax             mpg           engineSize
##  Min.   :  0.0   Min.   :  0.30   Min.   :0.000
##  1st Qu.:125.0   1st Qu.: 47.10   1st Qu.:1.200
##  Median :145.0   Median : 54.30   Median :1.600
##  Mean   :120.3   Mean   : 55.17   Mean   :1.663
##  3rd Qu.:145.0   3rd Qu.: 62.80   3rd Qu.:2.000
##  Max.   :580.0   Max.   :470.80   Max.   :6.600
##      price             age
##  Min.   :   450   Min.   :-40.000
##  1st Qu.:  9999   1st Qu.:  1.000
##  Median : 14495   Median :  3.000
##  Mean   : 16805   Mean   :  2.912
##  3rd Qu.: 20870   3rd Qu.:  4.000
##  Max.   :159999   Max.   : 50.000
```

c. Data Quality aspects

Since the data is recorded from the internet there is the possibility of it containing invalid information or missing values.

To begin with, we check the data for missing values. However, in this speficic case, there are none.

```
dim(car_data) ==
  dim(car_data[complete.cases(car_data),])
```

```
## [1] TRUE TRUE
```

Next up, we check plausibility of some of the extreme cases of numerical values. To keep it short, we only included one exemplary output her and then summarize the findings.

```
options(width = 60)
head(car_data[order(car_data$age),], 5)
```

```
## # A tibble: 5 x 11
##   brand model  year transmission mileage fuelType   tax
##   <chr> <chr> <dbl> <chr>          <dbl> <chr>    <dbl>
## 1 Ford  Fies~  2060 Automatic      54807 Petrol     205
## 2 Audi  Q7     2020 Semi-Auto         10 Diesel     145
## 3 Audi  Q5     2020 Semi-Auto         10 Petrol     145
## 4 Audi  Q5     2020 Semi-Auto         10 Petrol     145
## 5 Audi  A4     2020 Semi-Auto         10 Petrol     145
## # ... with 4 more variables: mpg <dbl>, engineSize <dbl>,
## #   price <dbl>, age <dbl>
```

```
# year 2060 is an error

head(car_data[order(-car_data$age),], 5)
```

```
## # A tibble: 5 x 11
##   brand model  year transmission mileage fuelType   tax
##   <chr> <chr> <dbl> <chr>          <dbl> <chr>    <dbl>
## 1 Merc~ M Cl~  1970 Automatic      14000 Diesel     305
## 2 Vaux~ Zafi~  1970 Manual         37357 Petrol     200
## 3 BMW   5 Se~  1996 Automatic      36000 Petrol     270
## 4 Ford  Esco~  1996 Manual         50000 Petrol     265
## 5 Audi  A8     1997 Automatic     122000 Petrol     265
## # ... with 4 more variables: mpg <dbl>, engineSize <dbl>,
## #   price <dbl>, age <dbl>
```

```
# the oldest cars seem realistic
```

Most of the extreme values in the dataset were realistic. Some entries contain questionable combinations of age and mileage, unrealistically high or low MPG values or “0” engine sizes. In those cases, some filtering should be done.

d. Visual Exploration of data properties and hypotheses

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

There are a few things we can learn from these 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 3**, this suspicion is confirmed. The vast majority of cars in the dataset is indeed less than five years old. There is 1 entry where the cars year is seemingly bigger than 2020, that is 2060, which will have to be dealt with in later steps.

The correlation plot in **Figure 4** shows some pretty good correlation between ther predictor variables and the price, so we might be able to create a solid regression using the data available.

From the view point of a human estimating the value of a used car, the most influential attributes should be age, mileage and brand/model as well as general condition, which is however not part of our dataset. Looking at the correlation matrix, we see that there is indeed a significant correlation between price and age as well as mileage. For model and brand, the correlation is much lower.

4 DATA PREPARATION REPORT

a. Potential for derived attributes

One adjustment that was already made right after importing the data was deriving the variable **age** from **year**. Our thoughts behind this decision were that as time progresses, a car’s value usually decreases. Therefore, if we only use the year, we would have to discard all existing data sooner or later because it would be outdated. The age of the car at the time of the listing is a much more stable attribute in that regard. One assumption that we had to make here is that all the listings were collected in 2020, otherwise our age assignment would be wrong. Other than that, we could not think of any variables where it would have made sense to derive new attributes.

b. Potential for additional data sources

An attribute that we were especially missing in the dataset was **horse power**. Whilst we have different engine sizes in terms of stroke volume included, in practice we often see engines with the same size produce different amounts of maximal performance. Pricing is of course also higher for stronger engines (within a car model’s options). Therefore, having additional horse power data in the set would have been desirable. Unfortunately, in many cases, it is impossible to add that information at a later point in time, since there is no reference to the original car listing included. In some cases, deriving the horse power from the other attributes might be possible. Usually, there will, however, be multiple choices, making it impossible to assign a value with 100% certainty.

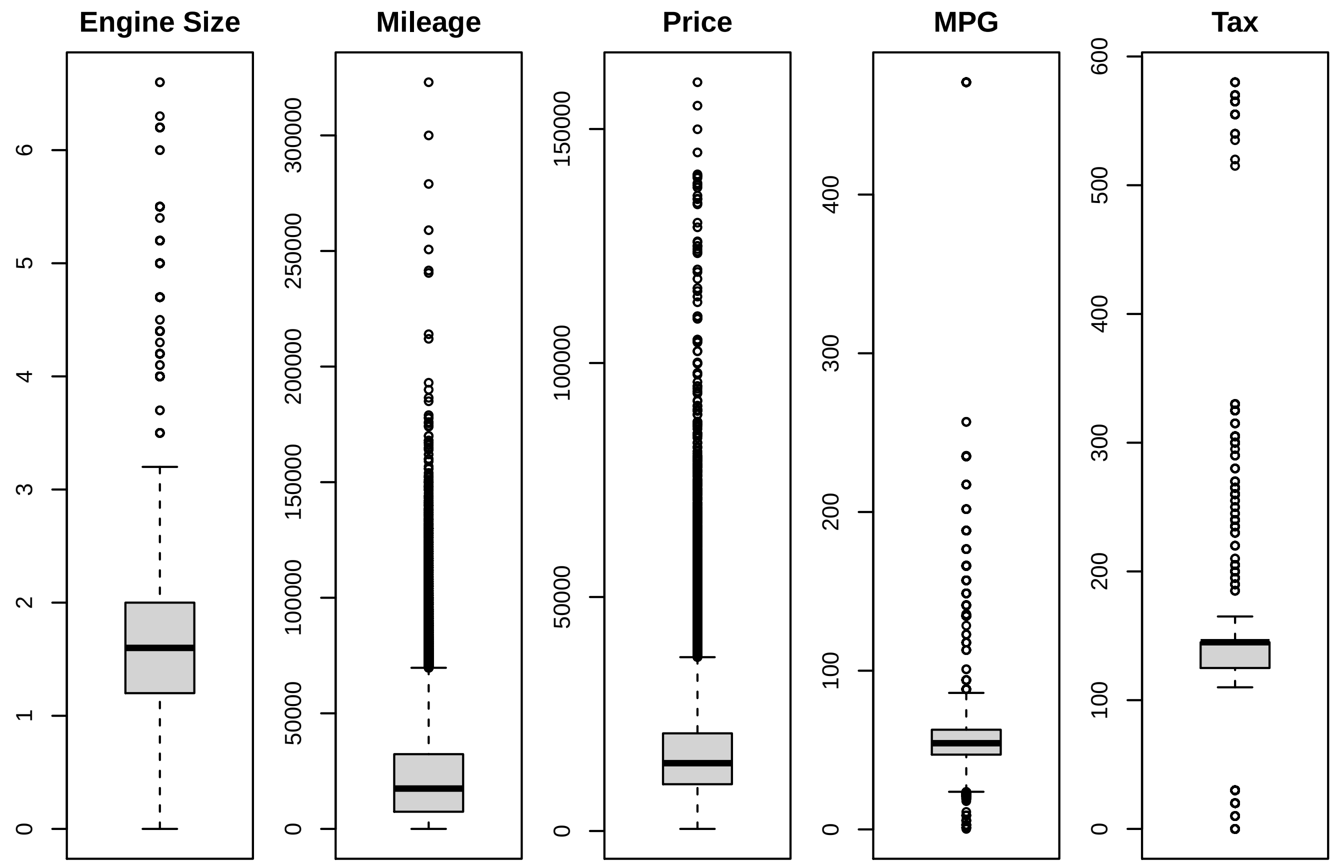

Figure 2: Boxplots on the distribution of the numeric attributes

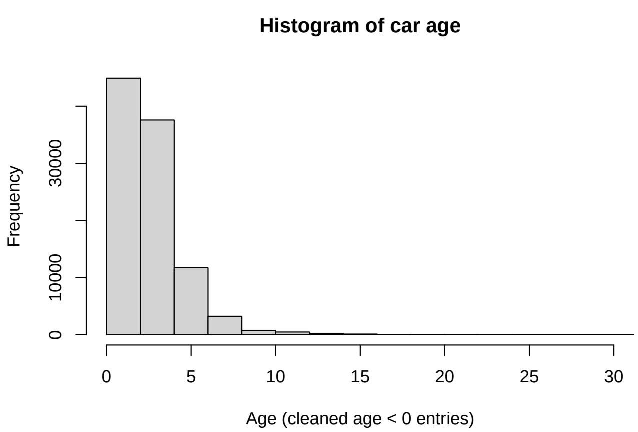

Figure 3: Histogram of car age.

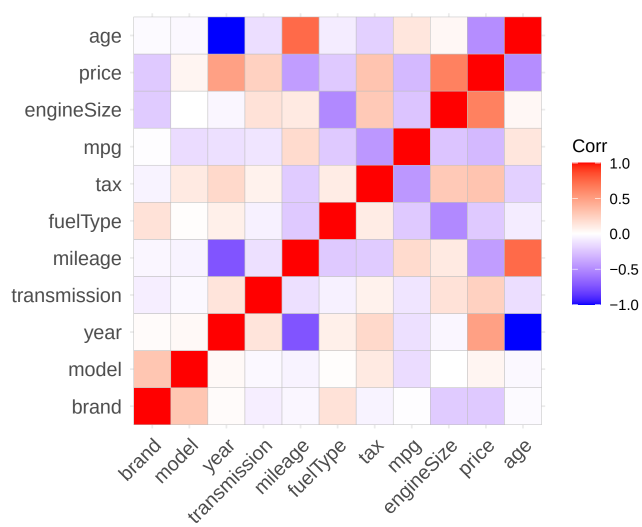

Figure 4: Pairs of all numeric attributes

Another very important aspect is the **condition** of the car. As previously mentioned, without knowing about the amount of damage that has already been done to a car, it is virtually impossible to judge its value. **Color** would be another interesting aspect since black or white cars are easier to sell than pink ones and therefore worth more. For both of these aspects, it is unfortunately again impossible to obtain values for the existing data set.

Other Pre-Processing steps

In the following section, we will prepare the dataset for modeling but applying the needed corrections identified before.

Removing Outliers. In 2c) we identified several entries that cannot be valid data. Next, we remove those entries from our data frame.

First, there is one entry with a negative age.

```
car_data <- car_data[car_data$age >= 0, ]
```

Next, we filter a few rows that seem unrealistic in terms of mileage and age.

```
car_data <- car_data[!(car_data$mileage < 1000 & car_data$age > 5), ]
```

The filtering becomes a little more interesting for miles per gallon. A quick Internet research produced the following results. In 2020, new cars with a combustion engine should be able to achieve around 25 miles per gallon on average. Top performers among hybrid cars can manage up to around 60 miles per hour. Everything significantly higher than that is currently not possible. On the lower end, we looked up some high performance sports cars. Even for the most powerful cars like the Bugatti Chiron or the Lamborghini Aventador, fuel economy scores were around 10 miles per gallon.

```
nrow(car_data[car_data$mpg > 60, ])
```

```
## [1] 34863
```

Unfortunately, the car listings in the dataset do not seem to agree with this information. Around one third of the car listings show MPG values higher than 60. Upon further inspection, we came to the conclusion that the MPG values from the listings might correspond to manufacturer ideal values that are practically unobtainable in real world use. For petrol cars, the the mean of the MPG values is about 50, which is way higher than expected.

```
summary(car_data[car_data$fuelType == "Petrol", ]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.30   43.50   51.40   50.83   58.90  166.20
```

For hybrid cars, the quartiles were higher, as expected.

```
summary(car_data[car_data$fuelType == "Hybrid", ]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.10   70.60   76.40   89.05   85.60  470.80
```

Our strategy to clean the data here therefore was to not remove overall outliers, but outliers per fuel type.

```
types <- distinct(car_data, car_data$fuelType)[, 1]$`car_da
for (type in types) {
  outliers <- boxplot(car_data[car_data$fuelType == type, ]
  if (length(outliers) > 0) {
    car_data <- car_data[-which(car_data$mpg %in% outliers
  }
}
```

This outlier removal procedure effectively filtered out 0 entries from the data frame.

c. Describe other pre-processing steps considered, specifying which ones were applied or not applied due to which reason. (e.g. data cleansing, transformations, binning, scaling, outlier removal, attribute removal, transcoding, ...) at a level of detail that ensures reproducibility of changes to the data. (Code may be supplied as supplement to the submission in case you produce your own code)

5 MODELING

Insert content here.

a. Identify suitable data mining algorithms and select one of these as the most suitable for your experiments, providing a brief justification. The goal is to estimate the price for a car based on it inputed characteristics.
b. Identify the hyper-parameters available for tuning in your chosen algorithm and select one that you deem most relevant for tuning, providing a brief justification.
c. Define and document a train / validation / test set split, considering where necessary appropriate stratification, any dependencies between data instances (e.g. time series data) and relative sizes of the respective subsets.
d. Train the model on the training set and comparing the performance on the validation set to identify the best hyper-parameter setting, explicitly documenting all parameter settings (avoid stating simply to have used "default parameters", focus on reproducibility of the results you report). 188.429 Business Intelligence (VU 4,0) – WS 2020/21 Assignment 2: Data Analytics
e. Report suitable performance metrics supported, where possible, by figures/graphs showing the tuning process of the hyper parameter.

6 EVALUATION

Insert content here.

a. Apply the final model on the test data and document performance.
b. Re-train the model with identical hyper-parameters using the full train and validation data and again apply it on the test data, documenting the performance
c. Identify and document

i. state-of-the-art performance from the literature using the same (albeit potentially slightly differently pre-processed) data set from the literature.
ii. the expected base-line performance of a trivial acceptor / rejecter or random classifier

d. Compare the performance achieved with the benchmark and baseline performances (c.f. Section 1e – Data Mining success Criteria) according to different metrics (i.e. overall, but also on per-class level (confusion matrix), micro/macro precision/recall in the case of classification tasks, regression errors in certain parts 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 stat 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