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.

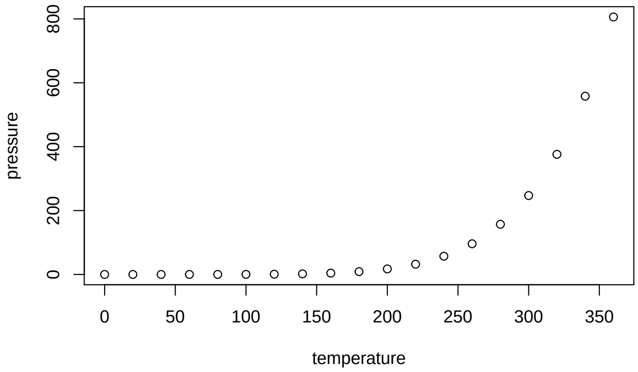

Figure 2: This is how great Tribute gets over time

2 BUSINESS UNDERSTANDING

Scenario

The

a. Define and describe a scenario in which a business analytics task based on the data set you identified should be solved
b. Define and describe the Business Objectives
c. Define and describe the Business Success Criteria
d. Define and describe the Data Mining Goals
e. Define and describe the Data Mining Success Criteria

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

Small plot example.

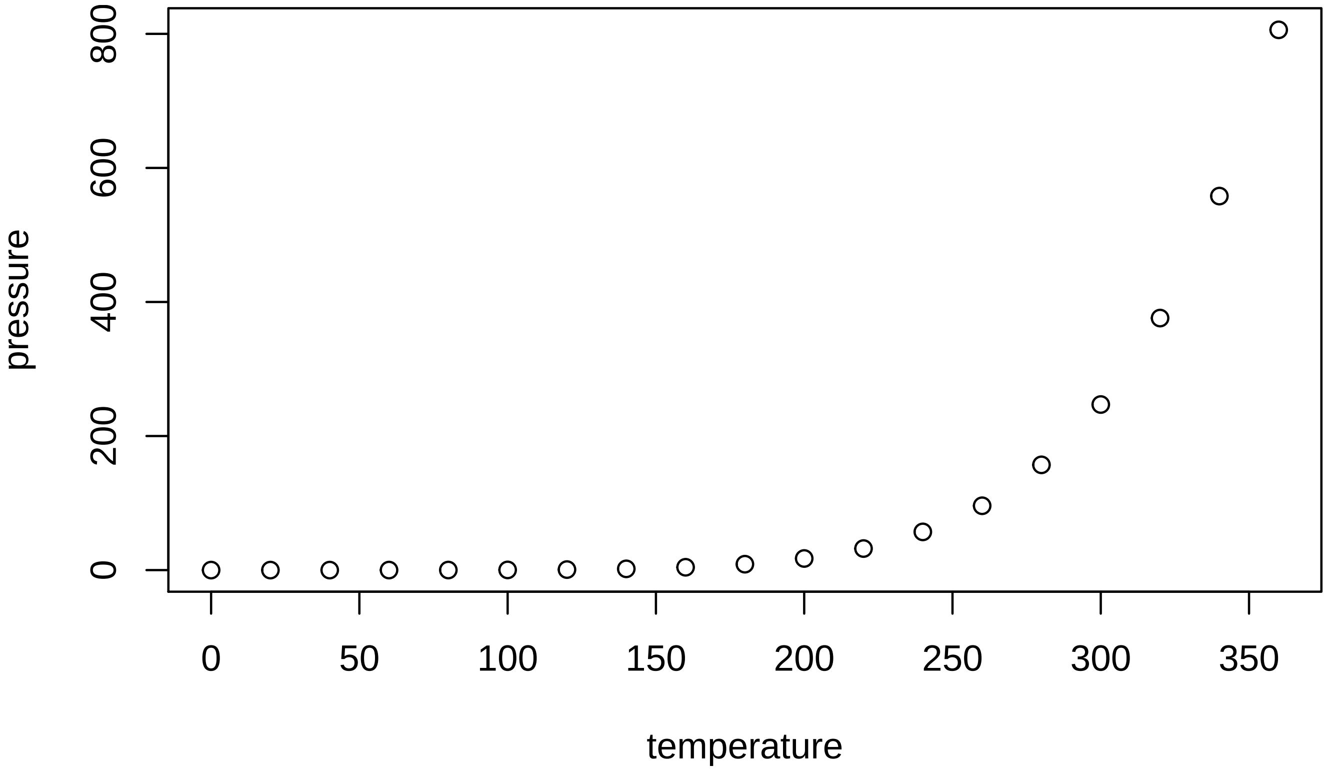

Figure 3: This is a two-column plot of how great Tribute gets over time

Table 1: The favorite iris' of Tenacious D.

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

Data Understanding

Insert content here.

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

Table example.

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)
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)

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.

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.

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.

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

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