Introduction to Data Science

Gerrit Nocker
Kilian Richard Lamprecht

December 1, 2025

Setting up

`https://github.com/lamkilian/used_car_prices`
`https://www.kaggle.com/competitions/playground-series-s4e9/overview`

Business Understanding

Identifying Business Goals

Background

When selling or buying a car, identifying a reasonable price for a given car can be challenging, especially for people that are not particularly familiar with cars.

Business Goals

The learned model might be used by potential car sellers to determine a baseline for their asking price by predicting a price based on their vehicles specifications. It may also be used by buyers to determine if a price is reasonable for a given car.

Business Success Criteria

If the model can accurately predict the prices on the test dataset, it indicates that it will produce accurate prices for additional cars that might be queried by sellers or buyers.
The primary measure of business success will be this accuracy on the test set.

Assessing your Situation

Inventory of resources

Kilian Lamprecht and Gerrit Nocker will work on this project. We will work on the datasets provided in the kaggle competition.

The main software tool used for this project will be python, employing different libraries with most data engineering done using the pandas library.

Requirements, assumptions, and constraints

The deadline for this project pdf is Monday the 8th of December 2025, though changes to the code may be done on the following three days. We plan to have the code feature complete by this deadline. Since the data we use is publicly available on the kaggle website and was generated from a deep learning model, no privacy considerations need to be made.

Risks and contingencies

Beyond time constraints, no realistic events could conceivably endanger the completion of the project.

Terminology

Regression: a machine learning process where a mathematical formula is learned based on labeled samples and is then used to predict numerical values for unlabeled samples.

Costs and benefits

The costs include the time that is invested by the two team members, the electricity cost of working on our electronic devices.
The benefits include potential use for car buyers and sellers from more informed pricing and buying choices, and practical knowledge gained by the team members. As we are students and the electricity cost should be negligible, the benefits clearly outweigh the costs.

Defining your data-mining goals

Data-mining goals

The main deliverable will be a model that accurately predicts used car prices. Other results include a cleaned up dataset after feature engineering, graphs that may aid the understanding of the data, and a poster that cleanly presents these to an interested audience.

Data-mining success criteria

The main measure of success is model accuracy on the test set. the poster should also illustrate how this accuracy compares to other attempts by other kaggle competitors.

Data Understanding

This section will summarize the findings presented in the `data_analysis.ipynb`-file. This file also shows how we were able to come up with our findings.

Gathtering Data

Since our data comes from a Kaggle competition, we can not write a lot about this. Our data was artificially created by a deep learning model. We don't have any further requirements for the data since it comes from this competition.

Describing Data

Our dataset contains 13 features. One of these features is the price that has to be predicted for the test set. The features are `id`, `brand`,`model`, `model_year`, `milage`, `fuel_type`, `engine`, `transmission`, `ext_col`, `int_col`, `accident`, `clean_title` and as mentioned `price`.
The `id` is just a counter that uniquily identifies every data point. `model_year`, `milage` and `price` are numerical values, which is why we are faced with a regression problem as mentioned earlier.
We have 188533 different data points.

Exploring the Data

Firstly we looked at how many values are missing for each feature. We found that for `fuel_type` 5083 values were missing, for `accident` 2452 values were missing and for `clean_title` 21419 values were missing. The other features had no missing values.
Since the features with missing values were all categorical we have to evaluate individually how to deal with them. More one that later.

Feature Engineering

Many of our categorical feature allow us to do some feature engineering. Firstly there is the `engine`. We can extract a lot of features out of it, like the liters it consumes and the type of engine it is. Moreover it holds information if there were some special features added to the engine like a turbo or other things.
Another challenge we faced was that some features were named differently eventhough they were the same for example `I` and `Straight` both refer to the same type of engine. In order to get this right we had to repeatitly look at the values of the engine category and filter the values accordingly.
Furthermore we found that a lot of the missing values in the `fuel_type` feature corresponded to the engine being an electrical engine. A lot of the missing values were also for cars of the brand 'Tesla' which famously only builds electrical vehicels. Therefore we were able to fill up a lot of the missing values in the

`fuel_type` feature. Only 124 of the missing values were not from vehicels that are obviously electric. These values we just filled up with the majority value from the feature.
For the accident category we just assumed that if the value was none, that just nothing was reported and it had no accident. Therefore we just used the value `None reported`.
For the `clean_title` we found that the value was either `Yes` or it was `NaN`. Due to this reason we assumed that a `NaN` value was equivelant to a "No".
To make better use of the features we used one-hot-encoding so that our model can use the data more effectivly.

Verifying Data Quality

As mentioned earlier the data was created artificaly, therefore there is no real point to be made for data quality. The way we dealt with missing data was described in the previous section.

Planning your project

- Preparing the Data
  - Kilian Lamprecht 8 hours
  - deal with missing values, inconsistent categories
  - reduce unneccessary detail
- Feature Engineering
  - Kilian Lamprecht 10 hours
  - transform given features into usable features, e.g. one-hot encode categorical features to make them applicable to regression
- Data Visualization and exploration
  - Gerrit Nocker: 12 hours
  - develop graphs do visualize the data and find patterns
- Build and evaluate Models
  - Gerrit Nocker and Kilian Lamprecht each 5 hours
  - try different regression learners, tune the hyperparameters
- visualize the learned models
  - Gerrit Nocker 5 hours
  - visualize some of the learned functions on the data, selecting specific features or principal components

- design the poster pdf
  - Gerrit Nocker and Kilian Lamprecht 7 hours
  - assemble an overview of the project for the poster
  - this should include before-and-after comparisons of data preparation and feature engineering, graphs of the data, the learned models, and comparisons to competing models on kaggle.