Capstone Proposal

Price calculation for used cars

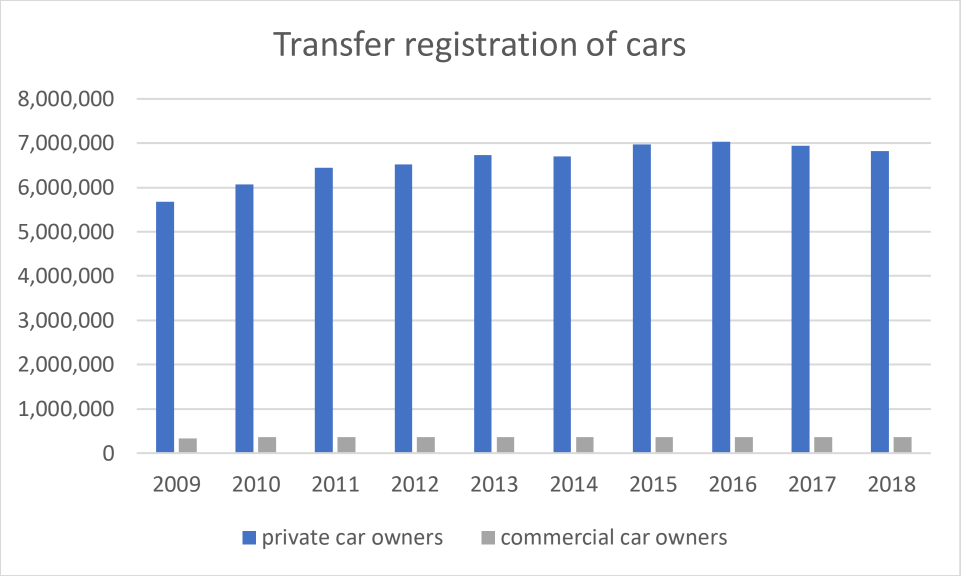
Ute Blume
January 11th, 2020

Domain Background

The used car market in Germany

The used car market offers private car owners the opportunity to sell their vehicle [1]. This can be done either through a dealer or online. Overall, the used car market is considered lucrative because good prices can be achieved there. It is accordingly popular, even if the sale through dealers is preferred to online sales.

The used car market is robustly positioned overall [2]. The transfer registration of cars covers all used car acquisitions in Germany. The share of commercial owners remains constant at 5.5 percent between 2009 and 2018.



However, vehicle experts predict times that are difficult to predict for the used car market [3], even if an existential crisis is not in sight. The reasons given for this are the use of diesel technology and the discussion about driving bans on older diesel models, which have a direct impact on the residual values of the vehicles. There are also other factors, such as the

[1] https://www.welt.de/motor/news/article201686698/Weiterhin-stabil-Gebrauchtwagenmarkt.html
[2] https://www.kba.de/DE/Statistik/Fahrzeuge/Besitzumschreibungen/Halter/z_u_halter.html
[3] https://www.schwacke.de/neuigkeiten/ausblick_2020_gebrauchtwagen/

current climate debate and the increasing use of Car-Sharing models, that can have an impact on used car sales.

Prices are also an emotional issue

The other day Inga (a good friend of mine) wanted to sell her 3-series BMW on an online platform on the internet. In her early 30s, her living conditions have changed. She has married her longtime boyfriend both have bought a beautiful house and are planning to start a family. In the process, the household budget has dropped sharply, and the sale of the beloved BMW would provide scope for new investments. And all this against the background that a family with children will not have enough space in the car!

So much for the hard facts. She was able to quickly research the current value of the car via the Internet and she published an offer. However, since the car was associated with many memories and experiences for her, it had a high sentimental value. She demanded a higher price than previously researched. She had the quiet hope that she would not have to part with the car at too high a price, or that a much higher price would help her get over the “pain of separation".

So, what do you think? Was she able to sell the car? In the end, yes! Although the price was negotiated a little bit down during the sale by the buyer, in the end the car was sold at a much higher price than the list price determined on the Internet!

Problem Statement

It would be great to have a model that can predict the price of used cars. Other thought that, too and had begun different studies to predict the price of used cars. They focused on developing countries that adopt the lease culture instead of buying a new car due to affordability[4]. They stress the importance of finding the right influencing factors.

This is a regression problem where the model expects the price as output variable. Several input variables that describe the characteristics of a car are examined to see whether they influence the price. These input variables are e.g. the brand, the kilometers, the age of the car, the gearbox or the fuel type. Another interesting aspect is to find out if there are regional differences in Germany.

Datasets and Inputs

For my analysis I will use the “Used cars database” provided by Kaggle. You can find the link here. The dataset contains 371,528 observations and 20 features, so there is a wide range to explore how they influence the price. The data was scraped in 2016 with Scrapy from Ebay-Kleinanzeigen, the leading online market in Germany.

[4] https://www.researchgate.net/publication/321180640_How_much_is_my_car_worth_A_methodology_for_predicting_used_cars_prices_using_Random_Forest

It contains a lot of different columns:

- dateCrawled : when this ad was first crawled, all field-values are taken from this date
- name : "name" of the car
- seller : private or dealer
- offerType
- price : the price on the ad to sell the car
- abtest
- vehicleType
- yearOfRegistration : at which year the car was first registered
- gearbox
- powerPS : power of the car in PS
- model
- kilometer: how many kilometers the car has driven
- monthOfRegistration : at which month the car was first registered
- fuelType
- brand
- notRepairedDamage : if the car has a damage which is not repaired yet
- dateCreated : the date for which the ad at ebay was created
- nrOfPictures : number of pictures in the ad (unfortunately this field contains everywhere a 0 and is thus useless (bug in crawler!))
- postalCode
- lastSeenOnline : when the crawler saw this ad last online

I also want to find out if there are regional differences that influences the price. The used cars dataset provides the feature postalCode with 8080 different categorical numbers. These are too much, so I want to reduce their number by transforming the postal code to a federal state. So in addition to the used car dataset I will use a look-up-table which can transform the postalCode feature to a federal state.

Solution Statement

The solution statement will be to create a model that can predict the price of a used car as it is published as an offer on the online-Market. Therefore, I will try different machine learning algorithms like Linear Regression, Random Forest Regression and Gradient Boosting Regression. I will fine tune the best algorithm by using grid search to improve the result.

Benchmark Model

As a benchmark model I will use the dummy regressor provided by scikit learn. The aim is to beat its score!

Evaluation Metrics

To compare the performance of the different algorithms, I will use R^2, Mean Squared Error (MSE) and Mean Absolute Error (MAE) as provided by scikit learn. These metrics are commonly used to compare regression results.

Project Design

The analysis will be divided into five parts:

I. Data Import
Import of the used car dataset
II. Data Wrangling
In part II I'm going to have a look at every feature to asses, if it's needed to be modified, replaced or removed from the dataset. The datast contains 20 different features so I will first need to find out, if I need them all for my analysis.
III. Exploratory data analysis (EDA)
In this part I want to build intuition on the target variable price. Can I find patterns in the data?
IV. Creating the models
After preparing the final dataset for the machine learning algorithms (one-hot-encoding the data, building a training and a testing set and normalizing the data) I will create the models for the different algorithms to choose the best one.
V. Discussion and Applicability
A discussion whether or not the model should be used in a real-world-setting and how it can be imporoved

References:

- https://www.kaggle.com/orgesleka/used-cars-database
- https://www.kba.de/DE/Statistik/Fahrzeuge/Besitzumschreibungen/Halter/z_u_halter.html
- https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwil-qXznfnmAhVCEVAKHYj2DVAQFjAAegQIAhAC&url=https%3A%2F%2Fexcel-karte.de%2Fwp-content%2Fuploads%2F2016%2F12%2FListe-der-PLZ-in-Excel-Karte-Deutschland-Postleitzahlen.xlsx&usg=AOvVaw0nh73_g6XUelEcL0Hh5qWl
- https://www.welt.de/motor/news/article201686698/Weiterhin-stabil-Gebrauchtwagenmarkt.html
- https://www.kba.de/DE/Statistik/Fahrzeuge/Besitzumschreibungen/Halter/z_u_halter.html
- https://www.schwacke.de/neuigkeiten/ausblick_2020_gebrauchtwagen/
- https://www.researchgate.net/publication/321180640_How_much_is_my_car_worth_A_methodology_for_predicting_used_cars_prices_using_Random_Forest