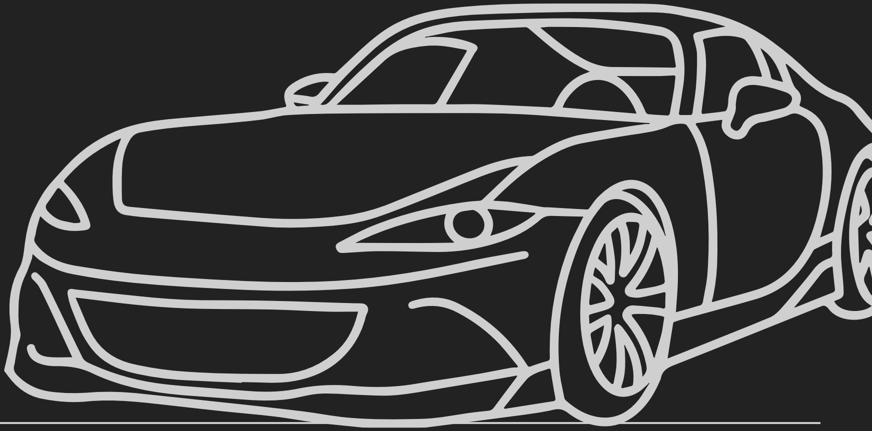
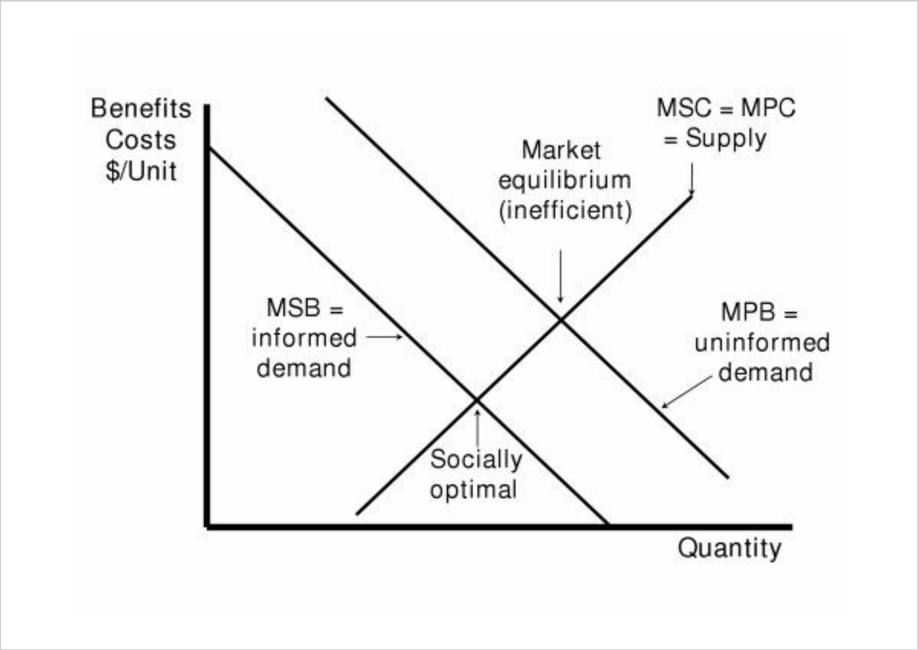
Reduction of information asymmetry in the used car market using the Random Forest method

Article review

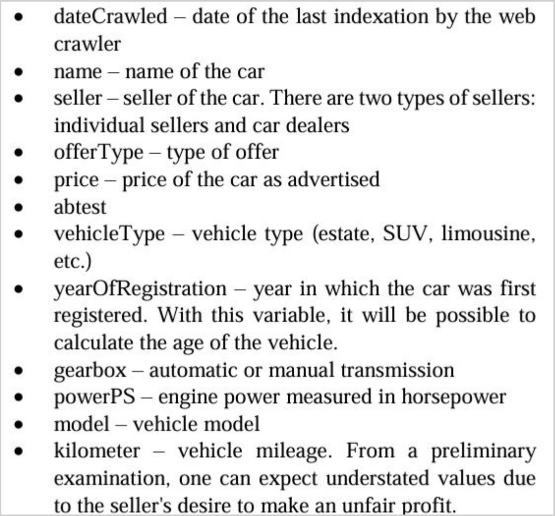


Information asymmetry

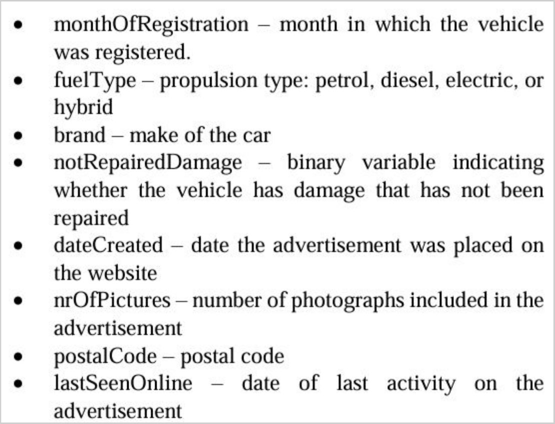
And how to handle it in the used car market



Dataset, methods and results

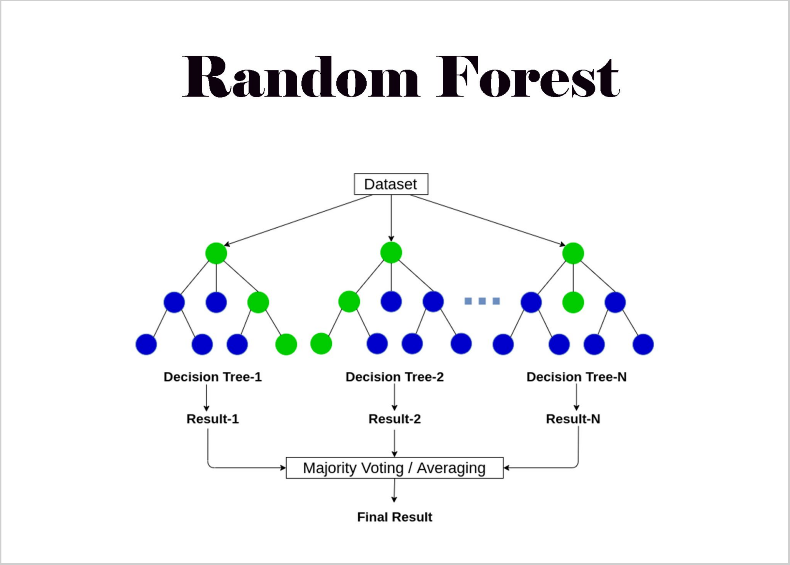


- dateCrawled – date of the last indexation by the web crawler
- name – name of the car
- seller – seller of the car. There are two types of sellers: individual sellers and car dealers
- offerType – type of offer
- price – price of the car as advertised
- abtest
- vehicleType – vehicle type (estate, SUV, limousine, etc.)
- yearOfRegistration – year in which the car was first registered. With this variable, it will be possible to calculate the age of the vehicle.
- gearbox – automatic or manual transmission
- powerPS – engine power measured in horsepower
- model – vehicle model
- kilometer – vehicle mileage. From a preliminary examination, one can expect understated values due to the seller's desire to make an unfair profit.



- monthOfRegistration – month in which the vehicle was registered.
- fuelType – propulsion type: petrol, diesel, electric, or hybrid
- brand – make of the car
- notRepairedDamage – binary variable indicating whether the vehicle has damage that has not been repaired
- dateCreated – date the advertisement was placed on the website
- nrOfPictures – number of photographs included in the advertisement
- postalCode – postal code
- lastSeenOnline – date of last activity on the advertisement

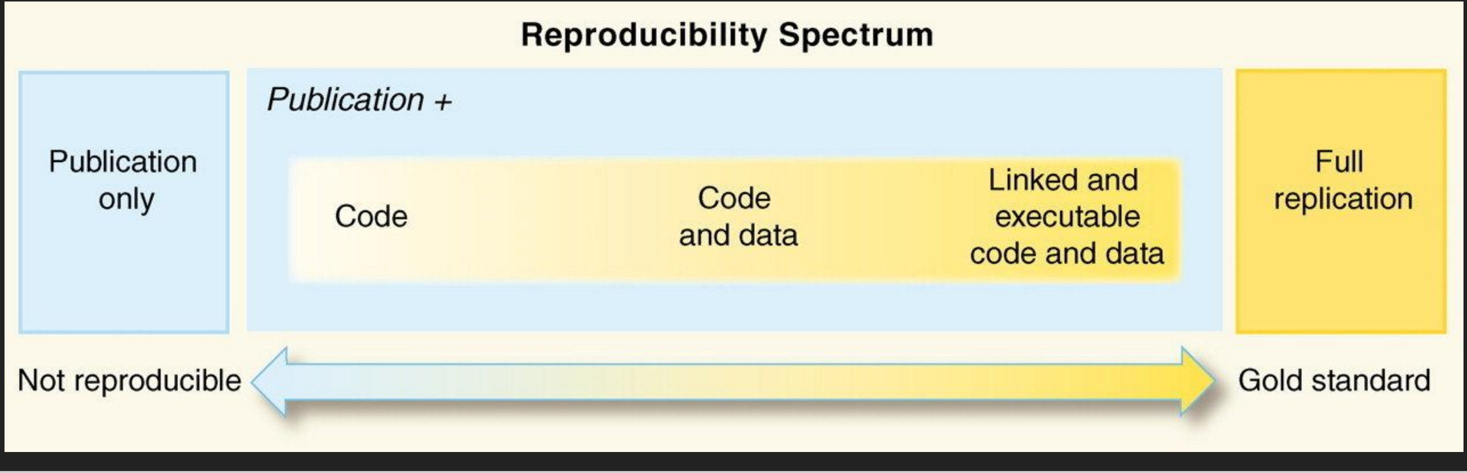
371,528 private advertisements listed on the German version of the website ebay.com (kleinanzeigen.de)



Final test set R^2 = 0.78

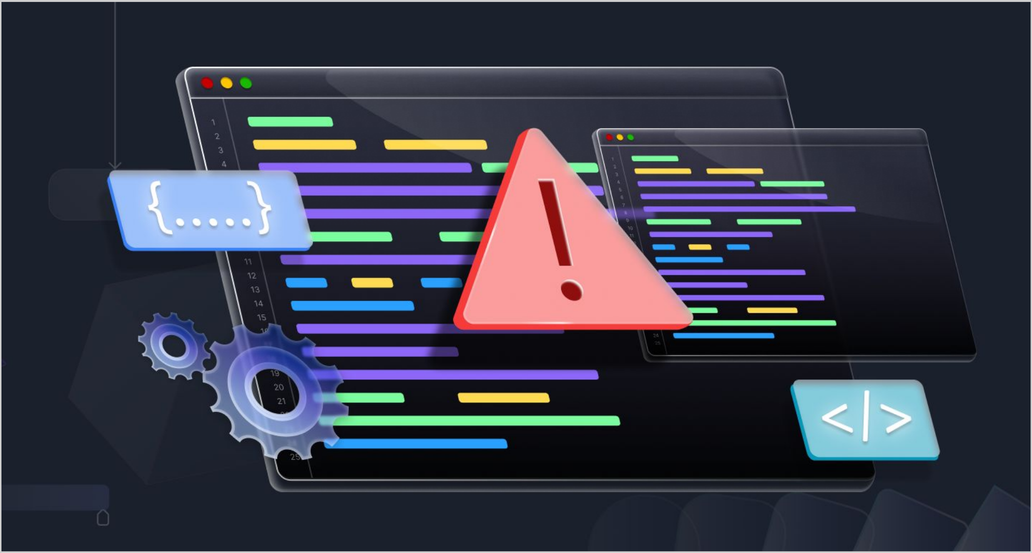
Why is it not reproducible?

- Most of the code is missing
- No shared codebase, software requirements
- Expired link to the dataset, no data gathering methodology description
- No description of preprocessing and feature engineering
- No description of sampling for the modelling phase (full/partial sample? random_seed?)



Other problems

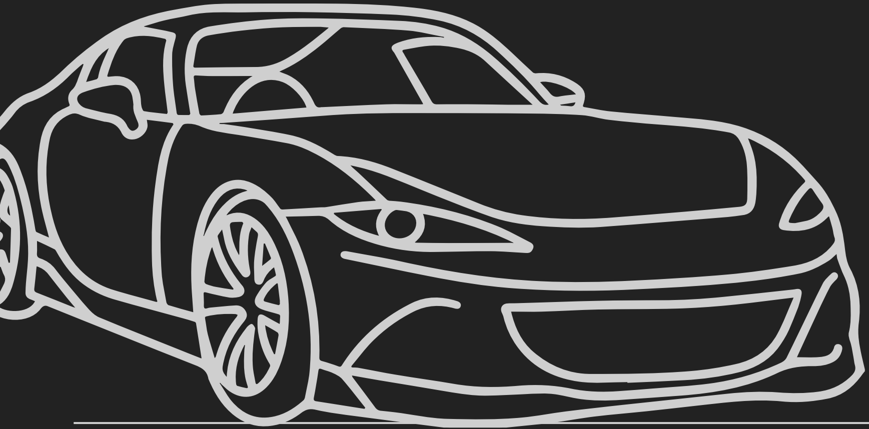
- Overall logic
- Omitted facts
- No literature review
- Model chosen without any reasoning
- No baseline model
- Random hyperparameter tuning
- Overfitting
- Uninformative performance metrics
- Other methodological issues



Thanks!

Any questions?

Piotr Bugajski
Michał Woźniak



Credits: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution