University of Warsaw
Faculty of Economic Sciences

Dawid Szyszko-Celiński
Album N°: 443709

Car prices prediction – a data-centric approach in machine learning

Magister (master) degree thesis
Field of the study: Data Science and Business Analytics

The thesis written under the supervision of
Assoc. Prof. Katarzyna Kopczewska
from Department of Statistics and Econometrics
WNE UW

Warsaw, June 2023

Oświadczenia kierującego pracą

Oświadczam, że niniejsza praca została przygotowana pod moim kierunkiem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

* Oświadczam, że mój udział w artykule naukowym, który stanowi część pracy dyplomowej wynosi ..0....%, zaś suplement do pracy został napisany samodzielnie przez dyplomanta(ów).

***** *skreślić jeśli nie dotyczy*

Data
10.06.2023

Podpis kierującego pracą

*Oświadczenie autora pracy***

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

* Oświadczam, że mój*** udział w artykule naukowym, który stanowi część pracy dyplomowej wynosi .100..% (nie mniej niż 60%), zaś suplement do pracy został napisany przeze mnie***.

***** *skreślić jeśli nie dotyczy*

Data
11.06.2023

Podpis autora pracy
Dawid Szyszko-Celiński

*** każdy ze współautorów studentów składa oświadczenie osobno*
**** w przypadku współautora studenta należy oświadczyć wkład merytoryczny i procentowy*

Summary

The thesis covers the comparison of the standard and data-centric approaches in Machine Learning based on the problem of car price prediction. The data was gathered via the web scrapping method and resulted in 52,361 unique observations. Then the preprocessing and data-centric methods were applied to clean the data and obtain additional features. The dataset was split by level of complexity and by size. Last but not least the machine learning models were applied to predict car prices and the comparison between datasets was performed. Ultimately, the conclusions were reached: it is worth applying data-centric approach and extending a number of observations in machine learning projects and it is sometimes worth ensembling models. The best model to predict car prices was ensembled gradient boosting with neural networks.

Key words

machine learning, data-centric approach, car price prediction, gradient boosting
(uczenie maszynowe, podejście dano-centryczne, predykcja cen samochodów, wzmocnienie gradientowe)

Field of the thesis (codes according to the Erasmus program)

Economics (14300)

Thematic classification

Machine Learning

The title of the thesis in Polish

Predykcja cen samochodów – metody uczenia maszynowego oparte na strategii dano-centrycznej

TABLE OF CONTENTS

1 Introduction 5
1.1 Characteristics of the Polish vehicle market 6
1.2 Data-centric approach 7
1.3 Research Theses 7
2 Literature Review 8
2.1 Data-centric approach in Machine Learning 8
2.2 Car prices prediction techniques 12
2.3 Machine Learning Models Overview 14
3 Research part 16
3.1 Data gathering process 16
3.2 Data preprocessing and Explanatory Data Analysis 18
3.2.1 Numeric variables distribution 20
3.2.2 Categorical variables preprocessing 24
3.3 Data-centric measures used in the research 26
3.4 Modelling 26
3.4.1 Data-centric approach compared to standard datasets 27
3.4.2 Number of observations 30
3.4.3 Improving both the quality and quantity of the data 32
3.4.4 Ensembling models 33
3.5 Conclusions 34
3.6 Future improvements 34
Summary 35
References 36
Appendices 38
Appendix 1 38

Scientific Article

Car prices prediction – a data-centric approach in machine learning

Dawid Szyszko-Celiński

Abstract

This paper aims to investigate supervised machine learning approaches for predicting car prices using scraped data from the Polish offer portal. Various machine learning algorithms will be employed to make predictions, including linear regression, random forest, gradient boosting, and neural networks. The primary objective is to compare the "standard" and "data-centric" approaches in machine learning and to assess whether the additional effort invested in exploring and extracting data significantly improves the prediction quality. The research will also compare approaches and models on different dataset sizes to determine the best method and model for predicting car prices. The research is constructed to analyse both the Polish vehicle market and data-centric approaches, moreover, it aims to promote the idea of such an approach and encourage machine learning developers to incorporate it in their projects. The second aim is to find the models and techniques most suitable for predicting car prices.

Keywords: Machine learning, data-centric approach, neural networks, random forest, linear regression, model comparison, car price prediction, text analysis

1 Introduction

Predicting car prices is a well-known regression problem in machine learning. Although many datasets are available online [3,4,5], it is challenging to find real data from the Polish vehicle market that includes crucial features like offer description, that could be used to present data-centric approaches by extracting extra variables. Considering this fact authors opted to scrape data from the Polish offer portal in several batches, resulting in 52,361 unique observations available for further analysis. Car price prediction on the vehicle market is critical from a business perspective, as it is essential for buyers, sellers, and offer portals to understand current market conditions and predict fair prices based on car parameters that maximize the probability of transactions. With the analytical tool, all parties could make better decisions and operate on the prices that would be acceptable for others. From the perspective of the offering portal, it might be used to suggest or compare car prices and help users to make more informed decisions and be treated as a competitive advantage.

Currently, machine learning researchers from all over the world seek the tools and approaches that would increase the quality of the predictions. For many years much more effort was put to develop and update the models with little to no interest in the data part of the projects. Data gathering and preprocessing was an important time-consuming part but without tendencies to dig deeper and extract more knowledge about the researched topic. Lately, the tendency to experiment and implement data-centric approaches is raising, however, there is still little research put into that matter. That is why our research paper considering the Polish vehicle market will include this approach to verify if the data-centric approach can be used on real problems. Both Polish vehicle market analysis and data-centric approach are of big importance as they have real-life application possibilities that can improve not only the analysis and knowledge of the Polish vehicle market but most importantly can change the approach of researchers to put more emphasis on the way the Machine Learning projects are developed. We strongly believe that our research based on the vehicle market analysis will bring added value to the machine learning community by showing the advantages of such an approach. The concept is fairly new but we believe it is crucial in terms of future machine learning development and hopefully, this paper will promote and encourage to explore the depths of the data-centric approach.

In the next parts of the chapter, there will be more detailed characteristics of the Polish vehicle market, a data-centric approach and at the end there will be presented research theses that would be of most importance in this paper as they will be used to verify and analyse the data-centric approach usability on the example of the Polish vehicle market.

In the second chapter of the paper, the literature review will be performed putting an additional emphasis on three extremely important topics: currently existing knowledge and research about data-centric approaches in machine learning, car prices prediction techniques and approaches based on previous studies and various markets from all over the world and the last part would focus on models that are well established for car prices prediction problems.

In the last chapter of the paper, the research part will be performed. First, the authors of the paper will describe the design and process of data-gathering, and then an emphasis will be put on data preprocessing and Explanatory Data Analysis, which is one of the most important steps of the data-centric approach. Later the data-centric approach steps and the modelling part will be outlined. The next very important step will consist of a comparison between the standard approach and the data-centric approach which will be performed by verification of the research theses. Last but not least the conclusions will be drawn as well as the future possibilities of research improvements.

1.1 Characteristics of the Polish vehicle market

The Polish vehicle market has been heavily influenced by joining European Union structures, which opened the country to foreign markets, allowing easy import of used cars from Western European countries, primarily Germany. Based on statistics from 2004 to 2018, approximately 12 million used vehicles were imported to Poland. Each year, starting from 2004, the ratio of used registered cars to all registered cars ranged from 70% to 80%, with the share of used cars slowly decreasing due to an increase in income and expanding lease sector. In 2018, over 1.4 million passenger cars were registered, of which 500 thousand were new cars, resulting in a share of used cars registered to all registered cars of approximately 65%. [1] As such, the used car market still has a significant share, and advertising service portals will continue to connect buyers and sellers. It is noteworthy that new cars are also often advertised and the future of advertising portals may change over time.

1.2 Data-centric approach

The increasing amount of data produced globally has prompted a growing interest in the field of big data. According to IDC projections, the amount of data will grow exponentially to 175 zettabytes by 2025. [13] Therefore, the data itself is of paramount importance in the current big data era. While efficient models have already been developed, the rapid growth of data requires a principled approach to managing the analysis process. Without structured workflows that handle the data, it will be hard to make more accurate predictions. In the past, the lack of data was a significant challenge. Nowadays, however, there is a sufficient amount of data, and the focus has shifted to ensuring that it is of high quality.

Before delving into an in-depth analysis of the subject, it is worth providing characteristics of the data-centric approach in a few brief statements. One of the biggest differences is that the data-centric approach places more emphasis on the data itself, not only on the technicalities of the models. This approach is advocated by Andrew Ng, one of the most influential machine learning experts, who proposed a competition aimed at promoting it. The competition aimed to improve the quality of predictions by applying techniques such as fixing incorrect labels, adding examples that represent edge cases, applying augmentation, adding new features, or combining some of them. [2] Although choosing a suitable model is crucial to construct high-quality predictions, it is essential to note that managing data quality, extending the training sample, and extracting additional features may also significantly improve the prediction power. Unfortunately, many machine learning engineers and researchers focus mostly on choosing the best model and sometimes neglect the importance of in-depth data exploration. The choice of a suitable model is truly crucial to construct high-quality predictions but it is worth noting that more attention should be paid to the data, which is the fuel of each model engine.

Therefore, this article will be dedicated to conducting a comparison of different models that are driven by various combinations of data. The research is performed to compare the quality of prediction based on different sizes and sets of variables of training samples for each of the models, focusing mostly on the data and not exclusively on the models. The article is also aimed at comparison of the predictive power based on a "simple" approach and a "data-centric" approach. Lastly, the article compares both factors and chooses the best approach and model. The modelling process is based on the scraped data and is built to predict car prices.

1.3 Research Theses

Thesis 1: The quality of machine learning data is at least as important as the choice of the model. After the model is sufficiently good, it is better to focus on gathering additional data or extracting additional knowledge from the data that the model uses. This study assumes that a decrease in Root Mean Squared Error for the test sample at least at a 5% level using a data-centric approach compared to the standard approach is significant enough to state that the approach is meaningful and gives good results.

Thesis 2: Extending the training sample with twice as many examples increases the prediction performance by at least 2.5%. This study tests different datasets on 5 thousand, 10 thousand, 20 thousand and around 41 thousand observations to see if the differences between algorithms are significant within each dataset and between the same algorithm used on different datasets. Each dataset is divided into training and testing samples (75% training and 25% testing).

Thesis 3: Ensembling a few best-performing models will give better results than a single best-model prediction with at least 5% better RMSE.

Aim: The article aims to choose the best model to predict car prices based on the data from the Polish vehicle market.

2 Literature Review

2.1 Data-centric approach in Machine Learning

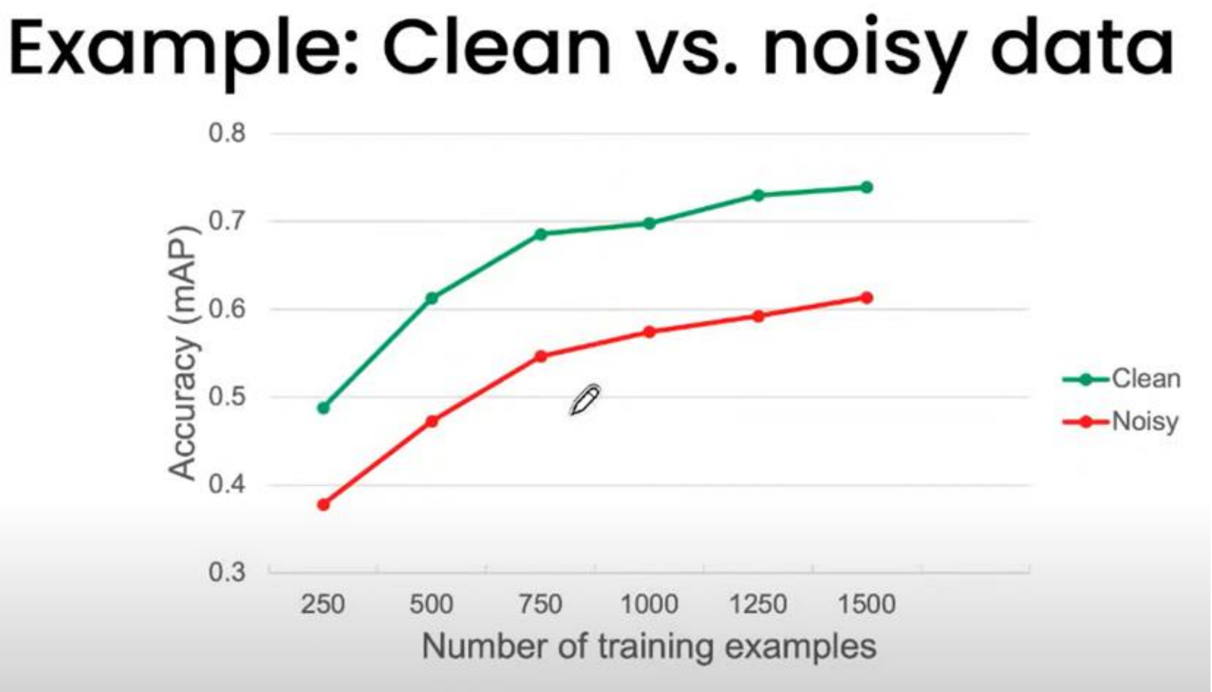
A significant body of current literature focuses on the "data-centric" approach in machine learning. This section provides a concise overview of some established perspectives regarding this topic.

Whang et al. in their article, "Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective" [13], define data-centric AI as “the centre of a fundamental shift in software engineering where machine learning becomes the new software, powered by big data and computing infrastructure”. The authors mainly focus on the applications of deep learning and identify three subtopics: data discovery, data augmentation, and data generation. Data discovery involves finding existing datasets in data lakes or on the web and indexing them. Data augmentation is a popular method for generating realistic data, labelling it, and expanding the available training sample. Improving existing data involves cleaning the data or improving its labelling. The authors note that it is a well-established fact that approximately 90% of efforts in machine learning projects are dedicated to algorithms, with only 10% focused on data preparation. However, in reality, the data preparation step takes up to 90% of the time spent on the projects. The authors also highlight that top data-driven companies such as Google and Microsoft are shifting their focus toward data-centric approaches, wherein the primary goal is to develop tools that improve data quality and models for predictive analytics. One of the most important and challenging issues raised in the article is data fairness, ensuring that there is no bias in the data. The second issue that has recently emerged is data poisoning, wherein malicious intent negatively affects model training. As a result, the model may fail or make predictions skewed toward a certain output. It is crucial to identify such phenomena and take appropriate measures to clean data of bias and poisoning, particularly since the real use cases of data are often generated by humans who may have prejudices based on factors such as race, gender, or age.

The article "A Survey on Bias and Fairness in Machine Learning" [22] provides good examples of bias issues. The problem is especially important when machine learning models are used for high-stakes decision-making such as hiring new employees, granting loans, or law enforcement. The authors outlined that the bias problem may be direct or indirect. Direct discrimination is related to sensitive attributes such as race or gender. Such bias is relatively straightforward to address, as it is often against the law to use such variables in certain decision-making models, and they are easy to identify. The second type of discrimination is indirect and cannot be easily tracked. The authors provide an example of a zip code variable, which at first glance is a non-sensitive attribute. However, it may correlate with the racial, political preference or age distribution of the population in certain residential areas.

Andrew Ng, in the video "A Chat with Andrew on MLOps: From Model-centric to Data-centric AI" [20], describes artificial intelligence systems as a combination of code (models and algorithms) and data. The author notes that over the past few decades, machine learning experts have focused on improving only the first part of AI systems - the models. However, he encourages a shift in mindset to improve the second part - the data. In the video, Ng presents a practical use case of a data-centric approach in defect detection on steel sheets. The steel company requested a detection accuracy of 90%, comparable to human abilities. The baseline model used to solve the problem achieved an accuracy of 76%. To improve the accuracy of the models, machine learning teams took two approaches - model-centric and data-centric. The model-centric approach showed no improvement in accuracy, while the data-centric approach increased the accuracy up to 93%, surpassing the requested threshold. Ng also mentions solar panel and surface inspection problems whose results were improved by introducing a data-centric approach. The author also draws attention to the data quality based on the example of labelling consistency in speech recognition and image recognition problems and remarks that the data should be consistent. One of the most important components of the data-centric approach is to clean the noise and collect new data or make data augmentations that will increase the training sample. The second point is especially important for small or heavily unbalanced datasets.

In the video, Ng presents the visualisation of the difference between the dataset with a data-centric approach (clean) and without (noisy). One of the main research theses of our paper focuses on the comparison of model prediction of car prices concerning both the number of training examples and the quality of features (which may be compared to a “Clean” - dataset with extended features and “Noisy” - data with standard features). The expected outcome for each of the tested models should follow the trends that are depicted in Figure 1, but in the case of RMSE, the values should decrease with the increase in the number of observations and the “clean” dataset should have lower values than the “noisy” dataset. The outcomes of the plot will be replicated in the research part of this paper.



Example: Clean vs. noisy data

Figure 1 – Comparison of accuracy measure regarding size and the quality of the data
Source: https://www.youtube.com/watch?v=06-AZXmwHjo

The essence of data-centric approaches in the field of artificial intelligence is to systematically improve the quality of data, as it is crucial for the performance of AI models. In this regard, several articles have emphasized the importance of continuously updating and monitoring the data used for AI applications, as well as the need to automate the data collection and update process.

One of the articles that highlight the significance of operating on high-quality data is “What can Data-Centric AI Learn from data and ML Engineering?” by Polyzotis and Zaharia [21]. The authors argue that in real business cases, the situation is constantly changing, and relying on data that is only representing reality at a given point in time may be insufficient. The authors present the relevance of such an approach based on the example of identifying defects in an assembly line, where changes in machines, products, materials, software updates or sensors can significantly affect the prediction of models. Thus, the authors recommend continuous updates and automation of the data collection and updates process, which may include error messages, automatic labelling, and data monitoring process. The authors also opt for a data monitoring process which can be performed by including database-like schemas of the data. It may contain the structure of the data, and ranges for each feature, providing insightful messages when the data point is out of range.

Similarly, in the article “The principles of Data-Centric AI (DCAI)” [25], the authors describe data-centric AI as a systematic and iterative approach that can deal with data issues. They emphasize the ability to capture dynamic changes in data and make AI systems evolve. The effectiveness of the system is determined both by the performance of the model and the quality of data. Each iteration of the process should be started with error analysis to examine data fit. This would show how a well-trained model represents real-world use cases and check if data is consistent and fully devoted to the solved problems. After the analysis, the necessary steps should be taken to improve the model for instance performing data augmentation. The next step is to assess the modified system via data benchmarking. This part is meant to show the difference in data quality between two consecutive iterations. The authors of the article propose six principles of data-centric AI, including systematic improvement of data fit and consistency, mutual improvement of model and data through iteration, human-centeredness of ‘data work,’ AI as a sociotechnical system, and continuous and substantive interactions between AI and domain experts. The summary of each principle is as follows:

1. Systematic improvement of data fit - this principle puts the main focus on the adequate coverage and balance of the data. In other words, the data that is used for modelling should include important variables and the samples from each class ought to be evenly distributed and have no biases. That can be reached by additional cleaning of the existing data, extending data with new observations or making data augmentation.
2. Systematic improvement of data consistency - this principle is based on the accuracy of annotations and how well they reflect the real-world problem. Usually, the labelling process is manual and time-consuming. The ideas to achieve consistency of the data without mentioned limits include the usage of automated tools such as weakly supervised labelling that gives pseudo labels. The authors also mention that this process even though automated should include a human supervisor and the documentation of the errors should be created to update the taxonomy.
3. Mutual improvement of model and data through iteration - emphasizes the continuous monitoring and iteration of the data-centric model that prevents data drift, which may occur when the input data has changed over time and the model predictions may be no longer robust. In such cases, the new data, that follow different circumstances, should be taken into consideration and applied within new iterations ensuring the model will be better at replicating the current reality.
4. Human-centeredness of ‘data work’ - this principle outlines an important matter of remembering that the data created by humans cannot be analysed independently but in a way that the outliers and biases in the data are taken into account. The authors mentioned that creating AI systems is not purely a mathematical and technical matter and advised working with domain experts during the implementation and monitoring of the AI projects.
5. AI as a sociotechnical system - the principle reminds us that many AI tools are commonly used as a decision system in social and important matters such as criminal justice, healthcare or Human Relations. Taking that into consideration the AI developers should put efforts to ensure that models are not only fully optimised but also embrace human needs. The authors mention that during the design phase of the AI system, there should be considered strategies which implement ethics. Such an approach includes for instance the right to privacy or compliance with the law regarding people`s right to data. There is also mentioned a need for auditing the outcomes and implementing explainable AI so that people could understand why such decisions are made.
6. Continuous and substantive interactions between AI and domain experts - the principle focuses on the importance of domain knowledge in the design, implementation and maintenance of the system. In the article, the authors provide an example of medical

> data, that is regulated by law in terms of privacy. To keep up with changing regulations domain experts must be present. What is more in many complicated cases only people with certain knowledge would be able to understand and interpret the results of an AI system. The authors also highlight that data annotators should be treated as collaborators not only as “service providers”. Such an approach could improve data quality and in each iteration potentially resolve some data issues, resulting in better performance of models.

In conclusion, the data-centric approach emphasizes the systematic improvement of data quality, which is crucial for the performance of AI models. The articles reviewed provide various recommendations and principles for achieving high-quality data, including continuous updates and monitoring, automation of data collection and updates of the process, data cleaning and augmentation techniques, auditing and continuous verification of data, and human-centeredness and sociotechnicalality of AI systems. Researchers commonly mention that there is not enough attention paid to data and most of the time developers improve models rather than data.

2.2 Car prices prediction techniques

The prediction of car prices has received considerable attention in the research community, leading to numerous studies utilizing different approaches and methodologies. These studies have applied a range of machine learning algorithms, including Linear Regression (LR) [6,7,9,10], K-nearest neighbours (KNN) [6,7 8,10], Decision Trees (DT) [6,7,9], Naïve Bayes [6], Radom Forest (RF) [7,11], Gradient Boosting [7,10], Support Vector Machines (SVM) [11], and Neural Networks (NN) [11].

One such study was conducted by Sameerchand Pudaruth in the article “Predicting the price of used cars using machine learning techniques” [6], which utilized data from daily newspaper advertisements in Mauritius in a one-month timespan to predict the prices of used cars. The study included attributes such as brand, model, cubic capacity, mileage in kilometres, production year, exterior colour, transmission type, and price. Then the sparse columns and records with missing crucial information were removed. Initially, more than 400 observations were collected but researchers decided to keep only the three most popular car makes in the analysed country. After preprocessing, algorithms such as Multiple Linear Regression (LR), K-nearest neighbours (KNN), Decision Trees (DT), and Naive Bayes (NB) were applied to predict the prices of cars. The LR and KNN algorithm's performances were measured by mean error and for Nissan cars, the result was Rs51,000 while for KNN it was about Rs27,000. For DT and NB algorithms the data were recoded into categories and analysed as a classification problem. Both algorithms gave around 60-70% accuracy measure. The authors pointed out a main limitation of the study which was an insufficient number of observations to perform more accurate predictions.

Another study by Gajera, Gondaliya, and Kavathiya [7] utilized a dataset containing 92,386 records to predict car prices. The available attributes included kilometres travelled, year of registration, fuel type, car model, fiscal power, car brand, and gear type. Five algorithms, KNN, Random Forest (RF), LR, DT, and Extreme Gradient Boosting (XGB), were implemented to predict prices after preprocessing the data. First, the data was explored, some low-frequency categories in certain variables were dropped and numerical variables were constrained to reasonable ranges. The results of the study pointed out that the lowest Root Mean Squared Error (RMSE) and highest R-squared were obtained for algorithms in such a hierarchy: RF, XGB, DT, LR, and KNN. The lowest RMSE was around 3700 (Euros) and the highest was around 7770 (Euros). The highest R-squared was more than 93% and the lowest was below 70%. The authors mentioned further improvements and limitations of the study including gathering more data and predictor variables that could positively affect the model accuracy such as the number of doors, colour, etc.

Samruddhi and Kumar in their paper “Used Car Price Prediction using K-Nearest Neighbor Based Model” [8] focused on using the KNN algorithm to predict used car prices. The data used for the analysis was obtained from Kaggle and contained 14 variables: Serial number, Name, Location, Mileage, Fuel Type, Engine transmission, Kilometres Driven, Power, New Price, Year, Seats, Owner Type, and Price. Preprocessing steps were performed to prepare the data for KNN implementation (Euclidean distance metric and k-values ranging from 2 to 10), with cross-validation techniques used to minimize overfitting effects. The KNN algorithm achieved a prediction accuracy of 85%, whereas the baseline LR model had an accuracy of 71%. The authors suggested exploring more advanced machine learning techniques in future work.

Venkatasubbu and Ganesh in the article “Used Cars Price Prediction using Supervised Learning Techniques” [9] utilized a dataset containing 804 records with attributes such as price, mileage, make, model, type of car, car`s body type, number of cylinders, fuel capacity, number of doors, and binary variables indicating whether the car had cruise control, upgraded speakers, and leather interiors. The authors proposed Lasso Regression, Multiple Regression, and Regression Tree algorithms for the modelling part. Multiple Regression gave the best results, while Lasso Regression had the worst, with mean percentage errors for all algorithms around 3.5%. The authors recommended increasing the number of observations and applying more advanced techniques such as Random Forest.

The research conducted by Nasiboglu and Akdogan [10] utilized data gathered via web scraping techniques. The authors were able to obtain data for car price prediction, with 100,000 observations and 12 attributes. LR, KNN, RF, GB, and XGB algorithms were used to model the data, with hyperparameter tuning performed for each model. The best model for each car brand was chosen based on RMSE and Mean Absolute Error (MAE) measures. The modelling part was performed for each car brand by applying mentioned algorithms. Then for each brand, the best model was chosen based on the goodness of fit measures. Gradient Boosting was chosen as the best in 19 different vehicle brands, XGB in 11, RF in 7, Ridge in 5, Lasso in 2, and Elastic Net in 1. Researchers concluded that more advanced techniques gave better results and proposed using many models for non-homogeneous data structures like car brands.

The study, conducted by Gegic et al. and published in an article entitled "Car price prediction using machine learning techniques," examined the performance of three algorithms: Artificial Neural Networks, Support Vector Machines, and Random Forest. The authors scraped data from a Bosnia and Herzegovina advertising website during the winter season, yielding a dataset comprising standard features, that were mentioned in previous studies, as well as numerous binary variables reflecting the presence of certain attributes such as navigation, alarm systems, and sensors. The data was then cleaned by removing sparse attributes and recoding categorical variables, resulting in a sample of 797 observations that could be used for classification modelling. The accuracy results for the outcomes of a single model were below 50% for all previously mentioned algorithms, therefore authors decided to ensemble multiple machine learning algorithms. The combined algorithms' accuracy was above 90%, which is a significant improvement.

In summary, the mentioned studies that focused on cars prices prediction were based on similar machine learning models like Linear Regression, K-nearest neighbours, Decision Trees, Naïve Bayes, Radom Forest, Gradient Boosting, Support Vector Machines, and Neural Networks. Studies that were conducted with simpler algorithms in the final remarks mentioned that the next steps for improvement would be an application of more advanced ones or combining the outcome of a few different algorithms. In one of the research, the authors proposed exploring and extracting information from already existing data or gathering some new attributes, which is a sign of a data-centric approach. In terms of datasets researches varied quite significantly. In some of them, data was gathered via web scrapping methods that gave the flexibility of the attributes gathered and the possibility to extend many observations. On the other hand, some research papers were based on smaller datasets, even gathered manually from newspapers, and depend on already existing resources such as the Kaggle website. Authors of papers that were using small datasets almost always mentioned that for further improvements it would be worth gathering more data, as such small datasets were insufficient. The authors of some studies also proposed applying individual models for each brand of car and put attention to changes in prices and the car market over time. Studies were mostly focused on certain countries' car markets and on exploring the differences between the prediction quality of machine learning models.

Some articles mentioned the issue of seasonality, which is why in the research part of this paper additional variable considering this issue will be included and investigated. All articles mostly focus on the used cars market, however in the data gathered in this research, new cars will be also included. What is more, the authors of this study could not find a paper that would analyse the Polish vehicle market with machine learning techniques.

2.3 Machine Learning Models Overview

Even though the main idea of the paper is to present a data-centric approach, the right choice of the model must be included. In this part, there will be performed the models' overview which will be taken into consideration during the research part.

Linear regression is a “statistical method that allows us to summarize and study the relationship between two continuous variables”. One variable is treated as the predictor and the second as a response variable. The algorithm aims to find “the best fitting line” which is referenced as a regression function. To find such a line one of the most popular methods is using the least squares criterion method. [15] The algorithm can be extended to “multiple linear regression” by adding more predictor variables and calculating such coefficients for all variables that will determine the best line of fit. In the case of car price prediction, there will be more than one variable that will explain the price of the car, so there should be used multiple regression. In the paper “Multiple Linear Regression Equation for Estimation of Daily Averages Solar Radiation in Chonburi, Thailand” [13] there is an example of how to use regression methods where multiple explanatory variables are required. Other ML methods such as Lasso, Ridge or ElasticNet can address the problem of overfitting, which may be the problem especially while using complex datasets. Based on the models that were used for car price modelling in the previously mentioned papers, linear models were rather used as a benchmark for better classes of the models.

The Random Forest algorithm is a well-known machine learning method, introduced by Breiman [16], that is a part of the supervised learning models category. The main concept is based on ensembling multiple decision tree models (base learners) and aggregating them to predict the outcome with better precision. The algorithm creates lower-level models with a random subsample of the training sample dataset and explainable variables. There is also a parameter that specifies the number of trees in the forecast, which may significantly increase the computational cost. That is why it is recommended to set the optimal trade-off between the computational complexity and accuracy of the prediction, but the number of trees should be high enough to ensure the stability of the model (usually over 500). One of the advantages of random forest is the ability to perform well on high-dimensional data, which is the case of the car prices prediction in the research part. [17]

A gradient boosting machine (gradient boosting of regression trees) is another algorithm that will be used in the further part of the paper. The idea of the algorithm is to estimate the predicted value step by step, but each step depends on the predictions that were made in the previous step. When the error of predictions in previous steps (calculated by RMSE or MSE) is small, the updates to the model will be negligible.[18] Considering this we may assume that after some optimal number of steps, the error term of the model should drop significantly.

Neural networks can be used for supervised learning problems such as car price prediction. The idea behind Artificial Neural Networks (ANN) is to recreate the processes of neurons in human brains in an artificial environment. This kind of algorithm uses connected perceptrons (artificial neurons) whose weights are updated during the learning (training) process. The update is based on output error values which are backpropagated. In each epoch, the model can predict with better precision taking into account what it learned from previous errors. One of the biggest advantages of ANN is the ability to capture non-linear and complicated patterns with high precision, especially when the dataset is complex and has many explanatory variables. However, this advantage comes with a cost of high computational demand and training such models may be expensive or time-consuming. [19]

3 Research part

The principal objective of this part of the paper is to present and explicate a data-centric approach, exemplified by the prediction of car prices. Additionally, three research questions will be addressed, pertaining to the quality and quantity of the data, as well as model ensembling. Finally, the optimal approach and model will be selected to predict car prices.

3.1 Data gathering process

Drawing on the insights gained from an extensive review of the literature, we opted to conduct data collection via web scraping, given its purported flexibility in acquiring additional attributes and expanding the number of observations. Some existing databases do not encompass data on the Polish car market and are confined to certain features. Moreover, web scraping aligns with data-centric machine learning approaches, as it facilitates the repetitive extension of the training set and in consequence, the model's predictions.

Data was procured from Lento.pl [23], a free advertising service that facilitates the sale and purchase of various commodities such as real estate, electronics, cars, and many others. Data used for research was amassed by scraping the posted offers in multiple batches, with the scraping algorithm executed approximately every month. A Python script was written to obtain data from the website, utilizing the Beautiful Soup package [24] functions to extract the most crucial traits of each available car offer. The process was designed to be data-centric, thereby conferring the possibility of extending the data sample. The code also included validation rules to prevent errors. Since the data gathering process was executed via web scraping, ethical guidelines were followed to ensure that the authors did not negatively impact the website's performance. To this end, there was a pause of 1 to 5 seconds between every request, thereby slowing down the process but making it less disruptive for the website. The process of data collection is depicted in the scheme below.

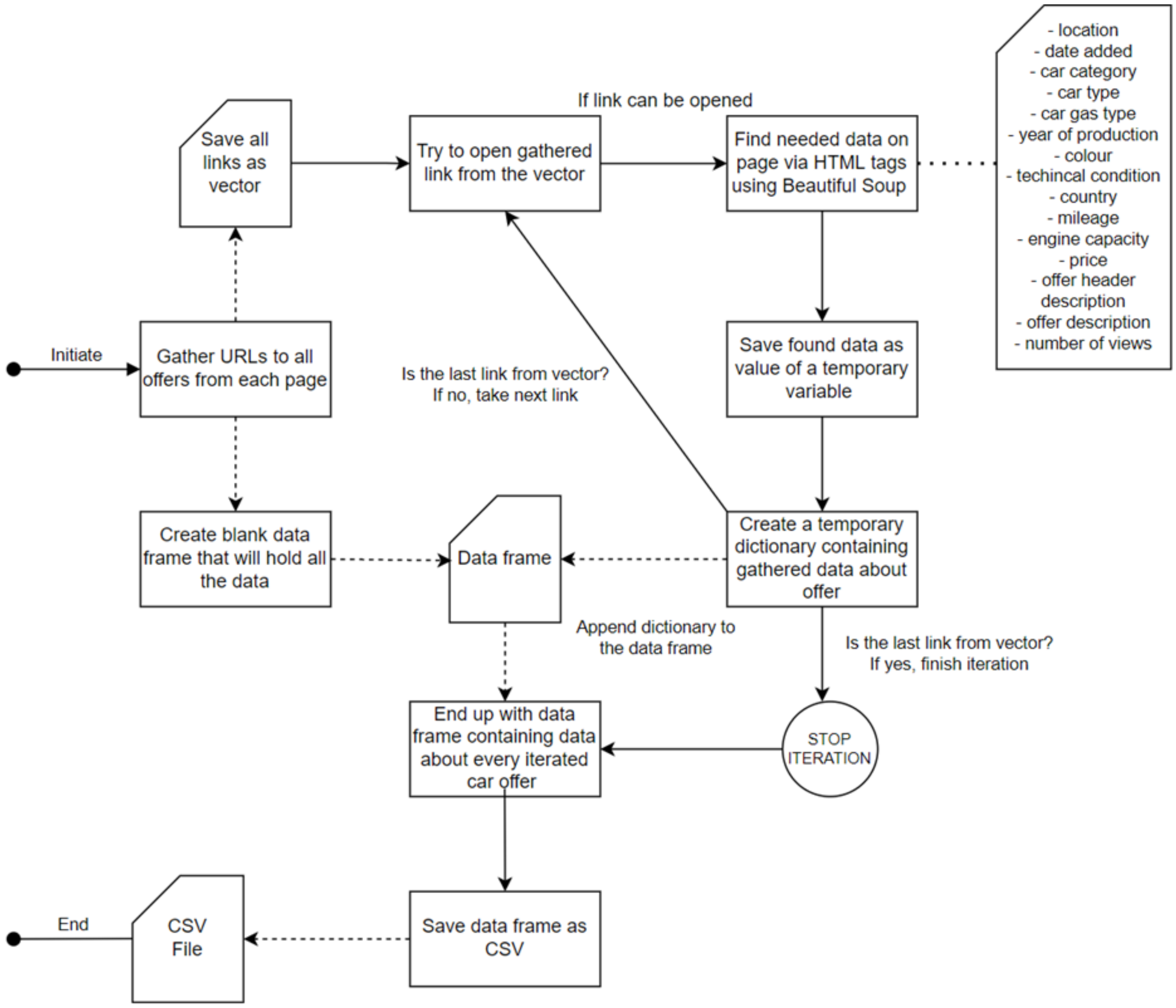

Initiate
Gather URLs to all offers from each page
Save all links as vector
Try to open gathered link from the vector
If link can be opened
Find needed data on page via HTML tags using Beautiful Soup
- location
- date added
- car category
- car type
- car gas type
- year of production
- colour
- techincal condition
- country
- mileage
- engine capacity
- price
- offer header description
- offer description
- number of views
Save found data as value of a temporary variable
Is the last link from vector? If no, take next link
Create blank data frame that will hold all the data
Data frame
Create a temporary dictionary containing gathered data about offer
Append dictionary to the data frame
Is the last link from vector? If yes, finish iteration
STOP ITERATION
End up with data frame containing data about every iterated car offer
Save data frame as CSV
CSV File
End

Figure 2 – Data gathering script process
Source: own elaboration

Initially, the script collated links to each offer posted between the day of gathering data and the last day of the previous data-gathering process. The number of web pages to select was manually adjusted each time, but this could be automated using a scheduler. As the author did not have access to a cloud environment, the data-gathering part of the research was conducted entirely on a local device, and some tasks were performed manually. The links were saved to a vector, and a blank data frame with the appropriate structure was created to hold the actual data. In the next step, the script attempted to open every link from the list. If links were expired or could not be opened, the script, in accordance with the validation rules, skipped such offers. If the connection was valid and the page opened, the script gathered the data and placed it as a new row in the temporary dictionary. Each offer on the website has a structured placeholder for essential information, making it relatively easy to acquire the necessary values. The data collected from each offer included the location of the offer, the date of offer posting, the car category (comprising the car brand and specific model), the car type (SUV, sedan, etc.), the car gas type (LPG, diesel, etc.), the year of production, the colour, the technical condition (undamaged, damaged), the gears type (automatic, manual), the country of origin, the mileage, the engine capacity, the price, the offer header, the offer's full description, and the number of views. The values of variables were extracted and saved to a temporary dictionary that was later appended to the dedicated data frame. The iteration continued until the last link was opened. The filled data frame was then saved as a CSV file and stored locally on the device.

3.2 Data preprocessing and Explanatory Data Analysis

The data preprocessing and exploratory data analysis (EDA) stage is critical in data-centric approaches. These processes ensure data quality by eliminating redundancy, removing outliers, recoding variables, obtaining new variables, and deleting unnecessary variables, among other tasks. A detailed description of the steps taken to ensure a data-centric approach will be presented.

The datasets collected from web scraping were first loaded and merged into one large dataset consisting of 52,361 unique observations and 21 columns. The offers data required extensive transformations that can be made regardless of the variables' values, to improve its quality. These actions included recoding one variable into multiple variables (car category divided into brand and model, date divided into a day, month, year, hour, minute and day of the week), removing extra spaces and labels (every car price contained “zł” at the end which is the polish currency shortcut, car capacity contained “cm3” - the engine capacity unit and mileage “km” label), dropping records that were not coded in ASCII, deleting unnecessary variables, combining year and month variables to create an inflation factor variable (Supposedly the variable would catch the changes of prices based just on the change in time), and setting the proper variable types.

While the option to clean the offer descriptions and gather data based on keywords was available at this stage, the task is time-consuming and computationally intensive. Therefore, it will be performed after the removal of outliers, saving time as many records with invalid values will not be computed during this process.

The next section involved EDA to detect and remove outliers based on the data distribution. Some of the categorical variable values were also recoded to reduce low-frequency categories and prevent the dataset from the "dimensionality curse" issue. Each variable category was changed to a binary variable using the one-hot-encoding approach, and without reducing the number of unique values in categorical variables it could result in thousands of sparse dimensions. Including many sparse variables can lead to consequences such as overfitting, inability to detect patterns and relationships in the data, multicollinearity, etc. [26] Therefore, reducing low-frequency categories is essential. Techniques such as variable selection and Principal Component Analysis (PCA) can be used to reduce dimensionality.

The EDA was performed on a full dataset consisting of 52,359 rows. Further analyses and modelling which will be based on the lower number of training samples (due to the division into smaller samples) have to follow the assumption of a priori knowledge about the distribution of variables. EDA on a significant number of training samples gives a better insight into the data, making it easier to identify patterns and significant behaviours. Therefore, the assumption is that the constraints, boundary conditions for outliers, and categories division are already known. In real-life use cases, the EDA process will be continuously monitored and updated.

Based on the simple statistics of numerical variables, visible in Table 1, some are definitely influenced by outliers (marked in red). Moreover, the Pearson correlation between variables, visible in Table 2, seems to be extremely low, even though the variables were expected to be correlated. Therefore, in the following steps, the examination of each variable will be made, and outliers will be removed.

	car production year	**car mileage**	**car engine capacity**	**price**	**inflation factor**
Count	52359.00	52359.00	52359.00	52359.00	52359.00
Mean	2012.22	83987.46	1966.63	67557.52	7.05
Std	7.23	**13977359.75**	3857.25	**600081.46**	3.77
Min	1926.00	**(2147483648.00)**	**1.00**	**0.00**	1.00
Q1	2008.00	69561.50	1400.00	18500.00	4.00
Q2	2013.00	141458.00	1685.00	40500.00	7.00
Q3	2018.00	202563.50	1998.00	79900.00	10.00
Max	2023.00	**1000000000.00**	**277188.00**	**123456666.00**	14.00

Table 1 – Data descriptive statistics before cleaning
Source: own elaboration

	car production	**car mileage**	**car engine capacity**	**price**	**inflation factor**
car production	1	**0.00873**	-0.020725	**0.068685**	0.002389
car mileage	**0.00873**	1	-0.010329	**0.000004**	0.002944
car engine capacity	-0.020725	-0.010329	1	**0.011159**	-0.033264
price	**0.068685**	**0.000004**	**0.011159**	1	-0.012509
inflation factor	0.002389	0.002944	-0.033264	-0.012509	1

Table 2 – Pearson correlation between numeric variables before cleaning
Source: own elaboration

3.2.1 Numeric variables distribution

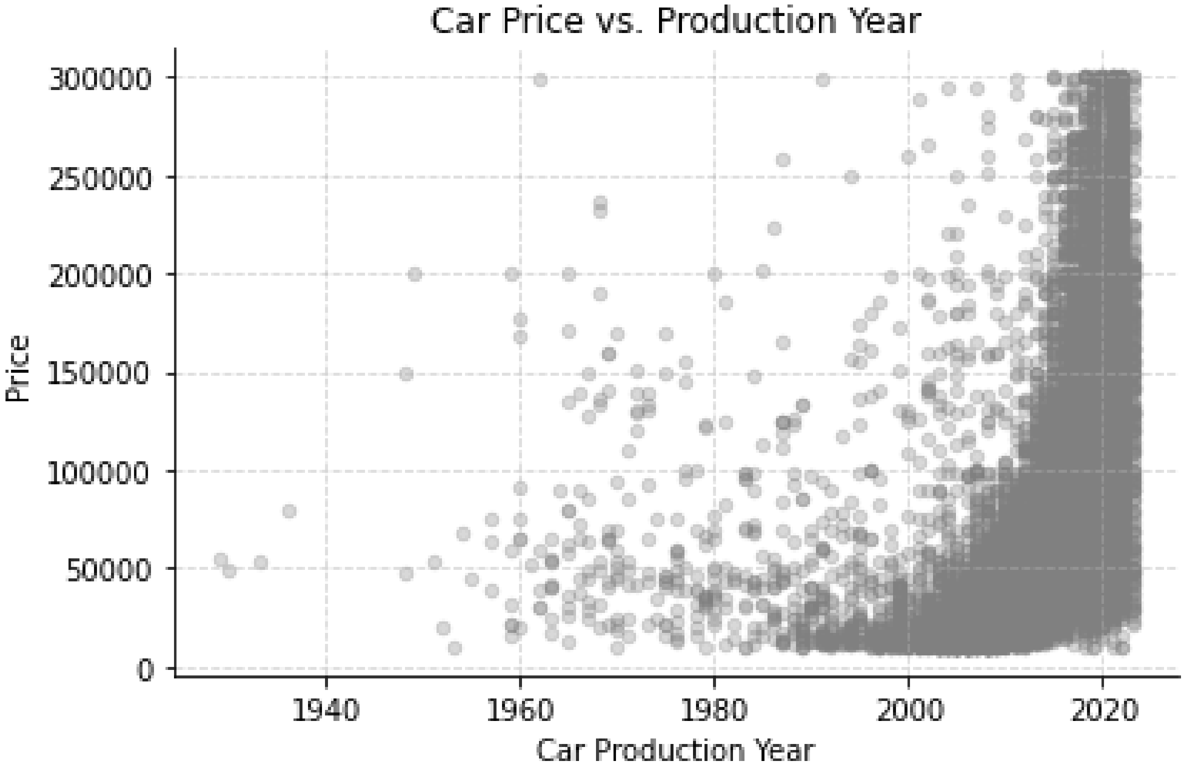


Figure 3 – Car price variable distribution after cleaning
Source: own elaboration

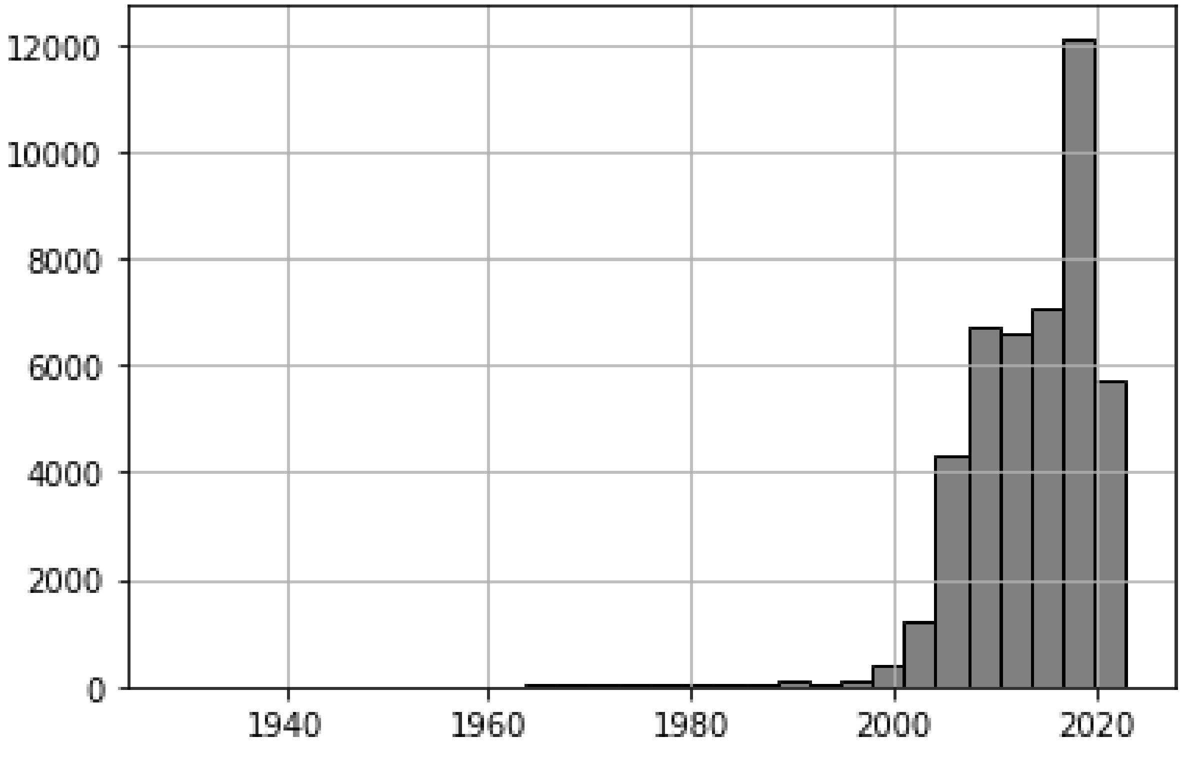


Figure 4 – Year of production variable distribution before cleaning
Source: own elaboration

The car price distribution was highly influenced by significant outliers. In order to perform a customized model, the authors of this article set price constraints to ensure that prices remain between 10,000 PLN and 300,000 PLN. The model aims to assist in pricing "standard" cars, hence the constraint. The upper-level values (between 200,000 and 300,000) were taken into account as many more expensive cars are typically leased by companies in the Polish market. The price distribution compared to the year of production showed that older cars are cheaper, while newer models are more expensive as it is assumed considering typical cars. The constraints were applied to cater to the most common cases, assuming the aim is not to look for extremes in terms of price, year of production, mileage, and engine capacity. This led to the removal of many outliers and errors in the data, and the model was customized for the prices of the cars available to most people.

The year of production variable was the second variable examined. Based on the frequency plot, offers below the year 2000 were in the minority. Consequently, all offers where the production year was below the year 2000 were excluded. As the data was gathered in 2022 and 2023, cars produced before 2000 shall not be taken into consideration as they would be too old for sale or considered as “classical” cars that could bias the model because of their age and price reversed pattern.

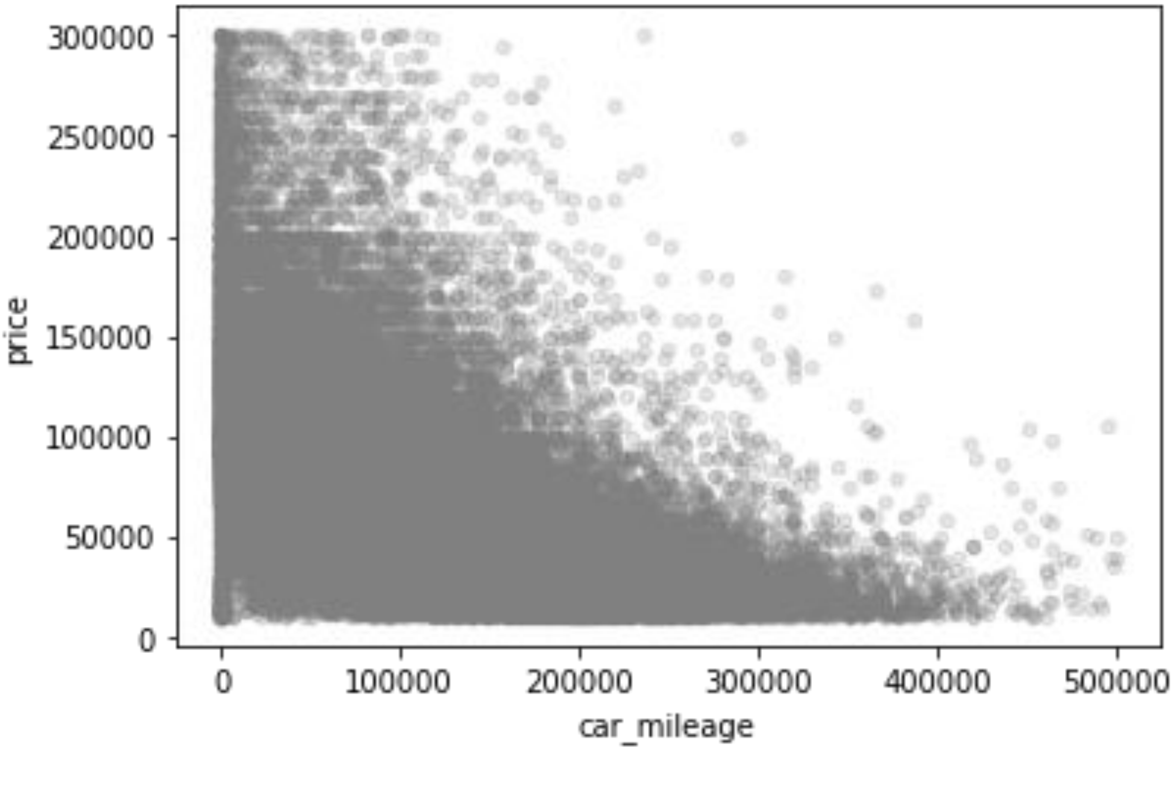


Figure 5 – Mileage variable distribution after cleaning
Source: own elaboration

The next examined variable was mileage, which was assumed to be under 500,000 kilometres. We believe that values above this threshold would present invalid entries or extremely used cars, that are not in the scope of the interest. The plot above shows the relationship between mileage and the prices of the car and it is in line with our expectations - the higher the mileage the lower the price.

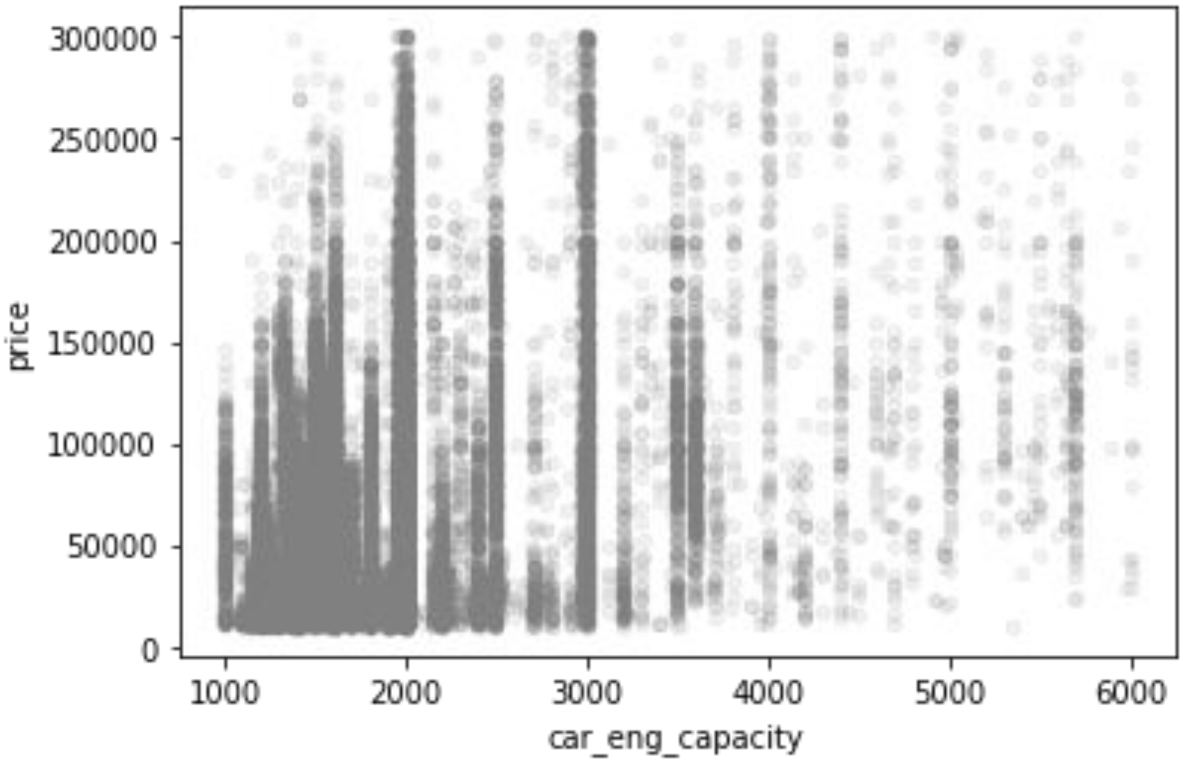


Figure 6 – Engine capacity distribution after cleaning
Source: own elaboration

The last basic numerical variable examined was engine capacity. The sensible ranges of engine capacities were assumed to be between 1,000 and 6,000. We believe that car offers that are above or below this range are invalid or present the wrong type of cars.

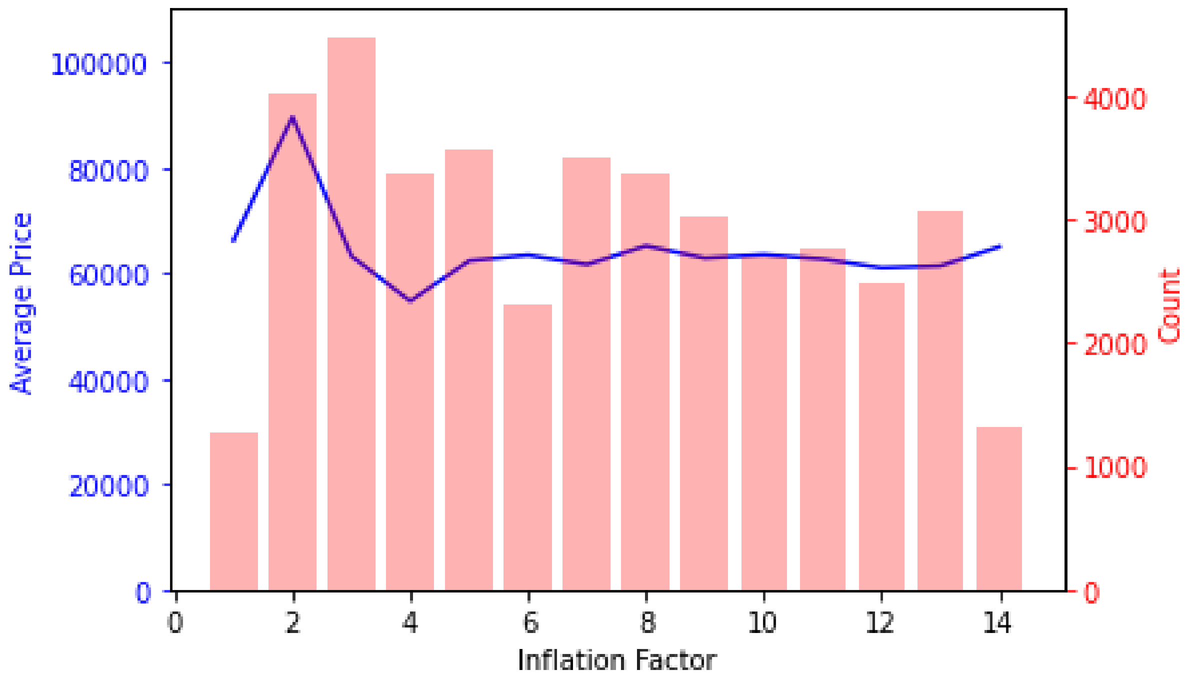


Figure 7 – Inflation factor analysis in time
Source: own elaboration

One of the newly extracted numerical variables "inflation factor" was also examined to verify if average prices have changed over time. The plot below shows average prices over time (from February 2022 (1 – first period) up to March 2023 (14 – last period)) and frequencies of offers in those periods. Starting from April 2022 (3), the average prices in the vehicle market appeared to remain stable with no observable effect of inflation, thereby suggesting that the inflation variable may be disregarded. It is worth noting that during the analysed period, Poland experienced a significant inflation rate ranging from 8.5% up to 18.5% year to year [28]. Nevertheless, other worldwide events such as the post-COVID era and the war in Ukraine starting in February 2022 may have influenced car prices in diverse ways. The inflation effect and other market factors might have worked in opposite directions, ultimately resulting in stable average prices. Notably, the graph highlights an outlier in March 2022, wherein the average prices were around 90,000 PLN. This observation could be attributed to the aforementioned war or other combined market conditions.

Following the application of numerical variable constraints and outlier removal techniques, the complete dataset was diminished to 41,274 observations. The fundamental statistics of the reduced dataset are presented as follows:

	car production year	**car mileage**	**car engine capacity**	**price**	**inflation factor**
Count	41274	41274	41274	41274	41274
Mean	2013.88	136257.72	1884.38	64843.16	7.08
Std	5.12	84317.83	661.57	52940.90	3.77
Min	2000.00	1.00	1000.00	10000.00	1.00
Q1	2010.00	70000.00	1499.00	25700.00	4.00
Q2	2015.00	137000.00	1796.00	48850.00	7.00
Q3	2018.00	194289.50	1999.00	85900.00	10.00
Max	2023.00	499000.00	6000.00	300000.00	14.00

Table 3 – Data descriptive statistics after cleaning
Source: own elaboration

	car production	**car mileage**	**car engine capacity**	**price**	**inflation factor**
car production	1	**-0.70**	0.00	**0.70**	-0.03
car mileage	**-0.70**	1	0.00	**-0.59**	0.10
car engine capacity	0.00	0.00	1	**0.35**	-0.15
price	**0.70**	**-0.59**	**0.35**	1	-0.06
inflation factor	-0.03	0.10	-0.15	-0.06	1

Table 4 – Pearson correlation between numeric variables before cleaning
Source: own elaboration

Following the data-cleaning process, the correlation coefficients have been observed to be in line with the expected outcome, indicating an improvement in the accuracy and reliability of the data. Upon closer examination, it has been noted that the inflation factor exhibits a relatively stable trend over time and holds little significance in terms of its correlation with the price. As such, it has been deemed unnecessary to include this variable in the subsequent analyses. This outcome not only highlights the robustness of the dataset but also reinforces the notion that inflation has had a limited impact on the prices of the variables under study.

3.2.2 Categorical variables preprocessing

The following stage of the data analysis involves inspecting the categorical variables. As stated earlier, certain infrequent categories may be combined to reduce the dimensionality.

For the car brand variable, the brands with a frequency lower than 1% of the entire dataset were recoded to the value 'Other.' Furthermore, the value 'Samochody,' which translates to 'Cars', was not considered specific enough and was also recoded to 'Other.' Additional variables were extracted from the variable to adopt a data-centric approach. The first variable, 'car_brand_country,' was intended to divide the car brands by the origin of the brand. The second variable, 'car_prestige,' was selected arbitrarily, as certain brands tend to be regarded as 'premium,' whereas others do not exhibit such tendencies in the general public opinion.

The 'location' variable comprises 1,150 unique location names where the offers were posted. Since most of the location frequencies are low, with values below 5, only those that were frequent at a minimum level of 1% of the entire dataset were included. Most of the remaining groups consist of locations based in major Polish cities or are located within a sensible range near them.

The car model was also recoded to reduce the number of categories, as there were initially 516 unique models. The 1% frequency threshold was applied to this variable as well, reducing the number of categories to 19.

For other categorical variables, expert knowledge was applied to reduce the dimensionality by merging values into grouped categories if necessary.

The subsequent stage involved adopting a data-centric approach to extract additional data from the textual description of the offers. Since the data is textual, necessary steps, such as removing stop words and special characters, performing lemmatization, and extracting keywords, needed to be taken. Since the text data was in Polish, some parts required the use of additional sources or extra libraries that could handle the language, such as the spacy package [29] and a list of custom stopwords [27]. The extraction process involved changing the text into tokens, performing lemmatization on each token, and then checking if the token was within the keywords dictionary. If it was, the intersection of the document (offer) and keyword was labelled as '1'. In the later stages, the data frame that contains the data would be appended to the actual data and used as additional features in the final dataset. The list of keywords with their translations and explanation can be found in the Appendix 1.

The last step involved performing one-hot encoding of categorical variables and appending the extracted keywords from the text as additional features. Finally, the data was split into subsamples consisting of 5 thousand, 10 thousand, and 20 thousand observations, and the full dataset (around 41 thousand observations). Each of the datasets was then split into training and testing sets, with a 75% training and 25% testing split.

Summing up the EDA process involved removing outliers and creating sensible ranges of variable values, deleting unnecessary variables, creating new variables, removing Polish stopwords and special characters from descriptions such as dots and commas, performing lemmatization, extracting key binary variables from the description text by checking if the word was present or not and one-hot encoding the categorical features. The steps presented in this part of the research are crucial in terms of a data-centric approach as they solely focus on data consistency, correctness, quality and quantity.

3.3 Data-centric measures used in the research

1. Data-gathering process - it can be used to extend the number of observations over time which can lead to better prediction quality. What is more the process can be automated and it takes into account the errors that may occur during the scrapping process.
2. Obtaining additional variables from the text - extending the variables set to include specific descriptions of offered cars. It can possibly include important information that may be useful to increase the prediction performance.
3. Categorize brands per country and prestige - the additional effort put into creating new variables based on the existing ones may as in the previous example lead to better quality predictions.
4. Setting constraints on the numerical variables, so that the limits are known and within sensible ranges. It prevents data inconsistency and ensures that the initial idea of the model (predict ‘standard’ car prices) remains.
5. Remove not important categorical variables - by including a 1% threshold for the values frequencies of categorical variables and later one-hot encoding the ‘curse of dimensionality’ is minimised as the sparse variables are omitted.

3.4 Modelling

In terms of the modelling section, Linear Regression, more specifically it’s variant - Ridge regression was chosen as the baseline model for comparison with other models' predictions. Additionally, more sophisticated models such as Random Forest, Gradient Boosting, and Neural Networks, will be employed as suggested in the literature. It should be noted that the models have been fine-tuned with the help of the Google Cloud Platform virtual machine by applying grid search hyperparameter tuning for each model. Due to the limited resources, the optimization was limited only to a combination of a few values of each parameter. The more advanced hyperparameter tuning shall be performed in the future enhancements of the research.

As previously stated, the models will be executed on different dataset sizes. Furthermore, the models will be tested using three approaches: a dataset containing only the numerical variables (denoted as ‘Num’), the data that was already available after cleaning (without textual binary variables and extra variables obtained from the dataset, denoted as ‘Sim’), and the data augmented by data-centric methods (denoted as ‘Adv’).

For each model, a comparison will be made to determine how well it performs on each dataset. Each of the presented models was first tuned on the given dataset. Finally, the combinations of two more advanced models will be ensembled to determine whether the combined model performs better than the individual ones.

The results of the research, measured by RMSE, are presented in the table below:

Dataset	Mean Test	Reg Lin	RF	GB	NN	GB + NN	RF + NN	RF + GB
5k Num	63 165	31 034	26 519	26 780	30 650	27 363	27 349	26 524
5k Sim	63 165	25 311	22 232	20 638	23 392	20 369	20 748	20 904
5k Adv	63 165	23 751	21 059	19 621	21 612	19 258	19 642	19 625
10k Num	65 060	31 246	24 821	24 909	27 253	25 383	25 289	24 656
10k Sim	65 060	26 048	20 902	19 295	22 525	19 693	19 733	19 277
10k Adv	65 060	24 642	20 188	18 569	21 070	18 392	18 764	18 747
20k Num	64 515	31 803	25 026	24 976	26 696	25 285	25 158	24 780
20k Sim	64 515	25 830	19 724	18 649	20 723	18 665	18 705	18 409
20k Adv	64 515	24 097	18 580	19 621	18 803	17 659	16 962	18 711
Full Num	64 729	31 779	25 304	25 046	26 940	25 318	25 540	24 953
Full Sim	64 729	26 187	19 277	17 914	20 509	18 158	18 160	17 809
Full Adv	64 729	24 640	18 610	16 713	17 917	16 439	16 755	16 870

Table 5 – Root Mean Squared Error for the combination of model and dataset on which the model was applied
Source: own elaboration

Based on the table the conclusion about the best-performing model may be reached. Taking into consideration all datasets combination Gradient Boosting outperforms other individual models in almost all cases. However, the best performance is reached by the ensembled Gradient Boosting and Neural Network model, which for the Full Advanced dataset obtained RMSE at the level of 16,439.

3.4.1 Data-centric approach compared to standard datasets

In this part, a comparison of the types of datasets within the same number of observation datasets will be made. The study's basic assumptions were that a decrease in Root Mean Squared Error for the test sample at least at a 5% level using a data-centric approach compared to the standard approach will be considered significant.

Dataset	Reg Lin	RF	GB	NN	GB + NN	RF + NN	RF + GB
5k Sim vs 5k Num	-18.4%	-16.2%	-22.9%	-23.7%	-25.6%	-24.1%	-21.2%
5k Adv vs 5k Sim	-6.2%	-5.3%	-4.9%	-7.6%	-5.5%	-5.3%	-6.1%
10k Sim vs 10k Num	-16.6%	-15.8%	-22.5%	-17.3%	-22.4%	-22.0%	-21.8%
10k Adv vs 10k Sim	-5.4%	-3.4%	-3.8%	-6.5%	-6.6%	-4.9%	-2.7%
20k Sim vs 20k Num	-18.8%	-21.2%	-25.3%	-22.4%	-26.2%	-25.6%	-25.7%
20k Adv vs 20k Sim	-6.7%	-5.8%	5.2%	-9.3%	-5.4%	-9.3%	1.6%
Full Sim vs Full Num	-17.6%	-23.8%	-28.5%	-23.9%	-28.3%	-28.9%	-28.6%
Full Adv vs Full Sim	-5.9%	-3.5%	-6.7%	-12.6%	-9.5%	-7.7%	-5.3%

Table 6 – Root Mean Squared Error comparison of model and datasets quality change
Source: own elaboration

The present research posited that the inclusion of additional features as part of a data-centric approach would enhance predictive accuracy by reducing the root mean square error (RMSE) by at least 5%. To assess the significance of this augmentation, comparisons were conducted on each dataset subset. The table presented above depicts the results of these comparisons. For each dataset size (5k, 10k, 20k, and around 41k, depicted as full), the initial comparison involved the dataset that contained solely numerical variables and the simple dataset comprising basic features after data gathering. The rows labelled "Sim vs Num" indicate the percentage change in RMSE between type “Num” and “Sim”. As anticipated, for each dataset size, there was a notable improvement that ranged from -10% to nearly -30% in RMSE. It is advisable to employ more than three numerical variables for relatively complex problems such as car price predictions. This comparison, though executed for illustrative purposes, effectively underscores the criticality of feature selection.

The most vital component of the comparison and the clue of the research involves types "Sim" and "Adv," which are visible in the rows labelled "Adv vs Sim" and presents the precent change in RMSE between “Sim” and “Adv”. The inclusion of extra variables for datasets containing 5,000, 20,000 and around 41,000 (full dataset) observations resulted in a substantial decrease in RMSE for almost all models or combinations of models (ensembled models are shown in the last three columns). While the addition of features yielded a reduction in RMSE for datasets containing 10,000 observations, it did not surpass the threshold for certain models, particularly for Random Forrest and Gradient Boosting. Two observations, highlighted in yellow, are on the verge of being significant and two observations highlighted in red have an increase in RMSE which was not expected as the sign of change was the opposite. In conclusion, the incorporation of supplementary features resulted in a significant decrease in RMSE for most datasets sizes that compared “standard” and data-centric approach (‘Adv’ vs ‘Sim) and most models (20 out of 28). The average change in RMSE across all 28 cases was -5.5%, which exceeds the presumed threshold in the first thesis.

The research findings suggest that the first thesis might be approved and draw a conclusion that additional effort required for data exploration and extraction is certainly worth considering, as it can substantially enhance prediction capability.

3.4.2 Number of observations

Dataset	Reg Lin	RF	GB	NN	GB + NN	RF + NN	RF + GB
10k Num vs 5k Num	0.68%	-6.40%	-6.99%	-11.08%	-7.24%	-7.53%	-7.04%
10k Sim vs 5k Sim	2.91%	-5.98%	-6.51%	-3.71%	-3.32%	-4.89%	-7.78%
10k Adv vs 5k Adv	3.75%	-4.14%	-5.36%	-2.51%	-4.50%	-4.47%	-4.47%
20k Num vs 10k Num	1.78%	0.83%	0.27%	-2.04%	-0.39%	-0.52%	0.50%
20k Sim vs 10k Sim	-0.84%	-5.64%	-3.35%	-8.00%	-5.22%	-5.21%	-4.50%
20k Adv vs 10k Adv	-2.21%	-7.97%	5.67%	-10.76%	-3.99%	-9.60%	-0.19%
Full Num vs 20k Num	-0.08%	1.11%	0.28%	0.91%	0.13%	1.52%	0.70%
Full Sim vs 20k Sim	1.38%	-2.27%	-3.94%	-1.03%	-2.72%	-2.91%	-3.26%
Full Adv vs 20k Adv	2.25%	0.16%	-14.82%	-4.71%	-6.91%	-1.22%	-9.84%

Table 7 - Root Mean Squared Error comparison of model and datasets size change
Source: own elaboration

The second hypothesis of this study posits that a twofold increase in the number of samples will result in a minimum 2.5% reduction in RMSE. The comparison in each row of the table pertains to the current dataset's RMSE versus that of a dataset of the same type but with the number of observations halved.

The first outcome of the analysis, based on the numbers for the Linear Regression model, suggests that dataset extension will not improve the model's performance if it is not well-suited for the problem at hand. In fact, extending the observations for the Linear Regression model not only fails to improve the prediction but also worsens it. This could be due to the inclusion of edge cases in the extended dataset that follows the non-linear pattern and cannot be correctly predicted using such a simple model.

The second lesson learned from the data is that extending the number of observations does not yield positive outcomes for datasets lacking sufficient features, as seen in the case of datasets with the type "Num". While extending the dataset from 5,000 to 10,000 observations for type

"Num" does produce some improvements, the further extension of observations does not yield any significant results. The takeaway from this is that the extension of observations may be beneficial when moving from a small to a larger sample size for datasets with insufficient meaningful features, but extending beyond that point is largely pointless. Moreover, the pattern of improvements appears to decay as the sample quantity increases.

The most significant aspect of the study is examining how changes in the number of observations affect models designed for non-linear complex problems and datasets with enough features to model such problems. In this case, the focus is on all models except for Linear Regression and considering the types of datasets the chosen ones are "Sim" and "Adv." For 10,000 datasets, all models show a significant improvement. The decrease in RMSE ranged between 2.5% and slightly above 7.5%.

Extending the number of observations from 10,000 to 20,000 results in a percentage decrease in RMSE that exceeds the threshold in 10 out of 12 cases, ranging from 3.3% to 10.76%. This represents a much bigger magnitude of improvement compared to the previous step, however not for all the models.

The next extension, from 20,000 observations to the full dataset (approximately 41,000 observations), was less successful, with improvement observed in only 8 out of 12 cases. This may be due to the decaying pattern of improvement, where additional observations become increasingly insignificant as the sample size grows larger. The phenomenon can be compared to the utility theory, where the additional observation (unit) might significantly improve the prediction considering a small sample, but when the sample is big enough the extra observation impact might be negligible.

In conclusion, extending the number of observations may not have an impact on prediction improvement if the initial model is unsuitable for the problem or if the dataset lacks meaningful features. When considering only models that are appropriate and datasets with sufficient features, 30 out of 36 cases achieved the improvement threshold of 2.5%, with an average decrease of 4.72% when the dataset was doubled. That fact shows the importance of data quantity and possibly encourages to implement data-gathering solutions as a part of Machine Learning projects and approves the second thesis raised at the beginning of the paper.

Dataset	Reg Lin	RF	GB	NN	GB + NN	RF + NN	RF + GB
Full Num vs 5k Num	2.40%	-4.58%	-6.47%	-12.10%	-7.47%	-6.61%	-5.92%
Full Sim vs 5k Sim	3.46%	-13.29%	-13.20%	-12.32%	-10.85%	-12.47%	-14.81%
Full Adv vs 5k Adv	3.74%	-11.63%	-14.82%	-17.10%	-14.64%	-14.70%	-14.04%

Table 8 - Root Mean Squared Error comparison of model and datasets size change from 5k to the full dataset
Source: own elaboration

The above graph additionally shows the improvement of prediction power measured by RMSE when going from the 5,000 observations dataset to the full dataset. The maximum decreases in RMSE were as high as 17%, which highlights the importance and worthiness of dataset quantity extension. In this study the extension was performed as a long-term data-gathering process, however, it is worth noting that it could be done by finding additional datasets that are available or performing data augmentation techniques.

3.4.3 Improving both the quality and quantity of the data

Dataset	Reg Lin	RF	GB	NN	GB + NN	RF + NN	RF + GB
10k Adv vs 5k Sim	-2.64%	-9.19%	-10.03%	-9.93%	-9.71%	-9.56%	-10.32%
20k Adv vs 5k Sim	-4.80%	-16.43%	-4.93%	-19.62%	-13.30%	-18.25%	-10.49%
Full Adv vs 5k Sim	-2.65%	-16.29%	-19.02%	-23.41%	-19.29%	-19.25%	-19.30%

Table 9 - Root Mean Squared Error comparison of model and datasets size change from 5k to the full dataset
Source: own elaboration

The graph depicted above illustrates the impact of two modifications: firstly, an increase in the number of observations from 5,000 to a larger sample, and secondly, a transition from the "Sim" type to the "Adv" type. The findings demonstrate that all of the outcomes exhibit a reduction of at least 2.5% in RMSE, however, the improvements for the Linear Regression model were not exceeding 5%. Some of the models are showing a considerable improvement up to around 23.5%, especially when the sample size was augmented from 5,000 to the full dataset. Such results signify a noteworthy advancement and encourage to implement data-centric techniques.

3.4.4 Ensembling models

Dataset	GB + NN vs GB	GB + NN vs NN	RF + NN vs RF	RF + NN vs NN	RF + GB vs RF	RF + GB vs GB
5k Num	2.18%	-10.72%	3.13%	-10.77%	0.02%	-0.96%
5k Sim	-1.30%	-12.92%	-6.68%	-11.30%	-5.97%	1.29%
5k Adv	-1.85%	-10.89%	-6.73%	-9.12%	-6.81%	0.02%
10k Num	1.90%	-6.86%	1.89%	-7.21%	-0.66%	-1.02%
10k Sim	2.06%	-12.57%	-5.59%	-12.40%	-7.77%	-0.09%
10k Adv	-0.95%	-12.71%	-7.05%	-10.94%	-7.14%	0.96%
20k Num	1.24%	-5.29%	0.53%	-5.76%	-0.98%	-0.78%
20k Sim	0.09%	-9.93%	-5.17%	-9.74%	-6.67%	-1.29%
20k Adv	-10.00%	-6.08%	-8.71%	-9.79%	0.71%	-4.64%
Full Num	1.09%	-6.02%	0.93%	-5.20%	-1.39%	-0.37%
Full Sim	1.36%	-11.46%	-5.79%	-11.45%	-7.62%	-0.59%
Full Adv	-1.64%	-8.25%	-9.97%	-6.49%	-9.35%	0.94%

Table 10 - Root Mean Squared Error comparison of model and datasets size change from 5k to the full dataset
Source: own elaboration

The third thesis posited that ensembled models would outperform individual ones. To evaluate this, two separate models were combined, and their mean prediction was compared to both individual models across the same datasets, taking both data quantity and type into account.

Upon analysis, it became apparent that ensembled models performed poorly on small samples with insufficient features. Specifically, ensembled models for the "Num" type showed poor performance overall, likely due to poor predictions based on such datasets.

Out of 48 cases, 21 did not see the ensembled model's prediction improvement exceed the 5% threshold. On average, the ensembled models saw a decrease in RMSE of around 4.71% compared to the single models. It is worth noting that the first and last column which compares the ensembled model to the Gradient Boosting model, which gave the best performance of all models, shows poor performance.

These results suggest that ensembling models are a worthwhile consideration for improving predictions, however, if the single model is of good quality ensembling may bring little to no added value. The second point is that ensembling models on bad-quality data may also not lead to the improvement of predictions. Considering this the last thesis cannot be fully approved and it would need further investigation.

3.5 Conclusions

To summarize all the findings of the research a few main conclusions may be reached:

1. Putting in additional effort to improve the dataset quality and extracting new features can lead to a significant increase in prediction power.
2. Extending a number of observations can bring a significant increase in prediction performance, however, if the model is not well defined to the problem or the dataset is missing crucial features that are used for modelling.
3. Ensembling models can bring value-added, however, if one of the ensembled models is outperforming others it might not give satisfactory results.
4. The best performance model of car prices on the Polish vehicle market was obtained by ensembling Gradient Boosting and Neural Network models.

3.6 Future improvements

1. Gather more data - extend the dataset with extra observations. The improvement consists of regular execution of the data gathering process. It could also include writing new data-gathering scripts or requesting access to APIs from the other offer portals.
2. Extract more variables from the text - other more advanced variables might be extracted from the textual data. Expert knowledge might be also required to include all necessary and the most cost-influential keywords.
3. Perform more in-depth hyperparameter tuning - the hyperparameter tuning was performed on a limited range of parameter values due to the availability of virtual machine resources.
4. Perform the analyses using different models - the analysis can be performed on the models that were not included in the current research. It might be also beneficial to ensemble more than 2 models to verify if such an action will positively affect the final quality of prediction.
5. Assumed threshold of the thesis - the threshold of 5% and 2.5% in the theses might not be the most accurate. Additional research should be made to set the thresholds based on expert knowledge.

6. Categorical variables recoding - the categorical variables were recoded by applying the 1% frequency rule, however, more advanced techniques such as PCA might be used. This matter should be addressed in further enhancements of the study.

Summary

The main objective of the research paper was to analyse the problem of car price prediction in the Polish vehicle market with an emphasis on data-centric techniques. At first, the overview of the topic was done and the research theses were formulated. In the second chapter, the current literature about car price prediction, data-centric approaches and machine learning models was reviewed. In the last chapter the study was performed based on the real data scraped from one of the Polish advertising portals. The whole process of data collection and modelling was designed to include data-centric techniques such as extraction of new features both from already existing variables and text, setting constraints on ranges of variables and removal of sparse variables. Then the comparison of models was made considering the quality and quantity of the data. The first and second theses that assumed significant improvement in prediction quality after applying data-centric approaches and extending a number of observations were approved by the performed tests. There were also additional conclusions that techniques mentioned before will not be effective if the dataset lacks significant features or the models used to predict complex issues are inadequate. The third thesis which assumed that ensembling of the models will significantly improve the prediction quality could be only partially approved. Although, the results of combined models were usually better compared to the individual ones on average they did not pass the 5% threshold. The additional conclusion is that ensembling might not give a significantly better result if one of the models is outperforming others. Last but not least the best model to predict car prices in the Polish car market, based on gathered data, was chosen. The ensembled Gradient Boosting and Neural Network model gave the lowest Root Mean Squared Error for the biggest and best quality features dataset.

References

1. Kołsut, B. (2020). The import of used cars to Poland after EU accession. Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego[Studies of the Industrial Geography Commission of the Polish GeographicalSociety], 34 (2), 129–143. doi 10.24917/20801653.342.9
2. https://https-deeplearning-ai.github.io/data-centric-comp/ [20.02.2023]
3. https://archive.ics.uci.edu/ml/datasets/automobile [21.02.2023]
4. https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho [21.02.2023]
5. https://data.world/dataman-udit/cars-data [21.02.2023]
6. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764
7. Gajera, P., Gondaliya, A., & Kavathiya, J. (2021). Old Car Price Prediction With Machine Learning. Int. Res. J. Mod. Eng. Technol. Sci, 3, 284-290.
8. Samruddhi, K., & Kumar, R. A. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE), 4, 629-632.
9. Venkatasubbu, P., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. Int. J. Eng. Adv. Technol.(IJEAT), 9(1S3).
10. Nasiboglu, R., & Akdogan, A. (2020). Estimation of the second hand car prices from data extracted via web scraping techniques. Journal of Modern Technology and Engineering, 5(2), 157-166.
11. Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. TEM Journal, 8(1), 113.
12. Alvarez-Coello, D., Wilms, D., Bekan, A., & Gómez, J. M. (2021). Towards a data-centric architecture in the automotive industry. Procedia Computer Science, 181, 658-663.
13. Whang, S. E., Roh, Y., Song, H., & Lee, J. G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. The VLDB Journal, 1-23.
14. Mekparyup, J., Saithanu, K., & Dujjanutat, J. (2013). Multiple linear regression equation for estimation of daily averages solar radiation in Chonburi, Thailand. Appl. Math. Sci, 7(73-76), 3629-3639.
15. Li, L., Dong, J., Zuo, D., & Wu, J. (2019). SLA-aware and energy-efficient VM consolidation in cloud data centers using robust linear regression prediction model. IEEE Access, 7, 9490-9500.
16. Breiman, L. (2001). Random forests. Mach. Learn. 45, 5-32.
17. Duroux, R., & Scornet, E. (2018). Impact of subsampling and tree depth on random forests. ESAIM: Probability and Statistics, 22, 96-128.
18. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
19. Russell, S. J. (2010). Artificial intelligence a modern approach. Pearson Education, Inc.
20. https://www.youtube.com/watch?v=06-AZXmwHjo [06.03.2023]

21. Polyzotis, N., & Zaharia, M. (2021). What can data-centric ai learn from data and ml engineering?. arXiv preprint arXiv:2112.06439.
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
23. https://www.lento.pl/ [13.03.2023]
24. Richardson, L. (2007). Beautiful soup documentation. April.
25. Jarrahi, M. H., Memariani, A., & Guha, S. (2022). The Principles of Data-Centric AI (DCAI). arXiv preprint arXiv:2211.14611.
26. Altman, N., & Krzywinski, M. (2018). The curse (s) of dimensionality. Nat Methods, 15(6), 399-400.
27. https://github.com/stopwords-iso/stopwords-pl [18.10.2022]
28. https://stat.gov.pl/wykres/1.html [01.05.2022]
29. Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Appendices

Appendix 1

cb – the keyword referring to cb (citizens band) radio
radio – keyword referring both to standard radio and previously mentioned cb radio
elektryczny, elektrycznie – electric – keyword referring to electric items such as electric (power) window
podgrzewany, podgrzewane, podgrzewanie – heated – keyword referring to heated seats
sensor, czujnik – sensors
alufelgi, felgi – rims, aluminium rims
aluminium, aluminiowy – as in the previous example the keyword indicating aluminium elements, especially rims
ubezpieczony, ubezpieczenie – insurance
hak – hook
MP3, AUX, USB, CD, DVD, Bluetooth – keywords referring to audio in the car
tempomat – cruise control
opona, zapasowe, zima – keywords referring to tires such as winter or spare tire
airbag, poduszka – airbags
komputer – computer
GPS, nawigacja – GPS
LED, halogen – keywords referring to lightning
ABS, ESP – keywords referring to security systems
uszkodzony, uszkodzenie, uszkodzić, szkoda – damage
bezwypadkowy – a car that was not damaged
rysa – scratch – indicates if the car has scratches
aso – authorised service station – in the context of repairing
garaż – garage – in the context of car parking
anglik – a car with the steering wheel on the right side (commonly from Great Britain)
immobiliser - immibiliser
wspomaganie – assistance
alarm – alarm
leasing, kredyt – keywords connected with lease or loan
skóra, skórzać – leather
pies – dog – possibly indicated if the dog was driving inside the car
przeciwsłoneczny – sunscreen – in the context of windows
welurowy – velvet
bogaty – rich – might be usually used in the context of interior state
asystent – assistant
kamera – camera
klimatyzacja – air conditioning
vat – vat
negocjacja – negotiation

Supplement

The supplement of the Masters Thesis consists of the codes and their description that are available on google drive under the link:
https://drive.google.com/drive/folders/1a56DGumcRPv8Qv2_sKw-gHD3Jl3uwsvj

Please note that the data used for the analysis is also included

List of shorts

LR – Linear Regression

RF – Random Forrest

GB – Gradient Boosting

NN – Neural Networks

ML – Machine Learning

List of tables

Table 1 – Data descriptive statistics before cleaning 19
Table 2 – Pearson correlation between numeric variables before cleaning 19
Table 3 – Data descriptive statistics after cleaning 23
Table 4 – Pearson correlation between numeric variables before cleaning 24
Table 5 – Root Mean Squared Error for the combination of model and dataset on which the model was applied 27
Table 6 – Root Mean Squared Error comparison of model and datasets quality change 28
Table 7 - Root Mean Squared Error comparison of model and datasets size change 30
Table 8 - Root Mean Squared Error comparison of model and datasets size change from 5k to full dataset 31
Table 9 - Root Mean Squared Error comparison of model and datasets size change from 5k to full dataset 32
Table 10 - Root Mean Squared Error comparison of model and datasets size change from 5k to full dataset 33

List of pictures

Figure 1 – Comparison of accuracy measure regarding size and the quality of the data 10
Figure 2 – Data gathering script process 17
Figure 3 – Car price variable distribution after cleaning 20
Figure 4 – Year of production variable distribution before cleaning 20
Figure 5 – Mileage variable distribution after cleaning 21
Figure 6 – Engine capacity distribution after cleaning 22
Figure 7 – Inflation factor analysis in time 22