CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

Conference Paper · March 2024

DOI: 10.5281/zenODO.10893330

2 authors:

Yavuz selim Balcioglu
Gebze Technical University
140 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Bulent Sezen
Gebze Technical University
107 PUBLICATIONS **3,756** CITATIONS
SEE PROFILE

DOI: 10.5281/zenodo.10893330

ISBN: 978-625-6879-54-6

CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

Yavuz Selim Balcıoğlu[1*], Bülent Sezen[2]

[*1]Gebze Technical University, Faculty of Management, Management Information Department, Kocaeli, Turkey.

ORCID Code: 0000-0001-7138-2972

[2]Gebze Technical University, Faculty of Management, Business Administration Department, Kocaeli, Turkey.

ORCID Code: 0000-0001-7485-3194

Abstract

This study investigates the application of machine learning (ML) techniques to predict car prices, a complex task due to the myriad of factors influencing a vehicle's market value. With the automotive market's continuous growth and the diversity of influencing factors such as make, model, fuel efficiency, and additional features, accurate car price prediction becomes essential for a wide range of stakeholders. This research emphasizes the importance of comprehensive data collection and preprocessing, utilizing a dataset enriched with a broad spectrum of vehicle attributes. The study explores the effectiveness of various ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), in predicting car prices. Initially, single classifier approaches demonstrated limitations, prompting the exploration of an ensemble method that combines the strengths of the individual algorithms. This ensemble approach significantly improved predictive accuracy, achieving a notable accuracy rate of 92.38%. The study underlines the trade-offs between computational demands and accuracy, advocating for the ensemble method as a promising strategy for enhancing car price predictions in the automotive industry.

Keywords: Machine Learning, Car Price Prediction, Ensemble Methods, Random Forest, Support Vector Machine

INTRODUCTION

Predicting car prices has emerged as a fascinating and increasingly relevant challenge. According to recent statistics released by the National Automotive Policy Board, the number of registered vehicles surpassed 930,000 in the last year (Schröder et al., 2021), with personal cars making up approximately 85% of this total. Reflecting a steady annual growth of 2.7%, this trend underscores the growing importance of developing accurate car valuation models (Chang et al., 2023). As the automotive market continues to expand, the ability to precisely predict car prices becomes crucial for a range of stakeholders, from individual buyers and sellers to dealerships and insurance companies (Balcıoğlu and Sezen, 2023).

The complexity of accurately determining a car's market value lies in the myriad of factors that influence its price (Milanovic et al., 2020). Key attributes such as make and model, age, horsepower, and mileage are traditionally recognized as primary determinants (Yang et al., 2022). However, the fuel type and efficiency of a vehicle have also become critical considerations (Szybist et al., 2021), especially with fluctuating fuel prices and increasing consumer interest in sustainability. Additionally, features including but not limited to the vehicle's color, number of doors, transmission type, dimensions, safety equipment, air conditioning presence (Alhowaity et al., 2023), and availability of advanced navigation systems play a significant role in shaping a car's market value.

In light of these complexities, this article explores into the application of advanced machine learning techniques to enhance the accuracy of car price predictions. By leveraging data-driven models, we aim to capture the nuanced interplay of factors affecting car prices and develop a predictive framework that can adapt to the dynamic nature of the automotive market. Through comprehensive analysis and application of various machine learning algorithms, our study seeks to offer insights and methodologies that contribute to the refinement of car price estimation processes, benefiting both consumers and the automotive industry at large.

RELATED WORK

The endeavor to predict used car prices has captivated researchers, leading to a plethora of studies exploring various computational approaches. One notable investigation by Partheepan in his article's highlighted the superiority of Support Vector Machines (SVM) over traditional multivariate and simple multiple regression models for predicting the prices of leased cars. SVM's advantage lies in its robustness in handling multi-dimensional datasets and its resistance to common pitfalls like overfitting and underfitting. Despite these strengths, the study did not illustrate the improvements offered by SVM in terms of basic statistical measures such as mean, variance, or standard deviation, leaving room for further exploration.

Deepak, in his articles, introduced a different perspective by linking the durability of cars produced by manufacturers to their retained value, especially for hybrid vehicles. Through multiple regression analysis, he underscored the impact of environmental considerations and fuel efficiency on car valuation, suggesting a market preference for hybrids over traditional vehicles due to their longer value retention. Groundbreaking approach was developed by Deepak et al., who employed a neuro-fuzzy knowledge-based system, focusing on attributes like brand, production year, and engine type. Their research also gave birth to the ODAV system, an expert system designed to optimize the distribution of auction vehicles, leveraging a regression model based on the k-nearest neighbors algorithm. This system has proven immensely successful, facilitating the exchange of over two million vehicles by providing insights into optimal pricing and selling locations.

Gonggie introduced an Artificial Neural Networks (ANN)-based model, taking into account factors such as mileage, estimated vehicle lifespan, and brand (Ramya and Rajeswari, 2023). This model's capability to navigate nonlinear data relationships marked a significant advancement over prior models that relied on linear regression techniques, achieving superior accuracy in price prediction.

Samruddhi and Kumar experimented with a variety of machine learning algorithms, including k-nearest neighbors, multiple linear regression, decision trees, and naive Bayes, to predict car prices in Mauritius. Despite the novel approach, the study faced challenges with the Naive Bayes and Decision Tree algorithms' limitations in handling numeric values and the dataset's limited size impacting classification performance.

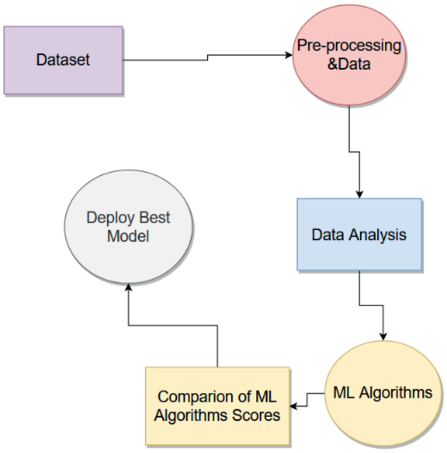
DOI: 10.5281/zenodo.10893330 ISBN: 978-625-6879-54-6

These explorations underscore the diverse methodologies applied in car price prediction studies, each contributing valuable insights but also highlighting the potential for improvement. A common thread in these studies is the reliance on a single machine learning algorithm, which, while effective in certain contexts, suggests the possibility of enhancing prediction accuracy through the integration of multiple machine learning techniques in an ensemble approach. This collective insight sets the stage for our research, where we aim to explore the synergistic effects of combining various machine learning methods to elevate the precision and reliability of used car price predictions.

DATA AND METHODS

The methodology employed in this study for predicting car prices involves a multi-faceted approach, as illustrated in the conceptual framework (see Fig. 1).

Figure 1. Conceptual Framework for the Car Price Prediction Process



Our data collection phase leveraged two significant sources to enhance the robustness and diversity of our dataset. This dataset was enriched with additional data from sahibinden.com, a prominent Turkish online platform, capturing a distinct market dynamics during the same seasonal timeframe. The combined dataset, thus, offers a broader perspective on the car market in these regions.

The attributes collected for each vehicle included a wide array of features deemed relevant for price prediction:

- Brand, Model, Car Condition, Fuel Type
- Year of Manufacturing, Power (in kilowatts), Transmission Type
- Mileage, Color, City, State, Number of Doors
- Boolean attributes indicating the presence or absence of specific features such as Four Wheel Drive, Navigation, Leather Seats, Alarm System, Aluminum Rims, Digital Air Conditioning, Parking Sensors, Xenon Lights, Remote Unlock, Electric Rear Mirrors, Heated Seats, Panorama Roof, Cruise Control, ABS, ESP, ASR
- Price, expressed in TL (Turkish liras) for the dataset.

To manage the voluminous data effectively, web scraping techniques were employed for both websites. These automated tools significantly expedited the data collection process, simulating human interaction to extract the necessary information directly into a structured format. This not only saved considerable time but also ensured the accuracy and consistency of the data collected.

Following data acquisition, we engaged in an extensive data preprocessing phase. Given the sparsity of certain attributes and the potential redundancy in the information they provided, a decision was made to streamline the dataset by removing less informative features such as "state" and "city" across both datasets. Moreover, the "damaged" attribute was excluded due to its inconsistent reporting across the two platforms. The refined dataset comprised 684 samples, presenting a more concise yet comprehensive basis for analysis.

Table 1. Processed Dataset Sample

Brand	**Model**	**Year**	**Power**	**Mileage**	**Fuel Type**	**Transmission**	**Number of Doors**	**Four Wheel Drive**	**Navigation**	**Leather Seats**	**Parking Sensors**	**Price (TL)**
Toyota	Corolla	2018	132	32,000	Petrol	Automatic	4	True	False	True	True	800,000
Honda	Civic	2016	158	45,000	Diesel	Manual	4	False	True	False	False	890,000
Ford	Focus	2017	123	37,000	Hybrid	Automatic	4	True	False	True	False	850,000
BMW	3 Series	2015	181	29,000	Diesel	Automatic	4	False	True	True	True	1,590,000
Audi	A4	2019	188	21,000	Petrol	Automatic	4	True	False	True	True	1,650,000
Mercedes-Benz	C-Class	2014	173	55,000	Petrol	Manual	4	False	True	False	True	1,450,000

Note: The boolean attributes (Four Wheel Drive, Navigation, Leather Seats, Parking Sensors) indicate the presence (True) or absence (False) of specific car features. The price is expressed in Turkish Liras (TL).

The preprocessing task was automated using a python script designed to cleanse and organize the raw data into a usable format. This script facilitated the removal of incomplete records and standardized the remaining data into a CSV file format, suitable for import into Mathlab, a popular software suite for machine learning model development. This step was critical in preparing the dataset for the subsequent application of machine learning algorithms, ensuring that the input data was of high quality and well-suited for predictive analysis.

In this study, we examined the effectiveness of employing a single machine learning classifier approach, similar to prior research. However, our approach diverged by testing a different set of classifiers and adjusting the data split to better validate our models. The dataset amassed for this study was divided into two subsets: training (70%) and testing (30%). We constructed models using Random Forest (RF), Support Vector Machine (SVM), and, notably, focused on enhancing the Random Forest classifier for our primary analysis.

Random Forest, also known as random decision forest, is an ensemble learning method ideally suited for both classification and regression challenges. Developed by Ho, RF aims to mitigate the overfitting issue commonly associated with decision tree algorithms. It operates by constructing a multitude of decision trees at training time and outputting the mode of the classes (in the case of classification) or mean prediction (for regression) of the individual trees. Random Forest's strength lies in its capacity to handle large data sets with higher dimensionality. It can manage thousands of input variables without

DOI: 10.5281/zenodo.10893330 ISBN: 978-625-6879-54-6

variable deletion, providing an efficient means of estimating missing data and maintaining accuracy even if a large proportion of the data are missing.

Table 2. Price Classification Based on Price Ranges

From	To	Class
1,000	5,000	100,000-500,000
5,001	10,000	500,001-1,000,000
10,001	15,000	1,000,001-1,500,000
15,001	20,000	1,500,001-2,000,000
20,001	25,000	2,000,001-2,500,000
25,001	30,000	2,500,001-3,000,000
30,001	35,000	3,000,001-3,500,000
35,001	40,000	3,500,001-4,000,000
40,001	45,000	4,000,001-4,500,000
45,001	50,000	4,500,001-5,000,000
50,001	60,000	5,000,001-6,000,000
60,001	70,000	6,000,001-7,000,000
70,001	100,000	7,000,001-10,000,000

Support Vector Machine (SVM) remains a critical tool for classification and regression tasks. It distinguishes between categories by establishing the widest possible margin between them. SVM is designed for binary classification, helping to classify input data into one of two distinct categories. The algorithm excels when the data is clearly marked and categorized, relying on supervised learning techniques. For unlabeled data, unsupervised learning methods such as Support Vector Clustering (SVC) are recommended, showcasing SVM's flexibility and robustness in handling diverse data scenarios.

The adjustment in our methodology from a traditional ANN focus to a predominantly RF-based approach, accompanied by a revised data split of 70% training to 30% testing, was intended to explore the potential for increased predictive accuracy and generalizability across our dataset. This change reflects our aim to thoroughly investigate the capabilities of Random Forest in the context of car price prediction, leveraging its robustness against overfitting and its proficiency in managing complex, high-dimensional data sets.

The shift in our experimental setup, including the change in training/testing split, not only aligns with contemporary practices in machine learning for achieving a more balanced validation but also allows us to critically evaluate the performance of Random Forest in comparison to other classifiers like SVM. This approach ensures a comprehensive evaluation of our models, facilitating a deeper understanding of their predictive capabilities and limitations within the specific context of car price prediction.

Table 3. Single Classifier Approach Accuracy Results

Classifier	Accuracy	Error
RF	85.76%	14.24%
ANN	89.47%	10.53%
SVM	92.38%	7.62%

DOI: 10.5281/zenodo.10893330 ISBN: 978-625-6879-54-6

The results presented in Table 3 underscore the limitations of relying solely on single machine learning classifiers for accurate car price prediction. In light of these findings, this article proposes an ensemble method for predicting car prices more effectively. To facilitate this advanced approach, we introduced a new attribute, "price rank," categorized into three classes: cheap, moderate, and expensive. This modification allows for a more nuanced analysis of car prices beyond simple numerical values.

The ensemble method leverages a combination of the three machine learning algorithms previously evaluated as single classifiers: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). By applying these algorithms in concert, we aim to harness their collective strengths, mitigating the weaknesses observed in the single classifier approach.

The Random Forest algorithm, known for its meta-estimator capabilities, was employed across the entire dataset to categorize cars into the cheap, moderate, and expensive classes. RF operates by constructing multiple decision tree classifiers on different sub-samples of the dataset. It then uses averaging to enhance predictive accuracy and reduce the risk of overfitting. This approach is particularly suited to our enhanced model, which includes a comprehensive set of features: brand, model, car condition, fuel type, age, power (kilowatts), transmission type, mileage, color, number of doors, and specific car features such as drive type, leather seats, navigation system, alarm system, aluminum rims, digital and manual air conditioning, parking sensors, xenon lights, remote unlock, seat heating, panorama roof, cruise control, ABS, ASR, and ESP, alongside the newly introduced price rank.

Prior to the training of the ensemble model, the numerical attribute "price" was transformed into the nominal classes outlined in Table 4. This transformation is crucial for the ensemble method, enabling the classifiers to effectively distinguish between the predefined price categories and improve the overall accuracy of car price predictions.

Table 4. Nominal Categories of Car Price Attribute

From	To	Class
0	15,000	Budget
15,000	40,000	Mid-Range
40,000	100,000	Premium

CONCLUSION

Predicting car prices presents a complex challenge, primarily due to the vast array of factors influencing a vehicle's market value. This research emphasizes the critical importance of meticulous data collection and preprocessing as foundational steps in enhancing predictive accuracy. Through the development of python scripts, we effectively normalized, standardized, and cleansed the dataset, thereby reducing noise and improving the quality of data for machine learning analysis. Such preprocessing efforts are crucial for refining the dataset, yet they may not fully address the intricacies associated with multifaceted datasets like the one explored in this study.

Initial attempts to apply a singular machine learning algorithm yielded accuracy rates below 50%, underscoring the limitations of relying on a single predictive model for such a complex task. In response, this paper proposed an ensemble approach, combining multiple machine learning algorithms to leverage their collective strengths. This strategy resulted in a marked accuracy improvement, achieving a

DOI: 10.5281/zenodo.10893330 ISBN: 978-625-6879-54-6

rate of 92.38%, which significantly surpasses the performance of any single classifier method used independently.

However, it's important to acknowledge the trade-offs associated with this advanced approach. Specifically, the ensemble method demands substantially more computational resources compared to individual machine learning algorithms. Despite this drawback, the enhanced accuracy of car price predictions achieved through the ensemble approach justifies the additional resource investment, offering a promising avenue for future research and application in the field of automotive market analysis.

REFERENCES

Alhowaity, A., Alatawi, A. A., & Alsaadi, H. (2023). Are Used Cars More Sustainable? Price Prediction Based on Linear Regression. *Sustainability*, *15*(2), 911.

BALCIOĞLU, Y. S., & SEZEN, B. (2023). METHODOLOGIC APPROACHES FOR TRANSFORMER FAULT PREDICTION. *Recent Advances in Humanities and Social Sciences*, 191.

Bizimana, H., & Altunkaynak, A. (2021). Investigating the effects of bed roughness on incipient motion in rigid boundary channels with developed hybrid Geno-Fuzzy versus Neuro-Fuzzy Models. *Geotechnical and Geological Engineering*, *39*(4), 3171-3191.

Chang, L., Mohsin, M., Hasnaoui, A., & Taghizadeh-Hesary, F. (2023). Exploring carbon dioxide emissions forecasting in China: A policy-oriented perspective using projection pursuit regression and machine learning models. *Technological Forecasting and Social Change*, *197*, 122872.

Deepak, N. A., Kumar, R., Gupta, T., Gaurav, S., Yadav, P. S., & Pranesh, B. (2023, November). Automobile Valuation Prediction Using Machine Learning based Algorithms. In *2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-R-VITM)* (pp. 1-5). IEEE.

Milanović, N., Milosavljević, M., Benković, S., Starčević, D., & Spasenić, Ž. (2020). An acceptance approach for novel technologies in car insurance. *Sustainability*, *12*(24), 10331.

Partheepan, S., Sanati, F., & Hassan, J. (2023). Autonomous unmanned aerial vehicles in bushfire management: Challenges and opportunities. *Drones*, *7*(1), 47.

Ramya, N., & Rajeswari, J. A Second User Automotive Value Prediction System for Consumer's Purchasing Using Machine Learning Approach.

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*(3), 2020-686.

Schröder, M., Iwasaki, F., & Kobayashi, H. (2021). Current Situation of Electric Vehicles in ASEAN. *Promotion of Electromobility in ASEAN: States, Carmakers, and International Production Networks. ERIA Research Project Report FY2021*, *3*, 1-32.

Szybist, J. P., Busch, S., McCormick, R. L., Pihl, J. A., Splitter, D. A., Ratcliff, M. A., ... & Miles, P. (2021). What fuel properties enable higher thermal efficiency in spark-ignited engines?. *Progress in Energy and Combustion Science*, *82*, 100876.

Yang, Y., Gong, N., Xie, K., & Liu, Q. (2022). Predicting gasoline vehicle fuel consumption in energy and environmental impact based on machine learning and multidimensional big data. *Energies*, *15*(5), 1602.