A Survey on Used Vehicle Price Estimation Systems using Artificial Intelligence Methods

Thuraya Alnajim, Nouf Alshahrani, and Omar Asiri

Faculty of Computers and Information Technology,
University of Tabuk, Tabuk, 71494, Saudi Arabia

Summary
Recently, the market of used car sales has been growing very fast in the Kingdom of Saudi Arabia, due to the high price of new cars that are not affordable by most people. Several Ecommerce websites offer third-party services between the buyers and sellers in the market of used cars. However, it is very helpful to have a previous information about the correct vehicle price for many buyers and sellers before making any decision of selling or buying a used car. Therefore, there is a great demand to develop an accurate vehicle price estimation system through the employment of Machine Learning and Deep Learning techniques. However, there are a large number of significant factors that affect the vehicle price, which make the vehicle price estimation is a difficult task. In general, the standard regression methods might not be efficient for high dimensional data. This paper aims to investigate the recent developed vehicle price estimation systems that are based on the employment of machine learning and deep learning techniques. The recent developed systems are discussed and criticized in details. In addition, we present a set of evaluation metrics to assess the efficiency of any vehicle price prediction system.

Keywords:
Vehicle price estimation, machine learning, deep learning, survey.

1. Introduction

Recently, the term Artificial Intelligence (AI) has been used widely in diverse types of applications and approaches [1, 2, 3]. In general, the AI landscape spreads across a collection of technologies including: Machine Learning (ML), Deep Learning (DL), Computer Vision, Natural Language Processing (NLP), and many others. These technologies allow the machine to understand human language, make a prediction, and learn from example [4].
Mainly, the AI system accepts data from an input unit (for instance, text, image, speech, etc.). The AI system then processes the received data through applying different rules and algorithms, predicting, interpreting, and acting upon the input data. After the processing task is completed, the system provides an outcome (for instance, success or fail). The final result is then assessed through analyzing, discovering, and obtain feedback. Finally, the AI system employs its assessments to maintain the input data, rules, and algorithms. Usually, this loop is repeated till the desired result is achieved [5].

Manuscript received January 5, 2025
Manuscript revised January 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.1.19

Recently, the used car price prediction issue has been considered as a high research interest, as it requires obvious effort and knowledge in the field. In 2021, almost 557,000 used cars were sold to customers in the Kingdom of Saudi Arabia [6], this number has increased in the last few years due to several reasons, including the increasing in exportation and transportation fees, as the prices of used cars have gone by 20% [7].

Therefore, there is a high demand for developing an efficient used car price estimation system using the advances in Machine Learning (ML) and Deep Learning (DL) technologies, where this kind of systems will benefit the car buyers in several ways: minimize the time needed for estimating such cars, minimize the required effort for checking up the vehicle, and save buyers' expenses [8].
According to the research study conducted in this paper, the problem of used vehicle predicting prices has not been studied intensively in the Kingdom of Saudi Arabia, although this kind of tasks is considered as a critical task for vehicle buyers and even for sellers.

This paper aims to discuss and analyze the problem of used vehicle price estimation in the Kingdom of Saudi Arabia, and summarize the recent developed AI-based vehicle price estimation approaches. The main contribution of this paper lies on the following aspects:

1. Research the recent developed AI-based used vehicle price estimation systems.
2. Discuss and analyze the recent developed systems.
3. Build an evaluation metrics for assessing the efficiency of any developed used vehicle price estimation system.

The remainder of this paper is structured as follows: Section 2 discusses the recent developed vehicle price estimation systems. In Section 3, we discuss and analyze the developed systems in terms of regression accuracy and the employed machine learning or deep learning models. Section 4 presents several evaluation metrics which are required to be taken into consideration when developing an efficient vehicle price estimation system. And Finally, Section 5 concludes the work presented in this paper and draws a list of future works.

2. Related works

This section discusses the recent developed machine-learning and deep-learning approaches for the used vehicle estimation prices. There are several artificial intelligence approaches that have been developed for the purpose of vehicle estimation, however, in this section, we discuss the most relevant and recent approaches.

2.1 Used Vehicle Price Estimation Systems

The work presented by Lavanya et al. [9] involves the design and development of vehicle price prediction system using machine learning. Three 3 different machine learning approaches have been employed: linear regression, decision tree, and random forest to predict the vehicles prices in Mauritius. The obtained results revealed that the random forest classifier offers the best means absolute error and mean square error among its competitors. On the other hand, authors showed that the lack of results referred to the low number of records that have been utilized.

Chandak et al. [10] developed a supervised machine learning model for estimating the value of a vehicle based on the employment of several attributes. The developed system consists of several machine learning models including the K-Nearest Neighbour (KNN) and decision tree. This work includes a correlation study among the employed attributes, for instance, the obtained results revealed that there is a positive correlation between the prices and the milage, year of registration and milage, whereas a negative correlation is presented between the prices and the year of registration. In conclusion, authors revealed that the KNN classifier offers the best mean square error.

The work presented by Gegic et al. [11] involves a development of a vehicle price estimation system using the employment of 3 different machine learning models: Artificial Neural Networks, Support Vector Machine, and Random Forest. The developed system ensembled multiple machine learning algorithm to increase the vehicle price estimation accuracy. The obtained accuracy was close to 92.38% using the developed ensemble system, where this is a significant improvement over employing single machine learning models.

The work presented by Gajera et al. [12] includes the design and development of a smart vehicle price estimation system using the employment of machine learning technology, through adopting 5 different supervised machine learning algorithms, including: Linear regression, KNN, Random Forest, XB boost and Decision tree. According to the obtained results, the presented results analyzed the performance of the employed machine learning models through assessing the root mean squared error rate for each one. Random Forest achieves the best vehicle price estimation with the minimum root means squared error of 3702.34.

Samruddhi et al. [13] proposed a supervised machine learning model using the K-Nearest Neighbor regression algorithm in order to analyze the prices of the used vehicles. The developed model has been trained with a dataset for 3 different ratios for k values from 2 to 10. The obtained accuracy result is close to 85% with Root Mean Squared Error of 4.01 and Mean Absolute Error rate of 2.01 with the K value of 4.

The work presented by Mammadov [14] involves the design and development of a linear regression model to predict the car prices in the United States (US) market, to help customers to understand the most significant features in the US automobile industry. In this work, several data analysis techniques have been employed, including data cleaning, exploration, data visualization, and feature selection through the adoption of the Recursive Feature Elimination (RFE).

A machine learning algorithm based on employment of linear regression model has been proposed by Asghar et al. [15] to estimate the price of cars. Different data cleaning and processing methods have been employed in order to overcome the problems exist in the employed vehicle price dataset. The obtained results showed an enhancement over the existing research working with RMSE of 0.919.

Karakoc et al. [16] developed two different artificial neural network models for car price prediction and tested them using a dataset from car sales website. The designed neural network model consists of two hidden layers with 30 and 25 neurons for the first hidden and second hidden layers respectively. The developed neural network model offers high prediction accuracy with 91.38%.

The work presented by Xia et al. [17] includes the design, development and an evaluation of ForeXGBoost model that takes the advantage of carefully-designed data filling algorithms for the recovery of missing data. A sliding window has been employed in order to extract the historical sales and production data features. ForeXGBoost can enhance the car prediction accuracy. In addition, this work analyzed the impact of different attributes and data correlation among features on the overall prediction accuracy.

Three different machine learning algorithms based on the kind of regression, have been implemented by Gegic et al. [18], in order to predict the price of used cars. Three different supervised machine learning algorithms have been employed as follows: linear regression, lasso regression and ridge regression. The obtained results from employing the aforementioned three machine learning models, are as follows: 83.65% for linear regression, 87.09% for the lasso regression, and 84.00% for the ridge regression models.

Liu et al. [19] considered the performance issue of a car prediction system, through analyzing the linear correlation between vehicle parameters, vehicle conditions, transaction

factors, and used car price. A car price prediction system based on the PSO-GRA-BPNN architecture has been proposed. The correlation between the new vehicle price, mileage, gearbox, displacement, fuel consumption, and registration time on used car prices is greater than 0.7. The proposed PSO-GRA-BPNN model archives low prediction error with 30.041% smaller than the error of the other three models (linear regression, random forest, and support vector machine).

The work presented by Hankar et al. [20] involves the employment of several regression methods based on supervised machine learning in order to predict the vehicle price of used cars through considering many factors including: milage, fuel type, mark, model, fiscal, and the production year of the vehicle. According to the experiments conducted, authors revealed that gradient boosting regression showed a high R-squared score and low root mean square error.

A comparative study on the performance of regression has been conducted by Monburinon et al. [21] based on supervised machine learning models using car market dataset collected from German E-commerce website. The presented work analyzed the performance of three different machine learning models: Gradient Boosted, Random Forest, and Multiple Linear Regression. As a result, the gradient boosted regression model achieves the best Mean Absolute Error (MAE) with 0.28, whereas the Random Forest Regression model offers MAE with 0.35, and the Multiple Linear Regression with MAE of 0.55.

Pudaruth [22] investigated the application of supervised machine learning methods to estimate the used vehicle prices in Mauritius. Different machine learning techniques have been employed, including: k-nearest neighbors, naïve bayes, and decision trees, through the adoption of data from daily newspapers. After considering several experiments, the mean error for the linear regression was 51,000, whereas the mean error was around 27,000 using the KNN.

A used vehicle price prediction system is presented by Noor & Jan [23] using the adoption of supervised machine learning algorithms. Multiple linear regression models have been adopted that offered 98% estimation precision. In addition, a prediction system has been proposed, where the price is dependent variable, where the price attribute is derived from several factors, such as: vehicle's model, city, make, color, version, mileage, and power-steering.

The work presented by Peerun et al. [24] includes an investigation on assessing where it is reliable to predict the used vehicle's price using artificial neural networks. A dataset with 200 different cars has been adopted, from different sources, and fed into four machine learning models. As a result, authors found that Support Vector Machine offers slightly better results than the neural network or linear regression. However, authors recommend to employ a large dataset in order to conduct more experimentation with different network types and structures to perform better prediction accuracy.

Sun et al. [25], a vehicle price evaluation model has been developed based on big data analysis that exploits the advantages of widely circulated vehicle data and a large number of vehicle transaction data to investigate the price data for each vehicle through adopting the optimized BP neural network algorithm. The optimized BP neural network model is employed to choose the best number of hidden neurons in the BP neural network, which enhances the convergence speed of the network topology, in addition to improving the prediction accuracy of the prediction model.

A modern data mining method has been employed by Listiani [26] that is independent of input dimension namely the Support Vector Regression, that has been used to overcome the problem of used vehicle price estimation. Authors compared the estimation accuracy against the statistical regression model. The work presented involves a fully automatic approach for turning and implementing the SVR.

The work presented by Wu et al. [27] involves an expert system for used vehicle prices using adaptive neuro-fuzzy inference system. The developed system consists of three main phases: data acquisition function, price estimator algorithm, and performance analysis system. Authors revealed that the main factors that affect the vehicle's price are: mark of the car, production year, and engine style. The developed system has been compared with a conventional artificial neural network model with back-propagation network. This work revealed that the proposed expert system enhances the vehicle's price prediction accuracy.

3. Discussion

As presented in the previous section, more than 18 research studies discussed the problem of used vehicle price estimation. However, the presented works differ in terms of efficiency, employed dataset, number of employed features, and the total number of records in the adopted dataset. Therefore, this section discusses the recent developed systems in terms of the total number of records in the selected dataset, total number of features, and the employment frequency of the ML model.

According to research work discussed in the previous section, there are 14 different vehicle price prediction datasets have been employed in the recent developed vehicle price prediction systems. Each dataset is different in size, parameters, and data types. Figure 1 presents the number of records that are exists in each dataset. As seen below, Alibaba Cloud Car Sales dataset is the one with the highest number of records with almost 1,000,000 records, whereas the E-commerce dataset includes the minimum number of records with 200 records in total.

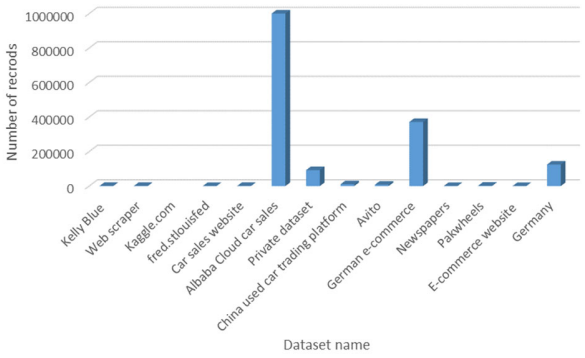

Fig. 1 Total number of records for each dataset

In general, used vehicle price prediction datasets with high number of records will achieve better prediction accuracy. Therefore, Alibaba Cloud Car Sales dataset may offer better prediction accuracy than other datasets. Then, it is significant to focus on a vehicle price prediction dataset with large number of records.

Through the ML model evaluation regimes, feature selection is considered as significant technique to reduce the model complexity and enhance the model performance in regards to model fit, generalization, and the prediction accuracy. Hence, it is significant to analyze the most significant features of the vehicle price dataset, to enhance the ML model accuracy and efficiency.

The number of features is also studies and analyzed for each dataset. The selected features in the training process have a high impact on the training accuracy. For instance, considering significant features in the training process will increase the prediction accuracy. Figure 2 shows the total number of features that have been employed for each vehicle price prediction dataset. As noticed, the Germany dataset includes the largest number of employed features, with almost 179 features, whereas Alibaba Cloud Car Sales dataset came in the next place with a total number of 33 features. The minimum number of employed features is exist in five different datasets: Kelly Blue book, Private dataset 2, Avito, Newspapers, and Pakwheels datasets.

The Germany vehicle price prediction dataset is considered as unreliable with comparison to other datasets. As most of the existing vehicle price prediction datasets employed less than 15 features in the training process. Moreover, the mean of the employed features in the recent developed vehicle price predictions systems is almost 8 features. Therefore, it is important to choose the most significant features that affect the vehicle price prediction function, in order to allow for better vehicle price prediction accuracy.

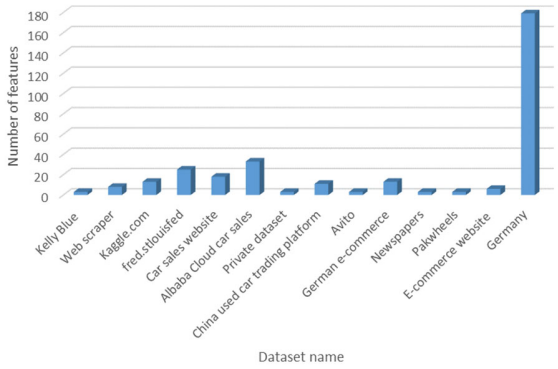

Fig. 2 The number of features for each dataset

The employment of machine learning or deep learning models have been studies and analyzed also. There are several machine learning and deep learning models have been employed with different requirements for the training process and complexity, and the prediction accuracy. Figure 3 shows the machine learning and deep learning models which have been employed in the vehicle price prediction. In addition, the frequency for employing each model is presented. As noticed below, Logistic Regression model has been employed in 6 different research works, with the maximum number of employments. On the other hand, XG, Naïve Bayes Ridge, and Lasso models have been employed only once.

Therefore, it is very important to select an efficient machine learning model with efficient parameters in order to allow for better used vehicle price prediction accuracy. An intensive research work is required to be accomplished in order to assess the performance of several machine learning and deep learning models, to pick the best model with the most efficient reliability and prediction accuracy.

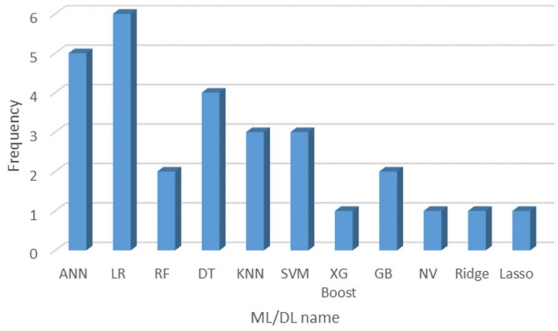

Fig. 3 The employment frequency of ML/DL models in the existing systems

4. Evaluation Metric for Used Vehicle Price Estimation Systems

As discussed above, several used car price prediction systems have been developed recently with various datasets, factors, accuracy results, and the employed ML model. Therefore, it is significant to study and analyze the performance of such system through adopting a comparison mechanism. For this purpose, we presented five main parameters which can be adopted in assessing the recent developed vehicle price prediction systems. The suggested parameters are as follows:

1. Employed ML/DL: many machine learning and deep learning models have been developed with different accuracy, memory requirements, and speed. Therefore, it is important to study and analyze the selected machine learning or deep learning models.

2. Car price dataset: as noticed in the literature, there are several car price datasets available online. However, it is significant to select an efficient car price dataset in order to develop an accurate car prediction system.

3. Dataset size: the car prediction dataset size is a significant for developing a reliable car prediction system.

4. Number of features: this presents the total number of features that have been employed to train the ML model.

5. Obtained results: the performance of each developed system has been studied and analyzed, in order to present the car price prediction accuracy.

Table 1 shows a comparison between the recent developed used vehicle price prediction systems. As presented below, the existing systems have been analyzed based on five main factors as stated above.

5. Conclusion and Future work

The used vehicle price estimation issue is considered as a significant task nowadays in the Kingdom of Saudi Arabia due to the high demand on used cars market. Therefore, it is important to study and analyze the recent developed systems in the same field. This paper discusses the used vehicle price estimation issue, and presents the recent developed approaches for used vehicle price estimation using the employment of ML and DL technologies. In addition, this paper presents a set of evaluation metrics that can be used to assess the efficiency of any ML/DL-based used vehicle price estimation system. For future work, we aim to develop an efficient and reliable used vehicle price estimation system through analyzing the efficiency of several ML and DL approaches, and the employment of an efficient feature extraction methods in order to enhance the system's accuracy.

References

[1] Latif, M.H. and Afzal, H., 2016. Prediction of movies popularity using machine learning techniques. *International Journal of Computer Science and Network Security (IJCSNS)*, *16*(8), p.127.

[2] Saad, O., Darwish, A. and Faraj, R., 2012. A survey of machine learning techniques for Spam filtering. *International Journal of Computer Science and Network Security (IJCSNS)*, *12*(2), p.66.

[3] Alsubait, T. and Alfageh, D., 2021. Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments. *International Journal of Computer Science & Network Security*, *21*(1), pp.1-5.

[4] Daugherty, P.R. and Wilson, H.J., 2018. Human+ machine: Reimagining work in the *age of AI. Harvard Business Press*.

[5] Suguna, S.K., Dhivya, M. and Paiva, S. eds., 2021. Artificial Intelligence (AI): *Recent Trends and Applications.*

[6] Saudi Arabia - sales volume of new motor vehicles 2021 | Statista, 2022. Available Online: https://www.statista.com/statistics/375695/motor-vehicle-sales-in-saudi-arabia/#:~:text=In%202021%2C%20almost%20557%20thousand,include%20Toyota%2C%20Hyundai%20and%20Mazda. Last accessed: 28/09/2022

[7] "Saudi Arabia's used cars sales market is poised to take-off." Consultancy-me. October 18, 2022. https://doi.org/https://www.consultancy-me.com/news/5464/saudi-arabias-used-cars-sales-market-is-poised-to-take-off.

[8] Muniasamy, A. and Alasiry, A., 2020. Deep learning: The impact on future elearning. *International Journal of Emerging Technologies in Learning*, *15*(1).

[9] Lavanya, B., Reshma, S., Nikitha, N. and Namitha, M., 2021. Vehicle resale price prediction using machine learning. *UGC Care Group I*.

[10] Chandak, A., Ganorkar, P., Sharma, S., Bagmar, A., and Tiwari, S., 2019. Car Price Prediction using Machine Learning. JCSE International Journal of Computer Sciences and Engineering. Vol. 7. No. 5.

[11] Gegic, E., Isakovic, B., Keco, D., Masetic, Z. and Kevric, J., 2019. Car price prediction using machine learning techniques. *TEM Journal*, *8*(1), p.113.

[12] Gajera, P., Gondaliya, A. and Kavathiya, J., 2021. Old Car Price Prediction with Machine Learning. *Int. Res. J. Mod. Eng. Technol. Sci*, *3*, pp.284-290.

[13] Samruddhi, K. and Kumar, R.A., 2020. Used Car Price Prediction using K-Nearest Neighbor Based Model. *Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE)*, *4*, pp.629-632.

[14] Mammadov, H., 2021. Car Price Prediction in the USA by using Liner Regression. *International Journal of Economic Behavior (IJEB)*, *11*(1), pp.99-108.

[15] Asghar, M., Mehmood, K., Yasin, S. and Khan, Z.M., 2021. Used Cars Price Prediction using Machine Learning with Optimal Features. *Pakistan Journal of Engineering and Technology*, *4*(2), pp.113-119.

[16] KARAKOÇ, M.M., ÇELİK, G. and VAROL, A., 2020. Car Price Prediction Using An Artificial Neural Network. *Eastern Anatolian Journal of Science*, *6*(2), pp.44-48.

[17] Xia, Z., Xue, S., Wu, L., Sun, J., Chen, Y. and Zhang, R., 2020. ForeXGBoost: passenger car sales prediction based on XGBoost. *Distributed and Parallel Databases*, *38*(3), pp.713-738.

[18] Gegic, E., Isakovic, B., Keco, D., Masetic, Z. and Kevric, J., 2019. Car price prediction using machine learning techniques. *TEM Journal*, *8*(1), p.113.

[19] Liu, E., Li, J., Zheng, A., Liu, H. and Jiang, T., 2022. Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network. *Sustainability*, *14*(15), p.8993.

[20] Hankar, M., Birjali, M. and Beni-Hssane, A., 2022, May. Used Car Price Prediction using Machine Learning: A Case Study. In *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)* (pp. 1-4). IEEE.

[21] Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S. and Boonpou, P., 2018, May. Prediction of prices for used car by using regression models. In *2018 5th International Conference on Business and Industrial Research (ICBIR)* (pp. 115-119). IEEE.

[22] Pudaruth, S., 2014. Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), pp.753-764.

[23] Noor, K. and Jan, S., 2017. Vehicle price prediction system using machine learning techniques. *International Journal of Computer Applications*, *167*(9), pp.27-31.

[24] Peerun, S., Chummun, N.H. and Pudaruth, S., 2015, June. Predicting the Price of Second-hand Cars using Artificial Neural Networks. In *The Second International Conference on Data Mining, Internet Computing, and Big Data (BigData2015)* (p. 17).

[25] Sun, N., Bai, H., Geng, Y. and Shi, H., 2017, June. Price evaluation model in second-hand car system based on BP neural network theory. In *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (pp. 431-436). IEEE.

[26] LISTIANI, M, "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application" Thesis (MSc), Hamburg University of Technology, 2009.

[27] Wu, J.D., Hsu, C.C. and Chen, H.C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications*, *36*(4), pp.7809-7817.

Table 1: A comparison among the existing vehicle price estimation systems

Research work	Authors	Employed ML/DL algorithm(s)	Dataset	Dataset size	# of features	Results
[9]	Lavanya et al. 2021	Linear regression, Random Forest, & Decision Tree regressions	Kelly Blue book	804	3	Random Forest RMSE: 5321.687 MAE: 1725.187
[10]	Chandak et al. 2019	K-Nearest Neighbour & Decision Tree	Private dataset	NA	NA	(KNN) RMSE: 5581.96 (Decision Tree) RMSE: 4961.64
[11]	Gegic et al. 2019	ANN, Support Vector Machine, & Random Forest	Web scraper	1,105	8	Best Accuracy: SVM 48.23% Minimum error: ANN 7.05%
[12]	Gajera et al. 2021	Linear regression, KNN, Random Forest, XG Boost, & Decision Tree	Private dataset 2	92,386	3	The best is the Random Forest with RMSE: 3702.34
[13]	Samruddhi et al. 2020	K-Nearest Neighbour	Kaggle.com	NA	13	Accuracy:85.00% RMSE: 4.73 MAE: 2.13
[14]	Mammadov 2021	Linear regression	fred.stlouisfed.org	205	25	RSME: 0.917
[15]	Asghar et al. 2021	Linear regression	fred.stlouisfed.org	205	25	RSME: 0.919
[16]	Karakoc et al. 2020	Artificial Neural Network	Car sales website	1,000	18	Success: 91.38%
[17]	Xia et al. 2020	ForeXGBoost	Alibaba Cloud car sales dataset	1,000,000	33	ForeXGBoost outperforms benchmark algorithms
[18]	Gegic et al. 2019	Linear regression, lasso regression, and ridge regression	Web scraper	1,105	8	Linear: 83.65% Lasso: 87.09% Ridge: 84.00%
[19]	Liu et al. 2022	PSO-GRA-BPNN	China used car trading platform https://www.iautos.cn	10,260	11	MAE: 0.475 R^2: 0.984
[20]	Hankar et al. 2022	Gradient Boosting	Avito	8,000	3	Gradient Boosting RMSE: 44516.20 R^2: 0.80
[21]	Monburinon et al. 2018	Gradient Boosting	German e-commerce website	371,528	13	MAE: 0.28
[22]	Pudaruth 2014	Multiple linear regression, Naïve bayes, k-nearest neighbour, and decision tree	Newspapers	400	3	MAE: 41,962
[23]	Noor & Jan 2017	Multiple linear regression	Pakwheels	2,000	3	Precision: 98%
[24]	Peerun et al. 2015	Support vector machine, Neural Network, Linear regression	E-commerce website	200	6	MAE: 30,605
[25]	Sun et al. 2017	Optimized BP neural network	E-commerce website	NA	NA	NA
[26]	Listiani 2009	Support Vector Regression	Germany	124,386	179	RMSE: 8.000
[27]	Wu et al. 2009	Adaptive Neuro-Fuzzy Inference System (ANFIS)	User car website in Taiwan	NA	3	NA