Machine learning modeling for reconditioned car selling price prediction

Fatema Abdullah[a], Md. Ataur Rahman[a], Mohammad Shidujaman*[,b], Mahady Hasan[b], Md. Tarek Habib[b]
[a]Department of Computer Science and Engineering, Daffodil International University; [b]Department of Computer Science and Engineering, Independent University, Bangladesh.

ABSTRACT

Almost 80% of the vehicles required for Bangladesh's road transportation industry are supplied by reconditioned cars. Using machine learning (ML) to predict car prices refers to using ML algorithms and techniques to make assumption about future car prices. This can be useful for a variety of purposes, such as helping car buyers and sellers make informed decisions, assisting car dealerships with inventory management, or providing insights for car manufacturers and other industry stakeholders. To predict car prices using ML, data is collected on a variety of factors that can affect the ongoing cost of a car, such as its make and model, age, mileage, condition, and location. This data is then fed into the Random Forest ML model, which uses statistical techniques to analyze the data and identify patterns and trends. The model performs 99.59% accurately in the tested portion of the data set and ensures that the model can then be used to make predictions on the future cost of an automobile based on these patterns and trends.

Keywords: Car Price, Accuracy, Prediction, Machine Learning, Random Forest Regression, Performance Metrics, R-Squared.

1. INTRODUCTION

A reconditioned car is a vehicle that has previously been owned by someone else and has been driven for a certain period. Reconditioned cars can be purchased from a variety of sources, such as dealerships, private sellers, or online marketplaces. A refurbished car is usually low-priced than a new car because the car has been driven and may have some wear and tear. A reconditioned car is a reconditioned car that has undergone repairs or renovations to bring it up to a higher quality standard. Reconditioning can include a variety of activities, such as fixing mechanical issues, cleaning and detailing the exterior and interior, and replacing worn or damaged parts. Reconditioned cars are typically more expensive than reconditioned cars that have not undergone any repairs, but they may also be in better condition and have a longer lifespan. Both used and reconditioned cars can be good options for people who want to save money on their vehicle purchases. However, it's important to carefully research and inspect a used or reconditioned car before buying it to ensure that it is in good condition and a good value. There are various machine learning (ML) algorithms and techniques for using and predicting car prices. The specific algorithm or technique used depends on the characteristics of the data and the purpose of the prediction. Overall, car price prediction using ML is a complex and multifaceted process that requires a deep understanding of ML algorithms and techniques, as well as knowledge of the specific characteristics of the car market.

*Mohammad Shidujaman:shidujaman@iub.edu.bd
Fatema Abdullah:fatema15-12146@diu.edu.bd
Md. Ataur Rahman:ataur15-13180@diu.edu.bd
Mahady Hasan:mahady@iub.edu.bd
Md. Tarek Habib: tarek.cse@iub.edu.bd

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik, Tao Shen, Proc. of SPIE Vol. 12779, 1277925 · © 2023 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2689745

Refurbished car prices can be difficult to predict for a multitude of reasons. For individuals who are looking to purchase or sell a reconditioned car, accurate price predictions can aid make informed decisions. Buyers can use price predictions to determine whether a particular car is a good value, while sellers can use them to set a realistic price for their car. For car dealerships, accurate price predictions can be useful in managing inventory and making informed purchasing decisions. By understanding the demand for different types of reconditioned cars, dealerships can stock the types of cars that are most likely to sell and avoid having excess inventory that is difficult to move. Accurate price predictions can provide valuable insights into trends in the reconditioned car market and help car manufacturers and other industry stakeholders understand the factors that are driving demand for reconditioned cars. This can inform business strategies and help companies to make more informed decisions. By providing more accurate predictions about reconditioned car prices, ML models can help to refine the efficiency of the reconditioned car market by reducing uncertainty and helping buyers and sellers to make more informed decisions. This can lead to more efficient resource allocation and overall market performance. Overall, estimating the price of refurbished vehicles can help individuals and businesses make informed decisions, provide insight into market trends, improve efficiency in the refurbished vehicle market, etc., which can be important for a variety of purposes.

As predicting the price of a quite used or reconditioned automobile is a quite troublesome fact, depending on various factors, prices can vary very often keeping pace with the ongoing economic situation. Using ML, although we cannot assure the predicted price would be the most accurate one or the fixed price it shall go on, we can assume the range of the car that is going to be expected to be sold. In this study, several machine learning (ML) models such as linear regression (LR), decision tree (DT) Regression, *k*-nearest neighbors (*k*-NN), least absolute shrinkage and selection operator lasso regression, random forest (RF) regression, and the extreme gradient boosting (XGBoost) algorithm have been tested and implemented to find out the best and most efficient one in terms of considering several performance metrics of these models.

To develop a machine learning (ML) model that properly forecasts the price of a reconditioned car in the future based on the car's attributes and other pertinent information, such as its make, model, age, mileage, condition, and location. This problem definition identifies the main goal of the project, which is to develop an ML model that can forecast the price of a reconditioned car. It also identifies the key inputs to the model, which are the characteristics and other relevant factors of the car, and specifies that the model should be able to forecast the future price of the car.

Other potential elements of a problem definition for this study might include the target audience or stakeholders for the model (e.g., car buyers and sellers, car dealerships, car manufacturers), the specific business or research question that the model is intended to address, and any constraints or limitations that the model will need to consider (e.g., the availability of data, computational resources, time constraints). This problem definition identifies the study's main goal, which is to develop an ML model that can evaluate the price of a reconditioned car. It also identifies the key inputs to the model, which are the characteristics and other relevant factors of the car, and specifies that the model should be enabled to distinguish the future price of the car.

One potential economic objective of a car price prediction model might be to improve the efficiency of the reconditioned car market by providing more accurate and reliable price predictions. This could help to reduce uncertainty for buyers and sellers and facilitate more informed decision-making, leading to more efficient resource allocation and overall market performance. Accurate price predictions could also help to reduce transaction costs by reducing the time and effort involved in negotiating prices and could help to reduce the risk of fraud or other unethical practices. Additionally, a car price prediction model could be used to help car dealerships and other industry stakeholders make more informed decisions about inventory management, pricing strategies, and other business practices. Overall, the economic objective of a car price prediction model could be to improve the efficiency and performance of the reconditioned car market and provide value to a range of stakeholders.

2. RELATED WORKS

Chandar et al. [1] sought to identify an effective soft computing method for stock prediction. For target prediction, the car's price is regarded as a dependent variable. A gradient-boosting regression model (GBRT) is used to forecast shared-car usage at the station level, and partial dependency plots (PDPs) are utilized to examine nonlinear connections between shared-car use and other variables. The resale value of the vehicle was predicted automatically in the work done by

Kiran et al. [2]. The only using algorithm is Linear Regression with an accuracy rate of 90%. use various machine learning techniques and algorithms to get a high accuracy rate and a low error percentage. Both Wang et al. [3] and Yadav et. al. [4] have investigated the cost of a reconditioned car using ML approaches by using the concept of object detection, such as car detection.

Shalini et al. [5] have used ML approaches are used to optimize prediction models, and two techniques are compared: one that will be implemented through ML techniques like LR, and one just through optimization algorithms like gradient descent and stochastic gradient descent. Making poor decisions can result in significant losses or possibly the closure of a corporation. A gradient-enhanced regression model (GBRT) was used to predict shared car use at the station level, and PDP was used to test for a non-linear association between shared car use and other predictors of the work done by Wang et al. [3].

By Narayana et al. [6] to provide a creative response to this problem the study focuses on the used automobile sales industry, one of the retail industries. Lu et al. [7] To predict the stock closing price of the following day, a CNN-BiLSTMAM method is suggested. Arefin et al. [8] developed a system to anticipate the price of used Tesla vehicles using machine learning (ML) is described. Samruddhi et al. [9] is yet another important piece of literature. Gajera et al. [12] create a statistical model that can predict the cost of a used automobile using machine learning methods. Listiani et al. [14] developed utilizing a Support Vector Machine (SVM), which has a higher level of accuracy than basic multivariate or multiple regression in predicting the price of a rented car. This is based on the fact that SVM works better with datasets that have more dimensions and are less prone to overfitting and underfitting. its weakness Research changes that with simple regression, more advanced SVM Regression is not shown Basic metrics like mean, variance, or value deviation. Kumar et al. [19] create a mathematical model that can be utilized to determine the worth of a second user's car based on data from prior buyers and a set of available choices. Satapathy et al. [20] combined data analysis with various machine learning techniques to build. Siva et al. [21] compared the support vector machine (SVM) classification algorithm's accuracy with the linear regression (LR) algorithm's accuracy in predicting car prices and demonstrated that the LR algorithm's accuracy performance parameters are superior to those of the SVM algorithm.

3. METHODOLOGY

The architecture of the car sales price prediction system is shown in Fig. 1. Here, the user must utilize the web application to respond to the questions. The data collected from a car-selling shop named Car View in Bangladesh is a collection of data on reconditioned car prices. By using a regression method on the prepared data, the output will be decided following the input. The output produced by the model will be an outcome. The results can be got through specific formats to be watched through the web application. We collected information on 2500 reconditioned cars, of which 30% of the dataset was used as testing data and 70% as training data. The next part will go through the data-gathering and preparation techniques we use. We utilized six classifiers that use machine learning, including *k*-NN, LR, DT regression, LAR, RFR, and XGBoost. We have repeatedly tested the performance of the classifiers. Performance is evaluated using data that has been processed using thirteen features, after which we apply feature selection strategies to the dataset. Accuracy is calculated in the processed dataset after features are selected using Random Forest and Principal Components Analysis (PCA) techniques. We estimated the classifiers based on accuracy and other measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R-Squared score (R^2).

The following diagram in Fig. 2 gives a description of these procedures.

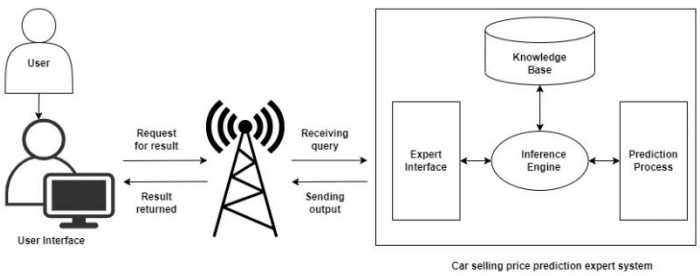

Fig. 1. Architecture of the car selling price prediction system.

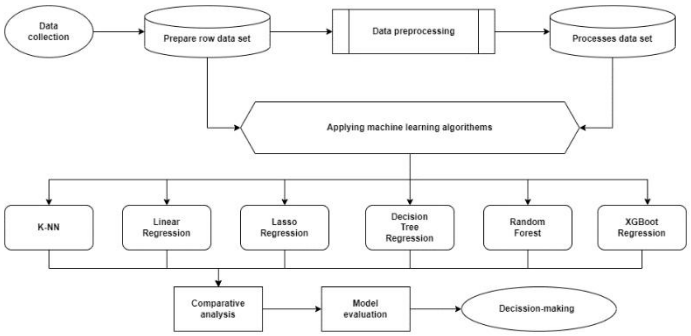

Fig. 2. Methodology applied for predicting the car selling price.

On the processed dataset, which had thirteen features, six machine learning techniques were applied. We not only calculated the accuracy of several algorithms but also calculated Mean Absolute Error (*MAE*), Root Mean Squared Error (*RMSE*), Mean Squared Error (*MSE*) and R-Squared Score (R^2). The specific classifier is quantified in the case of model evolution based on the test data set. The *MAE* is determined by dividing the total absolute error by the sample size. The *MSE* measures the root mean squared difference between the estimated and true values. The *RMSE* measures the variance between the sample values that an estimator or model predicts and the observed values. R^2 works by measuring the amount of variance in the predictions explained by the data set.

$$\text{Mean Absolute Error, MAE} = \frac{\Sigma_{i-1}^{n}|y_i - x_i|}{n} \tag{1}$$

$$\text{Mean Squared Error, MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

$$\text{Root Mean Squared Error, RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \tag{3}$$

$$\text{R2-Score, R2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

$$\text{Accuracy} = \frac{\Sigma_{i=1}^{w}|\hat{y}_o - \hat{y}_i|}{n} \tag{5}$$

3.1 Data set description

We have collected the data from the car-selling shop named Car View is a collection of data on reconditioned car prices and characteristics. Some properties of this data set are as follows:

- The number of observations: The data set includes more than 2500 observations, each representing a different reconditioned car.
- Variables: The dataset includes many different variables, including the car make and model, chassis number, registration year, car transmission, car mileage, car engine condition, and fuel. cars, doors, sizes, and prices of cars in Bangladesh.
- Data types: The variables in the data set are a mix of categorical (e.g., make and model, condition) and numerical (e.g., age, mileage, price) data.
- Missing values: There are a small number of missing values in the data set, which will need to be addressed before the data can be used in an ML model.
- Data sources: The data in the Car Price data set was collected from a variety of sources, including online classified websites, car dealerships, and private sellers.

Overall, the Car Price data set is a large and comprehensive collection of data on reconditioned car prices and characteristics that could be useful for developing an ML model for predicting reconditioned car prices.

3.2 Data set preprocessing

Web portals are used to collect the data [10, 16, 17]. The basis for predictions is historical data gathered from daily newspapers [11]. The 2005 Central Edition of the Kelly Blue Book was used to get the data [13, 18]. The data used was acquired from a school of information and computer science that has access to several databases [15].

We have collected the data from the car-selling shop named Car View. We collected 2500 data from the car-selling shop. All of the data that we collected included some text, some numeric, some noisy values, and some missing data.

From the dataset, we first check if there are any missing values, then we convert the text data into a numeric form using level encoding. Missing values were filled in using the imputer and median. The noisy values in the dataset are then evaluated and sorted using a box plot. Then the processed data set is generated by normalizing and applying algorithms to it. This procedure is shown below with the aid of Fig. 3.

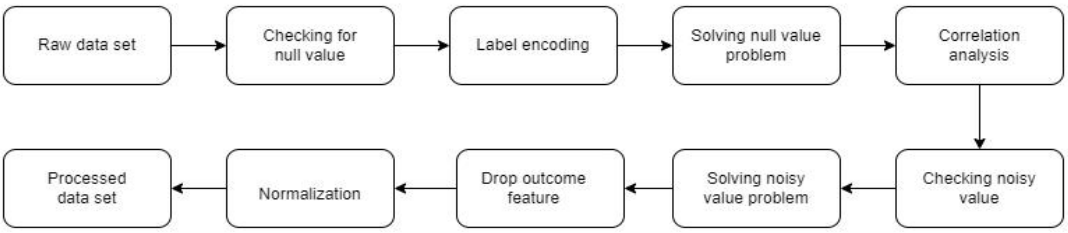

Fig. 3. Steps of data preprocessing of gathered data.

3.3 Applied regression algorithms

_k_-Nearest Neighbors

_k_-NN is a simple supervised machine learning algorithm. _k_-NN algorithms can be used to solve classification and regression problems. The _k_-NN algorithm checks that similar things exist in a near neighborhood [22]. Between query points and other points, Minkowski distances are calculated using (6).

$$\left(\sum_{i=1}^{k} (|x_i - y_i|)^q\right)^{\frac{1}{q}} \tag{6}$$

Linear Regression

Regression and classification are both possible using linear functions. As demonstrated, a linear classifier is produced by passing a linear function's output through a threshold function [22]. Linear regression formula

$$Y = a + bX \tag{7}$$

Lasso Regression

The feature selection and prediction method known as lasso regression helps keep parameter restrictions and decrease coefficients to zero to decrease the number of variables. Finding a subset of features in lasso regression that reduces the prediction error for a response variable is the goal [25]. The sum of the squared errors (SSE_{Lasso}) for the lasso regression is given by

$$\text{SSElasso} = \sum(y - \hat{y})^2 + \gamma\Sigma|\beta| \tag{8}$$

Decision Tree Regression

A decision tree represents a function that "decides" a single output value from a vector of input attribute values. Discrete or continuous values can be used for input and output [22]. DT is an ML algorithm used to model the relationship

between a dependent variable and multiple independent variables. For now, we will focus on situations where the inputs have discrete values and the result has exactly two potential values; This is a Boolean classification, where each input example will be classified as true or false [22] as shown in the decision tree diagram in Fig. 4.

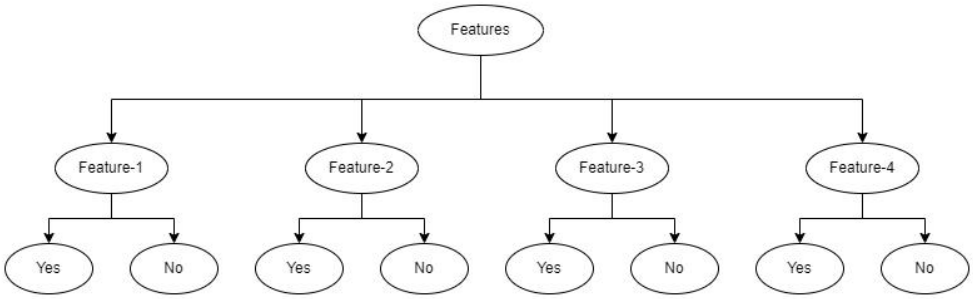

Fig. 4. Decision tree working principle.

Random Forest

For prediction purposes, RF builds a sizable collection of decoded trees. By providing randomness in tree development, it reduces the association between trees. The division variable is implemented randomly. [23]. In 1996, Freund and Schapire proposed the use of AdaBoost. It produces a classifier by combining several subpar classifiers. In each iteration, it sets the classifier weights and modifies the data. [23]. We can determine each dataset's error rate using (9).

$$\text{error}(\text{Mi}) = \sum_{j=1}^{d} \omega_j \times \text{err}(X_j) \tag{9}$$

XGBoost Regression

One of the best algorithms for supervised learning is XGBoost. which can be inferred as it flows, it includes basic target and learner functions. The loss function is included in the objective function and describes the difference between the actual and predicted values using the term regularization. Indicates how far the actual value is from the predicted value. While using ensemble learning in XGBoost to predict a single value, a variety of models known as base learners are considered [24].

$$\hat{f}(x) = \sum_{m=0}^{M} \hat{f}_m(x) \tag{10}$$

4. EXPERIMENTAL EVALUATION

k-NN, LR, lasso regression, DT regression, RF, and XGBoost regression are the six algorithms that we have implemented and tested on our collected data set, particularly for this study. A comparison of these models' performance metrics is shown in Table 1:

Table 1. Performance comparison of regression models

Model	MAE	MSE	RMSE	R^2	Accuracy
k-NN	121002.13	87991687487.76	45896.66	0.756	75.58%
Linear regression	154003.12	57781687487.95	24896.66	0.914	91.36%
Lasso regression	202004.16	64991687487.85	254934.67	0.915	91.47%
Decision tree	50292.61	44667877487.82	144934.67	0.994	98.02%
Random forest	2454371.23	4059363456.24	100234.46	0.996	99.59%
XGBoost	47786.06	24567575487.82	10024.46	0.995	98.85%

Comparing the results from Table 1, it is clear that the RF regression performs most efficiently in terms of error and testing accuracy. This algorithm shows promising enough performance to be selected compared to other models tested in this study. The predicted level against the actual levels of RF can be visualized as shown in Fig. 5.

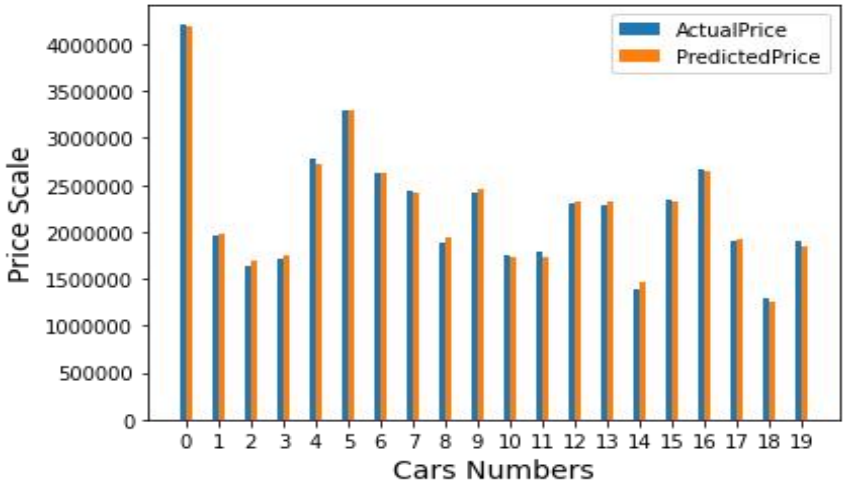

Fig. 5. Predicted versus actual levels of RF regression.

5. CONCLUSION AND FUTURE WORK

Predicting the price of reconditioned cars using ML can be a challenging task due to some limitations. One major limitation is the availability of data. The accuracy of ML models depends on the quality and quantity of the trained dataThe model's predictions might not be accurate if the data used to train it is incomplete, noisy, or biased. The impact of market conditions is another restriction. Many elements, like customer tastes, market demand, and prevailing economic situations, can impact the price of reconditioned cars. These factors can change over time, making it difficult to predict the price of a reconditioned car with high accuracy. Additionally, unforeseen events such as accidents, natural disasters, or changes in government regulations can impact the price of a reconditioned car in unpredictable ways. Finally, human factors such as the seller's motivation and the buyer's negotiating skills can also influence the price of a reconditioned car, which may be difficult to capture in an ML model. Despite these limitations, ML can still be a useful tool for predicting the price of reconditioned cars, but it is important to be aware of these limitations and to use caution when interpreting the model's predictions.

In the future, several directions could be taken to improve the accuracy and reliability of reconditioned car price prediction using ML. One possibility is to gather more data, including data on a wider range of makes and models, as well as data on market conditions, consumer preferences, and other factors that can impact the price of reconditioned cars. Another possibility is to develop new ML algorithms or modify existing algorithms in ways that improve their ability to handle the complexity and variability of reconditioned car pricing. In conclusion, predicting the price of reconditioned cars using ML can be a valuable tool for buyers, sellers, and other stakeholders in the reconditioned car market. While there are limitations to the accuracy and reliability of these predictions, advances in data gathering, ML algorithms, and other technologies have the potential to improve the performance of these models and make them more useful for a wider range of applications.

ACKNOWLEDGMENT

The authors of this paper acknowledge the car-selling shop Car View for providing the data set.

REFERENCES

[1] Kumar Chandar, S., Fusion model of wavelet transform and adaptive neuro fuzzy inference system for stock market prediction. Journal of Ambient Intelligence and Humanized Computing, 1-9 (2019).

[2] Kiran, S., Prediction of resale value of the car using linear regression algorithm. Int. J. Innov. Sci. Res. Technol, 6(7), 382-386 (2020).

[3] Wang, T., Hu, S., Jiang, Y., Predicting shared-car use and examining nonlinear effects using gradient boosting regression trees. International Journal of Sustainable Transportation, 15(12), 893-907 (2021).

[4] Yadav, A., Kumar, E., Yadav, P. K., Object detection and used car price predicting analysis system (UCPAS) using machine learning technique. Linguistics and Culture Review, 5(S2), 1131-1147 (2021).

[5] Shalini, L., Naveen, S., Ashwinkumar, U. M., Prediction of Automobile MPG using Optimization Techniques. In 2021 IEEE Madras Section Conference (MASCON), pp. 1-6. IEEE (2021).

[6] Narayana, C. V., Likhitha, C. L., Bademiya, S., Kusumanjali, K., Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business. In 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1680-1687. IEEE (2021).

[7] Lu, W. J., Li, J. Z., Wang, J. Y., Qin. L. L., A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 33, 4741-4753 (2021).

[8] Arefin, S. E., Second Hand Price Prediction for Tesla Vehicles. arXiv preprint arXiv:2101.03788 (2021).

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

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

[11] Sameerchand, P., Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764 (2014).

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

[13] Venkatasubbu, P., Ganesh, M., Used Cars Price Prediction using Supervised Learning Techniques. Int. J. Eng. Adv. Technol.(IJEAT) 9, no. 1S3 (2019).

[14] Listiani, M., Support vector regression analysis for price prediction in a car leasing application." Unpublished. https://www. ifis. uni-luebeck. de/~ moeller/publist-sts-pw-andm/source/papers/2009/list09. pdf (2009).

[15] Al-Turjman, F., Hussain, A. A., Alturjman, S., Altrjman, C., Vehicle Price Classification and Prediction Using Machine Learning in the IoT Smart Manufacturing Era. Sustainability, 14(15), 9147 (2022).

[16] Bukvić, L., Pašagić Škrinjar, J., Fratrović, T., Abramović, B., Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. Sustainability, 14(24), 17034 (2022).

[17] Benabbou, F., Sael, N., Herchy, I., Machine Learning for Used Cars Price Prediction: Moroccan Use Case. In Proceedings of the 5th International Conference on Big Data and Internet of Things, pp. 332-346. Cham: Springer International Publishing (2022).

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

[19] Kumar, S., Kaur, D., Parvez, A., Prediction of Prices Car Price Prediction with Machne Learning. In 2022 International Conference on Cyber Resilience (ICCR), pp. 1-4. IEEE (2022).

[20] Satapathy, S. K., Vala, R., Virpariya, S., An Automated Car Price Prediction System Using Effective Machine Learning Techniques. In 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 402-408. IEEE (2022).

[21] Siva, R., Adimoolam, M., Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm. ECS Transactions, 107(1), 12953 (2022).

[22] Russell, S. J., Artificial intelligence a modern approach. Pearson Education, Inc., (2010).

[23] Han, J., Kamber, M., Pei, J., Data mining concept and technique, 3rd Edition, Morgan Kaufmann, pp. 332–398 (2012)

[24] Avanijaa, J., Prediction of house price using xgboost regression algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 2151-2155 (2021).

[25] Shrivastava, S., Jeyanthi, P. M., Singh, S., Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. Cogent Economics & Finance, 8(1), 1729569 (2020).