# A Comparative Analysis of XGBoost Model and AdaBoost Regressor for Prediction of Used Car Price

S. Naveen Kumar Reddy* and S. Magesh Kumar†
*Department of Computer Science Engineering Saveetha School of Engineering,*
*Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, 602105, India*

Keywords: AdaBoost Regressor, Cars, Forecast, Machine Learning, Novel XGBoost Regressor, Vehicles.

Abstract: This study primarily aimed to estimate used car prices using the XGBoost Regressor to enhance accuracy in relation to vehicle data and benchmark this against the AdaBoost Regressor. The proposed methodology assessed the XGBoost and AdaBoost Regressor algorithms, represented by 10 samples split into two sets. This assessment employed an 80% Gpower and a 95% confidence interval. It emerged that the Novel XGBoost Regressor achieved an accuracy of 87.74%, surpassing the AdaBoost Regressor's 84.31%. Furthermore, SPSS statistical analysis confirmed the significance of both algorithms with p=0.000 (p<0.05). Ultimately, the results highlight the superior accuracy of the Novel XGBoost Regressor over the AdaBoost Regressor in predicting used car prices.

## 1 INTRODUCTION

The primary aim of this research is to develop machine learning models that accurately predict used car prices based on specific parameters or features. The dataset utilised comprises the selling prices of various car models from different locations (Yadav, Kumar, and Yadav 2021; Ramkumar G et al. 2021). As the number of private vehicles rises and the used car market evolves, used cars are becoming increasingly vital for buyers. For both parties, understanding the price of a used car is essential for a smooth transaction. Knowing the cost of used vehicles empowers buyers to negotiate confidently, while determining the residual value aids sellers in setting reasonable prices (Liu et al. 2022). The necessity arises as personal vehicles are integral for commuting between homes and workplaces. The decision to purchase a new car can be daunting, but deciding how to price a current car for sale can be even more challenging. The high cost of new vehicles often restricts consumer buying power. Various methods exist to estimate a car's price based on its market value (Asghar et al. 2021; Palanivelu et al. 2022). Machine learning could offer a solution. Here, regression algorithms predict used car selling prices using the Python module, Scikit-Learn, relying on historical car data (Vickram et al. 2016; C. Chen, Lulu Hao, and Congfu Xu 2017). Regarding used car price prediction, numerous studies have been published over the past five years. Springer Link released about 3822 articles, Google Scholar had 1690 related papers, ResearchGate published 40 pieces, and IEEE Xplore featured 31 articles. A notable study by Narayana et al. (2021) introduced a fair pricing system, forecasting vehicle prices based on features like model, age, fuel type, and seller type. AoQiang Wang et al. (2022) categorised the used car transaction cycle using six machine learning models and subsequently identified crucial factors influencing used car transactions. Gültekin and Organ (2020) employed a method known as ANN to predict vehicle prices in the used car market. Other studies have compared the efficacy of regression using supervised machine learning models and developed mathematical models for price prediction based on car attributes (Nitis Monburinon et al. 2018; Santosh Kumar Satapathy, Rutvikraj Vala, and Shiv Virpariya 2022). Estimating a used car's value is complex due to various influencing factors, such as its condition, mileage, year, and make.

A research gap has been identified, highlighting a lack of accuracy in current research regarding used car price predictions. This study aims to enhance

* Research Scholar
† Research Guide, Corresponding Author



Reddy, S. and Kumar, S.
A Comparative Analysis of XGBoost Model and AdaBoost Regressor for Prediction of Used Car Price.
DOI: 10.5220/0012510700003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 441-446
ISBN: 978-989-758-661-3

accuracy by deploying the Novel XGBoost Regressor, contrasting its performance with the AdaBoost Regressor.

# 2 MATERIALS AND METHODS

This research was conducted in the Lab of Artificial Intelligence at the Saveetha School of Engineering, part of the Saveetha Institute of Medical and Technical Sciences in Chennai. The study comprised two groups, each with a sample size of 10 sets, making a total of 20 sample sets considered for this research (Hankar Mustapha et al. 2022). Sample estimation was carried out with an alpha value of 0.05, a statistical power of 80%, and a 95% confidence interval. The sample size was determined using the ClinCalc software under supervised learning conditions. The tools utilised for this study's statistical analysis were version 26.0.1 of the SPSS software and the Jupyter Notebook.

Vehicle data was sourced from the well-known platform Kaggle.com, specifically from a dataset made available in Vijayaadithyan's 2022 Kaggle repository. This dataset includes 18 attributes, such as id, region, manufacturer, model, condition, odometer, and so forth. After the initial data collection, all missing and null values found within the vehicle dataset were eliminated through data cleaning and preprocessing steps. Ultimately, the algorithms were compared to determine which one was more effective.

## 2.1 Novel XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is a renowned machine learning algorithm, an enhanced and optimised version of the gradient boosting method, crafted for efficiency, speed, and model quality. Owned by the Distributed Machine Learning Community, this open-source library aims to refine and expedite established boosting techniques. It harmoniously integrates hardware and software capabilities. In our study, we use classification boosting with XGBoost. The boosting algorithm successively generates weak learner models and amalgamates their predictions to augment overall model performance. When an erroneous prediction occurs, misclassified samples are attributed higher weights, while correctly classified ones receive lower weights. The resulting ensemble model assigns greater weights to better-performing weak learner models. Boosting doesn’t alter previous predictors; instead, it adjusts subsequent ones by learning from past errors. Given boosting's greedy nature, it's prudent to set stopping criteria, such as model performance (early stopping) or specific iterations, like tree depth for tree-based learners, to evade overfitting the training data.

## 2.2 Procedure for the Novel XGBoost Regressor

1. Initialise required modules and libraries.
2. Import the target dataset.
3. Undertake preprocessing tasks, encompassing data cleaning and transformation.
4. Execute feature engineering to refine the dataset and follow up with exploratory data analysis (EDA) to gain insights.
5. Partition the dataset into training and testing subsets.
6. Validate the divided data to ensure its quality and distribution.
7. Deploy the Novel XGBoost Regressor to train on the training set and subsequently test it.
8. Evaluate and display the model's accuracy.

## 2.3 AdaBoost Regressor

Adaptive Boosting, commonly known as AdaBoost, is a renowned ensemble learning method in the realm of machine learning. Primarily effective for binary classification tasks, AdaBoost can also be adapted for multi-class classification or regression challenges. Typically, AdaBoost employs a one-level decision tree, often termed a "decision stump". The name "stump" stems from its simplistic nature, signifying a tree cut down to just a trunk.

The foundational concept of AdaBoost is "boosting". Its mechanism is delineated as follows:

a. Develop the inaugural predictive model using the original dataset.
b. Construct a subsequent predictive model that rectifies the shortcomings of the initial model.
c. Generate a third model addressing the imperfections of the second.
d. Continue creating models until either the set maximum number of iterations (or trees) is met or the model achieves absolute accuracy.

**Procedure For Adaboost Regressor**

1. Load the essential libraries.
2. Load the dataset.
3. Identify the X (features) and Y (target) variables.
4. Conduct preprocessing on the dataset, encompassing data cleaning and data transformation.

5. Split the dataset into training and testing sets.
6. Implement the AdaBoost Regressor model and fit the data.
7. Evaluate the model's performance.
8. Display the accuracy.

## 3 STATISTICAL ANALYSIS

The SPSS software was used for the descriptive statistical analysis of both the Novel XGBoost and AdaBoost Regressor models. In this dataset, the year and odometer are considered as independent variables, while lat, long, and price are treated as dependent variables. An independent samples t-test was employed, as cited by Cui et al. (2022).

## 4 RESULTS

The performance of the Novel XGBoost Regressor appears to be notably superior to that of the AdaBoost Regressor. As depicted in Table 1, the accuracy of the Novel XGBoost Regressor stands at 87.74%, whereas the AdaBoost Regressor achieved an accuracy of 84.31%.

Table 2 provides the group statistics of both the Novel XGBoost Regressor and the AdaBoost Regressor, highlighting a statistically significant difference between the two.

Table 1: Accuracy of Novel XGBoost Regressor and AdaBoost Regressor.

| S.NO | Novel XGBoost Regressor | AdaBoost Regressor |
|---|---|---|
| 1 | 89.45 | 86.40 |
| 2 | 86.37 | 85.32 |
| 3 | 87.50 | 83.66 |
| 4 | 88.10 | 84.20 |
| 5 | 89.00 | 81.58 |
| 6 | 87.00 | 83.36 |
| 7 | 85.53 | 84.53 |
| 8 | 86.70 | 82.15 |
| 9 | 89.20 | 85.71 |
| 10 | 88.60 | 86.22 |

The accuracy is utilised to gauge the overall performance in predicting used car prices.

In Figure 1, the Standard Deviation and Mean Accuracy for both algorithms are displayed: for the Novel XGBoost Regressor, they are 1.33167 and 87.74% respectively, and for the AdaBoost Regressor, 1.64992 and 84.31%. The figure contrasts the two algorithms (Novel XGBoost Regressor vs AdaBoost Regressor) on the X-Axis, while the Y-Axis showcases the Detection of Mean Accuracy, considering a range of +/- 2SD.

Table 2: Group Statistics of XGBoost Regressor (Mean Accuracy of 87.74%) AdaBoost Forest Regressor (Mean Accuracy of 84.31%).

| Algorithm | N | Mean Accuracy | Std.Deviation | Std. Error Mean |
|---|---|---|---|---|
| Novel XGBoost Regressor | 10 | 87.74 | 1.33167 | .42111 |
| AdaBoost Regressor | 10 | 84.31 | 1.64992 | .52175 |

Table 3: Independent Sample Test T-test is applied for the data set fixing the confidence interval as 95%. It shows that there is a statistically significant difference between XGBoost Regressor and AdaBoost Regressor with a two-tailed value p=0.000 (p<0.05).

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig | t | df | Sig(2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Accuracy | Equal variances assumed | .343 | .566 | 5.119 | 18 | .000 | 3.432 | .67049 | 2.02335 | 4.84065 |
| | Equal variances not assumed | | | 5.119 | 17.232 | .000 | 3.432 | .67049 | 2.01884 | 4.84516 |

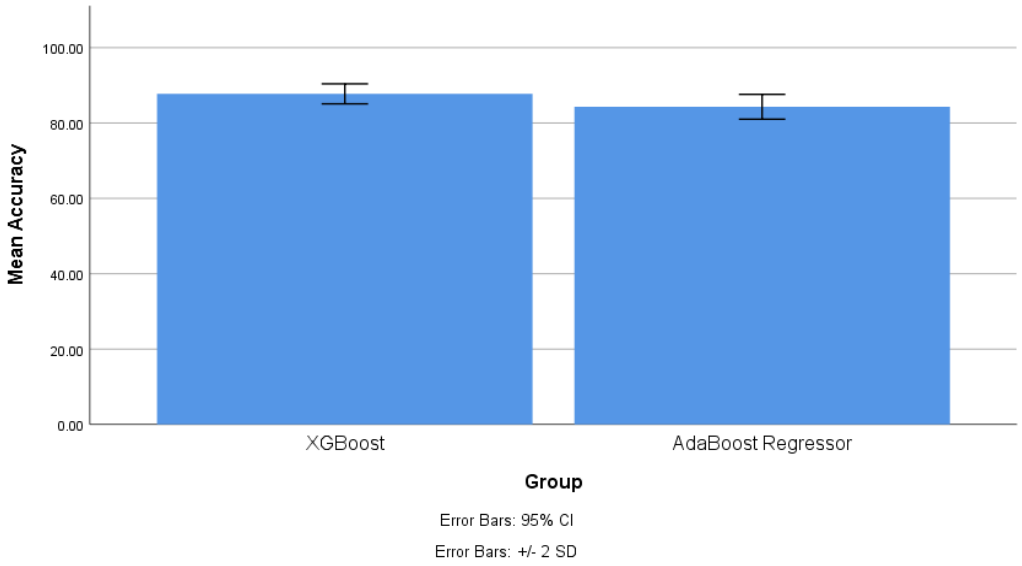


Figure 1: Comparison of Novel XGBoost Regressor and AdaBoost Regressor using a graphical representation. And the Mean Accuracy and Standard Deviation for Novel XGBoost Regressor (87.74% and 1.33167) and AdaBoost Regressor (84.31% and 1.64992). X-Axis: Novel XGBoost Regressor vs AdaBoost Regressor algorithms and Y-Axis: Detection of Mean accuracy +/- 2SD.

Lastly, Table 3 illustrates the categories associated with the Independent samples test, the test for equality of means, standard error differences, and the test for equality of variances. This table encompasses the Mean of Standard Error, Group statistics for Mean and Standard Deviation, each with a sample size of 10. The findings underscore that both the proposed method (XGBoost Regressor) and the comparative method (AdaBoost Regressor) have statistical significance, with a p-value of 0.000 ($p<0.05$), based on the statistical analysis conducted via SPSS.

## 5 DISCUSSIONS

The aforementioned study showcases the superior predictive accuracy of the Novel XGBoost Regressor in determining used car prices, standing at 87.74%, when compared to the AdaBoost Regressor, which demonstrates an accuracy of 84.31%. Enhanced performance was observed with the algorithm when exposed to a larger volume of training data. An Independent samples T-test analysis highlighted a p-value of 0.000 ($p<0.05$), signifying the statistical significance of both the proposed and comparative algorithms.

In alignment with these findings, several research publications have reinforced the outcomes of this study. One research piece by K. Samruddhi and R. A. Kumar in 2020 employed various training and testing ratios to examine the data, ultimately recommending an optimal model boasting an accuracy close to 85%. Pal et al. (2019) proposed the use of the Random Forest algorithm, attaining an accuracy of 83.63%. Experimental results from a study by Janke, K, and C in 2022 illustrated the superior performance of the Random Forest model relative to other models, evidenced by its R^2 score of 0.772584 (translating to an accuracy of 77.2%) and a Mean Absolute Error value of 1.0970472. Bharambe et al. (2022) implemented lasso regression, introducing a model with a commendable accuracy of 87%. Notably, no research articles have been found that contest the proposed algorithm's efficacy for predicting used car prices in the context of this study.

However, the study is not without its limitations. Recent pandemic-induced semiconductor shortages, which inadvertently augmented the price of used cars, are pinpointed as one of the research's primary constraints. Owing to the sudden shift in car pricing amidst the study's tenure, the current dataset may not aptly represent market-available vehicles, making it arduous to predict future car costs with precision. In the trajectory of future research, the incorporation of advanced machine learning techniques can be contemplated to elevate the model's accuracy and optimization levels. A real-time data model, seamlessly integrable within a mobile application and boasting widespread usage, might emerge as an optimal solution.

# 6 CONCLUSION

In the ongoing quest for developing optimal machine learning algorithms for predicting used car prices, the Novel XGBoost Regressor has emerged as a notably effective solution. When compared against the AdaBoost Regressor, the Novel XGBoost Regressor outperforms with a commendable accuracy rate of 87.74%. In contrast, the AdaBoost Regressor lags slightly behind with an accuracy of 84.31%.

Several aspects from the study underline this superiority:

- Statistical Significance: A rigorous T-test analysis for independent samples was conducted, leading to a p-value of 0.000 ($p<0.05$). This result alone provides a compelling argument for the statistical superiority of the Novel XGBoost Regressor over the AdaBoost Regressor. The low p-value highlights that the difference in their performances is significant and not just a random occurrence.
- Supporting Literature: Multiple research endeavors in the domain corroborate the findings. For instance, a study by K. Samruddhi and R. A. Kumar in 2020 reached a model accuracy close to 85%. Another investigation using the Random Forest algorithm, as suggested by Pal et al. (2019), secured an accuracy of 83.63%. These findings align with the current study's observations that advanced machine learning techniques, like the Novel XGBoost Regressor, tend to outdo traditional methods in predicting used car prices.
- Mean Accuracy and Standard Deviation: The analysis further delineated the Standard Deviation and Mean Accuracy of both models. While the AdaBoost Regressor recorded a mean accuracy of 84.31% with a standard deviation of 1.64992, the Novel XGBoost Regressor surpassed it with a mean accuracy of 87.74% and a slightly lower standard deviation of 1.33167. This not only highlights the higher accuracy of the Novel XGBoost but also suggests its consistent performance across different datasets.
- Limitations & Future Scope: Despite its impressive accuracy, the study did confront challenges linked to recent economic phenomena, such as the semiconductor shortage triggered by the pandemic, which affected used car prices. The dataset's potential inability to represent real-time market prices underscores the necessity for continuous algorithm updating and training.

In conclusion, while both models hold merit in their respective capacities, the Novel XGBoost Regressor exhibits a clear edge in the context of predicting used car prices. Its amalgamation of speed, efficiency, and accuracy makes it a potent tool for researchers and industry professionals, offering a blueprint for future advancements in the realm of machine learning-powered price predictions.

# REFERENCES


Santosh Kumar Satapathy, Rutvikraj Vala, and Shiv Virpariya. 2022. “An Automated Car Price Prediction System Using Effective Machine Learning Techniques.” https://doi.org/10.1109/CISES54857.2022.9844350.

Asghar, Muhammad, Khalid Mehmood, Samina Yasin, and Zimal Mehboob Khan. 2021. “Used Cars Price Prediction Using Machine Learning with Optimal Features.” Pakistan Journal of Engineering and Technology 4 (2): 113–19.

Bharambe, Prof Pallavi, Bhargav Bagul, Shreyas Dandekar, and Prerna Ingle. 2022. “Used Car Price Prediction Using Different Machine Learning Algorithms.” International Journal for Research in Applied Science and Engineering Technology 10 (4): 773–78.

Cui, Baoyang, Zhonglin Ye, Haixing Zhao, Zhuome Renqing, Lei Meng, and Yanlin Yang. 2022. “Used Car Price Prediction Based on the Iterative Framework of XGBoost+LightGBM.” Electronics 11 (18): 2932.

Gültekin, Sait, and Arzu Organ. 2020. “Price Estimation of Secondhand Cars Sold on the Internet with Artificial Neural Network Method.” Journal of Internet Applications and Management 11 (1): 49–61.

Janke, Varshitha, Jahnavi K, and Lakshmi C. 2022. “Prediction of Used Car Prices Using Artificial Neural Networks and Machine Learning.” In 2022 International Conference on Computer Communication and Informatics (ICCCI). IEEE. https://doi.org/10.1109/iccci54379.2022.9740817.

Liu, Enci, Jie Li, Anni Zheng, Haoran Liu, and Tao Jiang. 2022. “Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network.” Sustainability: Science Practice and Policy 14 (15): 8993.

Kishore Kumar, M. Aeri, A. Grover, J. Agarwal, P. Kumar, and T. Raghu, “Secured supply chain management system for fisheries through IoT,” Meas. Sensors, 10.1109/icesc51422.2021.9532845.

Palanivelu, J., Thanigaivel, S., Vickram, S., Dey, N., Mihaylova, D., & Desseva, I. (2022). Probiotics in functional foods: survival assessment and approaches for improved viability. Applied Sciences, 12(1), 455.

Pal, Nabarun, Priya Arora, Puneet Kohli, Dhanasekar Sundararaman, and Sai Sumanth Palakurthy. 2019.

"How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest." Advances in Information and Communication Networks, 413–22.

Nitis Monburinon, Prajak Chertchom, T. Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. 2018. "Prediction of Prices for Used Car by Using Regression Models.". https://doi.org/10.1109/ICBIR.2018.8391177.

AoQiang Wang, Qiancheng Yu, Xiaoning Li, Zekun Lu, XuLong Yu, and ZhiCi Wang. 2022. "Research on Used Car Valuation Problem Based on Machine Learning." https://doi.org/10.1109/ICCNEA57056.2022.00032.

Ramkumar, G. et al. (2021). "An Unconventional Approach for Analyzing the Mechanical Properties of Natural Fiber Composite Article ID 5450935, 15 pages, 2021. https://doi.org/10.1155/2021/5450935

Vijayaadithyan, V. G. 2022. "Car Price Prediction(used Cars)." https://www.kaggle.com/vijayaadithyanvg/car-price-predictionused-cars.

V. P. Parandhaman, "An Automated Efficient and Robust Scheme in Payment Protocol Using the Internet of Things," 2023s, pp. 1-5, doi: 10.1109/ICONSTEM56934.2023.10142797.

Mustapha Hankar, Marouane Birjali, and Abderrahim Beni-Hssane. 2022. "Used Car Price Prediction Using Machine Learning: A Case Study." . http://dx.doi.org/10.1109/ISIVC54825.2022.9800719.

Vickram, A. S., Kamini, A. R., Das, R., Pathy, M. R., Parameswari, R., Archana, K., & Sridharan, T. B. (2016). Validation of artificial neural network models for predicting biochemical markers associated with male infertility.

Chuancan Chen, Lulu Hao, and Cong Xu. 2017. "Comparative Analysis of Used Car Price Evaluation Models." . www.doi.org.10.1063/1.4982530.

K.Samruddhi, and Dr R.Ashok Kumar. 2020. "Used Car Price Prediction using K-Nearest Neighbor Based Model". www.doi.org.10.29027/IJIRASE.v4.i2.2020.629-632.

Yadav, Anu, Ela Kumar, and Piyush Kumar Yadav. 2021. "Object Detection and Used Car Price Predicting Analysis System (UCPAS) Using Machine Learning Technique." Linguistics and Culture Review