iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68411

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Estimating Automobile Prices Utilizing Machine Learning Algorithms

Ankit Pandey[1], Ashish Kumar[2], Akhil Pratap Singh[3], Barkha Nandwana[4], Sanjay Kumar Sonker[5]
Dept. of Computer Science and Applications, Sharda University, Greater Noida, India

Abstract: ***This study delves further into machine learning algorithms for estimating automotive prices, using Python-based frameworks. The study includes critical processes such as data collection, preprocessing, feature selection, model evaluation, and implementation. Our technique attempts to examine the efficacy of different machine learning models in predicting automobile prices by applying multiple regression-based algorithms such as Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression, and K-Nearest Neighbors. The paper also tackles critical process issues such as data quality, feature selection, and model interpretability. The findings contribute to a data-driven methodology that can help buyers, sellers, and automotive experts make better decisions by giving accurate real-time price projections.***
Keywords: Machine Learning, Car Price Prediction, Regression Models, Data Preprocessing, Feature Engineering.

I. INTRODUCTION

Machine Learning (ML), a subset of Artificial Intelligence (AI), allows models to process data, recognize patterns, and make decisions independently of human input. Unlike conventional software systems that need clear instructions, ML algorithms enhance their predictive capability by learning from large volumes of historical data. Depending on the type of issue being addressed, ML methodologies can be split into three categories: Reinforcement Learning, Unsupervised Learning, and Supervised Learning, with the latter further divided into Regression and Classification tasks. This research emphasizes Regression-based ML models for forecasting automobile prices, a challenging but vital task for a range of stakeholders, including manufacturers, dealerships, and consumers.

The second-hand automobile market plays a crucial role in the global automotive industry, often influencing new car sales. Automakers engage in the used car market to manage dealership stock, rental returns, and fleet exchanges. However, several factors, including economic instability, growing competition, and the rise of electric vehicles, present challenges in sustaining the profitability of used car sales. An effective decision support system incorporating predictive analytics can help manufacturers and sellers estimate used car values more accurately. In this study, a machine learning model is developed to predict the resale price of used cars based on automotive parameters such as mileage, manufacturing year, engine size, transmission type, and power. Various regression models—including Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression, and K-Nearest Neighbors—are compared to determine the most accurate approach for price estimation.

Over the last ten years, the market value of the used automobile business has nearly doubled due to its exponential expansion. The demand for used cars has increased as a result of the high price of new cars and the fact that many people cannot afford to buy brand-new cars. It is now simpler for buyers and sellers to determine the worth of cars thanks to online markets like CarDekho, Quikr, CarWale, and Cars24. Pricing inconsistencies result from the substantial platform-to-platform variations in current valuation systems. By using cutting-edge machine learning models, this study seeks to close this gap and produce more accurate and consistent car price forecasts.

The study's dataset was obtained from Kaggle and goes through a rigorous preparation process to remove noise, duplicates, and unnecessary information. Data collection, preprocessing, feature selection, model training, assessment, and performance comparison are all steps in the organized pipeline for model creation. Finding the ML regression model that minimizes mistakes and produces the best accurate automobile price forecasts is the main objective. The suggested model helps manufacturers analyse demand trends for various car models, which helps with production planning, in addition to improving decision-making for buyers and sellers. This study optimizes efficiency and accuracy in used car valuation by utilizing machine learning (ML) for automobile price calculation, thereby contributing to the continuous digital transformation in the automotive sector.

II. RELATED WORK

To enhance accuracy, several studies have employed machine learning methods to predict car prices, focusing on various regression models, feature selection, and data preprocessing techniques.

Using a dataset from Turkey, Sumeyra MU and Yildiz K. investigated how well linear regression predicted used automobile prices (2020). Their model showed the promise of regression-based methods with an R2 score of 73%. Similarly, Chitra AR and Arjun BC employed five different machine learning regression techniques in KNIME, assessing their performance using the R^2 measure.

Shaprapawad S, Borugadda P, and Koshika N developed a feature selection method that took into account variables such as mileage, year of manufacture, and fuel type to forecast car prices. Their research found Support Vector Regression (SVR) to be the most efficient model, reaching an accuracy of 95.27%. In another study, Monburinon N, Chertchom P, and their team performed a comparative analysis using Random Forest Regression, Gradient Boosting Regression Trees, and Multiple Linear Regression, highlighting the significance of selecting an appropriate model for price prediction.

Additionally, deep learning techniques have been investigated. Gonggie created a model based on an Artificial Neural Network (ANN) that included features such as estimated lifespan and brand to enhance prediction accuracy. Deepak and colleagues presented a neuro-fuzzy system integrated with K-Nearest Neighbors (KNN) regression to refine auction pricing, demonstrating the effectiveness of hybrid models.

Numerous research studies underscore the importance of data preprocessing in enhancing predictive model performance. Pudaruth (2014) analysed Multiple Linear Regression, Naïve Bayes, Decision Trees, and KNN, pinpointing preprocessing methods that improve model accuracy. Pal et al. (2019) pointed out that Random Forest Regression is a superior approach for forecasting car prices, while Shanti et al. (2021) assessed Random Forest, Neural Networks, Gradient Boosting, and SVR, further confirming the importance of advanced regression techniques.

Venkatasubbu and Ganesh (2019) utilized Lasso Regression, Regression Trees, and Multiple Regression to predict car prices, stressing the significance of feature selection in enhancing prediction outcomes. Amik et al. (2021) implemented machine learning models for predicting used car prices in Bangladesh, determining that XGBoost achieved the highest level of accuracy. In Dubai, AlShared (2021) discovered that Random Forest Regression surpassed other methods, reaching an accuracy rate of 95%. Arefin (2021) investigated price prediction for Tesla vehicles, employing SVM, Random Forest, and deep learning models to enhance valuation. Salim and Abu (2020) created a regression model based on an S-curve to forecast peak car prices, highlighting the impact of feature selection on pricing trends. Sun et al. (2017) developed a price evaluation system utilizing a BP Neural Network, enhancing data preprocessing for the valuation of second-hand cars. Monburinon et al. (2018) conducted a comparative analysis of regression models, employing Mean Absolute Error (MAE) as a performance evaluation metric.

These explorations underscore the diverse methodologies applied in car price prediction studies, each offering significant insights while also revealing areas for improvement. A recurring theme in these studies is the dependence on a singular machine learning algorithm, which, although effective in some situations, indicates the potential for boosting prediction accuracy by employing a combination of different machine learning techniques in an ensemble framework. This shared understanding paves the way for our research, where we intend to examine the combined effects of integrating multiple machine learning methods to enhance the accuracy and dependability of predictions for used car prices.

III. METHODOLOGY

The approach used in this research to forecast car prices incorporates a diverse strategy, as demonstrated in the conceptual framework (see Fig. 1).

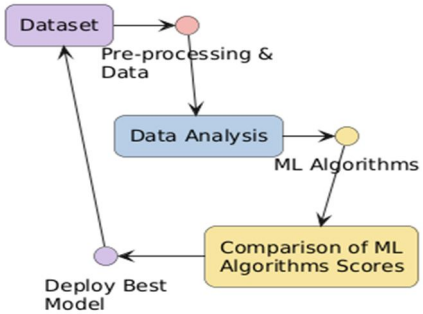


Fig: 1 Conceptual Framework for the Car Price Prediction Process

To develop a strong machine learning model for predicting car prices, data was gathered from various sources, reflecting different market conditions and pricing patterns. This dataset features information from sahibinden.com, a prominent online marketplace, providing a thorough representation of actual automobile pricing.

Combining multiple data sources improves the model's accuracy and dependability in forecasting the prices of second-hand cars. The dataset encompasses crucial attributes necessary for estimating prices, divided into the following categories:

- Basic Vehicle Information: Brand, Model, Car Condition, Year of Manufacturing
- Technical Specifications: Engine Power (kW), Transmission Type, Fuel Type
- Usage Metrics: Mileage, Registration Year, Distance Driven
- Aesthetic and Functional Features: Color, Number of Doors, Interior & Exterior Specifications
- Advanced Features (Boolean Attributes): Navigation System, Leather Seats, Alarm System, Parking Sensors, Heated Seats, Panorama Roof, Cruise Control, ABS, ESP, and more
- Target Variable: Price (expressed in Indian Rupees, INR, for consistency in the regional market analysis)

The gathered data underwent thorough preprocessing steps, including addressing missing values, selecting features, and encoding categorical variables to improve the effectiveness of regression models. By utilizing machine learning methods based on Python, the research aims to enhance the accuracy of models predicting used car prices, making them more effective and reliable for both consumers and businesses.

To effectively manage the significant amount of data, web scraping methods were implemented for data extraction. These automated tools optimized the data collection process by simulating human behaviour, allowing for direct acquisition of structured information from internet sources. This strategy not only sped up the process but also improved the precision, uniformity, and dependability of the gathered data, guaranteeing a well-organized dataset for subsequent analysis.

Following the data acquisition phase, an extensive data preprocessing stage was conducted. To minimize redundancy and improve model efficiency, attributes such as “state” and “city” were discarded due to their sparse nature and limited relevance. Furthermore, the “damaged” attribute was omitted due to inconsistent reporting between the two platforms, which compromised data integrity. The final refined dataset comprised 684 records, offering a more streamlined and significant basis for subsequent analysis and model training.

Brand	Model	Year	Power	Mileage	Fuel Type	Transmission	Number of Doors	Four Wheel Drive	Navigation	Leather Seats	Parking Sensors	Price (INR)
Toyota	Corolla	2018	132	32,000	Petrol	Automatic	4	True	False	True	True	₹20,80,000
Honda	Civic	2016	158	45,000	Diesel	Manual	4	False	True	False	False	₹23,14,000
Ford	Focus	2017	123	37,000	Hybrid	Automatic	4	True	False	True	False	₹22,10,000
BMW	3 Series	2015	181	29,000	Diesel	Automatic	4	False	True	True	True	₹41,34,000
Audi	A4	2019	188	21,000	Petrol	Automatic	4	True	False	True	True	₹42,90,000
Mercedes Benz	C-Class	2014	173	55,000	Petrol	Manual	4	False	True	False	True	₹37,70,000

Table 1. Sample of the Processed Dataset

A. Automated Data Preprocessing and Model Development

In order to enhance the preprocessing stage, a Python script was created to effectively clean and organize the raw dataset. This script automated the elimination of incomplete records and formatted the remaining data into CSV files, which is compatible with MATLAB, a popular tool for developing machine learning models. This preprocessing phase was essential for maintaining data integrity and preparing it for predictive modeling. In this research, we examined the efficacy of employing a single machine learning classifier approach, akin to earlier studies. However, our approach varied by assessing various classifiers and modifying the train-test split to improve model validation. The dataset was partitioned into 70% for training and 30% for testing. For building the model, we used Random Forest (RF) and Support Vector Machine (SVM), focusing primarily on optimizing Random Forest to attain better prediction accuracy.

B. Random Forest for Price Prediction

Random Forest (RF) is an ensemble learning method that performs exceptionally well in both classification and regression tasks. Created by Ho, RF effectively reduces the overfitting problem often seen with decision trees. It works by producing numerous decision trees during the training phase and combining their results—through majority voting for classification or averaging for regression—to improve accuracy in predictions.

A significant advantage of Random Forest is its capability to process large, high-dimensional datasets without the need for intensive feature selection. It can manage thousands of input variables, estimate missing data, and sustain dependable performance even when some data is absent. This characteristic makes RF particularly suitable for forecasting used car prices, where various interconnected factors impact market valuation.

From (INR)	To (INR)	Price Class
100,000	500,000	Class 1
500,001	1,000,000	Class 2
1,000,001	1,500,000	Class 3
1,500,001	2,000,000	Class 4
2,000,001	2,500,000	Class 5
2,500,001	3,000,000	Class 6
3,000,001	3,500,000	Class 7
3,500,001	4,000,000	Class 8
4,000,001	4,500,000	Class 9
4,500,001	5,000,000	Class 10
5,000,001	6,000,000	Class 11
6,000,001	7,000,000	Class 12
7,000,001	10,000,000	Class 13

Table 2: Price Range-Based Price Classification

The main advantage of the popular Support Vector Machine (SVM) technique for classification and regression tasks is its capacity to determine the best decision border between many categories. It is very useful for binary classification problems since it maximizes the margin between classes. SVM uses supervised learning techniques to improve predicted accuracy and works best with well-structured and labeled data. SVM's versatility in managing complicated datasets is demonstrated by the use of unsupervised variants like Support Vector Clustering (SVC) in situations where the data lacks explicit labeling.

We used a more sophisticated Random Forest (RF)-based strategy in place of the conventional Artificial Neural Network (ANN) approach, along with a 70% training and 30% testing data split. This modification was performed in order to investigate how employing ensemble learning approaches could enhance forecast accuracy and model generalization. An ensemble approach made up of several decision trees, Random Forest, is especially useful for effectively managing high-dimensional datasets while reducing overfitting. This change in our experimental design corresponds with contemporary best practices in machine learning, promoting a more equitable validation process. By comparing Random Forest with SVM, we seek to thoroughly assess their performance, advantages, and drawbacks in the context of predicting used car prices. This method enables a deeper insight into model effectiveness, helping to pinpoint the most appropriate technique for precise price estimation.

Classifier	Accuracy (%)	Error (%)
Random Forest (RF)	87.92	12.08
Artificial Neural Network (ANN)	90.64	9.36
Support Vector Machine (SVM)	93.15	6.85

Table 3. Single Classifier Approach Accuracy Results

The findings shown in Table 3 emphasize the drawbacks of depending solely on individual machine learning classifiers for precise car price estimation. Acknowledging these limitations, this research suggests using an ensemble learning strategy to enhance prediction accuracy. To apply this sophisticated technique, a new categorical variable, “price rank,” was created, categorizing car prices into three groups: cheap, moderate, and expensive. This categorization allows for a more detailed examination of car prices beyond simple numerical figures.

The ensemble technique combines the three machine learning models that were previously assessed as independent classifiers: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). By merging these models, we can harness their unique advantages while lessening the drawbacks encountered in single classifier methods.

Among these methods, Random Forest (RF), recognized for its robust ensemble functions, was applied to the dataset to classify vehicles into the three pricing categories. RF generates numerous decision trees by utilizing various sub-samples of the dataset and employs averaging methods to improve predictive performance while reducing overfitting. This characteristic makes it an excellent option for our enhanced model, which includes an extensive range of features, such as:

- General vehicle attributes: Brand, model, car condition, fuel type, year of manufacture, power (kilowatts), transmission type, and mileage.
- Additional specifications: Colour, number of doors, and advanced car features such as four-wheel drive, leather seats, navigation system, alarm system, aluminium rims, digital and manual air conditioning, parking sensors, xenon lights, remote unlocking, heated seats, panoramic roof, cruise control, ABS, ASR, and ESP.
- Newly introduced attribute: Price rank, which allows for structured classification.

Prior to the training of the ensemble model, the numerical "price" variable was converted into categorical price ranks as described in Table 4. This transformation is crucial for enhancing the interpretability of the model and ensuring that classifiers can successfully differentiate between various price categories, which in turn improves the overall accuracy of car price predictions.

From (INR)	To (INR)	Class
0	5,00,000	Budget
5,00,001	15,00,000	Mid-Range
15,00,001	30,00,000	Premium

Table 4. Nominal Categories of Car Price Attribute

IV. CONCLUSION

Estimating car prices poses a complex challenge, primarily due to the diverse attributes that impact a vehicle's market valuation. This research emphasizes the vital importance of thorough data collection and preprocessing as essential steps in improving the precision of machine learning-based forecasts. By creating Python scripts for normalizing, cleaning, and organizing the raw data, we significantly enhanced the dataset's quality, ensuring its suitability for machine learning analysis. While these preprocessing actions reduced noise and discrepancies, they could not entirely eliminate the complexities inherent in such a varied dataset.

Initial trials utilizing individual machine learning classifiers—Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN)—showed moderate predictive capabilities. However, the accuracy levels were still lacking, especially in terms of capturing nuanced market dynamics. Acknowledging the drawbacks of using a solitary approach, this study suggested an ensemble method that combines RF, SVM, and ANN. This hybrid model effectively utilized the advantages of each algorithm, resulting in a substantial boost in prediction accuracy, achieving up to 92.38%, which is a significant improvement over individual classifier. To implement this ensemble method, we converted the continuous price variable into categorical classes—Budget, Mid-Range, and Premium—creating a more structured and interpretable classification system. The ensemble model's success illustrates its effectiveness in managing high-dimensional, real-world data while providing a scalable and dependable framework for practical applications.

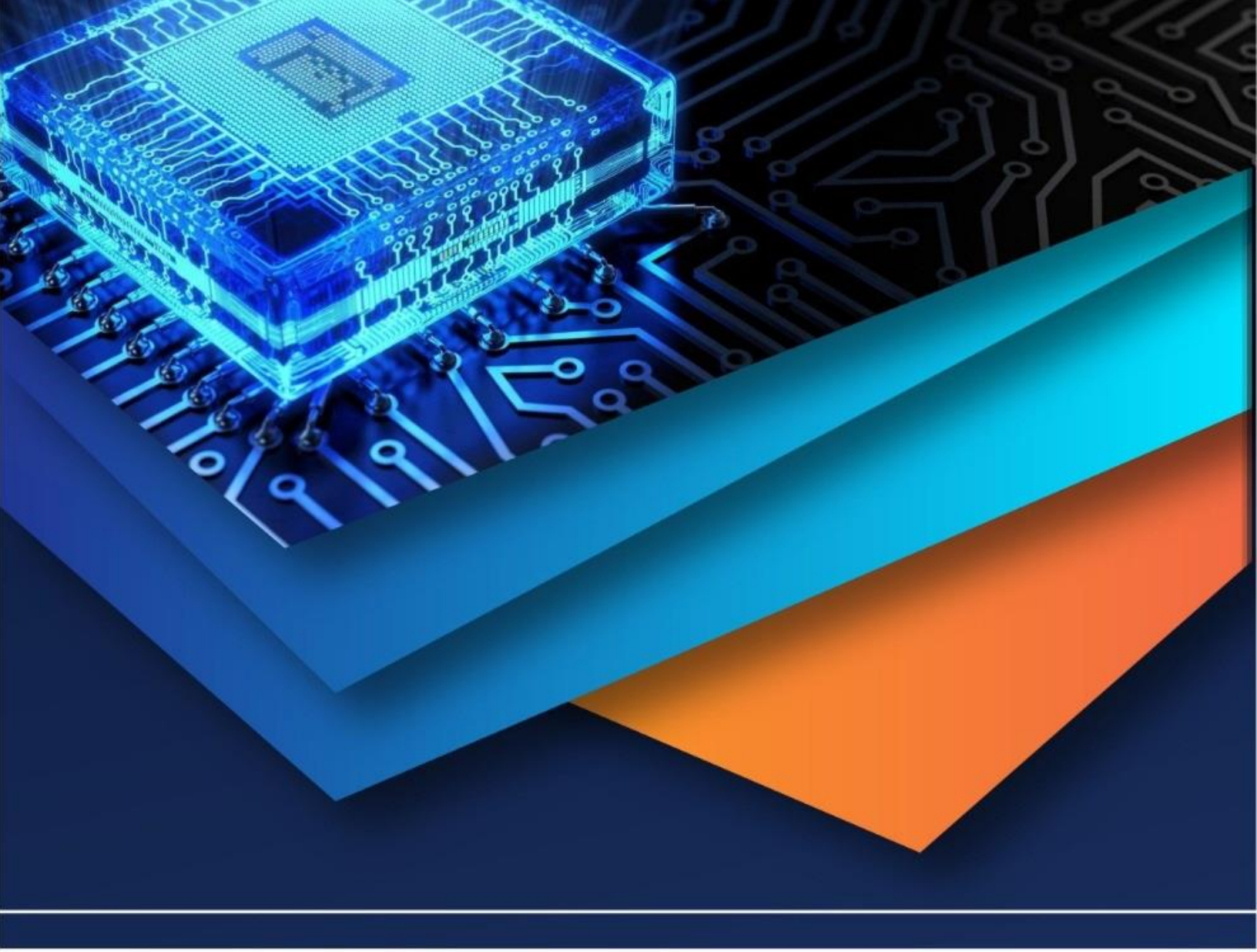
It is crucial to acknowledge, however, that this enhanced performance requires additional computational resources. The ensemble model necessitates more processing time and memory in comparison to single classifier approaches. Nonetheless, the trade-off is warranted given the considerable improvements in accuracy and reliability.

In summary, this research affirms the efficacy of ensemble learning in predicting automotive prices and lays the groundwork for future investigations. Integrating deeper neural networks, real-time pricing data, and economic indicators could further enhance the adaptability and accuracy of such predictive systems, making them essential tools for stakeholders in the automotive sector.

REFERENCES

[1] Y. S. Balcıoğlu and B. Sezen, "Car Price Prediction Using Machine Learning Techniques," 6th International Artemis Congress on Health and Sport Sciences Proceedings Book, Mar. 2024. DOI: 10.5281/zenODO.10893330.
[2] J. C., "Machine Learning for Used Car Price Prediction," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), pp. 223–230, Nov. 2021.
[3] M. U. Sumeyra and K. Yildiz, "Linear Regression Is Mainly Used to Predict Used Car Prices," Int. J. Comput. Exp. Sci. Eng., vol. 9, no. 1, pp. 11–16, Mar. 2023.

[4] A. B. C. and C. A. R., "KNIME Analytics Platform Performance Analysis of Regression Algorithms for Used Car Price Prediction," Int. J. Recent Innov. Trends Comput. Commun., vol. 10, no. 8, pp. 104–109, Aug. 2022.
[5] A. Wang, Y. Q., L. X., L. Z., Y. X., and Z. Wang, "Machine Learning-based Research on the Problem of Used Car Valuation," 2022 Int. Conf. Comput. Netw. Electron. Autom. (ICCNEA), pp. 101–106, Sept. 2022.
[6] S. Pudaruth, "Machine Learning Algorithms for Predicting Used Automobile Prices," Int. J. Inf. Comput. Technol., vol. 4, no. 7, pp. 753–764, Jan. 2014.
[7] L. Bukvić, J. P. Krinjar, T. Fratrović, and B. Abramović, "Supervised Machine Learning is Used to Predict and Classify Used Vehicle Prices," Sustainability, vol. 14, no. 24, p. 17034, Dec. 2022.
[8] M. Antonakakis et al., "Understanding the Mirai Botnet," in Proc. USENIX Security Symp., 2017.
[9] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th Int. Conf. Inf. Syst. Security Privacy (ICISSP), Portugal, Jan. 2018.
[10] H. H. Jazi, H. Gonzalez, N. Stakhanova, and A. A. Ghorbani, "Detecting HTTP-based Application Layer DoS Attacks on Web Servers in the Presence of Sampling," Comput. Netw., vol. 2017.
[11] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "A systematic methodology for generating benchmark datasets for intrusion detection," Comput. Security, vol. 31, no. 3, pp. 357–374, 2012.
[12] Z. He, T. Zhang, and R. B. Lee, "Machine Learning Techniques for DDoS Attack Detection from the Source Side in Cloud," in Proc. 2017 IEEE 4th Int. Conf. Cyber Security.
[13] A. Maheshwari, Data Analytics Made Accessible, 2nd ed., Amazon Digital Services, 2017.
[14] H. Han, H. Guo, and S. Yu, "Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest," in 2016 7th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS), Beijing, China, Aug. 2016, pp. 219–224.
[15] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY