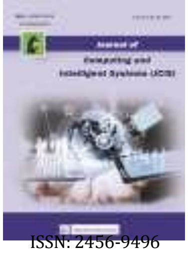
Original Research Article

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in



ISSN: 2456-9496

USED CARS PRICE PREDICTION USING MACHINE LEARNING

Poovizhi Magendiran [#1], **M. Sasikumar**[#2]

Received on 15 MAR 2024, Accepted on 03 APR 2024

Abstract — The research work addresses the critical need for precise pricing tactics in the approximately 40 million used cars that are sold annually. In response to the problem of inflated prices brought on by rising demand, we provide a workable approach that predicts used car prices through supervised learning. The model is chosen after careful consideration of the data, guaranteeing that it captures the essential elements affecting pricing dynamics. In order to estimate vehicle pricing, the machine learning algorithm takes into account a number of features, taking market trends and comparable automobile prices into account. It offers a data-driven method to improve pricing tactics in a competitive market, making it an invaluable resource for people and businesses involved in the used automobile trade.

The paper advances automotive pricing analytics by presenting a scalable system for used car price prediction. Through the utilization of all-inclusive features and the management of market fluctuations, the model enables stakeholders to make well- informed price decisions that optimize revenue.

Keywords - Used car, Price Prediction, Machine Learning, Automotive pricing analytics, Supervised Learning

1. Introduction

Customers who purchase new cars can therefore be sure that their financial investment is worthwhile. Used car sales, however, are rising globally as a result of rising new car prices and consumers' inability to purchase new vehicles owing to financial constraints. An intriguing and urgently needed problem to be solved is used automobile price prediction. Customers are often taken advantage of when used automobile costs are set unrealistically high, and many fall victim to this scam.

Consequently, in order to efficiently assess the car's worthiness based on a range of factors, a used car price prediction system becomes very essential.

Most automobiles are purchased on a lease, which is an arrangement between the buyer and seller, because of the high cost of cars and the migratory lifestyle of people in developed nations. After the arrangement is fulfilled, these cars are resold. Thus, resale has become a vital aspect of life in the modern world.

An automobile has characteristics, such as its age, make, country of origin, mileage, and horsepower. gasoline efficiency is particularly crucial because gasoline costs are rising. Additional elements include the fuel type, design, braking system, acceleration, number of doors, safety index, size, weight, height, paint color, customer feedback, and notable accolades the automobile manufacturer has received.

2. Related works

Several studies have been conducted on the topic of used car price prediction. Researchers typically use historical data to predict product prices. Pudaruth employed decision tree, k-nearest neighbors, Naive Bayes, and multiple linear regression techniques to forecast the values of vehicles in Mauritius [1]. The cars weren't brand-new. The prices derived from different strategies were found to be reasonably similar when the forecast results were compared. Nevertheless, it was shown that the Nave Bayes approach and the decision tree method were unable to categorize or predict numerical values. The small sample size does not produce excellent prediction accuracy, according to Pudaruth's research.

* Corresponding author: E-mail: [1]poomca12@gmail.com [2]sasikumar5731@gmail.com

[1] PG Department of Computer Science, Sacred Heart College, Thiruvalluvar University, Tirupattur, Tamil Nadu, India.

[2] PG Department of Computer Science, Sacred Heart College, Thiruvalluvar University, Tirupattur, Tamil Nadu, India.

Kuiper, S. (2008) introduced a multivariate regression model that helps with numerical value classification and forecasting [2]. It demonstrates how to forecast the price of General Motors (GM) vehicles manufactured in 2005 using this multivariate regression model. To estimate car prices, no specific information is required. The information available online is adequate for price prediction. The author of the paper forecasted the same car price and used variable selection techniques to help decide which factors should be included in the model based on greater relevance.

Pal et al. found a Random Forest technique for used car price prediction in 2019 [3]. Using the Kaggle data set, the study assessed used car price prediction; accuracy for test data was 83.62%, and accuracy for train data was 95%. Following the removal of outliers and superfluous information from the data set, the price, kilometers, brand, and type of vehicle were identified as the most pertinent features for this prediction. Because Random Forest is a complex model, its accuracy was higher than that of previous research using comparable data.

In 2022, Eesha Pandit et al. [4] Predicting the price of a used car is a difficult task since so many variables and attributes must be taken into account to produce reliable results. The first and most important step is pre-processing and data collection. The model was then created and explained in order to use algorithms and generate results.

Decision Tree Algorithm was found to be the best performer after the model was subjected to several regression algorithms. With the maximum R2 score of 0.95, the Original vs. Prediction line diagram explained that it presented ultimate correct indicators. The Decision Tree not only had deepest Mean Square Error (MSE) and Root Mean Square Error (RMSE) principles, but it also had highest in rank R2 score, that granted that the indicator faults were bottommost and the results were intensely correct.

According to Gegic, E. et al. [5] (2019), a model was developed to predict the price of used cars in Bosnia and Herzegovina. They engaged Artificial Neural Networks (ANN), Random Forests (RF), and Support Vector Machines (SVM) as machine learning techniques. The entire previously defined techniques were combined. The input dataset for the prediction process was collected from the website autopijaca.ba using a web scraper prepared using PHP programming language. Then, by comparing the relative performances of several algorithms, the approach that worked best with the provided data was identified. The final prediction model was used to create a Java application. Additionally, when the model was validated using test data, its accuracy of 87.38% was found.

Dholiya et al. [6] introduced a machine learning strategy for car resale in 2019. Giving the user a realistic estimate of the vehicle's potential cost is the aim of the Dholiya, M., et al. method. Depending on the user's particular vehicle needs, the web application system may also offer a list of potential options for various car types. It assists by providing pertinent information to enable the buyer or seller to make an informed decision. This system was trained using historical data collected over a long period of time and uses the multiple linear regression technique to generate predictions. The raw data was first gathered using the KDD (Knowledge Discovery in Databases) technique. After that, it underwent pre-processing and cleaning in order to identify useful patterns and derive some meaning from them.

Richardson's research was based on the notion that automakers are more likely to create durable vehicles [7]. Specifically, he used multiple regression analysis to show how hybrid cars are more likely to retain their value than conventional cars [9]. This could be due to its reduced fuel consumption as well as the escalating environmental and climate concerns. He gathered all of his information from a number of sources.

Listiani reported similar research that used Support Vector Machines (SVM) to predict the cost of leasing a car [8]. This study demonstrated that when a very large amount of data is available, SVM predicts prices significantly more accurately than multiple linear regression. Furthermore, SVM avoids over- and under-fitting problems and performs better when handling high-dimensional data [10]. Important features for SVM are found using a genetic approach. However, the method does not demonstrate why SVM is superior to simple multiple regression in terms of variance, mean and standard deviation.

The first section provides the background for the research work by eliciting the importance of the prediction using machine learning. A review of related works are presented in section 2. The section 3 describes the implementation of research and process methodology. The model evaluation was discusses in section 4. The section 5 describes the analysis of the results. The conclusion and the future enhancement of the research work were provided in section 6 and section 7.

3. Implementation of Research & Process Methodology

The project uses a Kaggle dataset that includes attributes such as make, model, year, mileage, fuel type, and transmission to forecast used car prices. Encoding categorical variables and dealing with missing data are examples of pre-processing. Feature relationships are visualized through exploratory data analysis (EDA).

Relevant features are chosen based on correlation with the target variable. The models that have been trained include Decision Tree, XGBoost, Random Forest, MLP, AdaBoost, Extra Trees, Gradient Boosting, and CatBoost Regressors. Metrics like Mean Absolute Error and accuracy ratings are used to assess the model's performance.

Models are compared in order to identify the most effective model for prediction. In order to evaluate the model's accuracy in predicting used car prices, test data is forecasted. The flow of this methodology is depicted in figure 1.

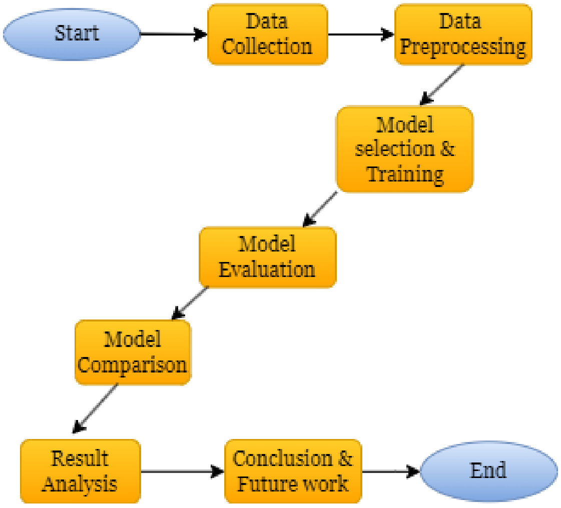


Figure- 1: *Workflow of Process Methodology*

3.1 Data Collection

The dossier accumulation process for this research project on secondhand car price guess complicated sourcing an inclusive dataset from Kaggle. This dataset was carefully curated to involve a diverse range of appearance essential for correctly envisioning and classifying the price ranges of secondhand trucks. Key attributes in the way that the car's name, year of manufacture, kilometres compelled, fuel type, broadcast type, proprietor type, mileage, diesel qualifications, capacity, places volume, and market action like the new price of the limousine were utterly resolved to specify a strong foundation for predicting shaping.

The choice of physiognomy was guided for one need for representativeness and pertinence to real-world synopsizes, taking everything in mind two together intrinsic vehicle traits and outside market influences. Machine learning methods were working to expand predictive models fit judging used car prices based on their physiognomy. Through preliminary data analysis, the impact of each feature on price was meticulously adjudicated, guaranteeing the reliability and strength of the predicting models.

In order to meet the indispensable demand for a trustworthy and effective used car price prediction application, this research uses the complex machine learning algorithms and a big dataset. A system like this is essential to enable sellers and purchasers to determine a car's market worth in an impartial manner.

3.2 Data Pre-processing

The research work was handled with the missing values, made sure variables were consistent, and encoded categorical features during the data pre-processing stage. Outliers were identified and dealt with, numerical features were scaled, and the data was divided into training and validation sets. These actions were essential for guaranteeing the dataset’s quality and preparedness for analysis and modeling, which in turn enhanced the project's machine learning algorithms' performance.

When the dataset was examined for missing values, it was found that a significant number of missing values were present in several of the columns. To fix this, the drop () function was used to remove columns with a large number of missing items, such "New Price." Likewise, superfluous columns such as 'Unnamed: 0', which plausibly functioned as an index, were eliminated in order to simplify the dataset.

This is the process methodology model figure-1's representation of the first pre-processing phase.

Training data with Null Values

```
Unnamed: 0              0
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 36
Power                  36
Seats                  42
New_Price            5195
Price                   0
```

Test set with Null values

```
Unnamed: 0              0
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 0
Engine                 10
Power                  10
Seats                  11
New_Price            1052
dtype: int64
```

The training dataset's missing values were handled, and the test dataset underwent the same process. To maintain uniformity in the data preparation process, dropna() was used to remove any rows in the test dataset that were missing values. Data from the 'Name' column was combined to create a new feature called 'Cars'. In order to improve the model's capacity for prediction, this feature attempted to gather information on the make and model of the vehicle.

To standardize textual information in columns like " Mileage," "Engine," and "Power," additional data cleaning was done. To improve uniformity, units like 'kmpl, CC, and bhp' were eliminated from these columns. Furthermore, a common value (112) was used to replace missing values in the 'Power' column that were indicated as 'null'.

Training data after pre-processing

```
Name                 0
Location             0
Year                 0
 Kilometers_Driven   0
Fuel_Type            0
Transmission         0
Owner_Type           0
Mileage              0
Engine               0
Power                0
Seats                0
Price                0
Cars                 0
```

Test set after pre-processing

```
Name                 0
Location             0

Year                 0
Kilometers_Driven    0
Fuel_Type            0
Transmission         0
Owner_Type           0

Mileage              0
Engine               0
Power                0
Seats                0
Cars                 0
```

3.3 Model Selection & Training

This research work aimed at predicting used car pricing, which employed a diverse array of machine learning models to discern the most effective predictor. The models evaluated include Decision Tree, XGBoost, Random Forest, MLP Regressor, AdaBoost, and Extra Trees. Among these models, the Gradient Boosting Regressor (GBR) emerged as the top performer, showcasing remarkable accuracy rates of 99.47% on the training set and 96.02% on the testing set. This outstanding performance was pivotal in the model selection process.

This approach to model selection was anchored in comparative performance analysis, wherein each model was rigorously assessed based on its ability to accurately predict used car prices. Notably, the GBR consistently outperformed its counterparts, demonstrating superior accuracy across both training and testing sets.

The training of the chosen model involved several crucial steps. Initially, the dataset underwent pre-processing to ensure its suitability for training purposes. Subsequently, the dataset was partitioned into distinct training and testing sets to facilitate robust model evaluation. The chosen model, GBR, was then fitted to the training data, allowing it to learn and internalize patterns within the dataset.

Finally, the efficacy of the trained model was assessed using the testing set, wherein its ability to accurately predict used car prices was thoroughly evaluated.

The exceptional performance exhibited by the GBR in this phase further validated and decided to adopt it as the primary predictive model.

In summary, through meticulous model selection and rigorous training procedures, the research work was identified the GBR as the optimal solution for predicting used car prices. Its exceptional accuracy and robust performance underscore its suitability for addressing the objectives of this research work effectively.

4. Model Evolution

In the dynamic landscape of the research, the journey towards identifying the optimal predictive model for used car pricing involved a multifaceted process of model evolution. With a keen eye on precision and reliability, the research work meticulously traversed through various stages, each designed to refine and enhance the predictive prowess of the proposed models.

The exploration began with an extensive array of machine learning techniques. Through rigorous experimentation, algorithms were deployed spanning from Decision Trees to sophisticated ensemble methods such as XGBoost and Random Forests. This diverse ensemble allowed us to capture a wide spectrum of modeling approaches, each offering unique insights into the complex patterns underlying used car pricing dynamics.

Performance measures like mean absolute error (MAE) and root mean square error (RMSE) were essential to the proposed model evaluation approach. These metrics functioned as guiding lights, shedding light on how well each model predicted used car prices. The research work able to make well-informed decisions in the following phases by carefully examining these measures and gaining important insights into the advantages and disadvantages of each strategy.

To ascertain the models' capacity for generalization, the research process embraced the robust methodology of cross-validation. By subjecting the proposed models to rigorous testing across diverse subsets of the data, we sought to unearth their true predictive potential beyond the confines of the training set. This meticulous validation process not only bolstered the confidence in the models' performance but also guarded against the perils of overfitting, ensuring their robustness in real-world scenarios.A pivotal juncture in the proposed model evolution journey was the hyperparameter tuning phase. Here, armed with a nuanced understanding of each model's intricacies, they embarked on a quest to optimize their configuration for maximum efficacy.

Through iterative adjustments and fine-tuning of hyperparameters, we endeavored to unlock the latent predictive power within the chosen algorithms, pushing the boundaries of performance to new heights.

Ultimately, the culmination of the proposed model evolution efforts bore fruit in the form of the top-performing predictive model. This model, meticulously refined through a synthesis of empirical analysis and methodical optimization, emerged as the pinnacle of predictive accuracy, surpassing its peers in both reliability and precision. In conclusion, the journey of model evolution epitomized a relentless pursuit of excellence. From the initial exploration of diverse modeling techniques to the rigorous validation and optimization processes, every step was guided by a steadfast commitment to delivering dependable and insightful predictions of used car prices. Through this iterative refinement, the research work forged predictive models that not only met but exceeded the exacting standards of efficacy and reliability, ushering in a new era of precision in the realm of automotive pricing analysis.

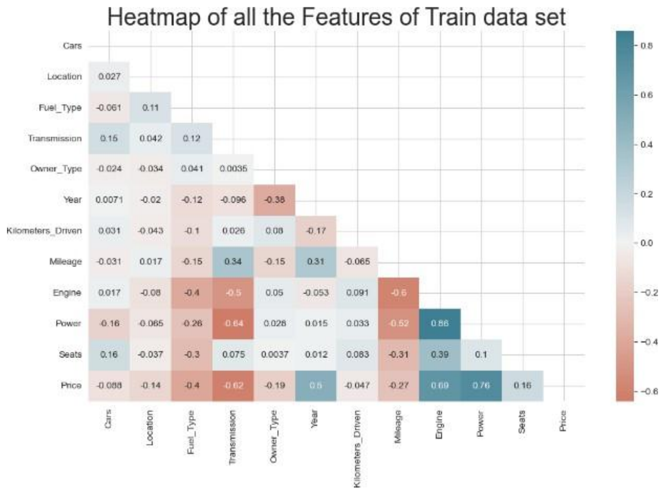


Figure-2: *Heatmap of Features*

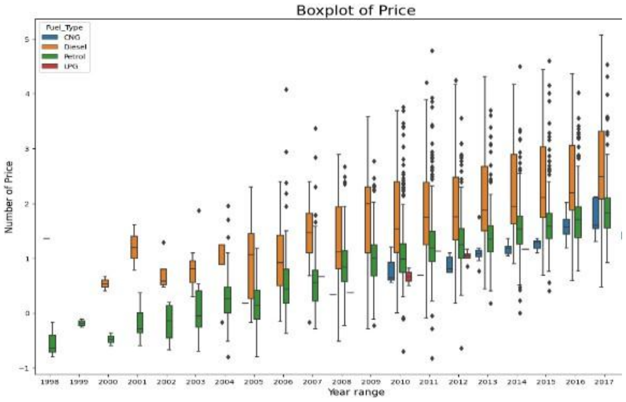


Figure-3: *Price vs fuel types*

In Figure 2, shows the heatmap illustrating the correlation between features in the dataset, providing insights into their influence on used car prices.

Figure 3 showcases the relationship between price and various fuel types, offering valuable insights into consumer preferences and market dynamics in the used car industry.

4.1 Model Comparison

The effectiveness of various predictive models was thoroughly compared in this research study on used car price forecasting.This comparison looked closely at important performance metrics like accuracy on training and testing sets, Mean Absolute Error (MAE), Absolute Average Deviation Percentage (AAD%), and Root Mean Squared Error (RMSE).

The RMSE served as a pivotal metric, indicating the average deviation between predicted and actual prices. Lower RMSE values suggested higher predictive accuracy, signifying minimal discrepancies between predicted and observed prices. Notably, the CatBoostRegressor (CBR) demonstrated the lowest RMSE of 64.79, closely followed by the GBR with an RMSE of 69.11. Additionally, the Random Forest Regressor exhibited strong performance with an RMSE of 79.08, showcasing its efficacy in predicting used car prices.

Accuracy scores on both training and testing sets were crucial in assessing the models' ability to generalize and avoid overfitting. While several models achieved high accuracy scores on the training set, it was essential to evaluate their performance on unseen data.

The CBR emerged as the top performer in this regard, achieving an impressive accuracy score of 96.50% on the testing set, closely followed by the GBR and XGB Regressor.

Table-2: *Model Comparison*

Model	Root mean squared Error	Trainin g Accura cy	Testing Accurac y
MLP	224.75	0.63	0.57
AdaBoost	149.30	0.82	0.81
Decision Tree	115.78	0.99	0.88
Random Forest	84.41	0.99	0.94
Extra Trees	80.36	0.99	0.94
XGB Regressor	73.52	0.99	0.96

The MAE provided valuable insights into the average magnitude of errors between predicted and actual prices. Lower MAE values indicated better performance, with smaller deviations from the true values. The CBR exhibited the lowest MAE of 43.62, followed by the GBR with an MAE of 47.05, showcasing their superior predictive accuracy.

Furthermore, the AAD% metric offered a percentage-based assessment of prediction accuracy, providing a holistic view of the models' performance. The CBR demonstrated the lowest AAD% of 64.79, closely followed by the GBR, underlining their robust predictive capabilities.

In summary, the comprehensive model comparison revealed the CBR and GBR as the top performers in accurately predicting used car prices. These models exhibited superior predictive accuracy, minimal error rates, and robust generalization capabilities, making them invaluable assets in automotive pricing analysis.

Table 2 compiles the relative efficacy of several machine learning models in predicting used car prices. Each model's effectiveness is evaluated using Root Mean Squared Error (RMSE) in addition to training and testing accuracies. This in-depth analysis helps identify the optimal model for accurate price prediction.

5. Result Analysis

The GBR algorithm was used in the study to predict used car prices because it is more accurate than other algorithms. The model was tweaked with parameters like 3000 estimators and a random state of 21 to optimize its performance.

The error table in the analysis, which is shown in Table 3, which depicts that the model performed well when evaluated. The average size of the differences between expected and actual prices was shown by the Mean Absolute Error (MAE) of 47.05, and the overall deviation of predictions from observed values was shown by the Root Mean Squared Error (RMSE) of 69.11. These metrics are essential markers of the accuracy and dependability of the model's used car price forecasting.

Furthermore, the accuracy scores on both the training and testing sets underscored the model's robust performance. With an accuracy of 99.47% on the training set and 96.02% on the testing set, the GBR demonstrated its ability to generalize well to unseen data, signifying its efficacy in real-world applications.

In summary, the GBR algorithm emerged as a standout performer in the research work, exhibiting superior accuracy and reliability in predicting used car prices. These findings highlight its potential to serve as a valuable tool for stakeholders in the automotive industry, aiding in decision-making processes and market analysis.

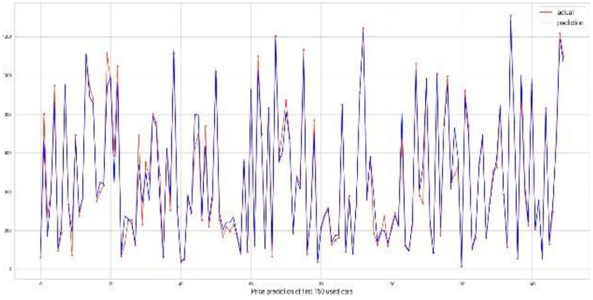


Figure-4: *Price prediction first 150 used cars*

Figure 4 represents the, "Price Prediction for First 150 Used Cars," displays the variance between predicted and actual prices for the initial subset of 150 used cars from the testing dataset. This visualization offers insights into the model's performance, aiding stakeholders in assessing accuracy and identifying areas for improvement. Table 4 depicts the sample training dataset used in the application for the prediction.

Table-3: *Error Table of XGB Regressor*

Error Table	
Mean Absolute	47.05
Mean squared	4775.90
Root Mean Squared	69.10
Accuracy on Training set	0.99
Accuracy on Testing set	0.96

Table-4: *Price prediction for used cars in Test Dataset with car id*

Car_id	Price
0	229.220
1	157.300
2	933.110
3	237.274
4	273.766
5	758.443
6	1032.597
7	122.656
8	166.785
9	751.605

6. Conclusion

In summary, this study employs a Kaggle dataset to effectively tackle the problem of used-car pricing prediction. The solution achieves remarkable accuracies of 96.02% for test data and 99.47% for train data by utilizing sophisticated machine learning techniques, including the Gradient Boosting Regressor model. This shows how reliable and efficient the created model is in correctly predicting car prices. The research work showcases a methodical approach to addressing practical issues in the automobile sector by means of thorough data gathering, analysis, and model assessment. The research contribution to the advancement of predictive analytics for used-car pricing is highlighted by the employment of sophisticated approaches. All things considered, this study shows how machine learning algorithms may be used to improve the way the automobile industry makes decisions, which will ultimately result in better informed and more effective transactions.

7. Future Enhancement:

The accuracy and resilience of the existing model for projecting car prices will be increased in the future by incorporating cutting-edge methods and algorithms. A completely automated and interactive system will also be created to function as an extensive database of pre-owned automobiles, complete with prices. With the help of this system's recommendation engine, customers will be able to easily access and contrast the costs of comparable cars.

The goal is to transform the process of predicting automobile prices and provide consumers with useful information for making well-informed decisions in the automotive sector by using cutting-edge technologies and improving user experience.

References

[1] Pudaruth, S. (2014) ‘Predicting the Price of Used Cars using Machine Learning Techniques’, International Journal of Information & Computation Technology, 4(7), pp. 753–764.

[2] Kuiper, S. (2008) ‘Introduction to Multiple Regression: How Much Is Your Car Worth?’, Journal of Statistics Education, 16(3). doi: 10.1080/10691898.2008.11889579.

[3] Pal, N. et al. (2019) ‘How Much is my car worth? A methodology for predicting used cars’ prices using random forest’, Advances in Intelligen Systems and Computing, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28.

[4] Eesha Pandit et al. (2022) ‘Prediction of Used Car Prices using Machine Learning Techniques’, International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 12

[5] Gegic, E. et al. (2019) ‘Car price prediction using machine learning techniques’, TEM Journal, 8(1),
pp. 113–118. doi: 10.18421/TEM81-16.

[6] Dholiya, M. et al. (2019) ‘Automobile Resale System Using Machine Learning’, International Research Journal of Engineering and Technology (IRJET), 6(4), pp. 3122–3125.

[7] Richardson, M. (2009) Determinants of Used Car Resale Value. The Colorado College

[8] Listiani, M. (2009) Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, Technology. Hamburg University of Technology.

[9] Yian Zhu, ‘Prediction of the price of used cars based on machine learning
Algorithms’, Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230917.

[10] Lucija Bukvi´c et al. ‘Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning’, Sustainability 2022,14,17034. https://doi.org/10.3390/su1424170