Resale Revolution: Pioneering the Future of Automotive Asset Valuation

Pranav Bafna, Chinmay Dalvi, Shruti Dawange, Aadhar Battulwar, Samarth Harne

Guided by: Dattatray Takale

Pranav Bafna, Vishwakarma Institute of Information Technology, Pune

Chinmay Dalvi, Vishwakarma Institute of Information Technology, Pune

Shruti Dawange, Vishwakarma Institute of Information Technology, Pune

Aadhar Battulwar, Vishwakarma Institute of Information Technology, Pune

Samarth Harne, Vishwakarma Institute of Information Technology, Pune

Dattatray Takale, Assistant Professor Vishwakarma Institute of Information Technology, Pune

**Abstract**

*A used vehicle's pricing on the market is determined by a variety of factors, making it difficult to determine whether the quoted price is accurate. In order to enable consumers to make well-informed decisions, the goal of this project is to create analysis models that can precisely forecast the cost of a used car depending on its qualities. Use a dataset of the sale prices of various brands and models in Indian cities to implement and assess different learning methodologies. Though they take a lot of processing, XG Boost and Gradient Boost clustering with linear regression get the best results. While it required a lot less training time than the previously stated techniques, conventional linear regression also produced acceptable results. It might be challenging to determine whether a used car is worth the listed price when you browse ads online. The actual value of an automobile can be influenced by a number of factors, such as the mileage, make, model, year, etc. Setting a fair price for a second-hand car presents challenges for sellers as well. The goal is to create models for used automobile price prediction based on available data.*

**Keywords**: Gradient Boost, XG Boost, Car, Data Science, Data Analysis, Random Forest, Dataset, Test Train Data, Algorithm, Prediction, Resale.

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**1. Introduction:**

Morocco's new automobile industry has been expanding quickly in recent years as a result of significant government investments made in the country's northern auto industries. But rather than taking out bank loans to buy new cars, a lot of clients choose to acquire inexpensive used cars. In the used automobile industry, several e-commerce websites, like Moteur 1 and Avito2, help buyers and seller’s deal.[1]

It is imperative that before making any decisions about the purchase or sale of an automobile, both buyers and sellers are aware of the fair market worth of the vehicle. Our machine learning algorithm has been created to help consumers estimate the pricing worth of used cars they wish to buy or sell. Based on a number of variables, such as mileage, year of production, fuel type, financial capability, car make, and model, this model forecasts the price of an automobile.

In order to forecast used automobile prices based on the previously described parameters, we deployed a number of supervised machine learning algorithms in this research. The online retailer avito provided the dataset that was used to train and evaluate the regression models. The gradient boosting regressor (GBR) had the lowest root mean square error (RMSE) and the highest R-squared score of all the tested models.[2]

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**2. Literature Survey:**

We gathered information from numerous websites and performed an analysis of IEEE papers. the use of supervised machine learning methods to forecast Indian used automobile prices. This prediction system uses historical data from daily newspapers. Predictions were made using a variety of methods, including decision trees, naive Bayes, k-nearest neighbours, and multiple linear regression analysis.[3]

Prediction of Car Prices analysed a large number of unique attributes for dependable and accurate prediction using machine learning techniques. Three machine learning techniques—Artificial Neural Network, Support Vector Machine, and Random Forest—were used to construct a model for estimating the cost of old automobiles in India. It was discovered that several current systems were not precise in estimating the cost of the car.[5]

A few researchers have used a variety of machine learning approaches in the literature to forecast car expenses depending on specified parameters. We looked into how our technique might be used to forecast how much automobiles will cost in India City. Predictions of car costs were made using methods such naive Bayes, decision trees, k-nearest neighbours, and multiple linear regression analysis. The study of multiple linear regression was used to find correlations between various features. To categorize new samples in KNN, only three attributes—make, year, and cylinder volume—were taken into account. automobiles with higher cylinder volume values have higher normalized values than automobiles with smaller values, according to this analysis. One notable flaw found was that naive Bayes and decision trees do a poor job of controlling continuous values with output classes. As a result, price classes were created for the price characteristic. Two different car models were used to test the prediction, and samples from vintage newspapers were used.[6]

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**3. Data Collection and Preprocessing**

Pre-processing is a Data Mining approach that entails formatting unprocessed data so that it may be easily understood. Real-world data frequently contains many erroneous facts and a deficiency of precise activity or trend data. As a result, this could lead to poorer data gathering and, eventually, poorer models made from the data. Pre-processing the data can help to solve such issues. Pre-processing, often known as encoding, is the step-in machine learning where data is modified to facilitate easier machine parsing. The algorithm can now appropriately understand the data as a result. The dataset in this project is pre-processed using the following procedures.[7]

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**4. Exploratory Data Analysis:**

Car Name was eliminated as it was considered unnecessary because it was included with the Model and Year columns. Follow these procedures to do an exploratory data analysis:[8]

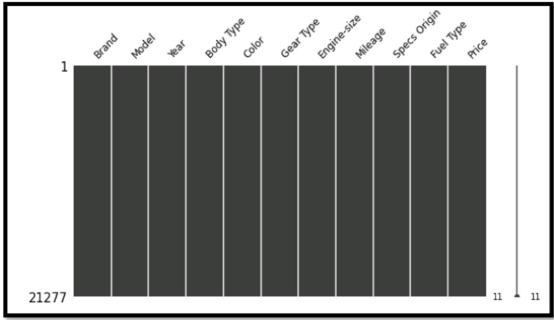
A missing value heap map and a missing matrix were used to plot the dataset in order to see the missing values. The dataset does not contain any sample that contains a null or missing value, as can be inferred from Figures 1 and 2.

Fig. 1. Missing Value Matrix

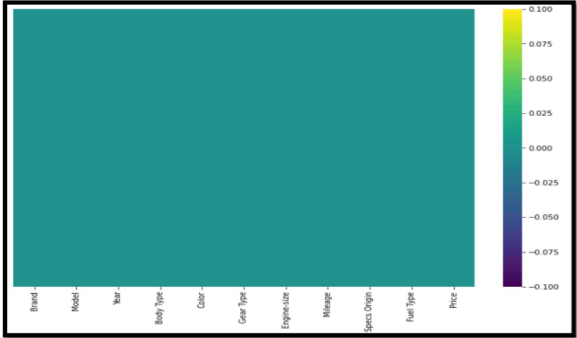


Fig. 2. Missing value Heap Map.

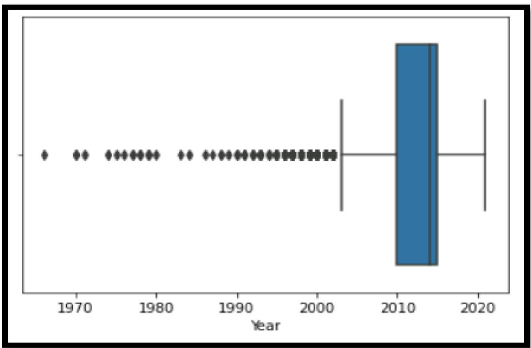
The heatmap matrix presented in Figure 2 indicates that there are correlations, but not collinearity, between the prices and the numerical attributes. On the other hand, a little correlation can be seen between manufacturing year and miles traveled.2. Box plot is used to identify the outliers of the integer attribute. Hence, Year attribute is visualized.[9]

Fig. 3. Year-box plot.

Recursive feature elimination (RFE) is the strategy used for feature selection after an EDA, keeping just the features that are most critical to automobile prices. In order of significance, the following features have been chosen: fuel type, model, mileage, mark, fiscal power, and year of manufacture. We have eliminated features that have no bearing on the price of the car, such as city, kind, and sector. Training and testing machine learning models precedes the label encoding of categorical features (fuel type, mark, and model) and standard feature scaling.[10]

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**5. Methodology:**

Data Extraction: Data Extraction using web scraping of the website with used car data ([www.cars24.com](http://www.cars24.com/)) and then proceed with data cleaning and feature selecting. After data cleaning, we will know how various features are distributed.[11]

Developing Model: Develop a regression model to predict the price of used car based on various models and cross-validate the results to find the best fit model for price prediction in the case of both test and train data.[12]

Designing Webpage: Designing a web page to demonstrate the working of the best fit model and provide an easy and interactive way for the user to predict the car price based on a few questions.[13]

The dataset was pre-processed following data collection in order to eliminate missing value samples, remove non-numerical components from numerical attributes, convert categorical values into numerical (if necessary), correct unit discrepancies, and, if necessary, remove attributes that don't affect price evaluations in order to simplify the model. Information Gaining insight into the data and determining what adjustments or alterations need to be made prior to constructing and implementing the model are crucial components of the model-building process. [14] A first examination of the data is necessary to have a more profound comprehension of its quality, concerning outliers and the skewness of the statistics. [15]

Qualitative To do thus, statistical analyses of numerical and categorical variables were conducted. knowing the primary characteristics that influence the pricing outcomes as well. This was accomplished by using a correlation matrix for each attribute to comprehend the relationships among the various components. [16] The procedures for gathering the dataset included bagging regressors, random forest regressors, logistic regression, pre-processing, encoding, normalization, missing values, and NULL cells conventions. Accuracy, MSE, MAE, and RMSE model evaluation were all examined in the experiments. Various data mining models were developed to estimate used car prices and values once the data was arranged and converted into a format that could be analysed by the data mining technique. Three models—the Bagging Regressor, the Random Forest Regressor, and the Logistic Regression model technique—are suggested to be constructed in this study. First, the data was divided into training and testing portions; varying ratios can be used to assess the portioning percentage and examine various outcomes.[17]

–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**6. Results and Discussion:**

**6.1 Used Car Analysis**

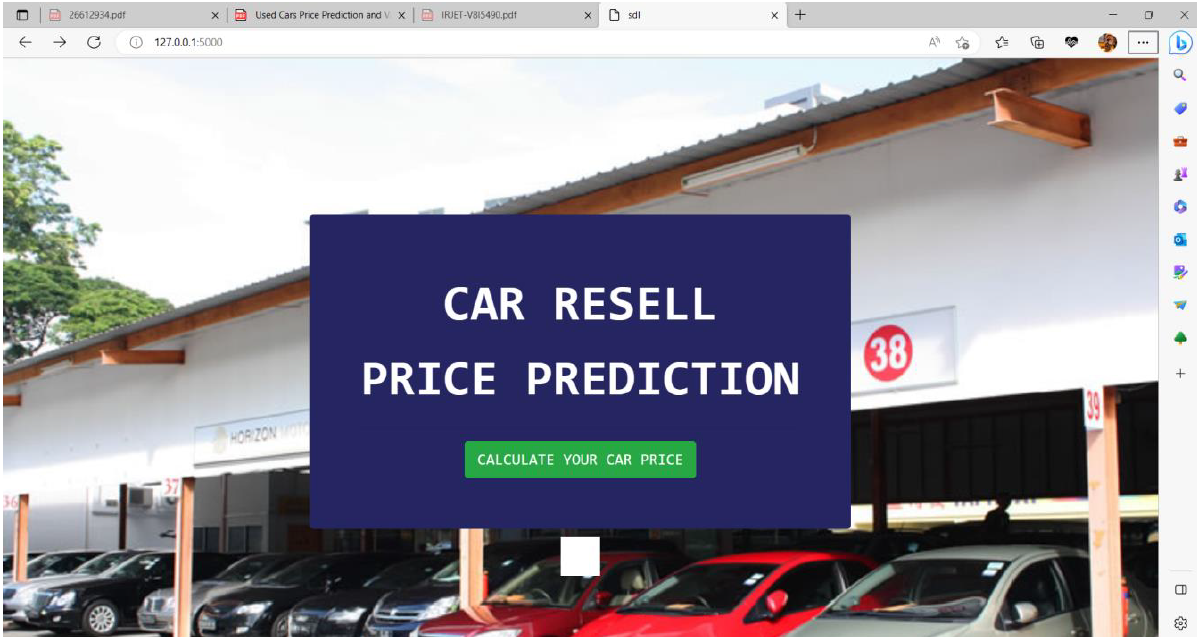
As the popularity of individual automobiles has grown, more and more people are searching for a cheap vehicle to buy, usually through the used car market. The sector will undergo a revolution in the upcoming years, with new business models and digitization driving markets to better meet customer demands and improve efficiency. The Manheim Used Vehicle Index saw an extraordinary rise in the value of used cars in the first 15 days of April 2021 compared to March, rising by 6.81%. In addition, the value has increased by 52.2% since April 2020, and the most recent trend indications point to a few more weeks of gain. This is brought on by tax refunds and stimulus payments as a result of the Covid-19 consequences, as well as a worldwide decline in auto output. (Index of Used Vehicle Value, 2021) In order to provide consumers with realistic car evaluations, it is necessary to have a technology that can anticipate used car pricing based on actual data collected from local websites. Price changes are typical and deceptive prices on other websites. The goal of this project is to provide a user-friendly interface that is realistic enough for consumers to assess car pricing before buying or selling. [18]

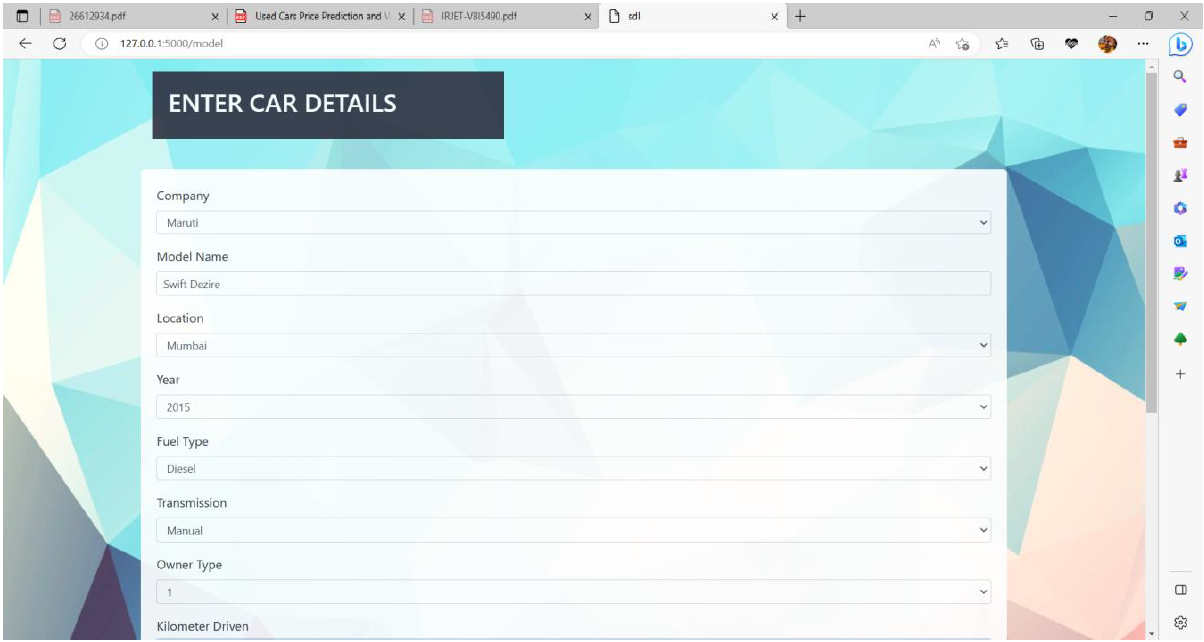
**6.2 Dataset Description**

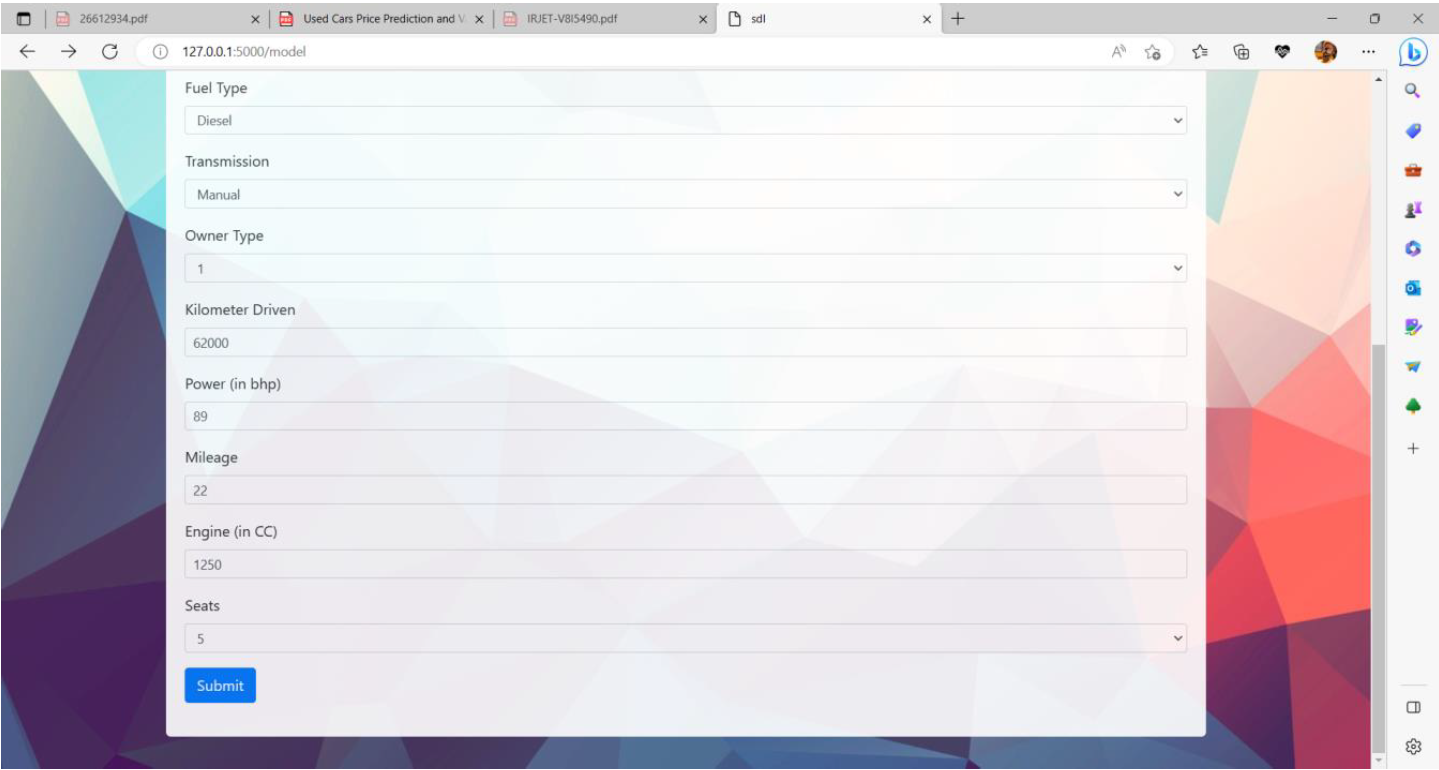
Through numerous runs and iterations, data was gathered and scraped from the Cars24 website using Parse Hub, a scraping application that is exclusively available for use. It was successful to get 21,277 rows of data from the website with 11 variables.[19]

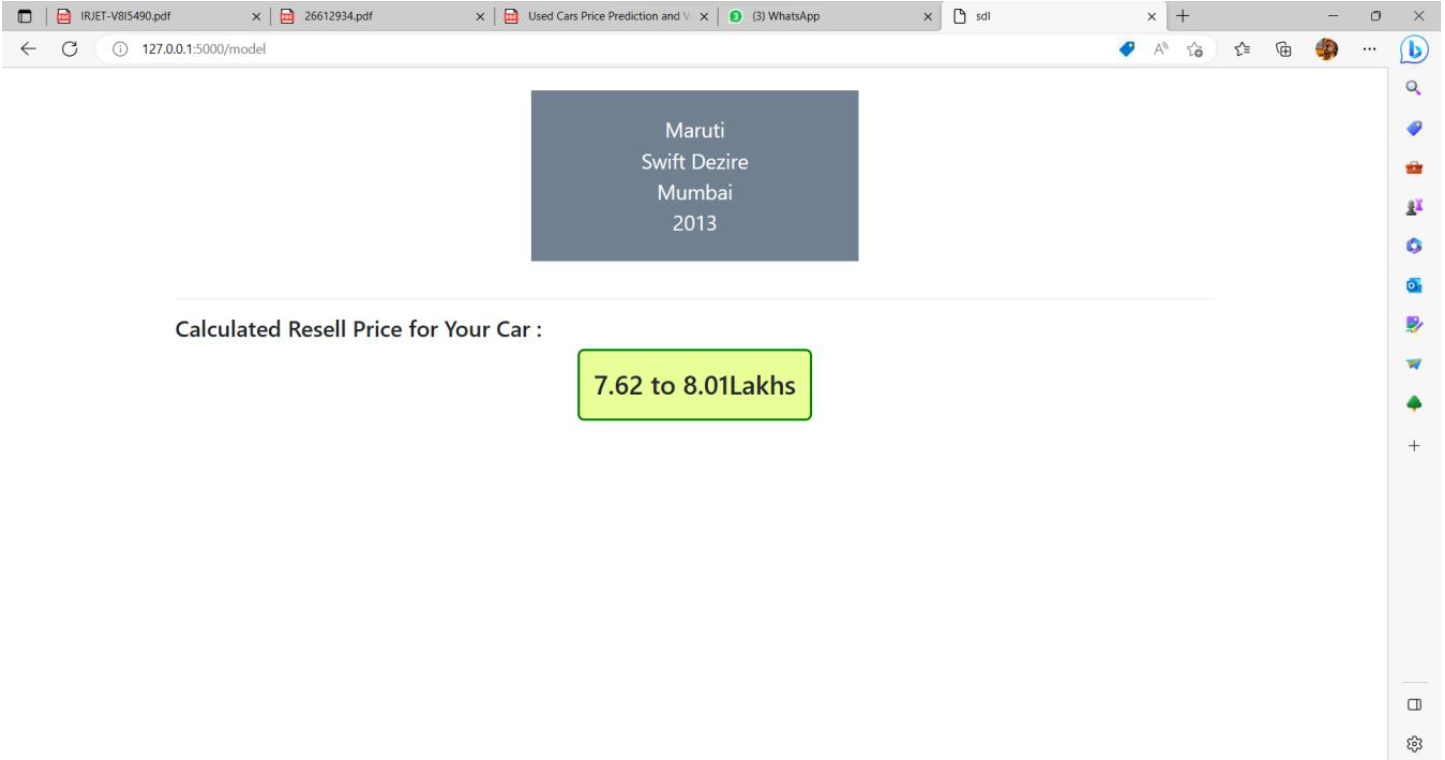
–––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**Output:**









**7. CONCLUSION:**

In order to forecast a reasonable price for the used autos, the suggested system performs admirably. Pricing for used automobiles in the Mumbai area are expertly projected by the system. Either the buyer or the seller will receive the true cost of the used car thanks to this approach. The implementation of a regression system for used car price prediction makes use of two well-known ensemble machine learning algorithms: XGBoost and Random Forest. Regarding accuracy, both methods are quite good and comparable. Through the employment of more trees, Random Forest avoids overfitting. XG Boost is a popular technique because it can prevent overfitting and manage missing values. The results show that XG Boost performs better than the Random Forest Algorithm. XGBoost operates quickly and provides a respectable level of accuracy. This technology is designed to be used exclusively in the Mumbai area. But if the supplied data is in the right format, it can also be expanded to other areas.

––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––––

**ACKNOWLEDGMENT**

We are grateful to the Computer Technology Department's faculty and technical personnel for their costly, valuable, and good advice in finishing this task. We are grateful to all of our classmates for their invaluable assistance with our class assignment. With the assistance of several business owners and lab technicians, we have made every effort to complete the project's work. We are also appreciative of our parents' well-wishes for our project's successful completion. We would like to express our gratitude to all of our friends and those who are either directly or indirectly involved in our initiative.

**REFERENCES:**

[1] S. Peerun, N. H. Chummun, and S. Pudaruth, “Predicting the Price of Second-hand Cars using Artificial Neural Networks,” \*The Second International Conference on Data Mining, Internet Computing, and Big Data\*, no. August, 2015, pp. 17–21.

[2] Lessmann, S., & Voß, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. International Journal of Forecasting, 33(4), 864-877.

[3] More, D. R., Dhumal, G. B., & Bhapkar, S. R. VECHICLE RESALE PRICE PREDICTION USING PYTHON VEHICLE RESALE PRICE PREDICATION USING MACHINE LEARNING.

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

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

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

[7] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753-764. [5] Liang Han, Zhaozheng Yin, Zhurong Xia, Minqian Tang, and Rong Jin, “Price Suggestion for

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

[9] Narayana, C. V., Likhitha, C. L., Bademiya, S., & Kusumanjali, K. (2021, August). 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.

[10] Hankar, M., Birjali, M., & Beni-Hssane, A. (2022, May). Used car price prediction using machine learning: A case study. In *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)* (pp. 1-4). IEEE.

[11] Saini, I. S., & Kaur, N. (2023, March). Comparison of Various Regression Techniques and Predicting the Resale Price of Cars. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 857-861). IEEE.

[12] Yu, H., & Wu, J. (2016). Real estate price prediction with regression and classification. *CS229 (Machine Learning) Final Project Reports*.

[13] Shah, K., Shah, H., Zantye, A., & Rao, M. (2021, July). Prediction of rental prices for apartments in Brazil using regression techniques. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 01-07). IEEE.

[14] Rane, M., Patil, M., Rane, N., & Amune, A. (2023, January). Random Forest Regressor Approach for Predicting Resale Value of used Vehicles (RFRVP). In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-6). IEEE.

[15] Hankar, M., Birjali, M., & Beni-Hssane, A. (2022, May). Used car price prediction using machine learning: A case study. In *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)* (pp. 1-4). IEEE.

[16] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.

[17] Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation*, *18*(1), 11..

[18] Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *The Computer Journal*, *41*(8), 537-546.

[19] Du, J., Xie, L., & Schroeder, S. (2009). PIN optimal distribution of auction vehicles system: Applying price forecasting, elasticity estimation, and genetic algorithms to used-vehicle distribution. *Marketing Science*, *28*(4), 637.

[20] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New model for residual value prediction of the used car based on BP neural network and nonlinear curve fit. In *2011 Third International Conference on Measuring Technology and Mechatronics Automation* (Vol. 2, pp. 682-685). IEEE.

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