Proceedings of the Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC-2023)
IEEE Xplore Part Number: CFP23V66-ART; ISBN: 979-8-3503-0009-3

2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-0009-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESC57686.2023.10192988

Predicting the Sale Price of Pre-Owned Vehicles with the Ensemble ML Model

M.Kathiravan
Computer Science and Engineering
Hindustan Institute of Technology and Science (Deemed to be University)
Chennai, India
*mkathiravan@hindustanuniv.ac.in

Ramya M
Computer Science and Engineering
St.Joseph Institute of Technology,
Chennai, India
mramya590@gmail.com

S. Jayanthi
Department of Computer Science and Engineering, Mohamed Sathak aj College of Engineering, Chennai, India
jaya_dhanush@yahoo.co.in

Vangala Vamseedhar Reddy
Computer Science and Engineering
Hindustan Institute of Technology and Science (Deemed to be University)
Chennai, India
kathirskec2009@gmail.com

Lokesh Ponguru
Computer Science and Engineering
Hindustan Institute of Technology and Science (Deemed to be University)
Chennai, India
kathirrec1983@gmail.com

N. Bharathiraja
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
bharathiraja@chitkara.edu.in

Abstract — **Car price forecasting is a popular study topic because it requires a lot of work and knowledge. Used car pricing forecasting is a major auto industry concern. Machine learning can accurately predict used automobile prices based on many characteristics. Many distinct qualities are considered for accurate predictions. The suggested model uses a dataset that contains vehicle brand and model, year of production, mileage, condition, and other factors that affect used car prices. This study used linear regression, GBT regression, and random forest regression to estimate secondhand car prices. Then, algorithm performance was compared to find which method better fit the data set. Thus, these methods outperform others.**

Keywords—*Linear Regression, GBT Regression, Random Forest Regression, Machine Learning.*

I. INTRODUCTION

Predicting car prices is fascinating and common. In 2014, the BiH Office of Statistics recorded 921.456 cars, 84% of which were personal [1]. In 2013, the number of cars rose by 2.7%, and it is expected to rise further in the future. This strengthens auto price predictions. Accurate automobile price forecasting requires expertise because vehicle prices depend on many factors. Most important are make, model, age, horsepower, and mileage. Vehicle prices depend on fuel type and mileage. Most important are make, model, age, horsepower, and mileage.

The type of fuel used in a vehicle and its fuel consumption per mile have a substantial impact on its price [2]. Due to the complexity of the data involved, it is difficult to predict the price of used vehicles for large data sets using machine learning techniques, despite the fact that existing authors perform well on small data sets. One of the primary benefits of machine learning is the ability to manage massive quantities of data and identify complex patterns and relationships between attributes and prices. Moreover, machine learning models can be trained to respond to market changes, such as shifting consumer preferences and supply and demand fluctuations. The accuracy of the extant systems is 54%, 43%, and 40%, according to recent publications. Later, the ensample-based method was implemented, in which additional algorithm combinations were supplied to enhance the model's precision. Using the random forest, Lasso, Ridge, and linear regression algorithms, the artificial neural network model [3] achieved an accuracy of 90%. Similar work was proposed [4] for the Chinese data set, in which the LightGBM model was demonstrated. Multiple regression model [5] performance was compared to that of all regression techniques. The purpose of the case study was to evaluate the algorithm's numerous outputs. Similarly, a literature review [6, 7] utilizing a hybrid machine learning algorithm was conducted.

More flaws exist in the preceding related work. The limitation is that only a small quantity of data was used for the prediction, and while some authors focused on fuel and model, others have focused on a few other variables. The ensemble-based approach, linear regression, and Random Forest with Gradient Boost algorithm are therefore proposed for large data sets to enhance the model's accuracy. Using techniques from machine learning, the proposed system detects used vehicle forecasts with greater accuracy than other algorithms. The proposed model is a useful instrument for estimating the price of a used automobile based on variables such as the make, model, year, mileage, condition, and location. The system is useful for buyers who want to determine the reasonable price of a used car they are considering purchasing; they can input the vehicle's pertinent features into the model, and the model will predict a price based on the patterns it has learned from historical data.

II. LITERATURE SURVEY

Is buying used cars more sustainable? In his master's thesis [8], he showed how an SVM regression model may better predict the cost of a rental car than multivariate regression or simple multiple regression. SVM is less prone to overfitting and underfitting and is better at handling

multidimensional datasets. This study failed to show a difference between basic regression and SVM regression in mean, variance, and standard deviation. Supervised Machine Learning for Used-Vehicle Price Prediction and Classification [9, 10]. multivariate regression analysis to prove that hybrid cars retain value longer. This promotes greater fuel efficiency and stems from environmental concerns about the environment and the climate. Similarly, the article from [11] was that the lifetime cost of driving a car, they considered the following factors: brand, year of manufacturing, and engine type. Similar outcomes were obtained by their prediction model and the straightforward regression model. Also, they created an expert system called ODAV (Optimal Distribution of Auction Vehicles) due to the strong demand for car dealers to sell the vehicles at the conclusion of the leasing year. This method provides information on where to find autos for the best prices as well as the best pricing. The price of an automobile was predicted using a regression model based on the k-nearest neighbors machine learning technique. An in-depth analysis of machine learning methods for forecasting car purchases based on consumer demands. He took into account the brand, expected car life, and kilometers' driven. The proposed model was developed to be able to handle nonlinear relationships in data, which was not possible with earlier models that used straightforward linear regression approaches. Compared to other linear models, the non-linear model has higher accuracy in predicting car prices [12,13]. The article [14] used modified decision trees, Light GBM, and XG Boost regression to predict car prices in Mauritius. The article [15, 16] used k-nearest neighbours, multiple linear regression analysis, decision trees, and nave bayes [19]. The dataset used to create a prediction model [20] was manually assembled from local newspapers over a month because time might affect car prices [21]. He examined the production year, brand, model, cubic capacity, mileage in kilometres, exterior color, price, and transmission type. The author found that Naive Bayes and Decision trees could not classify numerical values [22, 23].

III. SYSTEM ARCHITECTURE

Figure 1 depicts the suggested system. To begin, the raw data set that was acquired for forecasting the cars used samples. Because web scrapers are used to collect data, many samples have few or no properties. These samples are cleaned by a data. that cleans the samples after reading them from a scraped database, then saves the cleaned samples into a CSV file, and finally uses the CSV file to create machine learning models.

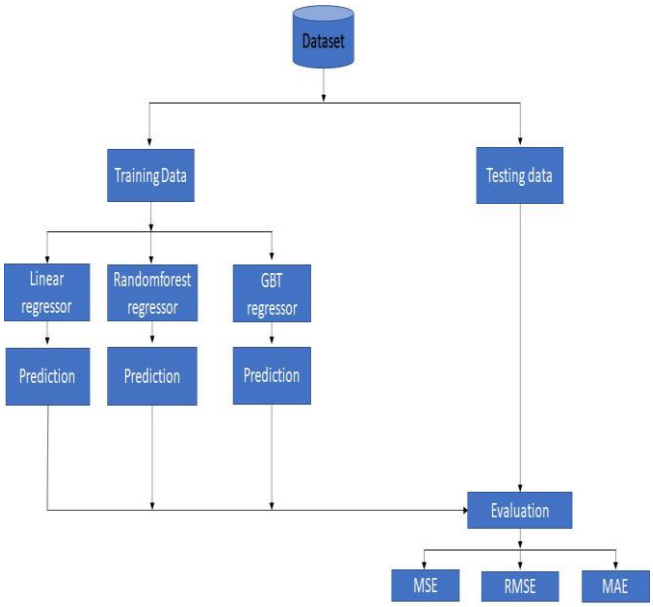

Fig. 1. Architectural Diagram

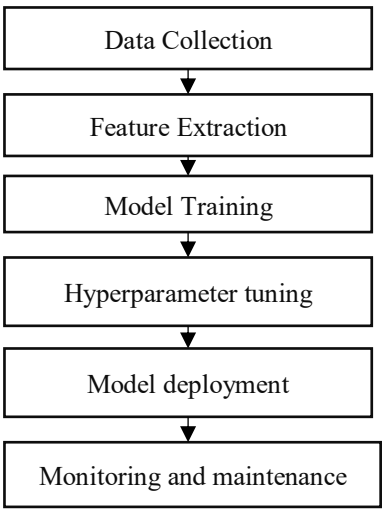

Fig 2: Design Diagram

The entire development of the suggested work is shown in Fig. 2. The process entails the following six actions: First, make sure Python is installed on your computer, and then download the Car Price Data Set. Bring in the necessary components, such as matplotlib, NumPy, etc. The second step is to pre-process the data that was just collected. This requires data cleansing, dealing with missing values, dealing with missing values and eliminating duplication. If not, data normalisation or standardisation will be required to provide each feature with a consistent range of values. Third, decide which features will have the biggest impact on the cost of the used car. The brand, model, year, miles driven, condition, geographical location, and any other pertinent details may fall under this category. Fourth, after settling on a set of features, pick a machine learning approach that can be used to estimate how much a second-hand car will cost. Linear regression, decision trees, random forests, and neural networks are just a few of the well-liked machine

learning methods that may accomplish this. Fifth, if the model's performance is subpar, consider changing its hyperparameters, features, or machine learning algorithms to see if it improves things. The sixth step describes deploying the model to a production setting if its performance has met your expectations. Making an online tool or application programming interface (API) that can estimate the value of a used vehicle based on its characteristics is one option. In order to evaluate something, it is necessary to collect data directly from the source. The used vehicle data collection in CSV format was downloaded from the website. Data is collected on a variety of levels, from an anonymous serial number to details like name, location, mileage, engine type, kilometres driven, and vehicle make and model.

IV. METHODOLOGY

Regression analysis is a statistical procedure that forecasts the future outcomes of a target and an independent variable. It can, for example, be used to determine the relationship between a car's price and its physical characteristics that may be suggestive of long-term viability. A linear regression analysis is carried out in a straight line and makes use of a best-fit line between two variables. The prediction analysis is divided into three sections. We begin by creating a training dataset. The training dataset is then used to fit a model. Finally, we connect the model's inputs to its predictions. To fit a model to a training dataset, we feed it all of the necessary information. As a result, the learning algorithm can figure out how the inputs and outputs relate to one another. A machine learning model can then be used to link the predictions to the model. We can define the input as an array of numbers, for example, one row with two columns, by specifying it as a list of rows with a particular number in each column. The model can be used in an application to directly tie the prediction's outputs and inputs to the given data. This enables us to do more efficient analysis. In our study, all of a car's independent properties, such as its make, number of doors, number of cylinders, manufacturer, and gearbox type, have an impact on the automobile's pricing (dependent variable). The three basic metrics used to evaluate a model are mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE). The MAE, or average error, is the easiest to understand. Although the RMSE is a more widely used metric, the MAE is more difficult to interpret, despite being commonly used to interpret linear models. When developed with square root formulas, it becomes more understandable. When analyzing a model, the RMSE should be the primary statistic. We test our model using all three measures; however, we rely more largely on the RMSE results for interpretation.

A. Algorithms

Linear Regression model: Linear regression is a linear method, also known as dependent and independent variables, that is used in statistics to represent the relationship between a scalar answer and one or more explanatory factors. The former, simple linear regression, is used when there is only one explanatory variable, whereas the latter, multiple linear regression, is used when there are several variables. This statement is more general than multivariate linear regression, which predicts numerous correlated dependent variables rather than a single scalar one.

Random Forest Regression: Random Forest is an ensemble learning-based regression model. It utilizes a decision tree model, namely, as the name implies, numerous decision trees, to produce the ensemble model, which collectively yields a forecast. The advantage of this model is that the trees are generated in parallel and are largely uncorrelated, resulting in good outcomes because each tree is less vulnerable to the unique faults of other trees. Bootstrap aggregation, often known as bagging, offers the randomness required to generate robust, uncorrelated trees, thereby ensuring uncorrelated behaviors. This model was chosen to compare a bagging strategy with the following gradient boosting methods and to account for the dataset's various features.

GBT Regression Model: Gradient boosting, a different decision tree-based method, is frequently referred to as "a way of turning weak learners into strong learners." As with a traditional boosting method, this entails that observations are given varying weights, and depending on specific metrics, the weights of observations that are difficult to forecast are increased and then fed into a separate tree that is being trained. In this case, the metric is the gradient of the loss function. This model was created to take into account non-linear correlations between the attributes and the projected price by splitting the data into 100 sections.

V. IMPLEMENTATION AND RESULTS ANALYSIS

This study investigated one machine learning classifier technique from previous research. Classifier models were created using linear regression, random forest regression, and GBT regression on training (90%) and testing (10%) data sets. "Random forest" (RF), sometimes known as "random decision forest," refers to ensemble approaches. RF solves classification and regression issues. Linear regression models divide input data into two groups using the biggest region between them. This study trains the model for k values 2–10 using data from three ratios. 85% accuracy, 0.28 MAE, 0.39 RMSE Figure 3 shows automobile year, mileage, and engine distributions.

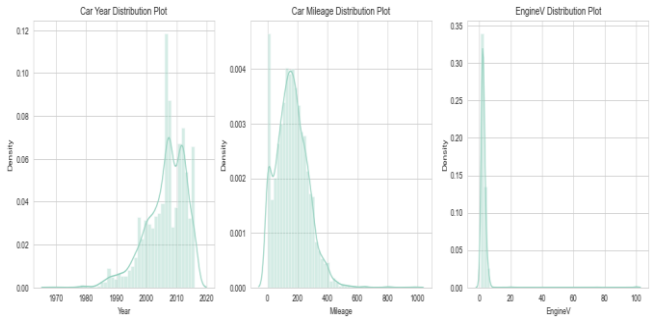

Fig 3: The car and mileage distribution

The importance of each feature, such as engine type, brand, body, engine, registration, and mileage, is presented in Fig. 4.

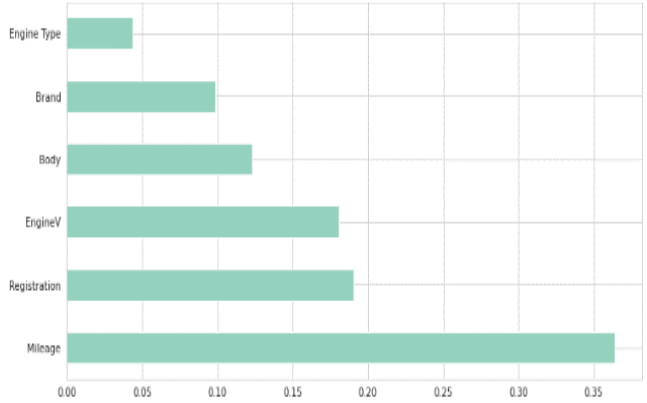

Fig 4: Represents feature importance

To do this, we will use a linear regression model. Linear regression was chosen as the initial model due to its simplicity and short training time. Without using a feature mapping step, we built the feature vectors from the features themselves. Regularisation was not used because the results showed an extremely small standard deviation. Figure 5 shows the values of the assessment metrics used to assess the linear regressor model. The evaluation metrics values for linear regressor model given below.

```
R_squared : 0.7726984972665856
RMSE : 0.42982065395637997
```

By employing a random forest regression technique Our random forest regression model will be trained using the Random Forest Regressor function in the SK-learn package. During training, the random forest algorithm generates many decision trees, each of which produces an average prediction. The results of the random forest regressor model's evaluation metrics are given below.

```
R_squared : 0.8088184799741465
RMSE : 0.3941931582408604
```

Data is given variable weights in a conventional boosting technique, and the weights of difficult-to-predict observations are increased before they are placed in a separate tree for training. In this case, we use the gradient of the loss function as our measure. Values for evaluation metrics were displayed in Figure 8 for the GBT regressor model.

```
R_squared : 0.8124884560131451
RMSE : 0.3903913012080388
```

Below is MAE, MSE, and RMSE values. Gradient boosting is used for classification and regression. It uses decision trees or other inefficient prediction algorithms to predict.

```
MAE: 0.2881124551546641
MSE: 0.15240536805890567
RMSE: 0.3903913012080388
```

These methods have also yielded encouraging percentile and generalisation results for a variety of datasets.

VII. CONCLUSION AND FUTURE ENHANCEMENTS

There are several things to consider when estimating car prices. Prediction requires data collection and preparation. This study normalized, standardized, and cleaned data to remove noise for machine learning algorithms. In complex data sets like this study's, data cleansing is insufficient to improve prediction performance. The data set's accuracy was 50% using a single machine algorithm. Thus, an ensemble of machine learning algorithms has been presented, which improves accuracy over a single method. Despite the system's success in the automotive price prediction challenge, we want to see how it performs on other data sets. OLX and eBay's used car data sets will improve our test data and validate the suggested methods.

ACKNOWLEDGMENT

Sincere thanks to Department of Computer Science and Engineering, Hindustan Institute of Technology and Science (Deemed to be University), Padur, Kelampakkam, Chennai, Tamil Nadu, India.

REFERENCES

[1] R. Dhaya “Flawless Identification of Fusarium Oxysporum in Tomato Plant Leaves by Machine Learning Algorithm” Journal of Innovative Image Processing (JIIP) 2, no. 04, pp. 194-201, 2020.

[2] C. Jin, "Price Prediction of Used Cars Using Machine Learning," *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, Chongqing, China, 2021, pp. 223-230.

[3] M. Kathiravan, S. Manohar, R. Jayanthi, R. Dheepthi, R. V. Sekhar and N. Bharathiraja, "Efficient Intensity Bedded Sonata Wiles System using IoT," *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2023, pp. 1360-1364.

[4] H. Zhang, "Prediction of Used Car Price Based on LightGBM," *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, Wuhan, China, 2022, pp. 327-332

[5] J. Varshitha, K. Jahnavi and C. Lakshmi, "Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning," *2022 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2022, pp. 1-4.

[6] S. Kumari and M. Kathiravan, "A Tour to Exact Visualization of Search Engine in Unique Boundaries Facilitating User Expectations to

Usage," *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2022, pp. 76-81

[7] M. Hankar, M. Birjali and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC)*, El Jadida, Morocco, 2022, pp. 1-4

[8] P. Ponmalar P and A. Christinal C, "Review on the Pre-owned Car Price Determination using Machine Learning Approaches," *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2022, pp. 274-278

[9] Deger A, Yusuf Yaslan and Mustafae .Kamasak," Emotion Based MusicRecommendationSystemUsingWearablePhysiologicalSensors", IEEETransactionsOnConsumerElectronics,Vol.14,No.8, May2018.

[10] M. Kathiravan, R. Logeshwari, S. Pavithra, M. Meenakshi, V. Sathya Durga and M. Vijayakumar, "A Cloud based Improved File Handling and Duplicate Removal using MD5," *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2023, pp. 1532-1536

[11] Krupa K S, Ambara G, Kartikey Rai, Sahil Choudhury, "Emotion aware Smart Music Recommender System using Two Level CNN", Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020), IEEE Xplore Part Number: CFP20P17-ART; ISBN:978-1-7281-5821-1.

[12] S. UmaMaheswaran, G. Kaur, A. Pankajam, A. Firos, P. Vashistha, V. Tripathi, and H. S. Mohammed, "Empirical analysis for improving food quality using artificial intelligence technology for enhancing healthcare sector," Journal of Food Quality, vol. 2022, 2022.

[13] B. Kaur and G. Kaur, "Heart disease prediction using modified machine learning algorithm," in International Conference on Innovative Computing and Communications. Springer, 2023, pp. 189–201.

[14] Ankita Mahadik, Prof. Vijaya Bharathi Jagan, Shambhavi Milgir, Vaishali Kavathekar, Janvi Patel," Mood based Music Recommendation System",International Journal of Engineering Research &Technology(IJERT)ISSN:2278-0181Vol. 10Issue06,June-2021.

[15] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems, The Adaptive Web: Methods and Strategies of Web Personalization, pp. 291-324, 2007.

[16] M. Kathiravan, M. P. K. Reddy, M. Malarvel, A. Amrutha, P. H. Reddy and S. Kavitha, "IoT-based Vehicle Surveillance and Crash Detection System," *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 2022, pp. 1523-1529

[17] Z. Hyung, J. S. Park, and K. Lee, "Utilizing context relevant keywords extracted from a large collection of user-generated documents for music discovery," Info. Processing and Management, vol. 53, no. 5, pp. 1185-1200, 2017.

[18] Chanda, Mona Lisa & Levitin, Daniel. (2013). The neurochemistry of music. Trends in cognitive sciences. pp. 179-93.

[19] Bharathiraja, N., Shobana, M., Manokar, S., Kathiravan, M., Irumporai, A., & Kavitha, S. (2022). The smart automotive webshop using high end programming technologies. In Intelligent Communication Technologies and Virtual Mobile Networks: Proceedings of ICICV 2022 (pp. 811-822). Singapore: Springer Nature Singapore.

[20] Pradeepa, K., Bharathiraja, N., Meenakshi, D., Hariharan, S., Kathiravan, M., & Kumar, V. (2022, December). Artificial Neural Networks in Healthcare for Augmented Reality. In 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1-5). IEEE.

[21] Murugesan, S., Bharathiraja, N., Pradeepa, K., Ravindhar, N. V., Kumar, M. V., & Marappan, R. (2023, March). Applying Machine Learning & Knowledge Discovery to Intelligent Agent-Based Recommendation for Online Learning Systems. In 2023 International Conference on Device Intelligence, Computing and Communication Technologies,(DICCT) (pp. 321-325). IEEE.

[22] Bhaskaran, S., Bharathiraja, N., Pradeepa, K., Kumar, M. V., Ravindhar, N. V., & Marappan, R. (2023, January). New Recommender System for Online Courses Using Knowledge Graph Modeling. In 2023 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[23] M. Kathiravan, S. J. Parvez, R. Dheepthi, R. Jayanthi, S. Gowsalya and R. V. Sekhar, "Analysis and Detection of Fake Profile Over Social Media using Machine Learning Techniques," *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2023, pp. 1164-1169.