International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Vehicle Price Prediction System using Machine Learning

Shorabh Gautam
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
20BCS1447@cuchd.in

Pranshu Kumar
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
20BCS1543@cuchd.in

Aniket Jaswal
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
20BCS1636@cuchd.in

Harshit Pant
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
20BCS7140@cuchd.in

Aditya Narayan Sharma
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
20BCS1460@cuchd.in

Monika
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India
Monikabishnoi3@gmail.com

Abstract— With the evolution of the world and working methods where people are needed to be present at their working locations at a moment's notice, a lot of people have now become heavily reliant on personal vehicles to suit their needs. But not everyone can afford to buy a new vehicle for themselves and therefore people are looking forward to buying used cars. But buying a worthy used car for the right price can be quite hectic and difficult. There exist a large number of attributes that must always remain under consideration so that the resulting inductions about the price of a used car are reliable and accurate. Prices depend on attributes such as the vehicle's model, year of manufacture, distance traveled, number of times the vehicle was serviced, color, interior, number of previous owners, etc. Owing to a large number of factors, Vehicle Price Prediction System becomes need of the hour so that one can effectively determine the true value of a used car. This research paper has developed a well-defined model which makes use of different machine learning algorithms like AdaBoostRegressor, MLP Regressor, Decision Tree Regressor, Random Forest Regressor, ExtraTrees Regressor, etc to make reliable predictions. This research paper has also designed a user interface that inputs data from a user and will go on to display the price of a car as per past values.

***Keywords*— Vehicle Price prediction, Decision Tree Regressor, Random Forest, AdaBoostRegressor, MLP Regressor**

I. INTRODUCTION

The market of used vehicles has seen a major boom in recent times due to the growing dependence on personal vehicles and the inability of a lot of people to buy new vehicles to suit their needs. This has given rise to a uniform and fair method of deduction of the price of a used car. Given the large variety of factors that are crucial in influencing a used car's market pricing, ensuring whether the asked price stands true is difficult. This requires a variety of distinct features and considerations for estimating a vehicle's price. The brand and model of the car, along with its mileage and fuel type, are its most important features in terms of determining its price. The price of a vehicle is significantly impacted by these characteristics. During the writing of this research paper, to determine as to which machine learning algorithm performs the best overall, this research has tried to compare the performance of an algorithm to other machine learning algorithms which include MLPRegressor, ExtraTreeRegressor, Random ForestRegressor , Decision Tree Regressor AdaBootRegressor as well. This research paper divided the working of a model into 3 parameters namely: Root Mean Square(RMS) Value, accuracy on training and testing data and calculated the records for each individual model and compared the records with other models. From this research, a conclusion can be derived that a Decision Tree based model is the best because of its higher accuracy and lower root mean square value suited technique in order to achieve higher precision in the predicted price of a used car's selling price.

II. LITERATURE SURVEY

Based on the research of author **[1]** ,he uses the provided Kaggle data set price based prediction on a used-car from the current value to its actual value. After going through different test train phases theauthor finalizes different classification methods like Random Forest, Logistic Regression, SVM Decision Tree and Extra Trees. After going through different combinations of algorithms, the author finds the best 2 algorithms which complement each other andare best in their prediction category.

Work done by **[2]** provides a pretty informative introductory paper on SVM also known as Support Vector Machine. By going through algorithms such as K-NN and rule-based classifiers while comparing them with SVM, the author is able to evaluate the effectiveness of these techniques. Meanwhile several different types of data sets chosen from the UCI Machine Learning Repository and other public Repositories were used for this comparison study, with this evaluation it is proven that SVM has to offer significantly higher classification accuracy comparatively to others.

Work by author **[3]** predicts the price of used cars in by employing these ML techniques: multiple linear regression, K-NN, Bayes classifier, and decision trees. The author makes use of historical information gathered from

Mauritius's daily newspapers. Applying the stated learning algorithms to this data yields results that are equivalent but with less accurate predictions. The primary distinction between categorizing price range and spam email, however, is that our motivation is primarily one-vs-the-other, whereas the goal of categorizing spam email is binary.

The study of author **[4]** can be used to employ machine learning algorithms on the basis of different sorts of attributes that define a vehicle like miles driven, transmission type, year of manufacture, the type of fuel needed, its mileage etc.

Work by author **[5]** The authors built a data-driven model to forecast used car pricing based on characteristics like manufacture year and mileage using data mining from a Croatian online vendor.

To forecast price patterns and evaluate accuracy, they used supervised machine learning techniques, particularly linear regression and classification algorithms.

Works of author **[6]** and **[7]** depict that algorithms that are being used like Decision Tree, Random Forest Regressor, MLP Regressor and AdaBoostRegressor in a vehicle price prediction system have proven to be a reliable and accurate method for predicting the valueof a used car. As technology continues to evolve, users can expect to see more sophisticated algorithms and methods being developed to tackle this problem in the future.

III. PROPOSED METHODOLOGY

This research paper proposes a methodology of using Machine Learning Algorithms to predict the prices of second-hand cars basedon the number of features mentioned below.

Step 1 This research paper collects the dataset for second-hand cars and further identifies the crucial factors that reflect the price.

Step-2 Then take off the entries with NA values resulting in discarding the features that are not relevant for predicting the price.

Step 3 Then the data is processed and ML Algorithms are applied such as Decision Tree, Random Forest, MLP Regressor, and many

more.

Step 4 Then the best algorithm with the highest accuracy is selected and uses the same for predicting the price of the car.

A. System Design

As id depicted in the Figure 1, the First process starts by gathering the dataset. Further, the next step is to Perform Data Preprocessing. Then various machine learning algorithms are applied which involveExtraTreesRegressor, AdaBoostRegressor, MLP Regressor, RandomForestRegressor, and DecisionTree Regressor. The best Algorithm having the least Root Means Square (RMS) value and maximum accuracy is selected. At the end the predicted price is displayed to the user in accordance with users input.

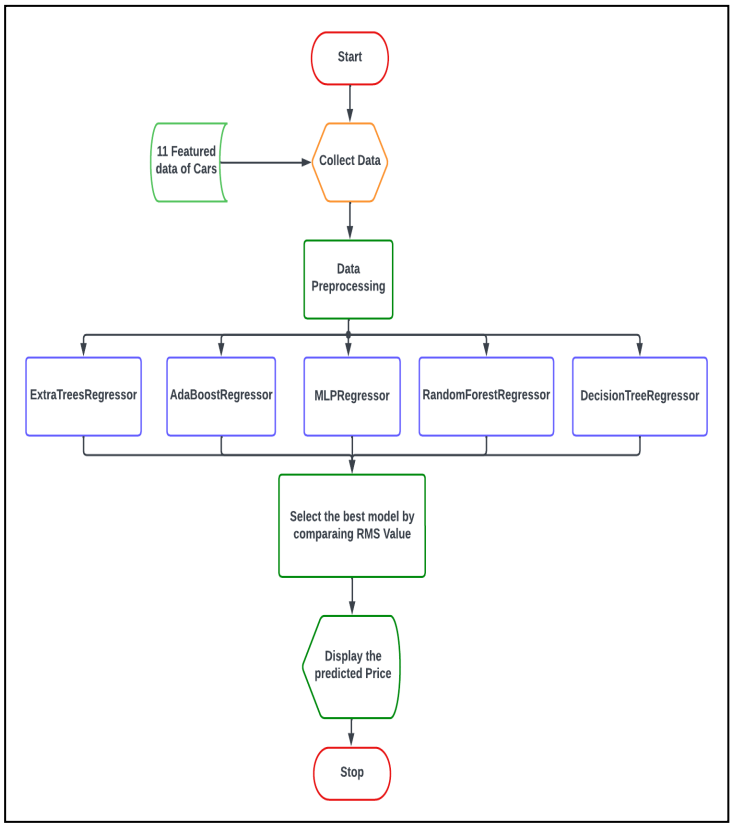


Fig. 1 System Flowchart

IMPLEMENTATION

A. DATASET COLLECTION AND DESCRIPTION

Collection of Dataset:- The dataset is being scrapped from different well-known second-hand car trading websites. This includes car247.com, autotrader.com, cardekho.com , carwale.com, cartrade.com and also from kaggle.com which is a dataset repository. Approximately 6000 entries were collected **[8].**

Description of the Dataset:- The dataset consists of 11 order-independent features and 5975 instances. A sample of the dataset can be obverse in Table I and II

Table – I

Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price
Manual	First	26.60	998.0	58.16	5.0	1.75
Manual	First	19.67	1582.0	126.20	5.0	12.50
Manual	First	18.20	1199.0	88.70	5.0	4.50

Table - II

Features are described as:-

1. Name(model): Full model name of the vehicle with their respective manufacturer name. (string)
2. Location: City from where the vehicle was purchased all cities are in India. (string)
3. Year: The year in which Vehicle was purchased. Date(yyyy)
4. Kilometer_Driven: Total kilometer vehicle is being driven. (integer)
5. Fuel_Type: Different type of fuel which is used in the vehicle. There are four categories of fuel in this particular dataset.
   - Petrol
   - Diesel
   - CNG
   - LPG
6. Transmission: This indicates the type of transmission which is used in the vehicle.
7. Owner Type: This indicates if the vehicle is first-handed or second or so on.
8. Mileage: Mileage of vehicle in km/kg.
9. Engine: capacity of engine in cc.
10. Power: Power of engine in bhp.
11. Seats: No passenger seats are available.
12. Cars: Full name of the car.

After cleaning the dataset using the panda's library Only 5975 entries were left. Then the Cleaning of the dataset is performed which includes the removal of null values, noise data, and outliers. All this removal process is done through the pandas library using Python. **[9].**

	Cars	Location	Year	Kilometers_Driven	Fuel_Type
0	Maruti Wagon	Mumbai	2010	72000	CNG
1	Hyundai Creta	Pune	2015	41000	Diesel
2	Honda Jazz	Chennai	2011	46000	Petrol

B. *ANALYSIS OF DATASET*

EDA (Exploratory Data Analysis) is done using the following scatterplot and boxplot that were generated through MatplotLib andthe Seaborn library **[10].**

Figure 3 depicts the scatterplot that obtained where the X-axis indicates power and Y-axis indicates the price, and the fuel type is marked with different colors. It can be seen, that petrol-powered vehicle has power in a range of 50-200 bhp and are sold from 1,50,000 INR to 5,00,000 INR. These vehicles are in the medium (4-seater) range which most second-hand buyers can afford. And In the case of diesel-powered vehicles power ranges from 100-350 bph with little bit higher segment of 2,00,000 INR to 6,50,000 INR and with increasing power prices increase more rapidly than petrol- powered vehicles.

The box-plot in Figure 4 also shows that the petrol engine with manualtransmission is the cheapest diesel engine with the automatic engine beingexpensive and the petrol engine being cheaper than a diesel engine. And According to the given data, diesel-powered vehicles are more likely to be purchased than petrol.

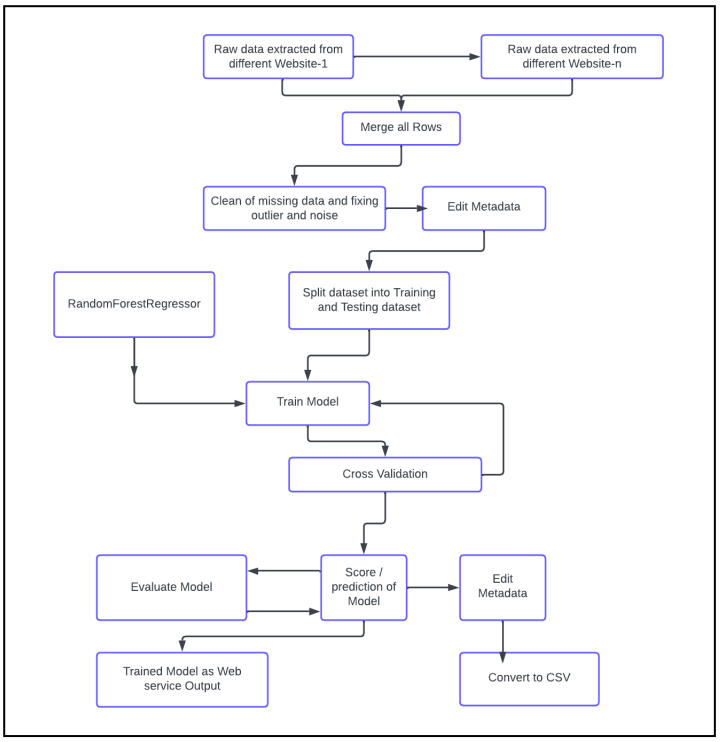


Fig – 2. IMPLEMENTATION FLOWCHART

Summarization of steps of implementation flowchart shown in Figure 2 is given below with an explanation:

Combine Dataset: This module takes all datasets collected with different data endpoints and normalizes them for similar structures and merges them all as one dataset.

Clean Missing Data: This module filters all missing value entries.

Edit Metadata: With this, the datatype of all attributes can be corrected and presented by the correct data type object.

Normalization of Data: Numeric attributes like total kilometers or prices are normalized for better visualization.

Split Data: This module splits the whole data into 70% and 30 % for training and testing models respectively.

Training Model: Using Python language and RandomForestRegressor implemented by libraries Keras, and scikitlearn.

IV. ALGORITHMS USED

DecisionTreeRegressor:- A decision tree Regressor is a well-known supervised machine learning algorithm that is used for regression problems. (Predicting prices or any continuous dependent variable) The model uses values and mean squared errorfor decision accuracy. This is the reason it is not good in generalization and is very sensitive to the change in training data. Even small changes in data can lead to a drastic effect on predictive accuracy [**11**].

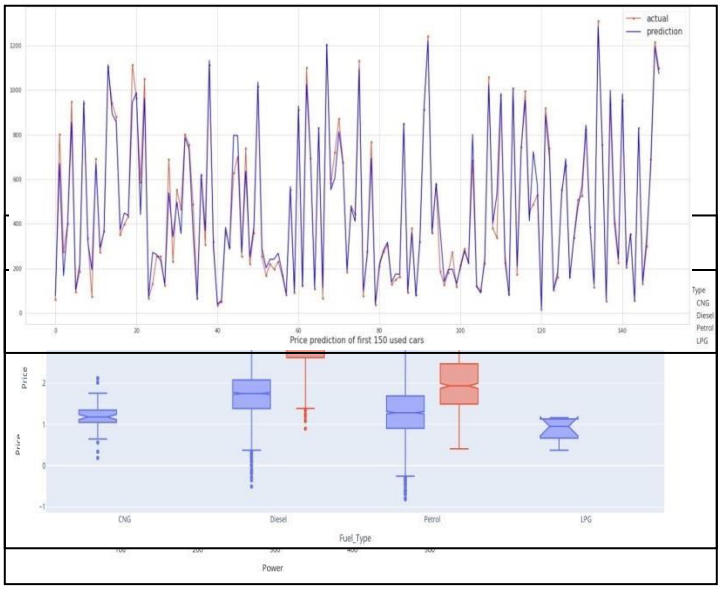
*MLPRegressor;-*Multi-layer perceptron is a feedforward artificial neural network that can be used to perform regression. This model is characterized by several layers of input nodes (where each layer can be any number of dimensions) connected as a directed graph with an output layer (dependent variable). It uses backpropagation for training the network model. Preerun, Saamiyah & Henna Chummun showed that the neural network produces the least possible mean absolute error among machine learning algorithms when it contains one hidden layer with two outgoing nodes.[**12**].

RandomForestRegressor:- The RandomForestRegressor is an prediction based machine learning method that is used to obtain various different decision trees in training set and output theirmean prediction. As demonstrated by Jehad Ali, Rehanullah Khan & Nasir Ahmad in their work, Random Forest is best suited for the type of dataset being dealt with. They achieve an accuracy of 96.13
% in prediction, about 29 % more accurately than other machine algorithms.[**13**].

*ExtraTreesRegressor:-*ExtraTreesRegressor is a supervised machine learning algorithm that uses decision trees. Random forest picks up the optimum split while extra trees regressor chooses it randomly. This research paper employed. a decision tree in order to deduce the prices on the given dataset. The results were very fruitful and yielded results with a success rate of 86.422%.

*AdaBoostRegressor:-*AdaBoostRegressor is used to boost the performance of ML models by merging multiple weaker classifiers to obtain a culminated better suited classifier. The

weaker classifiers in AdaBoostRegressor are decision trees having a single breakdown known as decision stumps.



AdaBoostRegressor algorithm is based on binary classification problems, it is one of the reasons it is being used it for the classification of different features used to predict the price of cars.

V. RESULT AND DISCUSSION

Model evaluation parameters: This research paper has used the following three parameters on actual price and predicted price to see how different algorithms works under the same dataset:

- Root Mean Squared Error
- Accuracy on training set
- Accuracy on testing set

model	Root Mean Squared Error	Accuracy on Traing set	Accuracy on Testing set
MLPRegressor	205.550645	0.689850	0.648143
AdaBoostRegressor	148.177027	0.832878	0.817152
DecisionTreeRegressor	115.421645	0.999993	0.889056
RandomForestRegressor	84.069592	0.991793	0.941142
ExtraTreesRegressor	80.262533	0.999993	0.946352
XGBRegressor	74.815814	0.994635	0.953386

Fig – 3. Accuracy of models

Also, a Comparison of the actual price and predicted price is depicted in Figure6 line chart for the first 150 entries.

Model ComparisonUpon calculation of aforementioned parameters for the used models, the following results were obtained:

Fig – 4. Scatterplot of dataset: Price vs Power

Fig – 5. Box-Plot of dataset: Price vs Fuel Type

Fig – 6. LineChart of Actual Price vs Prediction Price of 150 cars

Detailed analysis of the various algorithms and techniques that are being employed in the experiment goes on to show that RandomForestRegressor shows the lowest Root mean square error and the mean absolute error for our collected dataset. Henceforth this study can conclude that **RandomForestRegressor** is best suited for this study than other regression models.

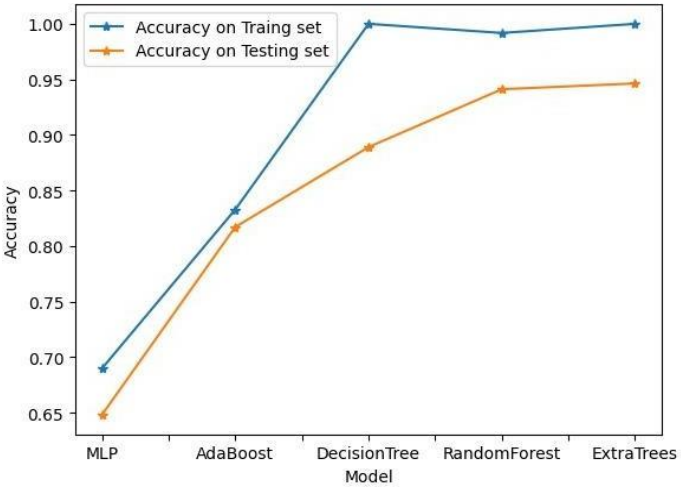


Fig 7- Model Comparison

In the above-mentioned graph, there are two lines representing accuracy of 5 different machine learning models based on their training and testing data. Although ExtraTree and RandomForest model shows the highest accuracy but RandomForest proves to be more stability over ExtraTree during training and testing data. As a result**, RandomForest** is selected for predicting the price of second-hand vehicles.

VI. CONCLUSION

Predicting car prices can be a daunting task as there are many features to consider for an accurate estimate. Data collection and forecasting are important steps in forecasting. In conclusion, a vehicle price prediction system can be a valuable tool for both buyers and sellers in the automotive industry. By

using machine learning algorithms to analyses historical sales data, vehicle features, and market trends, the system can provide accurate predictions of vehicle prices, helping buyers make informed purchasing decisions and sellers set appropriate prices for their vehicles. One must note that such a designed system may not always be perfect and may have limitations based on the quality of the data inputted and the accuracy of the algorithms used. Additionally, other factors such as the condition of the vehicle and the specific location where it is being sold may also impact the final price.Overall, a vehicle price prediction system can be a useful tool for those involved in the automotive industry, but it should not be the only factor considered when buying or selling a vehicle. Expert opinions, personal preferences, and individual circumstances should also be considered.

VII. FUTURE SCOPE

In the future, such machine learning models will link to many websites that can provide real data for price prediction. Users will also be able to add more historical data on car prices that can help improve the accuracy of machine-learning models.One can create an Android application as a UI to interact with the user. To achieve better performance, this paper plans to develop a deep learning model for communication integrity, use adaptive learning, and train data sets instead of whole data. used car market is booming, with buyers having a wide range of options, easy financing, convenient digital distribution channels, and a growing preference for personal mobility in the COVID-19 era. of attributes that need to be considered for accurate forecasting.Future work will collect more data and use more advanced techniques related to electric and internal combustion ngine vehicles. This paper can further be modified to use fitting AI calculations Kaggle dataset to estimate the cost of selling a used car.

REFERENCES

[1] S.E.Viswapriya, Durbaka Sai Sandeep Sharma and Gandavarapu Sathya kiran, “Vehicle Price Prediction System using SVM Techniques” (IJITEE-June 2020) https://www.ijitee.org/wp-content/uploads/papers/v9i8/G5915059720.pdf

[2] Durgesh K. Srivastava and Lekha Bhambhu, “Data Classification Method using Support Vector Machine”, (IRJET-April-2019 https://www.academia.edu/39681082/Automobile_Resale_System_Using_Machine_Learning

[3] Sameerchand Pudaruth, University of Mauritius, “Predicting the Price of Used Cars using Machine Learning Techniques”; 2014. https://www.researchgate.net/publication/319306871_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques

[4] Prashant Gajera *1, Akshay Gondaliya*2, Jenish Kavathiya*3 “OLD CAR PRICE PREDICTION WITH MACHINE LEARNING” March 2021 https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf

[5] Lucija Bukvić *,Jasmina Pašagić Škrinjar,Tomislav Fratrović andBorna Abramović “Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning” https://www.mdpi.com/2071-1050/14/24/17034

[6]Sayed Erfan Arefin, “Second Hand Price Prediction for Tesla Vehicles” ; January 2021 https://www.researchgate.net/publication/348403109_Second_Hand_Price_Prediction_for_Tesla_Vehicles

[7] Xiaona Song,”The AdaBoost algorithm for vehicle detection based on CNN features”; (August 2015) https://www.researchgate.net/publication/301372427_The_AdaBoost_algorithm_for_vehicle_detection_based_on_CNN_features

[8] Rashi Desai ‘Data inclusion and which data is important’ (Towards Data Science, September 2019) https://towardsdatascience.com/how-important-is-data-for-your-business-c15a35c6935e

[9] Wes Mckinney, “pandas: a Foundational Python Library for Data Analysis and Statistics”; (January 2011) https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_and_Statistics

[10] Arnav Oberoi, ”Visualizing data using Matplotlib and Seaborn libraries in Python for data science” , (March 2019) https://www.ijsrp.org/research-paper-0319.php?rp=P878342

[11] Wei-Yin Loh, “Classification and Regression Trees” ,(Research_Gate-January- 2021 . https://www.researchgate.net/publication/227658748_Classification_and_Regression_Trees

`[12] Yongfei Qin, [1] Chao Li, [1] Xia Shi, [1] and Weigang Wang , “MLP-Based Regression Prediction Model For Compound Bioactivity“ (Jul 2022) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9326362/

[13] Gerard Biau, “Analysis of Random Forest Model” (Journal of Machine Learning Research 13 (2012) https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf