Asian Journal of Engineering and Applied Technology
ISSN: 2249-068X (P) Vol.12 No.1, 2023, pp.35-39
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajeat-2023.12.1.3637

Car-Economics: Forecasting Prices in the Pre-Owned Market Using Machine Learning

Kandugula Sadhvik[1], Karamchedu Dhanush[2], Seelaboyina Jayaditya[3] and B. Ravinder Reddy[4]
[1,2&3]Undergraduate Scholar, [4]Assistant Professor,
Department of Information Technology, Sreenidhi Institute of Science and Technology, Telangana, India
E-mail: 19311a12d7@sreenidhi.edu.in, 19311a12g9@sreenidhi.edu.in, 19311a12h0@sreenidhi.edu.in, ravinderreddyb@sreenidhi.edu.in

(Received 5 April 2023; Revised 28 April 2023; Accepted 7 May 2023; Available online 16 May 2023)

Abstract - **The purchase price of a new car and a few additional costs are determined by the company that makes the vehicle. Used vehicle sales are increasing globally as a result of rising new car prices and people's unwillingness to afford them. This strategy is successful since the vendor often sets the price impulsively and the buyer normally isn't aware of all the features and the car's market value. According to research, figuring out a used car's fair market value is both difficult and important. Consequently, it is necessary to create a precise method for estimating used automobile cost. In this case, machine learning prediction approaches could be helpful. The model has to be trained on the massive dataset using techniques like random forest and decision tree before it can be used. Our primary objective is to develop a model that, given the information provided by the client, can accurately and dependably anticipate the selling price of a used car. In this project, our team designed a stunning User Interface (UI) that asks customers for comments and provides pricing estimates. The database stores both the user inputs and the predicted cost of the automobile.**

Keywords: **Machine Learning, Decision Tree, Random Forest, Database, Python**

I. INTRODUCTION

In decision-making processes, machine learning, which mainly deals with regression and classification problems, may be extremely useful. Regression and classification algorithms both aim to analyse unlabeled data through training. However, they are distinctive. The end result of the prior step allows us to categorise the data into different categories and then determine which category the newly received data belongs to. The latter's output is continuous, which enables us to calculate the probability of an event happening after taking consideration of the past. Regression analysis is used to forecast used automobile prices.

The manufacturer specifies the market price of a new automobile, with taxes imposed by the government compensating certain extra expenses. Customers may be convinced that their investment in a new automobile will be beneficial as an outcome. However, used automobiles are on the rise globally as an outcome of new car prices increasing and consumers no longer having the ability to afford to gather them. It involves an approach where a seller chooses a price at random while the buyer is unaware that the automobile and its current market value. In reality, he has no clue who the seller is or what the automobile should be sold for. We have developed an approach that will be highly efficient in dealing with this issue. Regression methods are utilised because they provide continuous results rather than categorical ones. Consequently, it will be easy to predict a car's true price rather than its price range.

We are able to generate accurate used car estimates using the data on our own platform. There are also those who designed a user interface that exhibits a car's cost determined on information provided by any user and seeks conclusions from any user. In order to access the transaction history, the user enters the used car's price, which is eventually stored in the database. In this procedure, both buyers and sellers offer an affordable price. Brand, Year, Fuel, Transmission, Engine, Max power, Seats, Mileage, Kilometres Driven, Owner, Seller Type, and Selling Price are among the details included in the dataset, which consists of 7887 entries.

II. METHODOLOGY

This system creates a machine learning model that takes the input parameters and factors in when estimating the price of the used car by taking into aspects like price, engine size, colour, advertisement date, the number of views, mileage in kilometres, power steering, alloy rims, gearbox, type of engine, registered city, city, version, model, make, and model year [1].

Since there are so many factors to take into consideration, estimating the cost of an automobile can be difficult. Important elements in the prediction process include data collection and analysis. To reduce unneeded noise for machine learning algorithms, clean, normalise, and arrange the data using PHP scripts was done for the goals of this study. Only around half of the data collection can be accurately analysed by a computer algorithm. It is advised to combine numerous machine learning methods since they produce an accuracy of 92.38%. This is a considerable advancement using only one machine learning technique. The disadvantage of the suggested approach is that it consumes significantly more computer resources than a single machine learning algorithm [2].

They seek to utilise the same statistical approach to create models that may be used in different countries. Their exact application makes it easier to understand and shows if any factors that influenced automobile prices had an impact on their results.

It also aligns with their understanding of the traits and data observations on the target variable. Pricing, which the estimator utilises to show information, has a favourable skew since most automobiles are cost-effective [3].

In contrast to multivariate regression or any other kind of simple multiple regression, an SVM-based regression model may be used to predict rental vehicle prices more precisely. This is because Support Vector Machine (SVM) is less prone to overfitting and underfitting and is better at managing datasets with more dimensions. This model's imperfections is that it was unable to demonstrate how to go from basic SVM regression to simplified SVM regression using basic statistics like mean, variance, and standard deviation [4].

To anticipate the price of a used car, they offered a model that would be created using ANNs (Artificial Neural Networks) [5]. Plenty of variables were taken into account, including markers, predicted vehicle life, and overall mileage. In contrast with previous techniques that used standard linear regression techniques, the new model was developed to deal with nonlinear data interactions. The non-linear model outperformed traditional linear models in terms of more precise automobile price predicting. Their projection model's outcomes matched those of the simple regression model. Due to the significant demand for this service, dealers have developed a special scheme called ODAV [6] (Optimal Distribution of Auction Vehicles) to distribute the automobiles at the conclusion of the lease year.

They employed naive Bayes, k-nearest neighbour, multiple linear regression, and decision trees to anticipate the value of used automobiles in Mauritius. Although there were fewer automobiles to be found in this research, they still get to the conclusion that decision tree and naive Bayes are unsuccessful for variables with continuous values [7].

III. MODELING AND ANALYSIS

The recommended strategy provides the most feasible price for used cars. To improve forecast accuracy, we'll integrate multiple approaches. The used car dataset is received by the system, which then prepares it for analysis. We still utilise feature extraction to obtain the data, regardless of how well the dataset is organised, how big it is, or how many ineffective attributes it contains.

Once learned, the dataset is then given to a machine learning technique that combines a decision tree or random forest. Both the predicted value and the user-provided input remain intact in the newly constructed file system. There are 7887 aspects in the entire set. Owner, Brand, Model, Year, Selling Price, Miles Driven, Fuel, Seller Type, and Fuel are some of the parameters.

TABLE I DATASET

Brand	Year	Fuel	Transmission	Engine	Max_Power	Seats	Mileage	KM_Driven	Owner	Seller_Type	Selling_Price
Maruti Swift Dzire VDI	2014	Diesel	Manual	1248	74.00	5	23.40	145500	First Owner	Individual	450000
Skoda Rapid 1.5 TDI Ambition	2014	Diesel	Manual	1498	103.52	5	21.14	120000	Second Owner	Individual	370000
Honda City 2017-2020 Exi	2006	Petrol	Manual	1497	78.00	5	17.70	140000	Third Owner	Individual	158000
Hyundai i20 Sportz Diesel	2010	Diesel	Manual	1396	90.00	5	23.00	127000	First Owner	Individual	225000
Maruti Swift VXI BSII	2007	Petrol	Manual	1298	88.20	5	16.10	120000	First Owner	Individual	130000

The proposed system's functioning is shown in Fig. 1. Preprocessing is the stage of the procedure that is most essential. The objective is to clean up the data so that a prediction algorithm may utilise it. There are two components to the model construction guidelines.

Every so often, the dataset is divided into training and testing sets, each of which contains 80% and 20% of the overall data. The training data is subsequently sent to the decision tree and random forest algorithms. Both the decision tree and the random forest are given the test set during the testing phase to determine their accuracy.

The more sophisticated model is used. A pickle file has been created for the provided model in anticipation of future implementations. Pickle supports binary protocols for the purpose of serialising and converting a Python object structure.

A Python object is pickled into a byte stream, and that byte stream is then unpickled back into a Python object (from a binary file or anything else that looks like bytes).

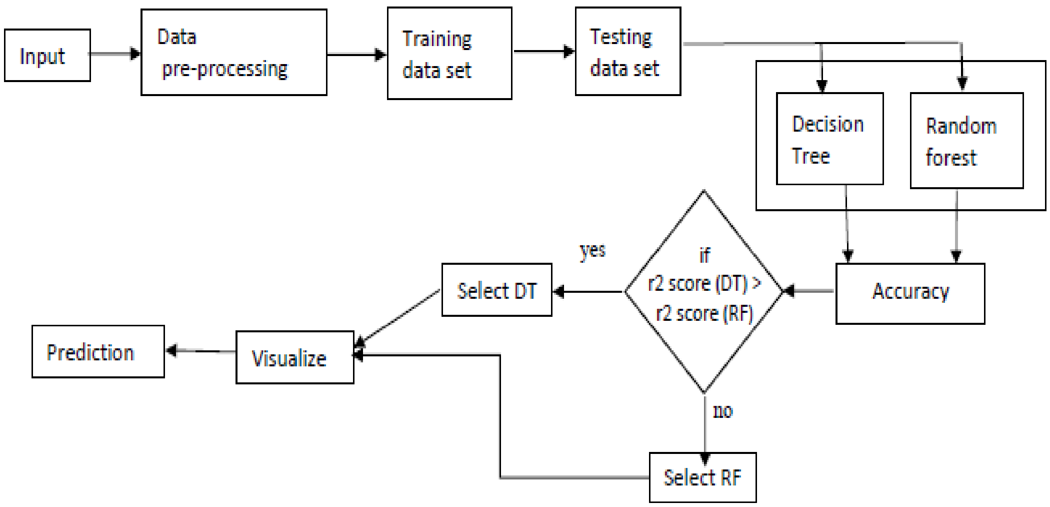
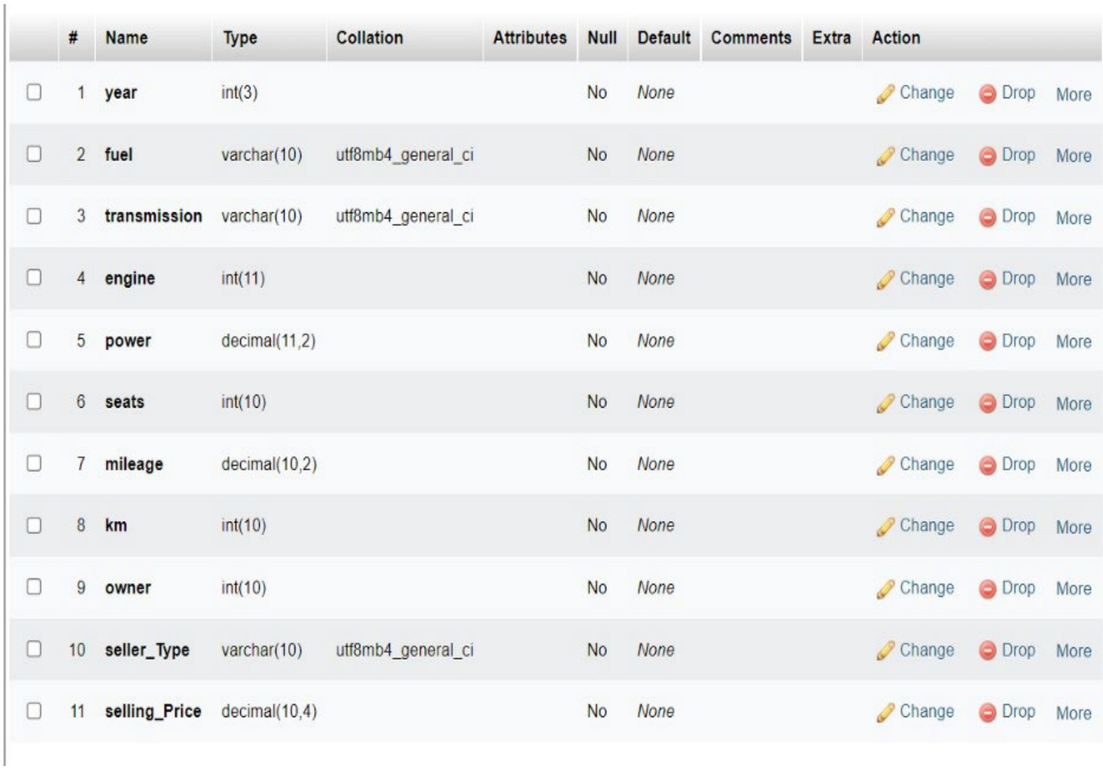


Fig. 1 The architecture of the proposed system

The HTML and pickle files are connected by a Python script to generate a webpage. Flask is a web framework used by Python. In order to create dynamic HTML pages, Flask makes use of the Jinja template engine and well-known Python concepts like variables, loops, lists, etc. HTML first incorporates aesthetic choices for the content before creating the semantic structure of a web page.

Data given by the user as well as the projected value that fits it have been recorded in a file system made by XAMPP. The values are kept by XAMPP in a database. Both the user-provided information and the predicted values are present in the relational database table that was built.

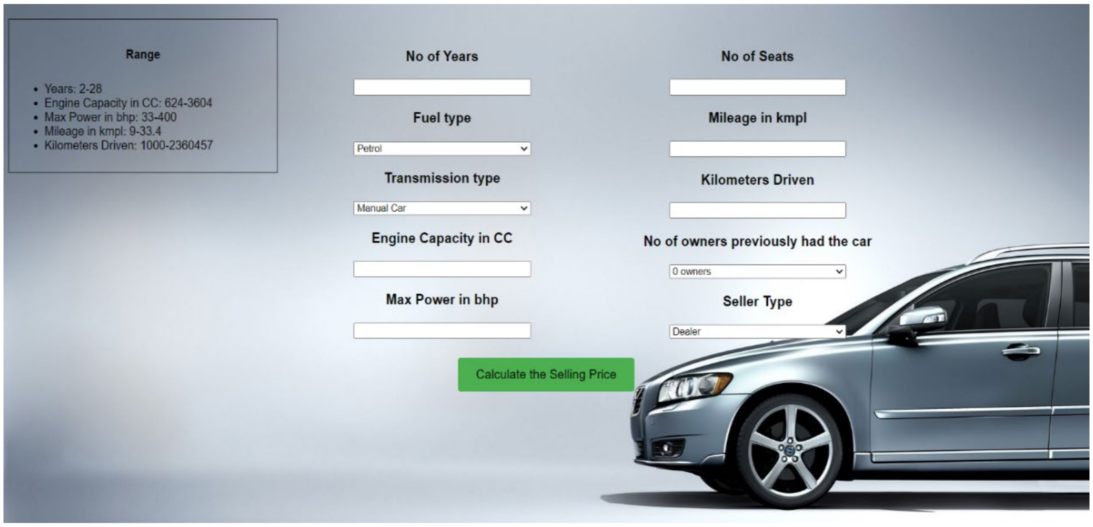


	#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
□	1	year	int(3)			No	None			Change Drop More
□	2	fuel	varchar(10)	utf8mb4_general_ci		No	None			Change Drop More
□	3	transmission	varchar(10)	utf8mb4_general_ci		No	None			Change Drop More
□	4	engine	int(11)			No	None			Change Drop More
□	5	power	decimal(11,2)			No	None			Change Drop More
□	6	seats	int(10)			No	None			Change Drop More
□	7	mileage	decimal(10,2)			No	None			Change Drop More
□	8	km	int(10)			No	None			Change Drop More
□	9	owner	int(10)			No	None			Change Drop More
□	10	seller_Type	varchar(10)	utf8mb4_general_ci		No	None			Change Drop More
□	11	selling_Price	decimal(10,4)			No	None			Change Drop More

Fig. 2 Table Stored on the XAMPP server

Figure 2 illustrates the table stored on the XAMPP server. The XAMPP server's functions and forms of data preservation are described in detail.

The application's web page is displayed in Figure 3. As seen in the associated Figure, the webpage loads when a user runs the application. With the assistance of the Flask web framework, which has been integrated in the Python script, the webpage acquires user input, transmits the information into the pickle file, extracts the value that is required from the pickle, and indicates the result of the user.



Range

- Years: 2-28
- Engine Capacity in CC: 624-3604
- Max Power in bhp: 33-400
- Mileage in kmpl: 9-33.4
- Kilometers Driven: 1000-2360457

No of Years

Fuel type: Petrol

Transmission type: Manual Car

Engine Capacity in CC

Max Power in bhp

No of Seats

Mileage in kmpl

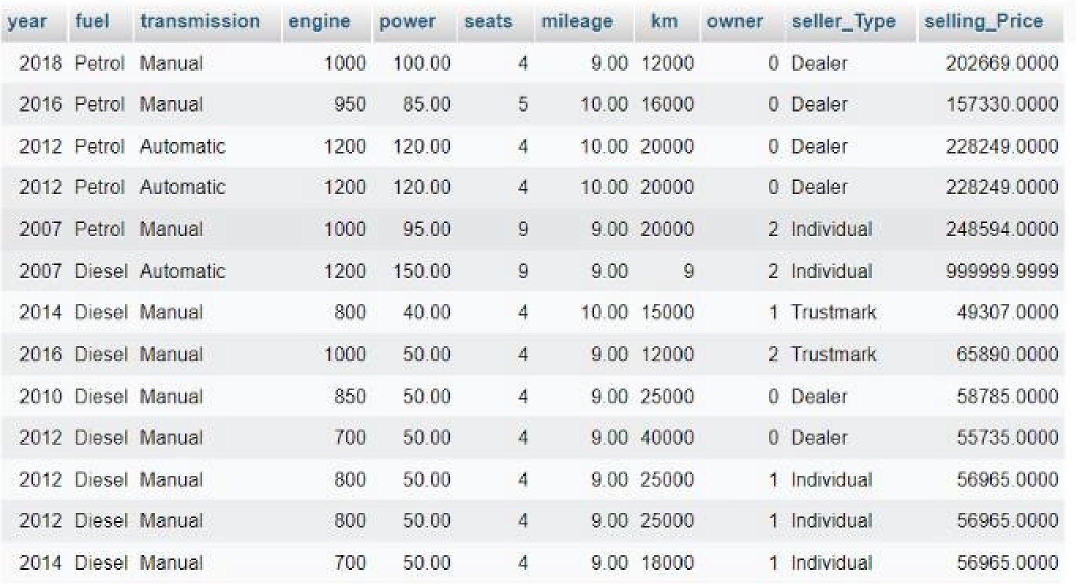
Kilometers Driven

No of owners previously had the car: 0 owners

Seller Type: Dealer

Fig. 3 Webpage of the application

The table that is stored on the XAMPP server is displayed in Figure 4. The user-entered values along with their results are shown below.



year	fuel	transmission	engine	power	seats	mileage	km	owner	seller_Type	selling_Price
2018	Petrol	Manual	1000	100.00	4	9.00	12000	0	Dealer	202669.0000
2016	Petrol	Manual	950	85.00	5	10.00	16000	0	Dealer	157330.0000
2012	Petrol	Automatic	1200	120.00	4	10.00	20000	0	Dealer	228249.0000
2012	Petrol	Automatic	1200	120.00	4	10.00	20000	0	Dealer	228249.0000
2007	Petrol	Manual	1000	95.00	9	9.00	20000	2	Individual	248594.0000
2007	Diesel	Automatic	1200	150.00	9	9.00	9	2	Individual	999999.9999
2014	Diesel	Manual	800	40.00	4	10.00	15000	1	Trustmark	49307.0000
2016	Diesel	Manual	1000	50.00	4	9.00	12000	2	Trustmark	65890.0000
2010	Diesel	Manual	850	50.00	4	9.00	25000	0	Dealer	58785.0000
2012	Diesel	Manual	700	50.00	4	9.00	40000	0	Dealer	55735.0000
2012	Diesel	Manual	800	50.00	4	9.00	25000	1	Individual	56965.0000
2012	Diesel	Manual	800	50.00	4	9.00	25000	1	Individual	56965.0000
2014	Diesel	Manual	700	50.00	4	9.00	18000	1	Individual	56965.0000

Fig. 4 The table in XAMPP server

IV. PERFORMANCE ANALYSIS

There are multiple approaches to evaluate the model's performance; nonetheless, we suggest employing approaches like R2 score and MAE. The predictive power of a model has been evaluated using the R2 score. The input independent variable predicts how much the output dependent characteristic(s) will fluctuate.

Mean Absolute Error, or MAE The property evaluates differences between two observations of the same event. The mean absolute error's size is consistent with the size of the data that have been seen. Since the precision of these statistics is scale-dependent, it is not feasible to contrast the series against various scales.

TABLE II PERFORMANCE ANALYSIS

Model	R2 Score	MAE
Decision Tree	0.9448	81408.25
Random Forest	0.9685	68440.84

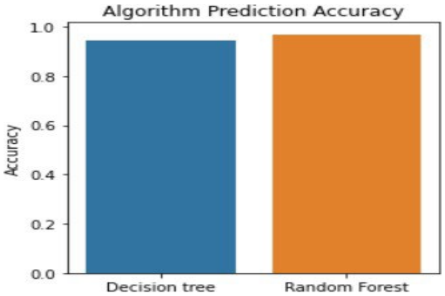


Fig. 5 Bar Plot of the Model Accuracy

We may conclude that the decision tree is less effective than the random forest based on the accuracy of the two models.

V. CONCLUSION

Determining the value of used cars on the open market is crucial when determining their price. The effort was created in response to growing new automobile prices, consumers' inability to purchase them, and an increase in used car sales globally. A system that effectively evaluates the car's worth based on a range of variables thus becomes necessary. The recommended strategy will make it simpler to determine a used car's pricing with accuracy.

REFERENCES

[1] Pedro Strecht, Luis Cruz, Carlos Soares and Joao Mendes Moreira, " A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance," in *Proceedings of the 8th International Conference on Educational Data Mining,* 2015.

[2] Enis Gegic, Becker Isakovic, Dino Keco, Zerina Masetic and Jasmin Kevric, "Car Price Prediction usingMachine Learning Techniques," *TEM Journal*, Vol. 8, No. 1, pp. 113-118, 2019.

[3] Huseyn, Mammadov, "Car Price Prediction in the USA by Using Linear Regression," *International Journal of Economic Behavior,* Vol. 11, No. 1, pp. 99-108, 2021.

[4] M. Listiani, *Support vector regression analysis for price prediction in a car leasing application*, (Doctoral dissertation, Master thesis, TU Hamburg-Harburg), 2009.

[5] Y. S. Park and S. Lek, *Artificial Neural Network,* 2015.

[6] Jie Du, Lili Xie, Stephan Schroeder, "PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution," *Marketing Science,* Vol. 28, No. 4, pp. 637-644, July-August 2009.

[7] Ketan Agrahari, Ayush Chaubey, Mamoor Khan and Manas Srivastava, "Car Price Prediction Using Machine Learning, Srivastava," *IJIRT,* Vol. 8, No. 1, pp. 572-575, 2021.