Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning

[1]Janke Varshitha, [1]K Jahnavi

[2]Dr. C. Lakshmi

[1]B.Tech Students in Electronics and Communication Engineering, SASTRA Deemed University, Thanjavur-613401,India

[2]Faculty in School of Electrical & Electronics Engineering, SASTRA Deemed University, Thanjavur-613401,India

2022 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-6654-8035-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCI54379.2022.9740817

Abstract—**With the extensive growth in usage of cars, the newly produced cars are unable to reach the customers for various reasons like high prices, less availability, financial incapability, and so on. Hence the used car market is escalated across the globe but in India, the used car market is in a very nascent stage and mostly dominated by the unorganized sector. This gives chance for fraud while buying a used car. Hence a high precision model is required which will estimate the price of an used car with none bias towards customer or merchandiser.In this model, A Supervised learning-based Artificial Neural Network model and Random Forest Machine Learning model are developed which can learn from the car dataset provided to it. This project presents a working model for used car price prediction with a low error value. A considerable number of distinct attributes are examined for reliable and accurate predictions. The results obtained agree with theoretical predictions and have shown improvement over models which use simple linear models. An ANN (Artificial Neural Network) is built by using Keras Regression algorithm namely Keras Regressor and other Machine Learning Algorithms namely Random Forest, Lasso, Ridge, Linear regressions are built. These algorithms are tested with the car dataset. Experimental results have shown that the Random Forest model with a Mean Absolute Error value of 1.0970472 and R2 error value of 0.772584 has given the less error among all the other algorithms.The work presented here has shown profound implications for future studies of Used Cars price Prediction using Random Forest and might one day help to solve the problem of fraudswith one hundred percent accuracy.**

Keywords; ANN, keras, Used car price prediction, Regression, Random Forest, Machine Learning, Ridge, LASSO, Linear regression

I. INTRODUCTION

The manufacturing rate of cars has been increasing notably during the past decade, with almost 90 million cars being manufactured in 2020. Car production rates have been expanding dynamically during the previous decade, with right around 90 million vehicles being delivered in the year 2020. The total number of vehicles that are sold in a year all throughout the planet is actually an enormous number. With the rapid growth of population rate, the individual's rate who wants buy a car is also increasing. So, the used car market presently comes into picture as a well-developing industry. There is a percentage of increase in the used cars purchase from the year 2019 to 2020 in India which is expected to increase by 10 percent by the year 2024.

A. Price Prediction of used Car

The most crucial part of used car industry is to predict used car price. To make this happen we need a model which can estimate the price of a used car to be sold with price that is optimal to both the car owner and the purchaser. So, car price prediction has been a research area with great interest, as it requires great knowledge in the field because price usually depends on many distinctive features and factors.There are many instances where buyers are getting cheated by the frauds. So, we cannot rely on price predicted by the individuals as it may contain bias. Hence in this project a model which predicts the price of used cars is developed using Artificial Neural Network which shows no bias to the owner or the purchaser.

B. Challenges

The challenges that are to overcome while designing the model will be:

- Bias towards either owner or purchaser.
- Unable to predict prices of new models or very old model cars.
- Data preprocessing before learning process.
- Attributes that are to be considered while building the model.

II. LITERATURE SURVEY

[1] The first paper is based on predicting the Price of Second-hand Cars using Artificial Neural Networks. In this paper they used various supervised techniques to estimate the price of used cars in Mauritius. They used the data from daily newspapers. Various Machine Learning algorithms like Support Vector Regression, Linear Regression and K-Nearest Neighbor were used.

[2] The second paper is Car's Selling Price Prediction using Random Forest Machine Learning Algorithm by downloading

978-1-6654-8035-2/22/$31.00 ©2022 IEEE

dataset at Kaggle. Two machine learning algorithms namely Random Forest and Extra Trees Regressor and was concluded that they were one of the best algorithms for regression problems.

[3] The third paper is Used Cars Price Prediction using Supervised Learning Techniques wherein error rates were calculated for models like Lasso Regression, Multiple Regression, Regression Tree and ANOVA based comparison was made for the models if they significantly differed from each other.

[4] The fourth paper is Prediction of Car Price Using Linear Regression. The data set was taken from an online ecommerce website quickr.com, in this paper comparison was made for various regression based models like Multiple Linear Regression, Random Forest Regression and was concluded by choosing the best algorithm.

III. METHODOLOGY

In any AI/ML models the first step is to pre-process the raw data based on the requirements of the project and then train and test the data based on the algorithm used which is shown in the Fig 1.

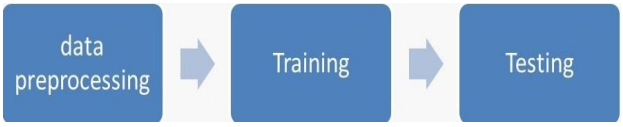

Figure 1. Work Flow of ML/AI model.

A. Data Attributes

The Dataset is collected from [5].

The Dataset contains the following columns:

- Name- Company of the car
- Year- Year in Which it was purchased
- Selling_Price – the price at which it is sold after using vehicle
- Present_price – the price when the car was purchased
- Kms_Driven – the total kilometers that the car was driven
- Fuel_Type – sort of fuel(petrol/diesel/CNG)
- Seller_Type – if it is being sold by individual or dealer
- Transmission – gear transmission of vehicle(whether it is automated/manual)
- Owner – Number of owners that used the particular car

In the pre-processing stage the following changes are made to the collected data:

- All the outliers are removed by using quartile deviation (the data points which are deviating by a factor 1.5 from Q1 and Q3 are removed).Year- Year in Which it was purchased
- All the attributes with string data type are converted into numerical data type by assigning numbers to the strings.Present_price – the price when the car was purchased

Figure 2. The figure represents correlation between different attributes of the dataset

- All the dummy and redundant data points are dropped from the dataset and the correlation between different attributes is shown in Fig 2.
- „Name" attribute is also dropped from the dataset
- All attributes except „Selling_price" is taken as input the model and „selling_price" is taken as input.

In the next steps training and testing is performed in Artificial Neural network, Random Forest, LASSO regression, Linear regression and Ridge Regression algorithms. After completing training and testing process one algorithm is selected based on the performance metrics And the methodology is followed as shown in Fig 3.

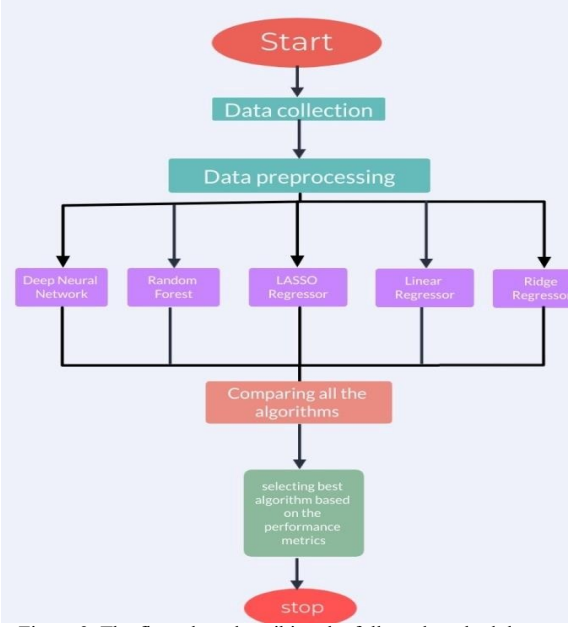

Figure 3. The flow chart describing the followed methodology

B. Deep Neural Networks

Deep Neural Network is a layered association of neurons which are associated to other neurons. These neurons pass a signal or message to different neurons dependent on given information as well as a structure of network which works based on feedback mechanism.

Firstly, input was given to the first layer which is consumed by the neurons in it and provides output to the next layers with output depending upon the activation function used. Each layer has at least one neuron or more neurons and each one of them has an activation function which chooses whether the neuron ought to be actuated or not by computing weighted aggregate and further adding bias to it.

Here the activation function used is Rectification Linear Unit (ReLU). The performance metrics (MAE and R^2 error) are noted down for different cases of layer number. In every case the epoch is fixed as 10,000; Neurons in hidden layer is fixed as 20; Input dataset is given as input layer i.e. first layer and selling price is taken as output in the output layer i.e., last layer of ANN.

TABLE I. MEAN ABSOLUTE ERROR AND R-SQUARED SCORE FOR DIFFERENT LAYERS

No.of .Layers	Mean Absolute Error	R-Squared Error
3	0.8064	0.8260
4	0.8022	0.8258
5	0.9446	0.7946
8	0.7863	0.8282
10	0.7660	0.8421

C. Linear Regression

Linear regression is a model which expects linear connection among input variables(a) and the result variable (b). A simple linear regression can be addressed as b=ma+c where m=slope and c=y-intercept.

D. Ridge Regression

.Ridge regression is a model tuning technique that performs L2 regularization. It adds regularization penalty to the model while training. It is an extension of linear regression.

$$\text{Min } (\|Y - X(\text{theta})\|^2 + \lambda\|\text{theta}\|^2)$$

Lambda is the penalty term. λ can also be denoted by an alpha parameter. We can control penalty term by modifying the alpha values. More is the value of alpha bigger is the penalty thereby reducing the magnitude of coefficients. The performance metrics obtained are: R^2 error – 0.8087027057393501 and MAE -1.143061531347656

E. Lasso Regression

The "LASSO" stands for Least Absolute Shrinkage and Selection Operator. It is an extended version of linear regression; the model is penalized for the sum of absolute values of the weights.

$$\text{Min } (\|Y - X(\text{theta})\|^2 + \lambda\|\text{theta}\|$$

λ is denoted by an alpha, the coefficient to penalize weights. The performance metrics obtained are: R^2 error – 0.8709167941173195 and MAE -1.0507413774170433.

F. Random Forest Algorithm

Random Forest is a classifier containing various decision trees on different subsets of the given dataset and takes the average to increase the accuracy of the dataset. The data from these trees are then combines to guarantee the most accurate predictions. While an independent decision tree has one result and a narrow range of groups, the forest guarantees a more precise outcome with a bigger number of groups and decisions. It has the additional advantage of adding randomness to the model by tracking down the best feature among random features. The performance metrics obtained are: R^2 error – 0.9170678286126046 and MAE -0.7461934065934066.

IV. CONCLUSION

The performance metrics of all the algorithms are mentioned below in the table II.

TABLE II. REPRESENTS MEAN ABSOLUTE ERROR AND R-SQUARED ERROR FOR DIFFERENT ALGORITHMS

Algorithm type	Mean Absolute Error	R-Squared Error
DEEP NEURAL NETWORK (with 10 layers)	0.766	0.842
LINEAR REGRESSION	1.152	0.837
LASSO REGRESSION	1.051	0.871
RIDGE REGRESSION	1.143	0.809
RANDOM FOREST	0.746	0.917

It can be concluded by saying that increased prices of new cars and the monetary lack of ability of clients to get them Used Car market is expanding globally. Therefore, there is an urgent need for a Used Car Price Prediction system that viably determines price of the car using a variety of features. The process of predicting used cars price involves high caution and great knowledge in the field of cars and their models. Among all the proposed models, Random Forest determines the price of a used car with minimum possible error.

V. FUTURE SCOPE

Although the model designed here is restricted to predict the price of used cars it can be extended to any electric gadget or household appliance as well. The model can be connected to real-time websites whose data can be scrapped and the model gets trained based on the dynamic dataset using reinforcement learning.

The model can be extended to get trained on clusters of data rather than on a small dataset. The accuracy of the model can be increased using large historical data. The model can be deployed on the web using API (Application User Interfaces) like Heroku, REST, Git, etc.,

REFERENCES

[1] .Peerun, Saamiyah & Chummun, Nushrah & Pudaruth, Sameerchand. (2015). Predicting the Price of Second-hand Cars using Artificial Neural Networks.

[2] Pandey, Abhishek and Rastogi, Vanshika and Singh, Sanika, Car's Selling Price Prediction using Random Forest Machine Learning Algorithm (March 1, 2020). 5th International Conference on Next Generation Computing Technologies (NGCT-2019), Available at SSRN: https://ssrn.com/abstract=3702236 or http://dx.doi.org/10.2139/ssrn.3702236.

[3] Ganesh, Mukkesh & Venkatasubbu, Pattabiraman. (2019). Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology. 9. 216-223. 10.35940/ijeat.A1042.1291S319

[4] Published in International Journal of Trend inScientificResearch and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4June 2021, URL:https://www.ijtsrd.compapers/ijtsrd42421.pdf Paper URL: https://www.ijtsrd.comcomputer-science/data-processing/42421/prediction-of-car-price-using-linear-regression/ravi-shastri

[5] https://www.kaggle.com/datasets

[6] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177..

[7] N. Sun, H. Bai, Y. Geng and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017, pp. 431-436, doi: 10.1109/SNPD.2017.8022758.

[8] .C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1680-1687, doi: 10.1109/ICESC51422.2021.9532845.

[9] .Vehicle Price Prediction using SVM Techniques S.E.Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya kiran International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-8, June 2020.