Machine Learning-Based Used Car Price Prediction

Sheikh Md. Samiul
Student, Dept. of Electrical and Computer Engineering
North South university
Dhaka, Bangladesh
Email: sheikh.samiul@northsouth.edu

Md. Abdullah Al Mahfuz
Student, Dept. of Electrical and Computer Engineering
North South university
Dhaka, Bangladesh
Email: abdullah.mahfuz@northsouth.edu

***Abstract*—A car price prediction has been a high-interest exploration area, as it requires conspicuous effort and knowledge of the field expert. The recent arrival of online doors has eased the need for both the client and the dealer to be better informed about the trends and patterns that determine the value of a habituated auto in the request. A considerable number of distinct attributes are examined for dependable and accurate vaticination. Using Machine Learning Algorithms similar as Lasso Regression, Multiple Retrogression, and Retrogression trees, we will try to develop a statistical model which will be suitable to predict the price of a habituated auto based on former consumer data and a given set of features. We'll also compare these models' vaticination delicacy to determine the optimal one.**

***Index Terms*—car price prediction, Supervised Learning, Regression, Batch Learning**

I. Introduction

The globally used car market is expected to reach USD 2.67 trillion by 2030. The market is anticipated to expand at a CAGR of 6.1 percent from 2022 to 2030 [1]. Rising technological advancements, such as implementing digital technology in the market and using artificial intelligence to improve the online buying experience, are expected to boost market demand over the forecast period.
The Rise of the new middle class and lower middle class in developing countries created a new demand for cars. But due to climate change, making new cars getting discouraging, and rising inflation, it is difficult to buy a new car. So used cars are the alternative consumers are looking for. But, Consumers do not know how to get the best car at a minimum price. This is why a price-predicting web application makes things easier for the consumer by simply entering some desired car features. Different websites have different algorithms to generate the retail price of used cars. Still, only some websites have a unified algorithm for determining the used car's price. We are creating a machine learning model where we are training a statistical model for predicting the price of cars, So one can quickly get a rough estimate of the price by entering so small details. The main objective of this paper is to use a prediction model to predict the retail price of a used car and compare its level of accuracy.

Machine learning (ML) is a field of inquiry devoted to understanding and building methods that learn, that is, methods that leverage data to improve performance on some set of tasks[2]. An ML model can predict the price of the car by learning and understanding the underlying pattern from a given dataset. So when a new data set is provided, the model can predict the price from the attributes. Many papers also used ML models to solve this problem as it is more efficient than collecting many car price samples.

Finally, We will compare these models' prediction accuracy to determine the optimal one.

This paper is about a machine learning model that predict the price of a use a car which brings forth the following contributions:

- A Dataset collected by the authors and a datset collected from the website Kaggle combining into a new dataset.
- A Web application is build for anyone to find the price of a use car by entering some attributes which the model will predict the price.

To the best of our knowledge, We are not the first to build such a model, but we made some unique addition to the algorithm model. And its presentation as a web application makes it easier for anyone without knowledge of machine learning models to predict the price of a used car.

II. Literature Review

We have used a data set from Kaggle for used car price prediction.The data set contains various features The literature survey provides few papers where researchers have used similar data set or related data-set for such price prediction.

Researchers have used many machine learning methods to estimate a used car's price. In one such paper [3], The authors describe a model that can calculate a price for a used car by evaluating its features considering the prices of other used cars by using a supervised learning method, namely Random Forest, to estimate the prices of used cars. Another Paper uses the same machine learning method [4], using a different data set. Nabarun Pal et al. found some experimental results, the training accuracy was out to be 95.82%, and the testing accuracy was 83.63%, where the model can predict the price of cars accurately by choosing the most correlated features.

Another approach was given by M. S. Richardson in his thesis work [5]. His theory was that car producers could produce more durable cars. He applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer than traditional cars. This is rooted in environmental concerns such as climate change and gives higher fuel efficiency. The researchers from Kalinga

University[6] carried out a study in which they used a multiple linear regression model in a data set consisting of 6000 used cars from a German e-commerce site, collected from Kaggle, by using different car body model types. The model-predicted outcomes give some small advancement in predicting car prices.

Enis Gegic et al. [7]conducted a car price prediction study. They applied three machine learning techniques (Artificial Neural Network, Support Vector Machine, and Random Forest). However, the mentioned techniques were applied to work as an ensemble. The data used for the prediction was collected from the web portal autopijaca.ba. Respective performances of different algorithms were then compared to find one that best suits the available data set. The final prediction model was integrated into the Java application. Furthermore, the model was evaluated using testing data, and an accuracy of 87.38% was obtained. Gonggie [8] proposed a model built using Artificial Neural Networks for the price prediction of used cars. He considered attributes like miles passed, estimated car life, and brand. He made the proposed model for dealing with nonlinear relations in data which was different from previous models that utilized simple linear regression techniques. The nonlinear model could predict the prices of cars with better precision than other linear models.

Abdulla Al Shared [9] proposed a car price prediction study. He used a supervised learning study with three regressions (Random Forest Regression, Linear Regression, and Bagging Regression) that have been trained, tested, and compared against a data set. Among all his experiments, the Random Forest Regression had the highest score at 95%. In addition to Random Forest Regression, Bagging Regression performed well with an 88% score, followed by Linear Regression having an 85% mark. A multiple linear regression approach was given by Noor et al. [10]; they used this model to evaluate the prediction using a collected data set(https://www.pakwheels.com/) of 2000 cars within the duration of one or two months. The collected data includes features like color, advertisement date, etc. There are multiple independent variables in the model but only one dependent variable whose actual and predicted values are compared to find the precision of results. The model has a precision rate of 98%. Another study done [11] by authors Pattabiraman Venkatasubbu and Mukkesh Ganesh used three regression(Lasso Regression, Multiple Regression, Regression Tree) using a benchmark data set of previous consumer data and a set of given features. These approaches are Lasso Regression with an error rate of 3.581%, Multiple Regression with an error rate of 3.468%, and Regression Tree with an error rate of 3.512%. An even more broad study carried out by a group of researchers from Maharaja Agrasen Institute of Technology [12] used eight different models (Linear Regression, Lasso Regression, Ridge Regression, Bayesian Ridge Regression, Random Forest Regression, Decision Tree Regression, XG Boost Regression, Gradient Boosting Regression) and seven different features with the initial price of the car being the most crucial feature. They found that the Decision Tree Algorithm have predicted the best possible price for a car while Gradient Boosting Regression was the second best.

III. Methodology

A. Dataset

The Dataset ”CAR DETAILS FROM CAR DEKHO” was taken from Kaggle. It primarily comprises eight features, four categorical, and four numerical attributes. These features are the car name, the year when the owner brought the car, the selling price in rupees, the fuel of the car, the seller type, the transmission type of the car, how much the vehicle has been driven so far in kilometers, and the current status of the owner since the owner first bought the car. The Dataset has 4340 records of used cars from various car manufacturers. Additionally, the authors collected and added 200 more records to the Dataset.

Due to the car name being hard to categorize, the authors divided the features into three different attributes the company name, the model name, and the edition/trim name.

B. Model Architecture

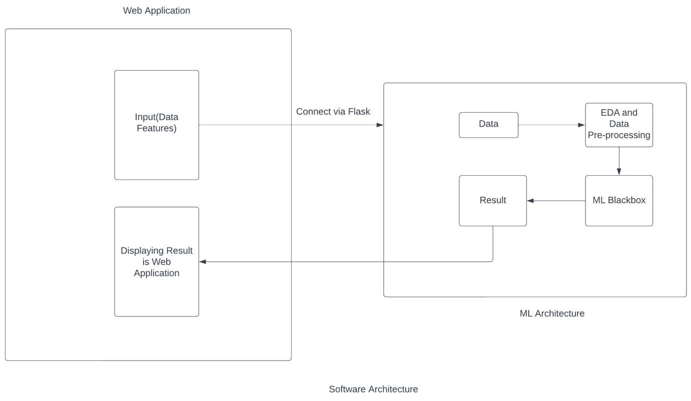


Fig. 1. Software Architecture

1) Software Architecture: The software architecture has two parts, the web application part, and the ML part. We connected both parts using Flask[17], which is a python web framework and is used to build web applications. Flask has the resources to deploy a machine-learning model in the form of a web app. So in the web application, there are two functions, it will take all the necessary user inputs, and then it will send it to the Ml model for predicting the price. The Ml model will make the prediction and will send it back to the web application to display the prediction.

2) ML Architecture: In the ML part, first, the data is read from the dataset. After that, EDA and pre-processing are done so that the raw data is well prepared and suitable for the machine learning model. It is the first and crucial step while creating a machine-learning model. After EDA and pre-processing, data will be sent to the ML black box, where the dataset will be trained and tested. And the ML black box will learn a model or pattern which, in the future, will be able to predict.The Black box in this project has five models: Linear Regression, KNN, Lasso Regression, Decision Tree, and Random Forest.

C. Model Algorithm

Lasso Regression[13]: Lasso (Least Absolute Shrinkage Selector Operator) regression is a regularization term of the "sum of the absolute value of the coefficients" added to the cost function. Lasso shrinks down the coefficients of redundant features to zero and thus also performs a direct feature selection.

$$\underset{\beta_0,\beta}{\arg\min}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\}+\lambda\sum_{j=1}^{p}|\beta_j|$$

So, the cost function for the lasso regression in generalized form is:

$$E(\beta)+\lambda R(\beta)$$

Here E() is the Error term, and R() is the Regularization term. If we use a high value of , we are putting a lot of premium on controlling the complexity of the model. It will not allow coefficients to be significant at all. If we allow close to zero, it implies no regularization and a high chance of overfitting. Here is the hyperparameter in the objective function that we minimize for regularized regression. In the lasso regression, we have the sum of absolute values of coefficients as a regularization term. This makes things inconvenient, and the total value is not differentiable when zero. Lasso's regularization term is helpful for feature selection. Lasso regression results in a sparse solution, meaning many of the model coefficients automatically become exactly zero, = 0.

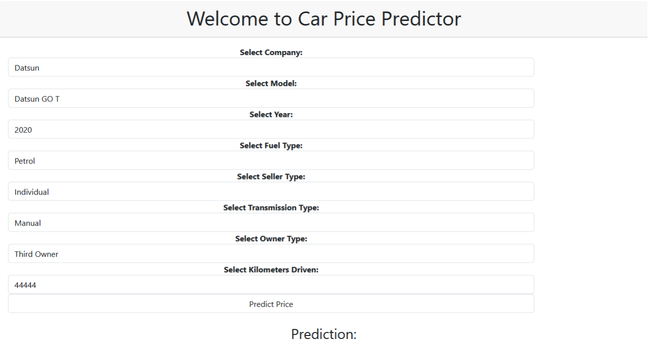
If * is the best model that we end up getting, which gives us the following:

$$\beta^* = argmin[E(\beta)+\lambda R(\beta)]$$

Here argmin evaluates values of for which the expression E()+ R() is minimum. The Sparsity(*) increases with an increase in , where the number of parameters in * defines the Sparsity(*) of a model that is precisely equal to zero. We sometimes have a large number of features in real-world problems, but we want the model to be able to pick up only the most useful ones.

D. Web Application

We have created a web application that takes some features as inputs and displays the predicted price.



Welcome to Car Price Predictor

Select Company: Datsun

Select Model: Datsun GO T

Select Year: 2020

Select Fuel Type: Petrol

Select Seller Type: Individual

Select Transmission Type: Manual

Select Owner Type: Third Owner

Select Kilometers Driven: 44444

Prediction:

Fig. 2. Web Application

IV. Models and results

A. The evaluation parameters

R-squared[14] measures how well your linear regression model fits the data. It calculates the strength of the relationship between your model and the dependent variable. There are several vital goodness-of-fit statistics for regression analysis. When R-squared represents, the accuracy of training data is called R2-Error. And when R-squared means the accuracy of testing data is called R2-Score.

Mean squared error (MSE)[15] measures the amount of error in the ML models. It calculates the average squared difference between the observed and predicted values. As model error increases, its value increases; when a model has no error, its value is Zero. The mean squared error(MSE) is also known as the mean squared deviation (MSD).

Mean Absolute Error (MAE)[16] is the average of absolute errors for a group of predictions and observations. MAE is a measurement of the magnitude of errors for the entire group.

B. Training and Testing Accuracy

Here is a table containing the accuracy rate of testing (R2 Score*100%) and training Data(R-Squared*100%)

Model	Testing Accuracy	Training Accuracy
Linear Regression	85.38	92.37
KNN	35.58	59.42
Lasso Regression	87.99	91.16
Decision Tree	77.21	88.06
Random Forest	84.26	88.06

C. *Graph of evaluation parameters*

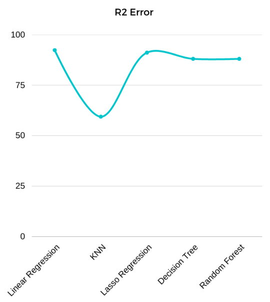


Fig. 3. Training Accuracy rate

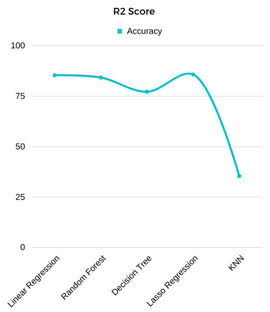


Fig. 4. Testing Accuracy rate

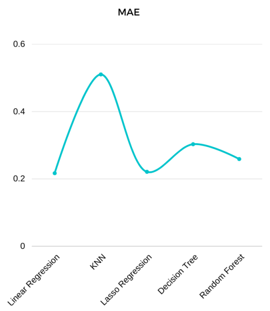


Fig. 5. MAE

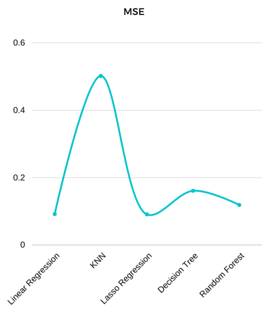


Fig. 6. MSE

V. Conclusion and Future work

Lasso Regressor produces better accuracy in this project than the other four models. Because the Lasso regression shrinks the dataset randomly toward the center point or mean point, Other models came close to the lasso regressor except for the KNN model, which failed to capture the basic underline of the dataset.

In this project, we build a model for car price prediction with an accuracy rate of 87.99%. So there is still room for improvement, and some of these are

- We will collect and add more data to the dataset for wide-range prediction and better accuracy.
- We will improve the interface of the website application.
- We can deploy this project with an android application.
- We will also explore new ML models and neural network models for this dataset.

Acknowledgment

The authors would like to thank Intisar Tahmid Naheen Sir from Department of Electrical and Computer Engineering, North South University, Dhaka for his guidance in this research.

References

[1] R. and M. ltd, “Used Car Market Size, Share Trends Analysis Report by Vehicle Type (Hybrid, Conventional, Electric), by Vendor Type, by Fuel Type, by Size, by Sales Channel, by Region, and Segment Forecasts, 2022-2030,” www.researchandmarkets.com. https://www.researchandmarkets.com/reports/5595868/used-car-market-size-share-and-trends-analysis? (accessed Jan. 04, 2023).

[2] T. M. Mitchell, Machine learning. New York: Mcgraw Hill, 1997.

[3] A. Pandey, V. Rastogi, and S. Singh, “Car’s Selling Price Prediction using Random Forest Machine Learning Algorithm,” papers.ssrn.com, Mar. 01, 2020. https://ssrn.com/abstract=3702236 (accessed Jan. 04, 2023).

[4] N. Pal, P. Arora, D. Sundararaman, P. Kohli, and S. Palakurthy, “How much is my car worth? A methodology for predicting used cars prices using Random Forest,” FICC. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf

[5] M. Richardson, “DETERMINANTS OF USED CAR RESALE VALUE In Partial Fulfillment of the Requirements for the Degree Bachelor of Arts,” 2009. [Online]. Available: https://digitalccbeta.coloradocollege.edu/pid/coccc:1346/datastream/OBJ

[6] C. Trivendra, “The Price Prediction for used Cars using Multiple Linear Regression Model,” International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 5, p. 1801, 2020, Accessed: Jan. 04, 2023. [Online]. Available: https://www.academia.edu/43313838/The-Price-Prediction- for-used-Cars-using- Multiple-Linear-Regression-Model

[7] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car Price Prediction using Machine Learning Techniques,” TEM Journal, vol. 8, no. 1, pp. 113–118, 2019, doi: 10.18421/TEM81-16.

[8] Shen Gongqi, Wang Yansong, and Zhu Qiang, “New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit,” 2011 Third International Conference on Measuring Technology and Mechatronics Automation, Jan. 2011, doi: 10.1109/icmtma.2011.455.

[9] A. Alshared, “Used Cars Price Prediction and Valuation using Data Mining Used Cars Price Prediction and Valuation using Data Mining Techniques Techniques.” [Online]. Available: https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220context=theses

[10] K. Noor and S. Jan, “Vehicle Price Prediction System using Machine Learning Techniques,” International Journal of Computer Applications, vol. 167, no. 9, pp. 27–31, Jun. 2017, doi: 10.5120/ijca2017914373.

[11] Pattabiraman Venkatasubbu and Mukkesh Ganesh, “Used Cars Price Prediction using Supervised Learning Techniques,” International Journal of Engineering and Advanced Technology, vol. 9, no. 1S3, pp. 216–223, Dec. 2019, doi: 10.35940/ijeat.a1042.1291s319.

[12] A. Datt Sharma, V. Sharma, S. Mittal, G. Jain, and S. Narang, “PREDICTIVE ANALYSIS OF USED CAR PRICES USING MACHINE LEARNING.” Accessed: Jan. 04, 2023. [Online]. Available: https://www.irjmets.com/uploadedfiles/ paper/volume3/issue-6-june-2021/12071/1628083486.pdf

[13] G. Doosa, “The Mathematical background of Lasso and Ridge Regression,” CodeX, Jun. 14, 2021. https://medium.com/codex/mathematical-background-of-lasso-and-ridge-regression-23b74737c817: :text=Lasso%20or%20L1%20Regression%3A (accessed Jan. 04, 2023).

[14] J. Frost, “How To Interpret R-squared in Regression Analysis,” Statistics By Jim, 2018. https://statisticsbyjim.com/regression/interpret-r-squared-regression/

[15] J. Frost, “Mean Squared Error (MSE),” Statistics By Jim, Nov. 12, 2021. https://statisticsbyjim.com/regression/mean-squared-error-mse/

[16] “Mean Absolute Error,” C3 AI. https://c3.ai/glossary/data-science/mean-absolute-error/: :text=What%20is%20Mean%20Absolute%20Error

[17] ”Flask is a lightweight WSGI web application framework,” https://palletsprojects.com/p/flask/

VI. Summary of the accuracy of the model from the papers mentioned in the literature review

Paper	**Model**	**Accuracy(%)**
Car’s Selling Price Prediction using Random Forest Machine Learning Algorithm[3]	Random Forest	Not calculated
How much is my car worth? A methodology for predicting used cars prices using Random Forest[4]	Random Forest	83.63
DETERMINANTS OF USED CAR RESALE VALUE[5]	Multiple Regression	Not calculated
The Price Prediction for used Cars using Multiple Linear Regression Model[6]	Multiple Linear Regression	Not calculated
Car Price Prediction using Machine Learning Techniques[7]	Artificial Neural Network(Cheap)	83.91
	SVM(Cheap)	86.96
	Random Forest	85.82
	Multiple Regression	92.38
New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit[8]	Artificial Neural Network	Not calculated
Used Cars Price Prediction and Valuation using Data Mining Used Cars Price Prediction and Valuation using Data Mining Techniques Techniques[9]	Random Forest	95
	Linear Regression	85
	Bagging Regression	88
Vehicle Price Prediction System using Machine Learning Techniques[10]	Multiple Linear Regression	98.61
Used Cars Price Prediction using Supervised Learning Techniques[11]	Lasso Regression	Not calculated
	Multiple Regression	Not calculated
	Regression Tree	Not calculated
PREDICTIVE ANALYSIS OF USED CAR PRICES USING MACHINE LEARNING[12]	Linear Regression	86.25
	Lasso Regression	86.59
	Ridge Regression	86.34
	Bayesian Ridge Regression	86.95
	Random Forest	85.76
	Decision Tree	95.44
	XG Boost Regression	89.58
	Gradient Boosting Regression	93.55