Used Vehicle Price Prediction

Muhammad Khaleduzzaman
ID: 180104102
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
180104102@aust.edu

Md. Tamzidul Islam
ID: 180104123
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
180104123@aust.edu

Md. Mubayeer Rahman
ID: 180104144
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
180104144@aust.edu

***Abstract*— The manufacturer sets the price of a new vehicle in the market, with some additional expenses paid by the government in the form of taxes. Customers who purchase a new vehicle may be sure that their investment will be worthwhile. Yet because new vehicles are becoming more expensive and consumers can no longer afford to acquire them, used car sales are rising everywhere. Hence, a system that accurately assesses the value of the vehicle utilizing a range of criteria is urgently needed for used car price prediction. The current system involves a procedure where a vendor chooses a price arbitrarily and the buyer is unaware of the vehicle and its current market worth. Neither the seller nor the price at which he ought to sell the automobile have any notion of the current value of the vehicle. We have created a model that will be quite successful in resolving this issue. Supervised machine learning algorithms are employed because, as opposed to categorical values, their output is a continuous value. As a result, it will be feasible to forecast the exact cost of a vehicle rather than its price range. A user interface that accepts input from any user and shows the price of a car based on their inputs has also been developed.**

Keywords—used vehicle; price prediction; supervised machine learning.

I. Introduction

Due to the numerous variables that affect a used automobile's market price, figuring out whether the quoted price of a used car is accurate is a difficult undertaking. This project focuses on creating machine learning models that can precisely forecast the price of a used vehicle depending on its qualities, allowing users to make educated decisions. Using a dataset comprised of the sale prices of various brands and models, we apply and assess several learning methodologies. The performance of different machine learning algorithms, including Linear Regression, Random Forest, and Naïve bayes will be compared, and the best one will be selected. We will calculate the cost of the car based on a number of factors. It's difficult to estimate a used vehicle's actual cost. The cost of every used car must be determined based on a variety of factors. One of the most noticeable qualities of an automobile is how long it has been in use. Other significant features include the build (model), origin (country of production), mileage (kilometers driven), horsepower, etc. The fuel type and economy are significant factors to consider for prediction models due to the rising cost of fuel. This type of prediction model would be beneficial to both purchasers and sellers, who might use it to estimate the worth of a vehicle they planned to sell.

II. Background Study

This is a supervised learning problem and can be solved using various machine learning techniques. We need to predict the selling price of a vehicle based on the given its features. Supervised Regression problems require labeled data where our target or dependent variable is the selling price of a vehicle. All other features are independent variables. Following are some algorithms that can be used for predicting the selling price.

Linear Regression: It is a linear strategy for modeling the interactions between a scalar response and dependent and independent variables in the field of statistics. In linear regression, model parameters that are unknown are estimated from the data and functions like the linear predictor are used to model relationships.

Random Forest: Random Forest is a supervised learning algorithm since it uses the ensemble learning approach for classification and regression. The trees in random forests are parallel to one another and do not interact as they grow. A meta-estimator called random forest compiles the outcomes of numerous predictions. Additionally, it aggregates different decision trees with the aid of various adjustments.

Naïve Bayes: An example of a machine learning (ML) method is the naive Bayes regression classifier, which is thought to be more accurate than more complex algorithms like univariate decision trees because it is based on the Bayes theorem conditional probability for prediction.

III. Literature Review

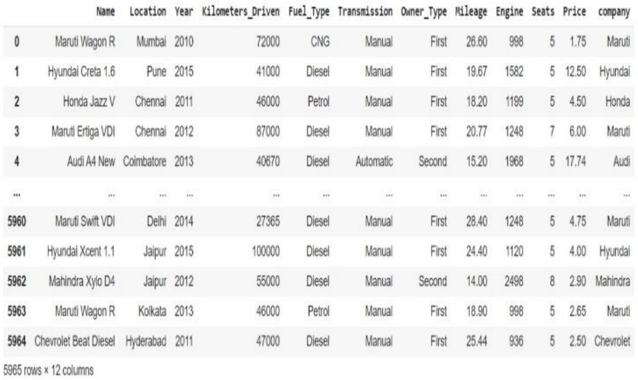
Predicting the Price of Used Car Using Machine Learning Techniques is the first paper. The use of supervised machine learning techniques to forecast the cost of used cars in Mauritius is examined in this paper. The forecasts are supported by historical information gathered from daily newspapers. The predictions were made using a variety of techniques, including multiple linear regression analysis, k-nearest neighbors, naive bayes, and decision trees.[1]

Car Price Prediction Using Machine Learning Techniques is the second paper. For the reliable and accurate prediction, a large number of unique attributes are examined. They used three machine learning techniques to create a model for predicting the cost of used cars in Bosnia and Herzegovina (Artificial Neural Network, Support Vector Machine and Random Forest).[2]

The third paper presents a second-hand car price evaluation model using BP neural networks. The price evaluation model based on big data analysis is put forth in this paper. It makes use of widely disseminated vehicle data as well as a sizable amount of vehicle transaction data to analyze the price data for each type of vehicle using the BP neural network algorithm that has been optimized. In order to determine the price that best fits the car, it aims to establish a model for evaluating used car prices. [3]

IV. Data Collection Process

This section performs the Selling price prediction using a dataset consisting of 6018 used car details. This dataset is prepared by Cardekho.com and available on Kaggle. After cleaning duplicate or null values, we got 5964 used car details.



	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Seats	Price	company
0	Maruti Wagon R	Mumbai	2010	72000	CNG	Manual	First	26.60	998	5	1.75	Maruti
1	Hyundai Creta 1.6	Pune	2015	41000	Diesel	Manual	First	19.67	1582	5	12.50	Hyundai
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199	5	4.50	Honda
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248	7	6.00	Maruti
4	Audi A4 New	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968	5	17.74	Audi
...	...	...	...	...	...	...	...	...	...	...	...	...
5960	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual	First	28.40	1248	5	4.75	Maruti
5961	Hyundai Xcent 1.1	Jaipur	2015	100000	Diesel	Manual	First	24.40	1120	5	4.00	Hyundai
5962	Mahindra Xylo D4	Jaipur	2012	55000	Diesel	Manual	Second	14.00	2498	8	2.90	Mahindra
5963	Maruti Wagon R	Kolkata	2013	46000	Petrol	Manual	First	18.90	998	5	2.65	Maruti
5964	Chevrolet Beat Diesel	Hyderabad	2011	47000	Diesel	Manual	First	25.44	936	5	2.50	Chevrolet

5965 rows × 12 columns

Figure 1: Used car details.

We have some categorical objects as well as continuous features here.

Name	object
Location	object
Year	int64
Kilometers_Driven	int64
Fuel_Type	object
Transmission	object
Owner_Type	object
Mileage	float64
Engine	float64
Seats	float64
Price	float64

V. Methodology

We built a mathematical model that could predict the price of a used car based on previous consumer data and the collection of characteristics by using the supervised machine learning techniques. This uses 3 different algorithms of Machine Learning There are also systems 2 main phases. They are as follows: 1. Training phase: Using the data in the data set, the system is trained to fit a model (line or curve) based on the algorithm selected appropriately. 2. Testing phase: the system is given inputs and is evaluated for functionality. The precision is examined. As a result, the data that is used to develop or validate the model must be appropriate. The system must use the proper algorithms to complete the two distinct tasks because it is built to detect and predict the price of used cars. Different algorithms were compared for accuracy before being chosen for further use. The person who was best suited for the job was picked.

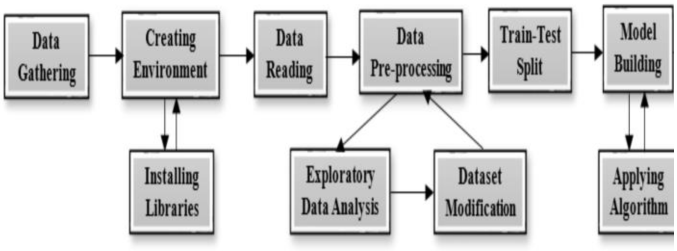


Figure 2: Flowchart of our study.

A. Linear Regression

By applying a linear equation to the observed data, the linear regression method attempts to model the relationship between two variables. The second party is regarded as the dependent variable. As an illustration, a modeler might use a linear regression model to compare a person's weight to their height. Finding a relationship between several continuous variables can be done with the help of linear regression.

Both many independent variables and a single independent variable are present.

y = m1X1+m2X2+……+b
m1, m2, m3 …. → slope
b → y intercept
X1, X2, X3 …… → independent variables
y → dependent variables.

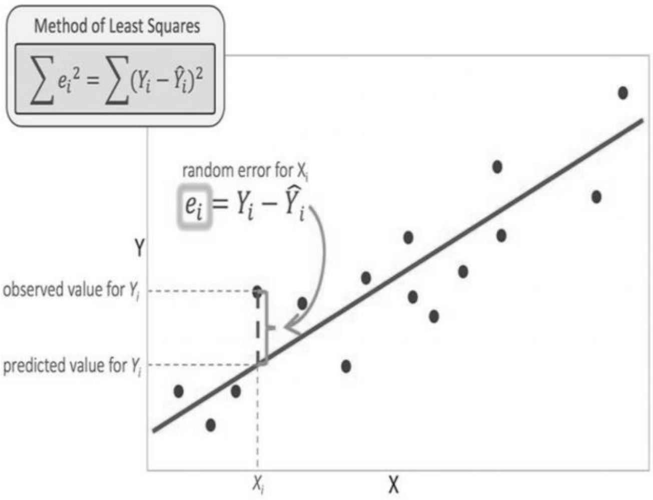


Figure 3: Linear Regression.

B. Random Forest

A supervised learning algorithm called Random Forest employs an ensemble learning strategy for regression and classification. Weak learners can form strong learners, which is the main tenet of ensemble methods. At training time, Random Forest builds several decision trees. On bootstrapped datasets, these decision trees are independently trained. By averaging the predictions made by each individual tree, the final predicted value is determined.

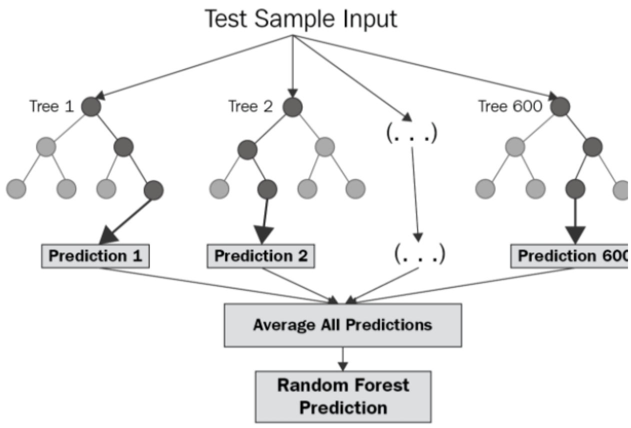


Figure 4: Random Forest.

C. *Naïve Bayes*

Naive Bayes Classifier (NBC) is a probabilistic based prediction technique with consideration of independence [14]. All attributes in the dataset are considered has no relationship. This algorithm uses a probability approach by summing the average for each attribute. The formula of Naive Bayes is as follows:

$$P(Y|X) = \frac{P(Y)P(Xi|Y)}{P(X)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \; (1)$$

Where:

P (Y | X) = Probability of data with vector X in class Y

P (Y) = Initial probability of class Y

P (|Y) = Probability X based on the condition of hypothesis H

P (X) = Probability X

VI. Result Analysis

The least square method was used to estimate the model, and the following Minitab results were obtained.

A. *Figures and Tables*

a) Positioning Figures and Tables:

R – Square Value	Test accuracy
Linear Regression	83.67%
Naïve Bayes	88.68%
Random Forest	89.01%

Table 1. Test Accuracy and Loss for ML models.

The percentage answer variance in a variable called R-square is explained by a linear model (Rsq). This means that a high R-square value indicates that the model is more suited to the data and hence produces more reliable results which is Random Forest. To get visual representation we can plot some figures to get clear idea.

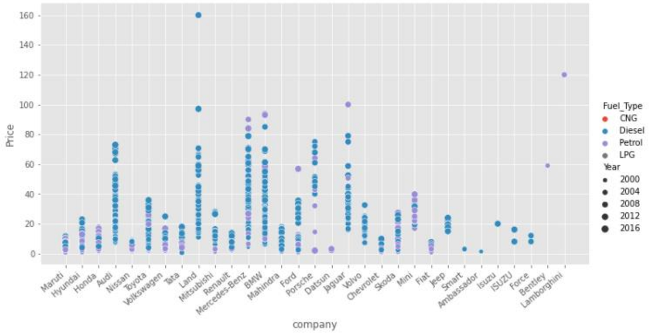


Figure 5: Price to Company graph with Fuel type and year.

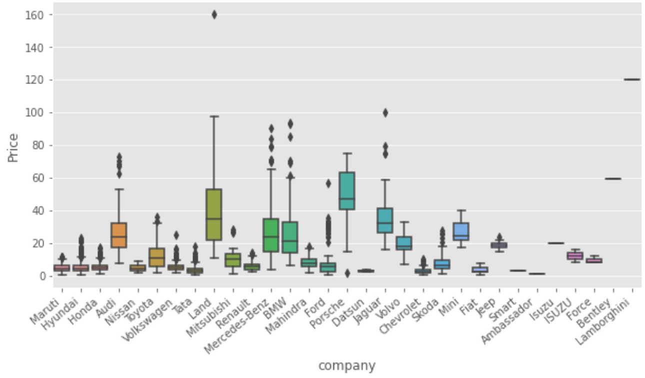


Figure 6: Price to Company only.

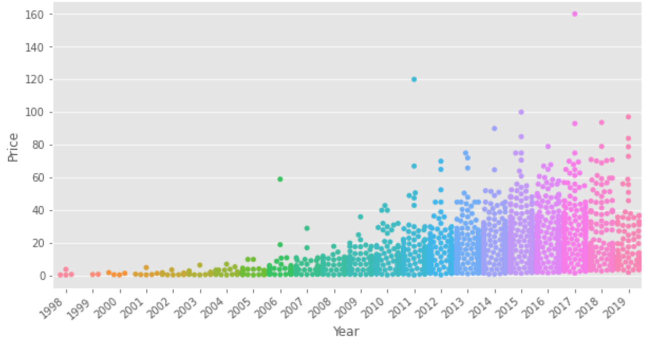


Figure 7: Year to price.

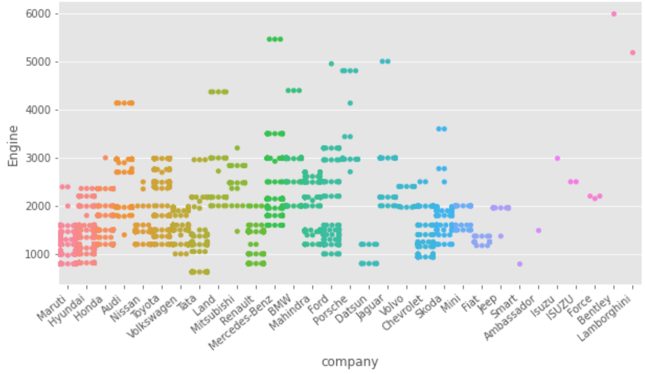


Figure 8: Company to engine.

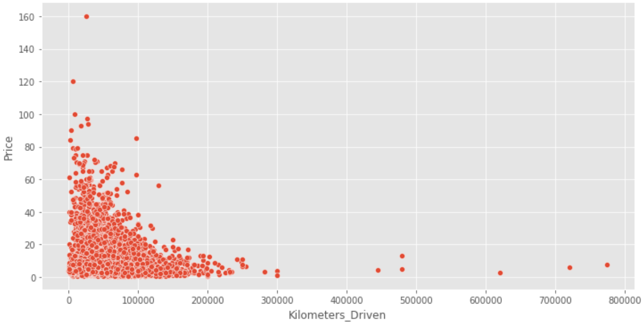


Figure 9: Price to Kilometers_Driven.

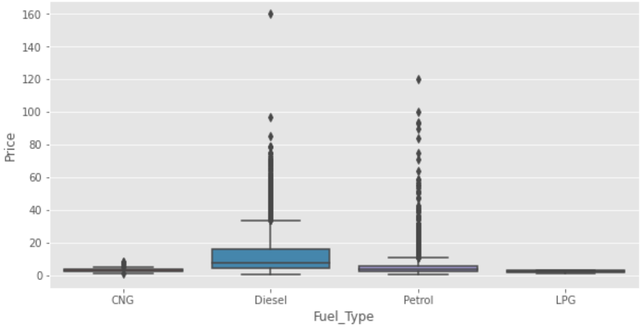


Figure 10: Price to Fuel_Type.

VII. Conclusion

Predicting prices of a used car is a challenging task because of a high number of features and parameters that should be considered to generate accurate results. The first and foremost step is data gathering and preprocessing data. Then a model was defined and created for implementing algorithms and generating results. We started with understanding the use case of machine learning in the Automotive industry and how machine learning has transformed the driving experience. Moving on, we looked at the various factors that affect the resale value of a used car and performed exploratory data analysis (EDA). Further, we evaluated the performance of the model using the R-squared score. Finally, we saw a Random Forest Regression model gave a higher value to predict the resale value of a used car.

References

[1] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”; (IJICT 2014).

[2] Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, ”Car Price Prediction Using Machine Learning”; (TEM Journal 2019).

[3] Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, “Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory”; (Hohai University Changzhou, China).

[4] Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, “Prediction of Prices for Used Car by using Regression Models” (ICBIR 2018).