

Price Prediction of Used Cars using Linear Regression

¹Amit Kewat, ²Nitesh Kanojiya

¹Research Scholar, MCA

Thakur Institute of Management Studies, Career Development & Research (TIMSCDR)

Mumbai, India

amit.kewat.4035409@gmail.com

²Research Scholar, MCA

Thakur Institute of Management Studies, Career Development & Research (TIMSCDR)

Mumbai, India

nkanojiya007@gmail.com

Abstract—Nowadays due to covid19 pandemic everyone wishes to have their own vehicle. So that we can avoid traveling in public vehicles and can distance ourselves socially from others. But getting a new car is not an easy task any more due to its high price and its depreciation and many other government taxes, then after that its additional cost on maintaining and repairing that car. So, the only choice we have left with is buying a car that is previously used. But sometimes the owner of car lists improbable price according to their demand. So, this research paper focuses on predicting the price of used cars. For predicting the price, the dataset which was used was downloaded from Kaggle. The dataset included some factors like name of car, company, year, kilometers driven, fuel type that are independent variables and price which is a dependent variable with the help of that we are going to apply supervised machine learning method called linear regression to it and with the help of the preprocessed dataset and linear regression algorithm it will help us in the finding the accurate price of used cars

Keywords—Price prediction; Kaggle; machine learning; train test split; linear regression

I. INTRODUCTION

Now a days getting a new branded car is not an easy task for any middle-class person because the initially price of the new cars is much higher which the customers cannot afford it and the other disadvantage is that the tax and insurance premiums of new cars are also higher, so now the only option the customers are left with is buying a second hand or we can say a used car. And nowadays there has been an increased in demands for buying a used car as well and on the other hand the advantages of buying a used cars is that used cars has low insurance rate, they cost less and they have also had a low depreciation. So, most of the customers prefer buying a used car rather than buying a new branded car. So, for that, finding the accurate price of used car has also become a challenge nowadays.

So, this research paper focuses on predicting used car's selling price based on some features such as Car Company, Car Model, Year of Purchase, fuel Type & Kilometers driven where these features are inputs and here there is another variable called as Price which can be called as target variable. In this paper we will be using a dataset which has been downloaded for Kaggle.com. And with the help of the downloaded dataset, we will pre-process or we can say we will be cleaning the dataset because the downloaded dataset had few inappropriate data such as nan values, their names are inconsistent and few more, so we are going to discard that data which are not appropriate for prediction of price.

So, after that we will be using a machine learning algorithm named linear regression model for predicting the accurate price of used cars. Machine learning is a field of study that gives computers the capability to learn from past data and to learn without being explicitly programmed. We are using linear regression because it is supervised learning. It consists of labeled data it says that we already have



input as well as output. And we have to predict the continuous value and linear regression is used to predict the price with the help of features (inputs) as independent variable and the Price (output) as dependent variable.

II. LITERATURE REVIEW

The first paper is how much is my car worth? A methodology for predicting cars prices using random forest. In this research paper the author has used supervised learning method called as random forest for predicting the price of used cars. The main features that are used by the authors are price, kilometers, brand and vehicle type. Using the random forest model on the pre processed dataset the result that was found that this model can help predict the car price accurately by using its features.

The second paper is used car price prediction using machine learning techniques. In this research paper the authors aim was to predict the price of cars in Bosnia and Herzegovina. The data which was collected for building model was from web portal autopijaca.ba using web scrapper. The dataset consists of following features like brand, model, car condition, fuel type, year of manufacture, colours, doors and miles. Using this dataset, the author has used three machine learning techniques like Artificial Neural Network, Support Vector Machine, and random forest. But this technique was applied to work as an ensemble. Further the model was evaluated using test data and the accuracy which was obtained was 87.38%.

The third research paper is Car's selling price prediction using random forest machine learning algorithm. Here to predict the price of used cars the author has used previous used car selling data which is downloaded from Kaggle and has used supervised machine learning techniques and used algorithms like random forest and extra tree regression and used some steps like importing the dataset, taking care of missing data, encoding categorical data, splitting the data into training set and test set. The result that is shown is that both algorithms were highly accurate in prediction. The prediction has been done using features like model year, showroom price, driven distance, owner type, fuel type, and transmission and seller type.

In the fourth research paper Used Car Price Prediction using k nearest neighbor based, the author has used a dataset which is retrieved from kaggle which includes some factors like fuel type, colour, model, mileage transmission, number of seats, engine, location, etc. For analyzing the price of used cars, the author has used KNN (K Nearest Neighbor) regression algorithm. Before that the author has pre-

processed the downloaded dataset like converting categorical values into numerical, removal of non-numerical part from numerical cells. After that the data was trained by model and accuracy was examined among different ratios between trained and test data. And after that the data was cross validate using the K fold method for accessing the performance of its model. The accuracy that was found is 85% using KNN algorithm.

The fifth research paper that has been used is Predicting the Price of used Cars using Machine Learning Techniques. In this research paper the author has created a research paper on predicting the price of used cars in Mauritius. The data that has been used for prediction is collected from daily newspapers such as L'Express and Le Defi. Different techniques of supervised machine learning like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees used for prediction.

In the sixth research paper Used Car Price

Prediction the author's goal was to predict the accurate price of cars based on its features. The process which has been followed for prediction is from collecting data then processing the data which includes data cleaning, data reduction, data transformation then using machine learning regression techniques because regression algorithms provides continuous values as output. Algorithm that were used are linear regression, ridge regression, lasso regression and after comparing them selecting the best model and then according to that displaying the predicted price.

The seventh research paper is Car Price Prediction using Machine learning. The main goal of this study is to discover the best predictive model for estimating the price of used car. The data was collected from kaggle which includes few features such as capacity of engine, distance traveled by car, year of manufacture, fuel type. For predicting the price machine learning algorithms were used such as linear regression and lasso regression.



The result of this algorithm will be analyzed and based on the efficiency and accuracy the best one of them can be used for prediction.

The eight research paper is Car Price Prediction. The author has collected dataset from cardekho.com and has used machine learning method named random forest regression algorithm. Price prediction were done on some factors of car like KM driven, fuel type, year, No. of owner. The aim of the author was to use machine learning algorithm to develop model for predicting used cars.

III. METHODOLOGY

A. Data Preprocessing

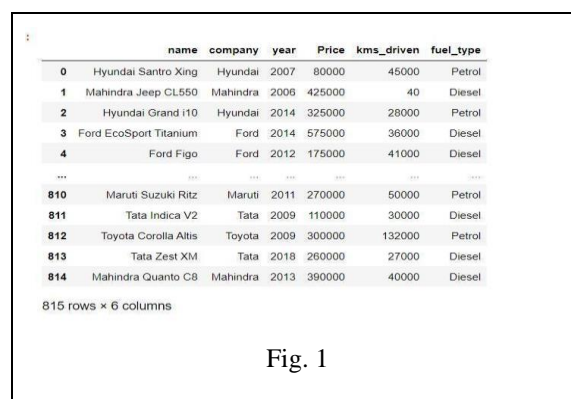
Preprocessing of data is an important task and it is very important for getting accuracy, because few data have some false values and some missing values so we have to identify that raw data and convert it into well format data. It can also call as cleaning the dataset.

So, here the raw data set was collected from Kaggle. The raw dataset consists of 892 rows and 6 columns. The 6 columns are the features of dataset which are:

„name“, „company“, „year“,

Price“, „kms_driven“, „fuel_type“. Since the raw dataset also has some irrelevant records that are not going to be useful for prediction of cars so we are going to pre-process or clean that data by discarding those records. The Preprocessing that we are going to apply on the dataset are as follows:

- year has many non-year values, so we will filter out only non-numeric values.
- year is in object. Change to integer.
- Price has Ask For Price. Filtered out rows with price set as= Ask for price.
- Price has commas in its price and is in object.so, we will remove commas and change price into integer.
- We will split the numeric and kms, now we will extract numeric values from it and change the object values into integer.
- Two rows contain fuel, and it contains nan values. So, we will extract only numeric values and we will drop rows having nan values
- fuel_type has nan values. So, we will drop rows having nan values.
- names are inconsistent. Keeping first three words of name.
- name and company had spammed data but because of previous cleaning those rows got removed.

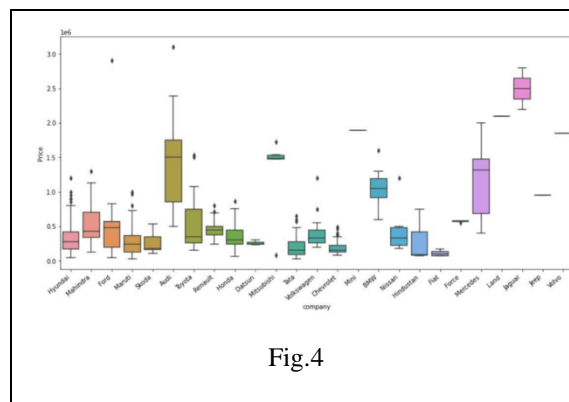
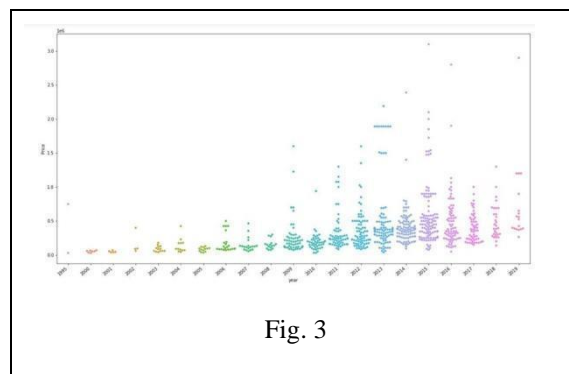
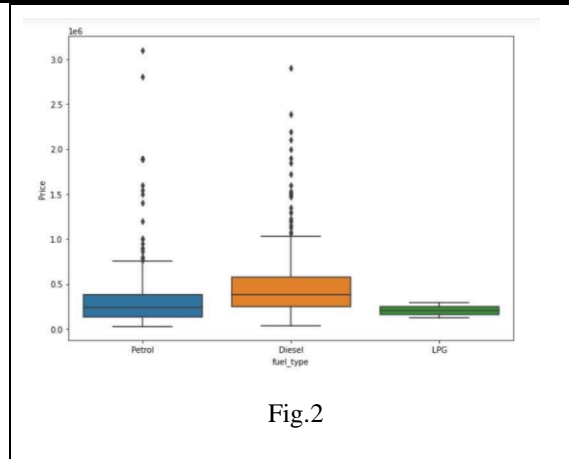


	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	Ford Figo	Ford	2012	175000	41000	Diesel
...
810	Maruti Suzuki Ritz	Maruti	2011	270000	50000	Petrol
811	Tata Indica V2	Tata	2009	110000	30000	Diesel
812	Toyota Corolla Altis	Toyota	2009	300000	132000	Petrol
813	Tata Zest XM	Tata	2018	260000	27000	Diesel
814	Mahindra Quanto C8	Mahindra	2013	390000	40000	Diesel

815 rows x 6 columns

Fig. 1





B. Exploratory Data Analysis

After preprocessing the data is analyzed or investigate through visualization method. EDA is all about visualization. It gives clear picture about

C. Training the model

So, after data preprocessing, we have to train the model so for that first thing we have to do is extract the features and target column from the cleaned dataset. So, from the cleaned dataset everything is our feature except the price column. Here the price column is the target variable. So, for that we are going to split them in the following manner,

- $X=[,name,company,year,kms_driven,fuel_type]$



- $Y = [„Price“]$
- Here, we have split the data using x and y where x contains all the features and y contains the target variable.

D. Applying Train Test split

We are going to split our dataset into train test split because, splitting the dataset into two parts where you can use some part for training and we can use some part of dataset for testing the model to which will help in getting the accuracy.

IV. FUTURE SCOPE

In future we can add large number of dataset and will use various machine algorithms and will try to and compare them with each other's algorithm for getting the accuracy.

V. CONCLUSION

This paper mainly focuses on predicting the price of used cars with the help of dataset which is gathered from kaggle. After preprocessing that gathered data using its factors like its independent variables and dependent variable a machine supervised learning.

Using the train test method by splitting x and y where x is the independent variable and y is the dependent variable and also applying the ratio i.e., test size. Applying test size means if we take test size as 0.3 it means our training data size will be 70% and test size will be 30%.

Training the dataset mean it is used for training the model and testing means it is used for finding the accuracy.

Applying linear regression algorithm: - In this work linear regression algorithm is used for prediction of cars. We have used linear regression because; linear regression is a regression problem. Regression is used when the expected output value or dependent variable is continuous in nature. Regression is used when modeling a relationship between independent and dependent variable.

We are applying linear regression for prediction because it is based on supervised learning. It has the task of finding the relationship between multiple independent variables(x) and one dependent variable(y).

An algorithm named linear regression is applied on it because it is a regression problem and a regression problem is always expecting a continuous value. This will help in predicting the price of used cars.

References

- [1] Praful Rane,Deep Pandya,Dhawal Kotak “Used CarPricePrediction”,2021
- [2] Ketan Agrahari,Ayush Chaubey,Mamoor Khan, Manas Srivastava, “Car Price Prediction Using MachineLearning”,2021
- [3] Abhishek Jha, Dr. Ranveer Singh, Manish, Imran Saifi, Shipra Shrivastav, “Used car price prediction”

