iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: V Month of publication: May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51760

Used Car Price Prediction Using Machine Learning

Raghav S[1], Bharath T R[2], Bharath G T[3], Ravi Prakash Hegde[4], Sohan Ramesh[5]
[1]Assistant Professor, [2, 3, 4, 5]Student, Department of Information Science and Engineering Sir M Visvesvaraya Institute of Technology, Bangalore, India and affiliated to Visvesvaraya TechnologicalUniversity, Belagavi, Karnataka, India

***Abstract:** Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend more purchase used cars. A primary objective of this project is to estimate used car prices by using attributes that are highly correlated with a label (Price). The aim of this project is to develop a model for predicting the prices ofused cars. The study uses two popular regression algorithms, Random Forest and Linear Regression. Thedata used in the project is obtained from various sources and includes information about the make and model of the car, its mileage, age, and other factors that affect its price. The study evaluates the performance of the models using metrics such as Mean Absolute Error, Mean Squared Error, and R-Squared. Theresults show that both Random Forest and Linear Regression models can accurately predict used car prices, with Random Forest performing slightly better. The study provides valuable insights into the factors that influence the prices of used cars and can assist car dealerships, buyers, and sellers in making informed decisions then the model will be integrated into a webpage for general public to use.*

***Keywords:** Machine Learning, Car price prediction, Linear Regression, Random Forest.*

I. INTRODUCTION

Used cars are famous in the automotive industry and are a price-powerful way for individuals to very own a vehicle. Accurate prediction of the used vehicle price may bebeneficial for both customers and dealers in making informed selections. In this paper, we can explore using linear regression and random forest algorithms for predicting theprice of used cars. Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models. Depending on various parameters we will determine the price of the car.

Regression Algorithms are because they provide us with continuous value as output and used not a categorized value because of which it will be possible to predict the actual price of a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the price of a car according to user's inputs. The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. Different websites have different algorithms to generate the retail price of used cars, and hence there isn't a unified algorithm for determining the price.

II. RELATED WORK

While most studies on used car price prediction use machine learning in paper[1]. "Used Car Price Prediction Using Machine Learning Techniques: A Comparative Study" by A. Yadav et al. This study compares the performance of several machine learning algorithms, including linear regression, decision tree, and k-nearest neighbours, in predicting used car prices and reports an accuracy rate of around 80%.

In paper[2]. "Predicting Used Car Prices using Machine Learning Techniques" by N. Nandwani et al. This study uses a dataset of over 3,000 car listings from a local marketplace to develop a machine-learning model for predicting car prices. The authors use a combination of linear regression and decision tree algorithms and report an accuracy rate of around 70%.

In paper[3]. "Used Car Price Prediction: An Empirical Study of Machine Learning Techniques" by H. M. Al-Bayati et al. This study evaluates the performance of several machine learning algorithms, including linear regression, support vector regression, and decision tree, in predicting used car prices. and report an accuracy rate of around 65%.

This paper[4] aims to compare the efficiency of different models' predictions to find the appropriate one. On the subject of used automobile price prediction, several previous studies have been conducted. To anticipate the value of pre-owned automobiles in Mauritius, Pudaruth employed naive Bayes, k- nearest neighbours, multiple linear regression, and decision trees.

However, because there were fewer cars observed, their results were not good for prediction. In his article, Pudaruth concluded that decision trees and naive Bayes are ineffective for continuous-valued variables.

In paper[5] Peerun et al. conducted a study to assess the neural network's performance in predicting used automobile prices. However, especially on higher-priced cars, the estimated value is not very close to the real price. In forecasting the price of a used car, they found that support vector machine regression outperformed neural networks and linear regression by a little margin.

Surprisingly, work on estimating the price of used cars is very recent but also very sparse. In her MSc thesis [6], Listiani showed that the regression model built using support vector machines (SVM) can estimate the residual price of leased cars with higher accuracy than simple multiple regression or multivariate regression. SVM is better able to deal with very high dimensional data (number of features used to predict the price) and can avoid both over-fitting and underfitting. In particular, she used a genetic algorithm to find the optimal parameters for SVM in less time. The only drawback of this study is that the improvement of SVM regression over simple regression was not expressed in simple measures like mean deviation or variance.

In another university thesis [7], Richardson works on the hypothesis that car manufacturers are more willing to produce vehicles which do not depreciate rapidly. In particular, by using multiple regression analysis, he showed that hybrid cars (cars which use two different power sources to propel the car,

i.e. they have both an internal combustion engine and an electric motor) are more able to keep their value than traditional vehicles. This is likely due to more environmental concerns about the climate and because of its higher fuel efficiency. The importance of other factors like age, mileage, make and MPG (miles per gallon) were also considered in this study. He collected all his data from various websites.

In paper[8] Du et al. This system not only estimates the best price for reselling the cars but also provides advice on where to sell the car. Since the United States is a huge country, the location where the car is sold also has a non-trivial impact on the selling price of used cars. A k-nearest neighbour regressionmodel was used for forecasting the price. Since this system was started in 2003, more than two million vehicles have been distributed via this system.

In paper[9], Gonggi proposed a new model based on artificial neural networks to forecast the residual value of private used cars. The main features used in this study were: mileage, manufacturer and estimated useful life. The model was optimised to handle nonlinear relationships which cannot be done with simple linear regression methods. It was found that this model was reasonably accurate in predicting the residual value of used cars.

III. OBJECTIVE

The main objective of this project are:

1) To develop an efficient and effective model which predicts the price of a used car according to user’s inputs.
2) To achieve good accuracy.
3) To develop a User Interface (UI) which is user-friendly and takes input from the user and predicts the price.

IV. TOOLS USED

A. Hardware

The minimum requirements needed to install andrun are as followed:

1) Operating System- Windows 7,8,10
2) Processor- Dual-core 2.4 GHz (i5 or i7 seriesIntel processor or equivalent AMD.
3) RAM- 8GB

B. Software

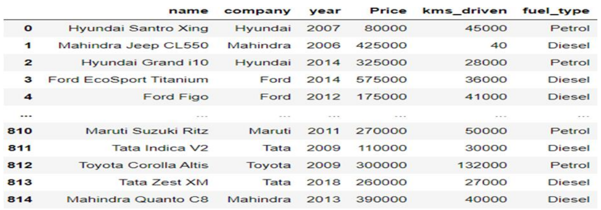
1) Python Flask
2) VS Code
3) PIP 2.7
4) Jupyter Notebook
5) Chrome

V. METHODOLOGY

The main goal of this method is to give users an accurate estimate of how much has to be paid for a given vehicle. The model may give the customer a record of possibilities for various cars based on the details of the automobile the customer wants.

The system assists in providing the customer with sufficient data to help him to reach a conclusion. The used car market is expanding at an exponential rate, and vehicle vendors may profit from this by offering incorrect prices to capitalise on the demand. As a result, a system that can predict the price of a car based on its parameters while also taking into consideration the costs of competing vehicles is necessary. Our system fills in the gaps by providing buyers and sellers with an estimate of the car's value based on the bestalgorithm available for price prediction.

1) *Data Collection:* To solve machine learning problems firstly we require raw data because without raw data we cannot do machine learning problems. The first step is to collect data on the used cars, including their year, company, model name, fuel_type, and kilometers_driven. This data can be obtained from various sources such as online car dealerships, car auction sites, or private sellers.



	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

Fig 1: Dataset

2) *Data Preprocessing:* The collected data needs to be cleaned and processed to ensure that it is suitable for use in the algorithms. This involves removing any missing or irrelevant data, encoding categorical variables, and scaling numerical variables. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.
3) *Data Splitting: The* preprocessed data is split into a training set and a testing set. The training set, which constitutes 80% of the data, is used to train the model, while the testing set, which constitutes 20% of the data, is used to evaluate the performance of the model.
4) *Model evaluation:* Testing the trained model on the testing set and evaluating its performance using metrics such as mean absolute error, root mean square error, and R-squared.
5) *Performance Comparison:* The efficiency of the linear regression and random forest models is compared to determine which algorithm is more efficient for predicting the price of used cars.
6) *Price Prediction:* The final step is to use the selected model to predict the price of a used car based on its features. The predicted price can be used by buyers and sellers to make informed decisions in the used car market.

In conclusion, the working model for predicting the price of used cars involves data collection, preprocessing, data splitting, applying linear regression and random forest algorithms, comparing their performance, and using the selected model for price prediction.

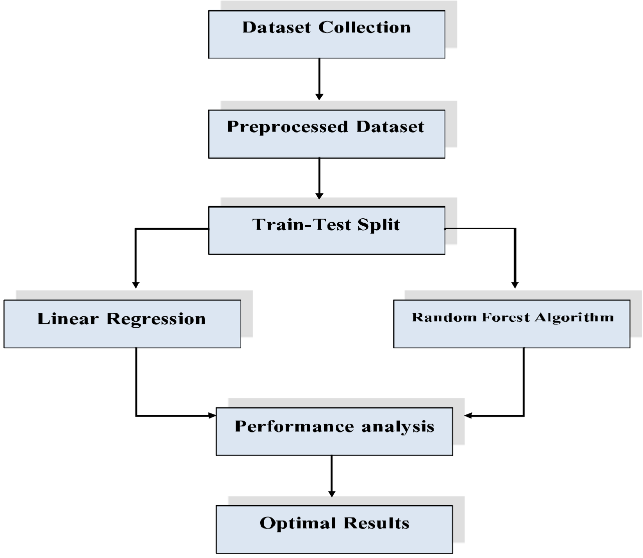


Fig 2: Process Flow Diagram

A. Linear Regression

Linear regression is a statistical model that analyzes the linear relationship between the dependent variable and one or more independent variables. In this study, we used linear regression topredict the price of used cars based on their year, company, model name, and fuel type, kilometres driven. We trained the model using a dataset of used car prices and their corresponding features. Linear regression fits a straight line or surface that minimizes the discrepancies between the predicted output and the actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. Then estimate the value of A (dependent variable) from B (independent variable).

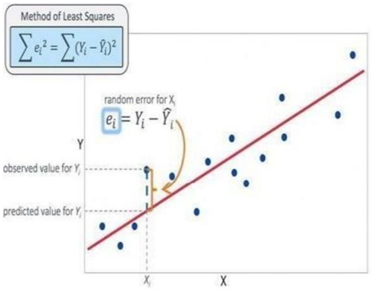


Fig 3: Linear Regression

Linear regression is useful for finding relationship betweenmultiple continuous variables. There are multiple independent variables and single independent variable

y = m1X1+m2X2+……+bm1, m2, m3 …. → slope

b → y intercept

X1, X2, X3 …… → independent variablesy → dependent variables.

The dataset was split into a training set and a testing set, with 80% of the data used for training and 20% for testing. The linear regression model achieved an efficiency of 81%on the test dataset, which means that the model accurately predicted the price of 81% of the cars in the testdataset.

B. Random Forest

Random is a popular machine-learning language algorithm that is classified under supervised learning techniques. It canbe used for both Classification and Regression problems in machine learning. It is based on the concept of ensemble learning. This is a machine-learning algorithm that uses decision trees to create a model. It creates multiple decision trees and combines their outputs to make a final prediction. The dataset was split into a training set and a testing set, with80% of the data used for training and 20% for testing. The random forest model achieved an efficiency of 87% on the test dataset, which means that the model accurately predictedthe price of 87% of the cars in the test dataset.

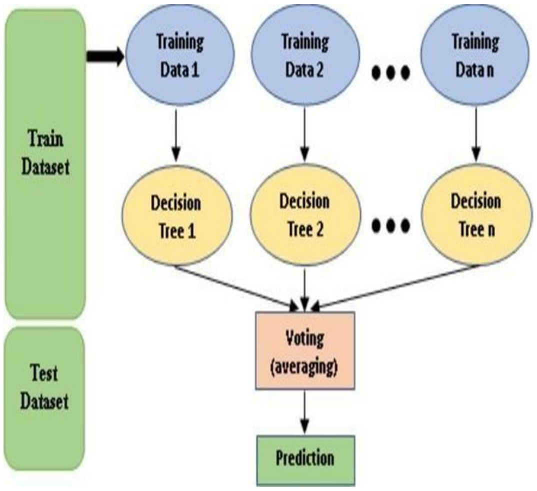
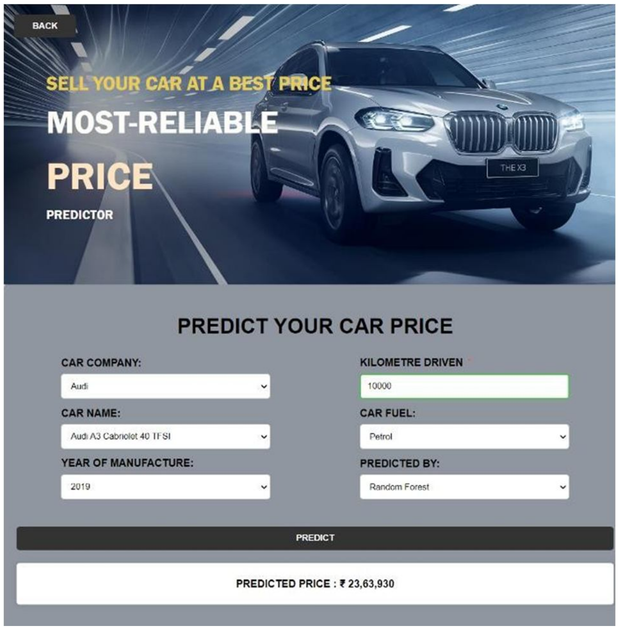


Fig 4: Working of Random Forest Algorithm

VI. IMPLEMENTATION

The tech stack used in the project contains Python libraries like NumPy, Pandas and MatPlotLib for dataset manipulation and visualization. PyTorch, Scikit-learn, and Pickle are used for model implementation and serialization. The Flask framework is used as the backend of the project. The HTML templates with CSS styling are used to predict the car price based on user's input.

EXPERIMENTAL OUTPUT:



PREDICT YOUR CAR PRICE

CAR COMPANY: Audi

KILOMETRE DRIVEN * 10000

CAR NAME: Audi A3 Cabriolet 40 TFSI

CAR FUEL: Petrol

YEAR OF MANUFACTURE: 2019

PREDICTED BY: Random Forest

PREDICTED PRICE : ₹ 23,63,930

Fig 5: Web interface to accept input from users.

VII. CONCLUSION AND FUTURE WORK

The model we were making is to predict the value of a second- hand car using machine learning techniques. We have collectedthe data of cars from Kaggle having attributes like different cars and their year, km driven, fuel type, model name, company etc. The data is then processed using different algorithms where we chose linear regression and random forest algorithms and compared them, getting the Random Forestalgorithm as the most accurate amongst them, so we have used Random Forest because irrespective of size it runs efficiently and gives more accuracy than any other algorithm. Further, it would be available in GUI as a Web application developed using Pythonflask making it user-friendly so that users could give input and get the price of a car according to it.

The main limitation of this study is the low number of records that have been used. In future work, we intend to collect more data and also in future, this machine learning model may bind with various websites which can provide real-time data for price prediction. Also, we may add large historical data on car prices which can help to improve the accuracy of the machine learning model. We can build an Android app as a user interface for interacting with users. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

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).

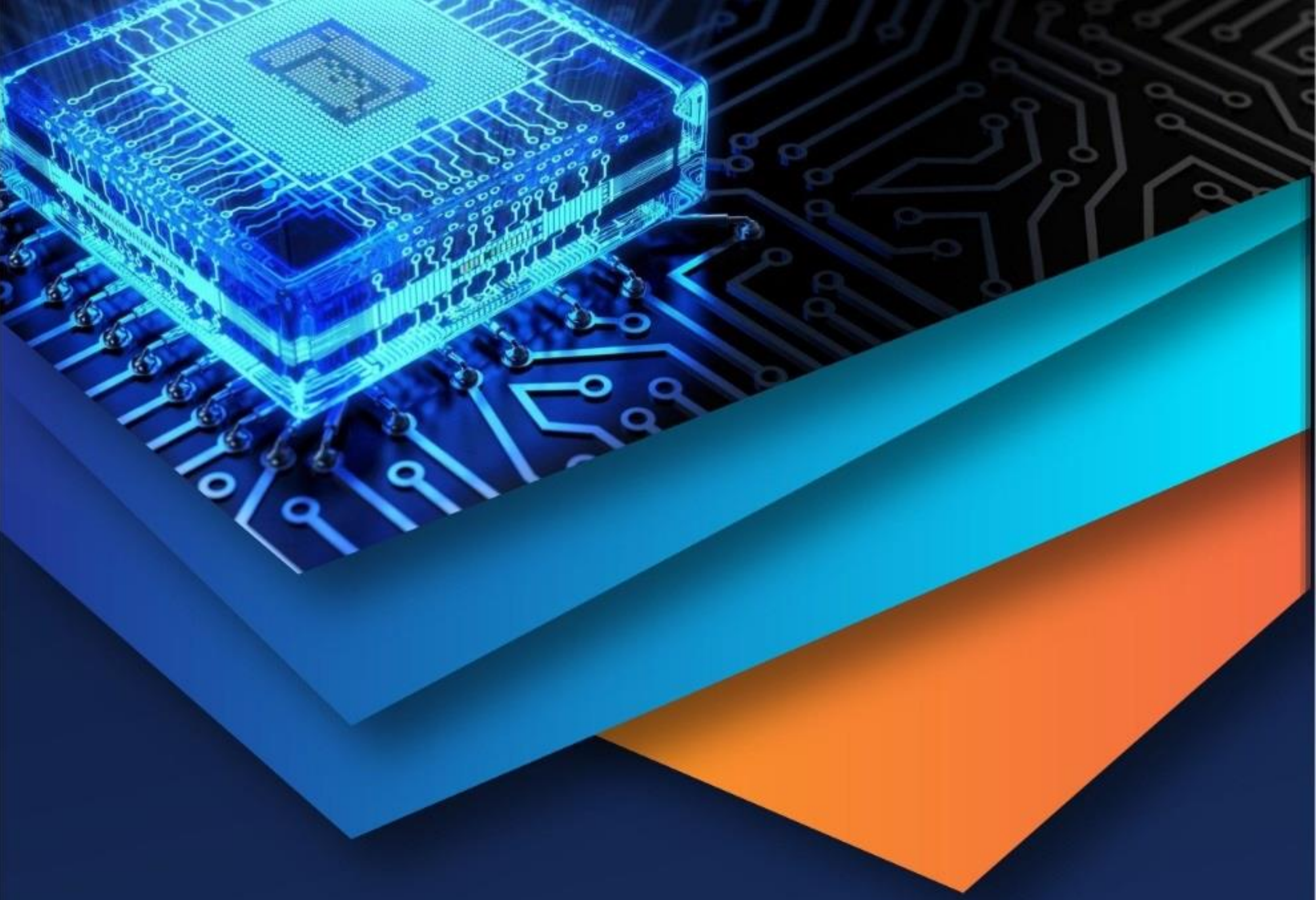
[3] Ning Sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, "Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory"; (Hohai University Changzhou, China)

[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)

[5] Jaideep A Muley, "Prediction of Used Cars' Prices by Using SAS EM", Oklahoma State University

[6] Nabarun Pal, "A methodology for predicting used cars prices using Random Forest", Future of Information and Communications Conference, 2018

[7] Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, "Prediction car prices using qualify qualitative data and knowledge-based system" (Hanoi National University)



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY