INTERNATIONAL JOURNAL OF RECENT TRENDS IN TECHNOLOGY AND ENGINEERING (IJRTTE)

Journal Home Page: https://ijrtte.com/

Regression and Prediction of Cars Using Machine Learning

Nisharani RA[1] and Privietha P [2]

[1]Student, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India.
[1]nrnisharani2@gmail.com, [2] priviethaprabhakar@gmail.com

Abstract. The mission entitled “Regression and prediction of automobiles the usage of Machine Learning” is an automobile dataset analysis. Regression evaluation a set of statistical strategies for estimating the relationships between an established variable and impartial variable. Data evaluation is carried out the use of panda’s library bundle in python. The subject is decided by way of a set of education documents. Readily accessible dataset from Kaggle internet site is used for facts analysis. In order to assemble a regression model, computing device gaining knowledge of algorithm is used Random Forest regression mannequin is designed and implemented. Dataset is break up as cut up as 70% for education and 30% for testing. The mannequin is nicely skilled and based totally on the training. The data are examined and accuracy is calculated. Each and each information evaluation are stated and tabulated for in addition identification. Random wooded area is a Supervised Machine Learning Algorithm that is used extensively in Classification and Regression problems. It builds selection bushes on unique samples and takes their majority vote for classification and common in case of regression. Predicting the rate of used automobiles is one of the big and fascinating areas of analysis. As an extended demand in the second-hand vehicle market, the commercial enterprise for each shoppers and marketers has increased. For dependable and correct prediction, it requires specialist expertise about the area due to the fact of the fee of the vehicles established on many vital factors. Decision Tree is one of the most frequently used, sensible techniques for supervised learning. Voting Regressor is an ensemble meta-estimator that suits a number of base regressors, every on the complete dataset to common the character predictions to shape a last prediction.

Keywords: Random Forest Regression, Deep Learning, Decision Tree, Voting Regressor.

1. Introduction

With an increasing number of flourishing extents of personal automobiles and the development of the used vehicle market, used automobiles have to grow to be the pinnacle precedence for buyers. The fee of a used automobile is a vital thing of a profitable transaction for each shoppers and sellers. For vehicle buyers, acknowledging the rate of used automobiles lets in for buying and selling with peace of mind; for auto sellers, evaluating the residual cost of used automobiles can assist them set costs reasonably. In different commodity markets, such as inventory markets, gold markets, and agricultural markets, rate forecasting has been a key focal point of research. Used cars, as a commodity, can be priced in the identical way. However, used vehicle transactions are a good deal extra complicated than different commodity transactions, as the sale charge is influenced no longer solely by using the primary points of the auto itself, such as brand, power, and structure, however additionally through the situation of the car, such as mileage and utilization time, as properly as a lack of at present on hand strategies figuring out which elements hit the sale charge most dramatically. At the equal time, on-line transactions additionally make it challenging to investigate the rate of used cars. Used motors are journey goods. Different from search goods, it is challenging for customers to make a buy choice based totally on the car configuration. The true person trip has a large influence on the purchase. This exacerbates the challenge of predicting used auto expenditures accurately. Therefore, how to display the characteristic variables that have an effect on

Volume: 02 Issue: 02 Year: 2023
Received: 07.04.2023 Received in revised form: 12.05.2023 Accepted: 06.06.2023
Available Online: 19.06.2023
Published by NTL Publisher

the fee of used automobiles and enhance the accuracy of fee prediction of used motors is of tremendous magnitude for honest transactions between customers and dealers and the sustainable and healthful improvement of the used automobile market.

Determining whether or not the listed rate of a used automobile is a difficult task, due to the many elements that force a used vehicle's rate on the market. The focal point of this venture is growing computer mastering fashions that can precisely predict the fee of a used vehicle based totally on its features, in order to make knowledgeable purchases. Implement and consider a variety of gaining knowledge of strategies on a dataset consisting of the sale expenditures of special makes and models. Will examine the overall performance of a number of laptops gaining knowledge of algorithms like Random Forest Regression, XGBoost Regressor, and Decision Tree Regressor and pick the great out of it. Depending on a number of parameters, will decide the rate of the car. Regression Algorithms are used due to the fact they furnish with non-stop fee as an output and now not a labeled fee due to the fact of which it will be viable to predict the proper rate an automobile as an alternative than the rate varies of a car. User Interface has additionally been developed which acquires enter from any person and shows the rate of an automobile in accordance to user's inputs.

2. Materials and Methods

2.1 Random Forest Algorithm

Random Forest algorithm is Ensemble-Bagging technique which operates with the aid of setting up more than one choice timber for the duration of the education phase. The Decision of the majority of the outputs (trees) is chosen with the aid of the random woodland as the ultimate decision. The most important gain of the usage of Random Forest is that it is a combination of each kinds of supervised getting to know issues i.e. Regression and Classification. The Random Forest algorithms are used in many desktop mastering functions such as:

For Remote Sensing such as ETM units used to gather photographs of earth's surface, Random Forest is the first preference as it presents Higher Accuracy in a much less coaching time [1]. For Multiclass Object Detection, Random Forest is used as it gives higher detection in tricky environments [2]. Some gaming consoles use this algorithm as it is used to tune physique motion and recreates it in the game. Random Forest algorithm is skilled to perceive the physique components and algorithm learns from it. Then it identifies the physique components of the customers such as hands, feet, face, eyes, nostril etc. "Random wooded area consists of an outsized wide variety of man or woman choice bushes that function as an ensemble the place every tree inside the random wooded area spits out a class prediction then the class with the essential votes turns into our model's prediction. [3]" Measurement of the relative significance of every characteristic on the prediction is some other benefit to the Random Forest Algorithm. Another excellence of the random wooded area algorithm is that it is elementary too. In Random Forest, every tree is picked from a random subset of features. This excessive stage of version consequences in decrease correlation amongst bushes and introduces extra diversification [4].

2.2 Decision Tree Regressor

Decision tree is one of the nicely recognized and effective supervised computers gaining knowledge of algorithms that can be used for classification and regression problems. The mannequin is primarily based on selection regulations extracted from the education data. In regression problem, the mannequin makes use of the fee as a substitute of category and imply squared error is used to for a choice accuracy. Decision tree mannequin is no longer properly in generalization and touchy to the adjustments in coaching data. A small alternate in an education dataset might also impact the mannequin predictive accuracy [5]

2.3 Library and Packages Details Numpy

NUMPY: NumPy is a Python library used for working with arrays. It additionally has features for working in area of linear algebra, Fourier transform, and matrices. It is an open supply venture and you can use it freely. NumPy arrays are quicker and extra compact than Python lists. An array consumes much less reminiscence and is handy to use. NumPy makes use of a good deal much less reminiscence to keep facts and it affords a mechanism of specifying the information types.

PANDAS: Pandas is a fast, powerful, bendy and convenient to use open supply facts evaluation and manipulation tool, constructed on pinnacle of the Python programming language. Pandas is effective and bendy quantitative evaluation tool, pandas have grown into one of the most famous Python libraries. It has an extraordinarily energetic neighborhood of contributors.

SEABORN: Seaborn is a Python facts visualization library based totally on matplotlib. It offers a high-level interface for drawing captivating and informative statistical graphics. It builds on pinnacle of matplotlib and integrates intently with pandas' statistics structures. Seaborn helps you discover and recognize your data [6].

3. Methodology

Data series is the highest step for any project. The machine is designed for used vehicles in the Mumbai region, for which the statistics of used vehicles is amassed the use of on 15-March- 2021. We used ensemble computing device studying strategies is used to enforce (Thereby produces upgraded Outcomes than a single model would. Can instruct an ensemble and in addition use it to make predictions. Hence, an ensemble is a supervised mastering algorithm.

Using distinctive ensemble methods, can mix quite a number of models, thereby, shifting on the direction of attaining higher accuracy. Suppose that you have designed an android application, earlier than making it public you desire to comprehend its ratings. Each of these fashions contributes to raise the overall performance of the ensemble. Stacking is an ensemble method that makes use of predictions outputted from more than one fashions to assemble a new model, which is similarly used to make predictions on the check set. Random Forest is a bagging algorithm whereas XGBoost is a boosting algorithm; these algorithms are used to put in force the proposed system. Random Forest algorithm is a famous supervised laptop gaining knowledge of algorithm that depends on the thought of ensemble getting to know and can be deployed for each classification and regression issues in laptop learning. The higher range of timber in the woodland makes a sturdy woodland that leads to correct and secure prediction. Figure 1 shows the flow diagram of the methodology.

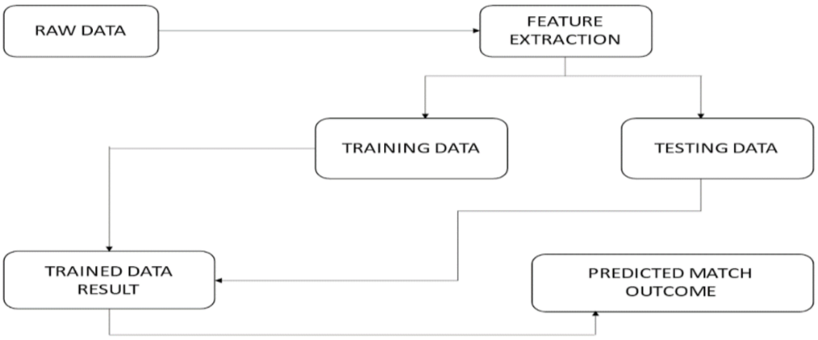
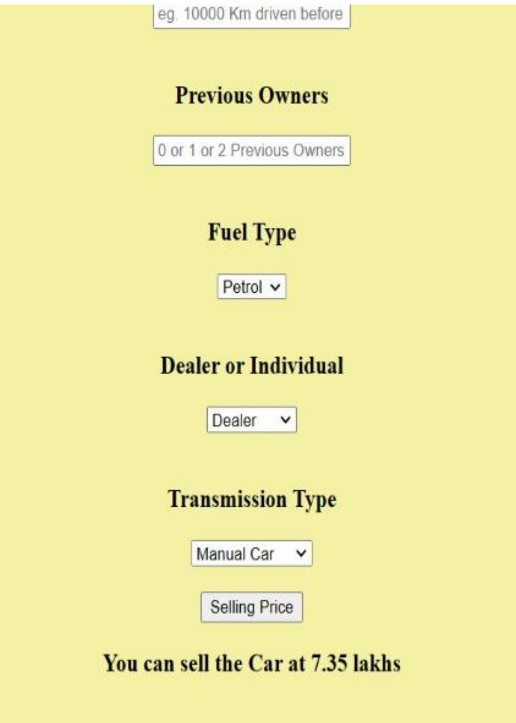


Figure 1: Flow diagram.

4. Implementation and Result Analysis

The gadget is applied in python the use of anaconda as IDE. Sklearn device is used. A python bundle sklearn. Ensemble. Random Forest Regressor is used to enforce the random wooded area algorithm. This assessment is crucial to discover out the accuracy stage of prediction that used to be produced by means of the model. Random woodland is a Supervised Machine Learning Algorithm that is used broadly in Classification and Regression problems. It builds choice bushes on one of a kind samples and takes their majority vote for classification and average in case of regression. Predicting the rate of used automobiles is one of the massive and fascinating areas of analysis. As an improved demand in the second-hand automobile market, the enterprise for each consumers and retailers has increased. For dependable and correct prediction, it requires professional information about the discipline due to the fact of the rate of the motors based on many necessary factors. Decision Tree is one of the most used, realistic strategies for supervised learning. Voting Regressor is an ensemble meta-estimator that suits numerous base regressors, every on the total dataset to common the man or woman predictions to structure a closing prediction. Figure 2 shows the result analysis in pickle.



eg. 10000 Km driven before

Previous Owners

0 or 1 or 2 Previous Owners

Fuel Type

Petrol

Dealer or Individual

Dealer

Transmission Type

Manual Car

You can sell the Car at 7.35 lakhs

Figure 2: Result Analysis.

4.1 Car Price Prediction

An auto fee prediction has been a high-interest lookup area, as it requires substantive effort and information of the discipline expert. A significant range of awesome attributes are examined for dependable and correct predictions. The foremost step in the prediction procedure is the series and pre-processing of the data. In this project, facts were once normalized and cleaned to keep away from pointless noise for laptop getting to know algorithms. Applying a single desktop algorithm to the records set accuracy used to be much less than 70%. Therefore, the ensemble of more than one computing device gaining knowledge of algorithms has been proposed and this aggregate of ML techniques good points an accuracy of 93%. This is a sizable enchantment in contrast to the single computer mastering technique approach. However, the downside of the proposed gadget is that it consumes a good deal greater computational sources than a single computer gaining knowledge of algorithm. Although this machine has finished astounding overall performance in the vehicle fee prediction problem, it can additionally be applied the usage of a superior laptop getting to know mannequin and with Deep getting to know strategies to enhance its effectively and accuracy. Moreover, as innovation has been expanded in cars and we can study Electric motors have won public interest and are favored via most than ordinary car. Figure 3 represents the Heat map showing Top correlation Features.

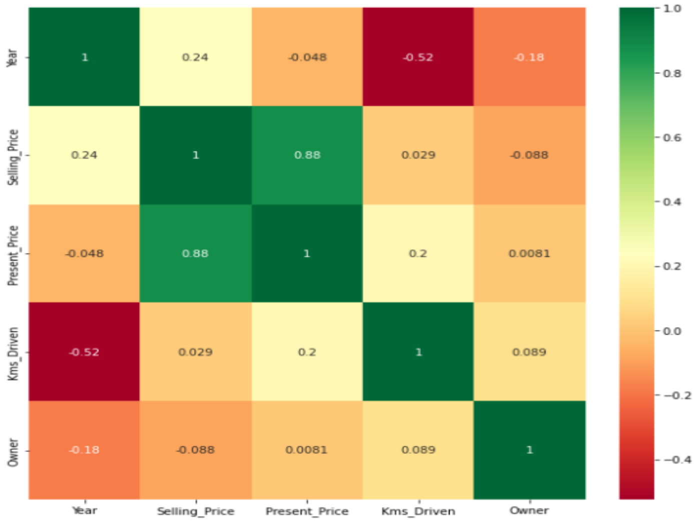


Figure 3: Heatmap showing Top correlation Features.

The distplot in parent 3 under indicates a regular distribution of the mannequin with check dataset, this proves the accuracy of this model. Hence, we can say that the prediction of this mannequin is extraordinarily accurate.

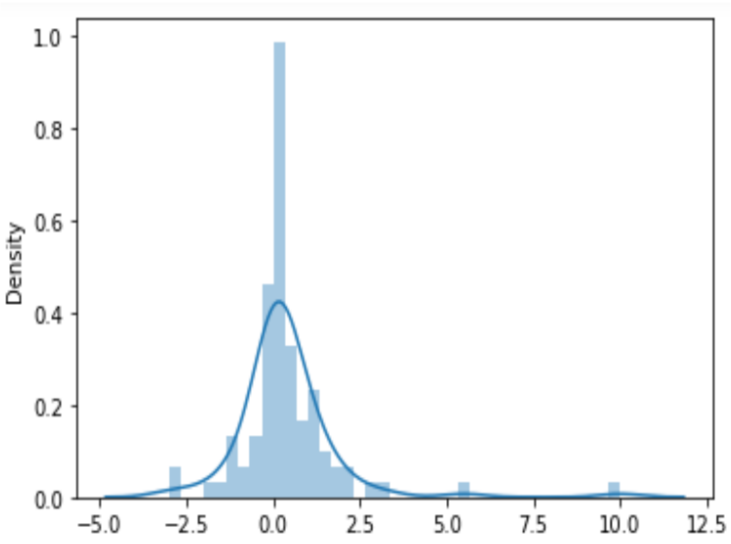


Figure 4: Distplot showing the distribution.

The scatterplot in Figure 4 and 5 indicates a linear distribution which ensures the accuracy of this mannequin so we can in the end say that prediction of the promoting fee the usage of accessible dataset is accurate.

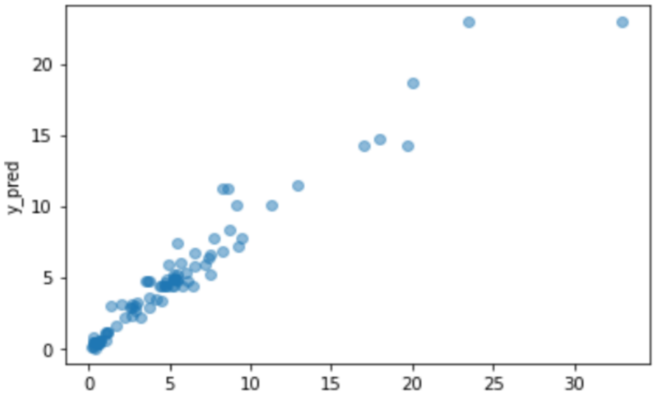
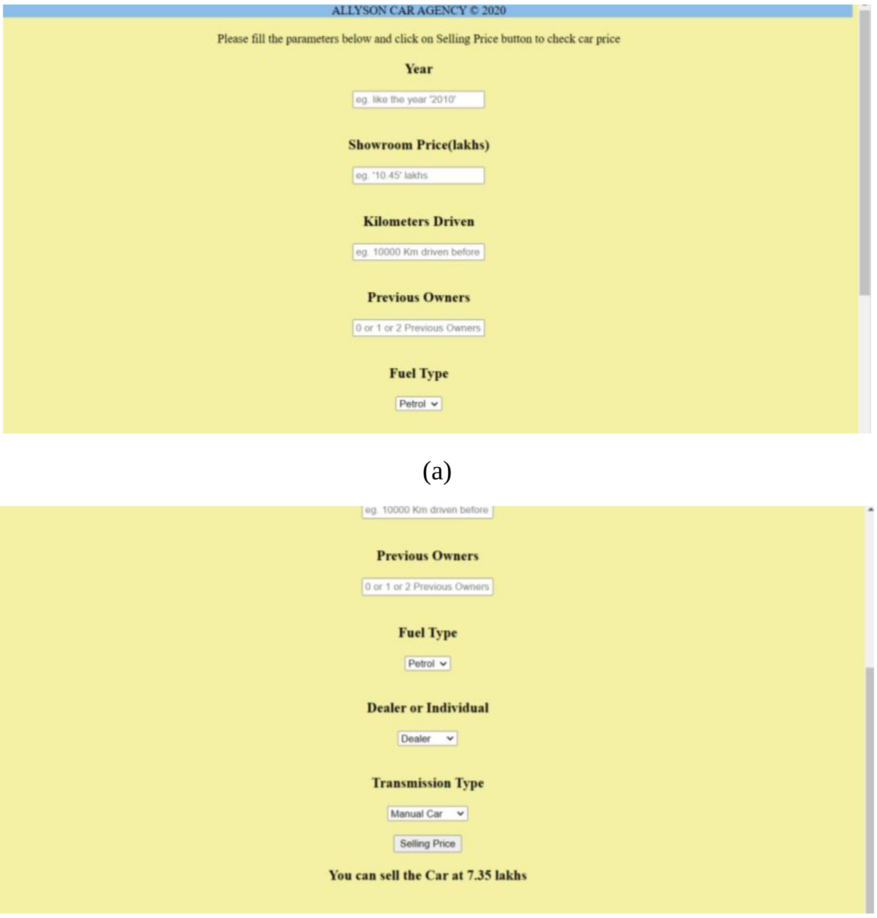


Figure 5: Scatterplot showing the distribution.

Finally, the usage of the flask we can install this mannequin as a web-based application. In addition, deployed this software the usage of the equal platform the use of flask as proven. Figure 6 is the screenshot of pickle version, the front end or the user end to test the data and to get the accuracy level.



ALLYSON CAR AGENCY © 2020

Please fill the parameters below and click on Selling Price button to check car price

Year

eg. like the year '2010'

Showroom Price(lakhs)

eg. '10.45' lakhs

Kilometers Driven

eg. 10000 Km driven before

Previous Owners

0 or 1 or 2 Previous Owners

Fuel Type

Petrol

(a)

eg. 10000 Km driven before

Previous Owners

0 or 1 or 2 Previous Owners

Fuel Type

Petrol

Dealer or Individual

Dealer

Transmission Type

Manual Car

You can sell the Car at 7.35 lakhs

(b)

Figure 6 (a), (b): Car Prediction.

5. Conclusion

This mannequin is primarily based on the desktop studying algorithms and we have been trying to predict the selling charge of the used automobiles primarily based on the dataset supplied at Kaggle. To predict this dataset, we used two computer studying algorithms i.e. Random Forest and Extra Tress Regressor. The prediction of this mannequin is in addition in contrast with the check dataset created with the aid of selecting random values from the authentic dataset and the assessment of the prediction is in addition evaluated the usage of distinct methods. After an entire contrast of the predictive model, we can conclude that the accuracy of this mannequin is very and Random Forest and Extra Tree Regression is one of the nice algorithms for regression problems. These two algorithms are enormously correct and speedy in prediction irrespective of the measurement of the dataset.

6. Future Enhancement

An automobile charge prediction has been a high-interest lookup area, as it requires substantive effort and information of the area expert. A significant variety of awesome attributes is examined for dependable and correct predictions. The fundamental step in the prediction system is the series and pre-processing of the data. In this project, facts were once normalized and cleaned to keep away from useless noise for laptop getting to know algorithms. Applying a single computing device algorithm to the records set accuracy was once much less than 70%. Therefore, the ensemble of a couple of computing device gaining knowledge of algorithms has been proposed and this aggregate of ML techniques features an accuracy of 93%. This is a widespread enhancement in contrast to the single laptop gaining knowledge of approach. However, the disadvantage of the proposed machine is that it consumes an awful lot of greater computational assets than a single desktop getting to know algorithm. Although this device has executed impressive overall performance in the automobile charge prediction problem, it can additionally be applied the usage of a superior desktop studying mannequin and with Deep gaining knowledge of methods to enhance its effectively and accuracy. Moreover, as innovation has been improved in motors and we can look at Electric motors have won public interest and are favored through most than a regular car.

References

1. Tomar, Ravi, Hanumat G. Sastry, and Manish Prateek. 2020. “A Novel Protocol for Information Dissemination in Vehicular Networks.” Pp. 1–14 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 11894 LNCS. Springer, Cham.
2. Bansal, Parnika, Bhawna Aggarwal, and Ravi Tomar. 2019. “Low-Voltage Multi-Input High Trans- Conductance Amplifier Using Flipped Voltage Follower and Its Application in High Pass Filter.” Pp. 52 in 2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019. IEEE.
3. Tomar, Ravi, Rahul Tiwari, and Sarishma. 2019. “Information Delivery System for Early Forest Fire Detection Using Internet of Things.” Pp. 477–86 in Communications in Computer and Information Science. Vol. 1045. Springer, Singapore.
4. Kumar, Shiwanshu and Ravi Tomar. 2018. “The Role of Artificial Intelligence In Space Exploration.” Pp. 499–503 in 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT). IEEE.
5. Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast Cancer Detection Using Machine Learning Algorithms. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 114–118.
6. Rohit, Sabitha, S., & Choudhury, T. (2018). Proposed method for e book suggestion primarily based on consumer k- NN. In Advances in Intelligent Systems and Computing.