IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 11, November 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

Used Bike Price Prediction Using Random Forest Method

Pratiksha Patil[1], Paritosh Nehete[2], Yash Jain[3]

B.E Student, Department of Computer Engineering, Smt. Indira Gandhi College, Ghansoli, Maharashtra, India[1]

T.E Student, Department of Information Technology, Dr.DY Patil Institute of Technology, Pune, Maharashtra, India[2]

S.E Student, Department of Computer Engineering, Medicaps University, Rau, Indore, Madhya Pradesh, India[3]

ABSTRACT: The used bike prediction is a much needed and high-interest research area due to The exponential growth of Steel cost and many other factors that increased the input cost value of bike production hence increasing the overall cost of the bike this increased sales of second hand used bikes in the post-covid-19 pandemic. In this paper, we have implemented random Forest algorithm as a supervised learning method for Predicting used bike prices. We Used a dataset of used bikes which is web scrapped from India's leading website droom.com for used vehicle sale using webscraper.io Multiple data science and machine learning algorithms were used but the most efficient was random Forest algorithm which is supervised learning techniques and the mixture of regressions and classification this regression model is most likeable to be used for numerical data analysis through which we were able to get an actual predicted price of the bike rather than getting a range of prices concluded with a training accuracy of 92 % and the testing accuracy was 84.01%%. The model can predict the price of bikes accurately by choosing the most correlated features. Random forest and the hypermeters provided by the sklearn library enabled us to predict the price of used bikes efficiently. Other algorithms will be implemented in the near future to improve the model's accuracy and speed.

KEYWORDS: Used bike price prediction, Machine learning, Random Forest, Regression, Decision tree.

I. INTRODUCTION

Machine learning is a growing technology that allows computers to automatically learn from past data. To develop mathematical models and generate predictions based on past data or information, machine learning makes use of different methods. Supervised learning is when a model is trained in a labelled database. A set of labelled data is one that has both input and output. Unchecked reading, also known as unchecked machine learning, uses machine learning algorithms to analyze and integrate non-label data sets. These algorithms detect hidden patterns or data collection without the need for human intervention. Learning to strengthen is part of machine learning. It is about taking the necessary steps to increase the income in a particular situation. It is used by a variety of software and equipment to determine the best course of action or approach to a particular situation.Ensemble Learning is the most popular and recommended machine learning method, in which multiple single models, also known as basic models, are integrated to create a more efficient guessing model.Bagging, also known as Bootstrap aggregating, is an integrated learning approach for improving the performance and accuracy of machine learning systems.It is used to cope with bias-variance transactions and reduces the predictive model's variability.Bag insertion is utilised in both retrospective and editing models to avoid data overload.Random Forest is a well-known machine learning technique that is used in conjunction with supervised learning. It can be used for both Scheduling and retrieval problems in ML.It is premised on the idea of integrated learning, which is defined as "the process of combining multiple dividers to solve a complicated problem and improve model performance.Random Forest is a system that takes measures to increase the accuracy of data prediction by integrating the number of decision trees in the various data sets offered."Random Forest includes two phases:the first is to combine the N-trees to build a random forest, and the second is to generate predictions for each tree created in the first phase.To evaluate the data, we are employing the Random Forest Method. This paper is organized as follows. A review work is presented in the following section. Section III describes how to operate while in phase IV, evaluates, and analysed it.. Finally, we conclude the paper with a conclusion on specific indications that point to future work.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

II. RELATED WORK

Wu, J. D., Hsu[1] used a neuro-fuzzy information-based system to predict the price of used cars. There are only three factors to consider: the design of the car, the year it was built, and the engine style in this study. The proposed system produced similar results compared to simple retrospective methods. Car dealers in the USA sell hundreds of thousands of cars annually for rent [2]. Most of these vehicles are returned at the end of the lease term and must be resold. Selling these cars at a fair price has a huge economic impact on their success. The concept of drawing was first introduced by Bertalmio et al. [1]. This approach is inspired by the actual painting process of the artists. Laplacian image smoothing information is distributed according to the isophote guidelines, which are measured by a 90-degree image gradient. Exemplar-based Methodology proposed by Criminisi et al. [2] used the patch of the best example to distribute the target clip including the missing pixels. This process uses a method that combines the dispersion of the structure and blends the texture and thus produces excellent results.Praful Rane[3], Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value because of which it will be possible to predict the actual price a car Enis Gegic.[4]Because the price of an automobile is frequently determined by a number of distinct features and elements, accurate car price prediction necessitates specialist expertise.rather than a car's price range.Bin Yu[5] Random Forest in particular shows that the procedure is consistent and adapts to sparsity, in the sense that its rate of convergence depends only on the number of strong features and not on how many noise variables are present.Sai Sumanth Palakurthy[6] They took dataset from kaggle,they Train data and test the data using Machine Learning Algorithm. In Researchers frequently forecast product prices based on historical data, such as Pudaruth [7], who predicted the worth of motorcycles in Mauritius, which were newer than secondhand bikes.To predict prices, you have implemented multiple line regression, k-nearest neighbours, Nave Bayes, and a decision tree method.Predictability comparisons arising from these strategies showed that prices from these methods are available closely related. Kuiper [8] did the same namely predicting the price of the bike and was introduced, Flexible selection techniques help determine which variables are most appropriate to be included in the model. He encouraged students to use different models and find out how model guessing functions. According to Nau [9], the following requirements should be Satisfaction with the effective descent model:

a. The model should collect useful data and source and how to collect this data should be known.
b. The model must be practical descriptive analysis data when patterns are common can be misunderstood.
c. There should be a comparative power throughout different models.
d. There must be a verification power to see whether the given model fulfils what is said consideration. If there is no compliance with the law stated thought, need for something else the model can be seen.
e. Based on the level of accuracy provided, it must have the ability to select the appropriate model based on.
f. There must be access to help to understand throughout the process [9].In the related work shown above, the authors proposed a guessing model based on a single machine learning algorithm. It is noteworthy, however, that the single- machine learning method did not yield significant predictive results and could be improved by combining various machine learning methods in the ensemble. Gonggi [10] has proposed a new model based on sensory networks to predict the remaining number of autonomous used vehicles. The main factors used in this study were: mile, manufacturer, and average useful life. The model is designed to handle indirect relationships that can be done with simple ways to change the line. It was found that this model was quite accurate in predicting the number of residual vehicles in use.

III. METHODOLOGY

1. Collect dataset: For implementing this model we needed a dataset for training and testing For the Random Forest algorithm and for implementing linear regression we used data set names used Bike prices in India web scrapped from droom.in with the help of different Data Scraper tools, India's finest online market for buying and selling automobiles and bikes. our dataset contains data of approximately 32000 used bikes Pan-India. The dataset contains bikes of numerous brands like Yamaha, Honda, Hero, Bajaj and Royal Enfield and a Variety of unique attributes like Age, Kilometre driven and Power etc necessary for promoting and selling the bike and other few irrelevant attributes like Brake, Stroke, Body type etc were removed for unnecessary inaccuracy cleaning this irrelevant and Nan value helped us for further data preprocessing and data transforming.
2. Clean data, Reduction data, Transformation data: Within a dataset, we cleaned the data by removing any incorrect, corrupted, improperly formatted, duplicate, or incomplete data. After that reduce the data and After that, we transformed the Data.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

3. Data Preprocessing: Before training any model in data science and machine learning data Preprocessing is that the most significant step Data processing is carried out through various steps like:

Step 1: Importing Libraries: Numerous Python libraries were used for data processing like Pandas for data analysis and data manipulation, NumPy for numerical analysis, Matplotlib and Seaborn for better visuals and graphical statistics, Plotly for graph plotting, For data training and testing we used machine learning library famously known as Scikit Learn in Python for Supervised Learning.

Step 2: Import the Dataset: Using the Pandas library in Python to import the dataset for further data manipulation after downloading and web scraping it from Kaggle.

Step 3: Identifying and handling the missing values: During the evaluation of this dataset, we discovered a number of NAN values, which we managed by using the dataset mode for that model with the same specifications as the other brands' models.

Step 4: Encoding categorical data: To train our model, we employed different methods, one of which was the pandas get Dummies() method. Categorical data refers to data that contains specific categories inside the dataset, such as owner type and model type, which encoded this data in encoded binary value.

4. Random Forest Method: This is a supervised learning technique that comprises of multiple decision trees and constructs them while in a training phase. Because of its flexibility to work with multiple variables, the Random Forest technique can manage large datasets. Ensemble-Bagging is a method used in the Random Forest algorithm. The Decision or voting of the majority of the trees is chosen by the random forest. Random forest method constitutes of both supervised learning method i.e regression and classification of Random forest consists of numerous individual decision trees that operate as an ensemble where each tree within the random forest spits out a category prediction then the category with the most votes becomes our model's prediction. Random forest algorithm set of rules is that is easy.

In Random Forest, each tree is picked from a random subset of capabilities. This high stage of various outcomes in decreases correlation amongst trees and introduces greater diversification.

Hyper Parameters in Random Forest:The hyperparameters help to increase the model's speed and its predictive power. Sklearn's built-in random forest functions are considered as hyperparameters. It is Random Forest Algorithm cross-validation method. Both regression and classification tasks can be carried out by this method. versatility, usability and efficiency. Various Hypermaters provides ease for model prediction.

Although random forest is generally used for classification, we converted the problem into a similar regression problem and implemented it as a regression model. Random forest is a type of ensemble learning approach that consists of a group of decision trees, usually hundreds or thousands in number. Individually trained on sections of the dataset, these trees aid in the learning of extremely surprising patterns by becoming exceedingly deep. However, this may result in an issue of overfitting. This is handled by averaging out individual tree forecasts with the purpose of reducing variation and ensuring consistency.

A. Training and Testing: With a 50:10:5 split ratio, we partitioned our input data into training, testing, and cross-validation. The splitting was done at random, resulting in a balance of training and testing data all over the entire dataset.To avoid overfitting and improve generality, this will be done.

B. Accuracy: The R2 coefficient of determination of the prediction is the model score. We can find the accurate data after training and testing our module. The model was fine-tuned so that just the most relevant elements were retained and the rest were deleted. The significant qualities are discovered by correlation, which measures their importance in estimating a vehicle's price. Overall, the random forest model was successful in capturing the intricacies of the data and producing accurate vehicle price estimates.

5. Data Analysation: After Pre-processing our final Used bike dataset Contained approximately 6 attributes for used second-hand bikes like 'model', 'price', 'city', 'age', 'kilometre driven' and 'brand' out of these attributes most important features required for our prediction models where

1. Price - Selling price of second used bike
2. Kilometer Driven- kilometre drove by bike
3. Bike Name- Bike name and model
4. Brand- Bike's manufacturing company
5. Age - Age of bike

Simply by reviewing our input data set, we can observe that used cars were valued at around €10000 on average, with an average kilometre reading of 105000 kilometres. After 10 years of use, 50 per cent of the bike in our database were sold.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

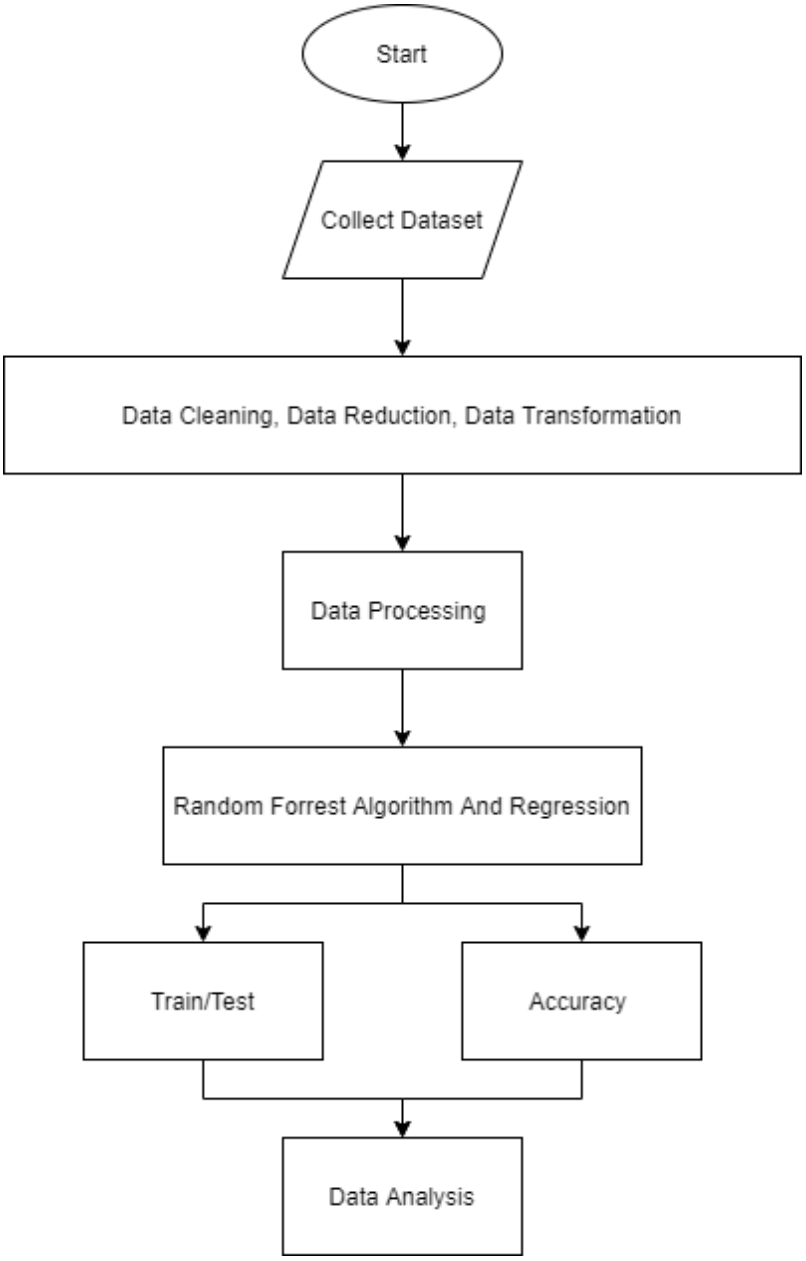

FIGURE NO 1
FLOWCHART OF USED BIKE PRICE PREDICTION METHODOLOGY

IV. RESULT AND DISCUSSION

The average prices of each bike type are shown in Figure 2. It demonstrates that bike types such as Bajaj Pulsor, Suzuki Gixxer, Royal Enfield have higher average prices. This data can be used to determine which vehicle types people are most likely to purchase.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

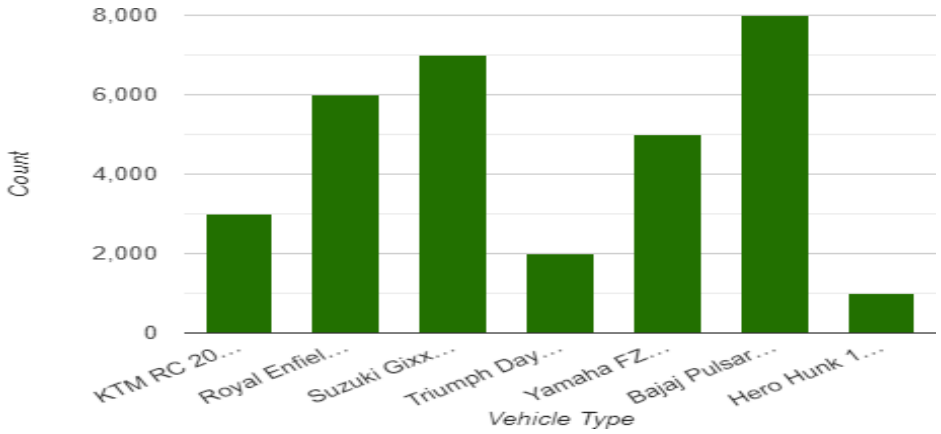

FIGURE NO.2

The top ten average prices by the brand are shown in Figure 3. The average price of a Royal Enfield is around Rs.100000 while the next highest is around Rs.80000, as shown in this bar graph.

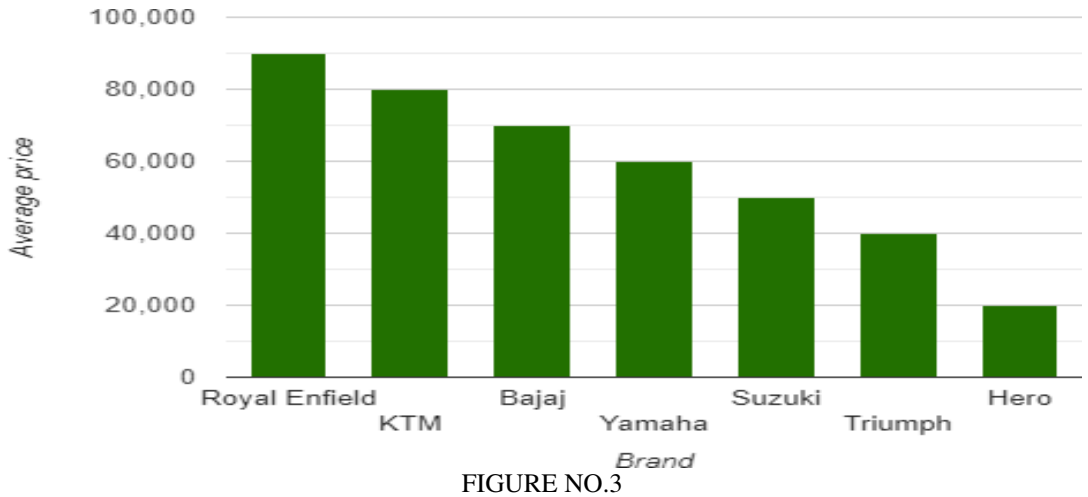

FIGURE NO.3

Figure 4 depicts the vehicle's age distribution. The year Of Registration characteristic and the current year are used to calculate the age (2018). It appears to be the number of years between the date of registration and 2018. The graph shows that the majority of the vehicles are reported as being between 3-8 years old. The graph shows that people don't sell their cars either too soon, which defeats the point of buying, or too late when the bike's value starts to drop drastically.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

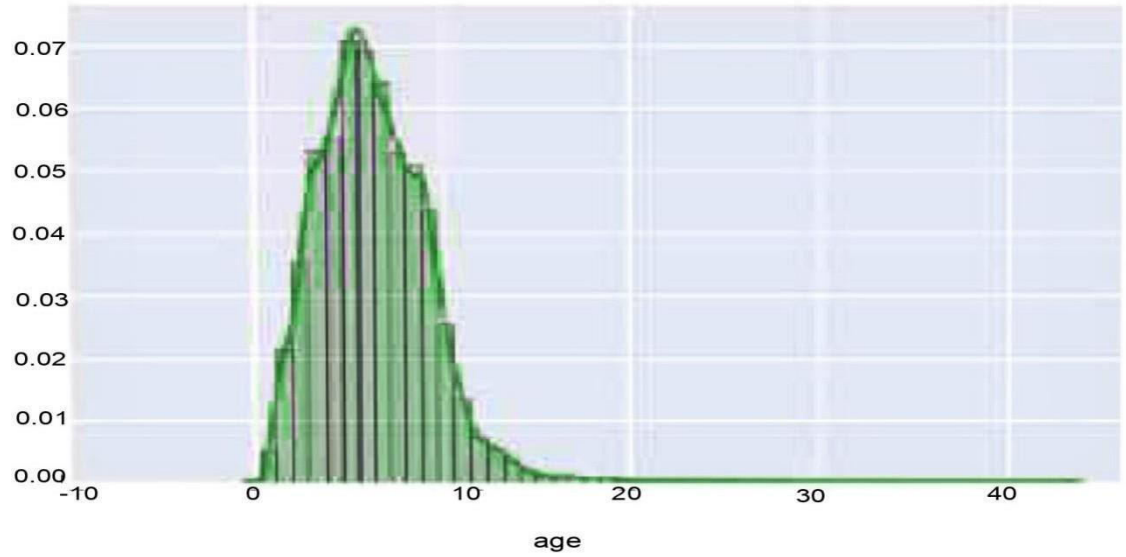

FIGURE NO.4

Figure 5 demonstrates that the average price of these bikes with damage that is repaired is greater than the average price of other cars. This may seem self-evident, yet it provides information about a vehicle's cost. as the car's value drops down drastically

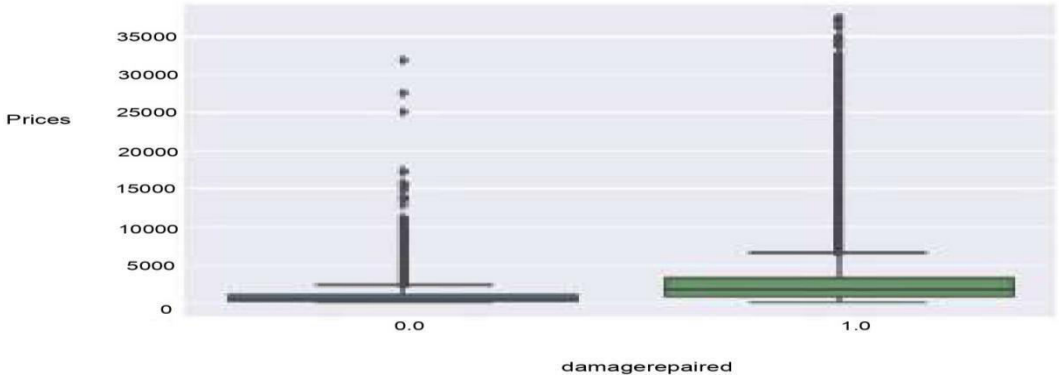

FIGURE NO.5

V. CONCLUSION

The accuracy of used-Bike price prediction is 84.01 Percent for test data and 92 percent for train data in this paper's study, which uses the Kaggle dataset. By filtering out outliers and unnecessary features from the dataset, the most relevant features used for this prediction are price, kilometre, brand, and vehicleType.In comparison to previous work using these datasets, Random Forest, as a sophisticated model, provides good accuracy.

REFERENCES

1. WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.
2. DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, Vol. 28, Issue 4, pp. 637-644.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011044|

3. Praful Rane"Used Car price prediction"International Research Journal of Engineering and Technology,Volume:08 Issue:04,April 2021
4. Enis Gegic"Used Car Price Prediction Using Machine Learning Techniques"TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16, February 2019.
5. Bin Yu"Analysis of a Random Forests Model"Journal Of Machine Learning Research 13(2012)1063-1095.
6. Sai Sumanth Palakurthy "How much is my car worth? A methodology for predicting used cars prices using Random Forest"Future of Information and Communications Conference (FICC) 2018
7. Pudaruth, S. 2014. "Predicting the Price of Used bikes Using Machine Learning Techniques ", International Information and Computer Technology Journal, 4 (7), page.753-764.
8. Kuiper, S. 2008. "Introduction to Multiplication: How Much Is Your bike? ", Journal of Statistics Education, 16 (3).
9. Nau, R. 2014. Refractory analysis notes,Textbooks, Duke University, Furqa School of Business, 26 Nov 2014
10. GONGGI, S., 2011. New model for residual value prediction of used cars based on BP neural network and n linear curve fit. In: Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details