International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

# Used Car Price Prediction by using Random Forest.

**Akshada Pawar[1], Kishori Kamble[2], Komal Survanshi[3], Namrata Mirje[4], Pradnya Chougule[5]**

[1] *Dept of Information Technology, Dr. J. J. Magdum College of Engineering, Jaysingpur, Maharashtra, India*
[2] *Dept of Information Technology, Dr. J. J. Magdum College of Engineering, Jaysingpur, Maharashtra, India*
[3] *Dept of Information Technology, Dr. J. J. Magdum College of Engineering, Jaysingpur, Maharashtra, India*
[4] *Dept of Information Technology, Dr. J. J. Magdum College of Engineering, Jaysingpur, Maharashtra, India*
[5] *Dept of Information Technology, Dr. J. J. Magdum College of Engineering, Jaysingpur, Maharashtra, India*

---

**Abstract -** The price of a new car in the industry is fixed by the manufacturer industry with some additional costs. So, buying a new car can be assured. But due to the increasing prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is a need for a Used Car Price Prediction system. Existing System includes a process where a seller decides a price and buyer has no idea about the car and it's value in the present day scenario. To overcome this problem we have developed a model which will be highly effective.[9] Using Random forest algorithm it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which require input from any user and displays the Price of a car according to user's inputs.

***Key Words***: Machine Learning, Random Forest, Prediction, Supervised learning, Kaggle, Heroku.

## 1.INTRODUCTION

Many platforms such as cars24.com and OLX.com provides users a platform where they can sell used cars, but what will be the price of the car this is the biggest question. Machine learning algorithms can be solution to this problem. Using history of previously used cars selling data can predict a price of the car. To find the price of used vehicles a model has been developed using of machine learning algorithm that is Random Forest. This technique were used not on the individual items but for the whole group of data items. This data group has been taken from some web portal and that same has been used for prediction.

Determine the listed price of a used car is a challenging task, due to the many factors. The main goal of this project is developing machine learning models that can accurately predict the price of a used car.

Deciding whether a used car is worth the posted price when you see listing online can be difficult. Several factors including model year, showroom price fuel type etc.. From the perspective of a seller. Based on existing data, the aim is to use machine learning algorithm to develop models.

The production of cars has been increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a popular in industry.

## 2. LITERATURE SURVEY

For predicting the price of used cars we used to do survey on different research papers.

The first paper is Predicting the Price of Used Cars using Machine Learning Techniques .

In this paper ,Author Sameer Chand collected historical data from newspapers and applied supervised machine learning techniques to predict the price of used cars.

They used four different Machine Learning techniques like multiple linear regression analysis, k-nearest neighbors, naive bayes and decision tree algorithm [1].

The Second paper is Car Price Prediction using Machine Learning Techniques. In this paper, they applied three machine learning techniques Artificial Neural Network, Support Vector machine and Random Forest. Performance of different algorithms are compared to find best result [2].

The Third paper is Used Cars Price Prediction using Supervised Learning Techniques. In this paper, they applied Multiple Regression and Regression tree to develop statistical model to predict the price of used cars. Error Data matrix is formed on test and train data to find result [3].

The Fourth paper is Vehicle Price Prediction using SVM Techniques, In this paper, they considered number of attributes to predict the result more reliable and accurate. The model is developed with Artificial Neural Network, Support Vector Machine and Random forest. To normalize, standardize and cleaning of data, PHP Scripts are built [4].

## 3. COMPARATIVE ANALYSIS

Random forests reduce the variance in decision. A single decision tree is a weak in prediction , but is fast to build. More trees give you more robust model. If you have more trees the slower performance will. We use two algorithms for comparative analysis that are Random Forest and Gradient Boost algorithms.

Random forest algorithm is built using a bagging method. Gradient Boost uses boosting technique. Decision trees are connected sequentially to obtain a strong learner.

The random forest model highly depends on decision trees. Boots rapping plays a key role in decision trees. GBDT does not use or need bootstrapping [11].

## 4. PROPOSED SOLUTION:-

The price of a new cars in the industry is fixed by the manufacturer with some additional costs incurred by the government in the form of taxes, So, customers buying a new car can be assured.

The proposed system which have machine learning code as background, in which we use Random forest algorithm which predicts the price of the car depending upon the actual price of car. The purpose for using random forest because the complexity of this algorithm is less. We have dataset which have price as columns name and current price car.

The user has to select the kilometers driven, the showroom price and how many years does the car has been driven. This creates a pattern, then input data passed to algorithm and algorithm predicts corresponding price.

4.1Random forest algorithm:-

Random Forest is a popular machine learning algorithm. In Classification and Regression problems Random forest can be used. for combining multiple classifiers to solve a complex problem and to improve the performance Random forest is used .

Random Forest is improve the predictive accuracy of that dataset. Instead of one decision tree, the random forest takes the prediction from each tree and it predicts the final output.

4.2Why use Random Forest?

- It takes less training time as.
- It predicts output with accuracy, even if the large dataset is there.
- It can also maintain accuracy when a large data is missing [7].

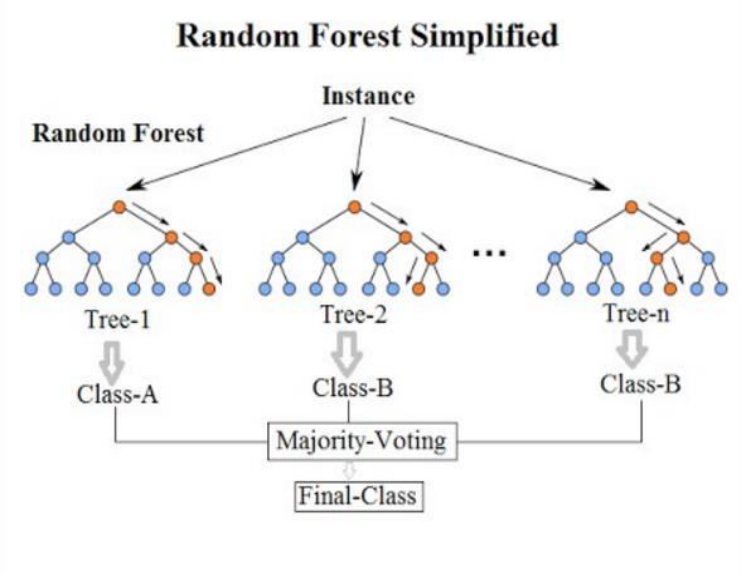



Fig 1-Overview of Random Forest [10]

4.3 Advantages:-

Lots of benefits of Random Forest Algorithm, one main is that it reduces the risk of overfitting and required training time. Additionally, it offers a high accuracy. Random Forest algorithm runs efficiently on large datasets.

## 5. METHODOLOGY

The data in the project is downloaded from Kaggle. Kaggle is the website where all types of datasets are available in .csv, .xlsx, .jpg, .txt, .png, .json, .xml, etc. format. In this dataset all the information of car is available

5.1Data wrangling

- In data wrangling some data explorations and visualizations were applied on data set before the process of data cleaning. It gives idea about how to deal with extreme values and missing values. After that data explorations was applied again to understand cleaned version of data.

5.2Data cleaning :

-In the process of cleaning firstly all unnecessary features are removed. In this 'city','color', 'image_url' columns are dropped..

- In next step null values are found and removed.

-In next step extreme values were dropped because they prevent prediction power of model.

(Extreme value=These characteristic values are the smallest of largest (max value) & are known as extreme values. ex- height of pepole - Too small and too large are extreme values of height)cars which has high price and very low price values were dropped because they create noise in data.

- In addition cars with extreme kilometer driven values were dropped and lastly cars from earlier than 1990 were dropped

In the next, missing values were found using is null().

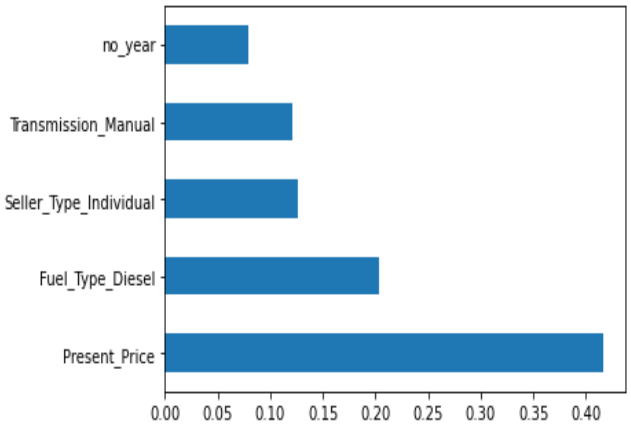
we use fillna() with method 'ffill' which handles missing values. Missing values are replaced with special value or an

aggregate value such as mean, median.

After this all the data is cleaned for analyzing.

5.3Variable importance-

Variable importance is calculate as the decrease in node. The node can be calculated by the number of samples and divided by the total number of samples. The higher the value the more important as the feature.[7]




At each split in each tree, the improvement in the split-criterion is important , and is accumulated over all the trees in the forest separately for each variable. The left plot of Figure shows the variable importance's computed.[8]

5.4Data splitting

a. Training set (70-80% of data): Model learns on this.

b. Validation set ( 10-15% of data): Model hyperparameters are tuned on this.

c. Test set ( 10-15% of data): Models' final performance is evaluated on this. If you done it right, the results on the test set give a good indication of how the model should perform in the real world.

-The error of prediction model is calculated by using accuracy metrics in machine learning. The MSE, MAE, RMSE are popular metrics for calulating Prediction error rates in regression analysis.

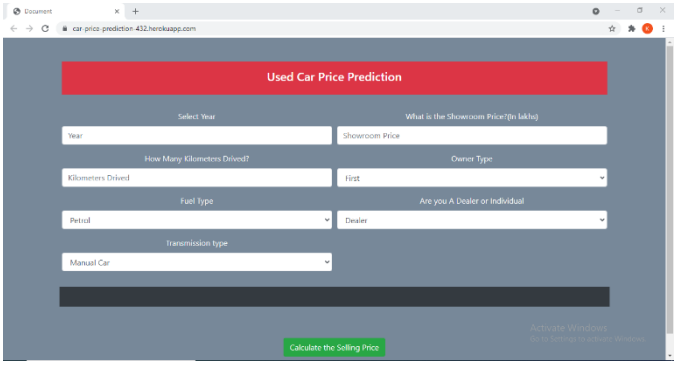
MAE(mean absolute error) It measures the difference between original and predicted values extracted by average absolute difference of data set.
-MSE - (mean squared Error) It measure the difference between original and predicted values extracted by squase of average difference of data.
RMSE (Root mean Squared Error) error rate is square root of MSE.

| MAE | MSE | RMSE |
|---|---|---|
| 0.8976 | 4.1004 | 2.0249 |

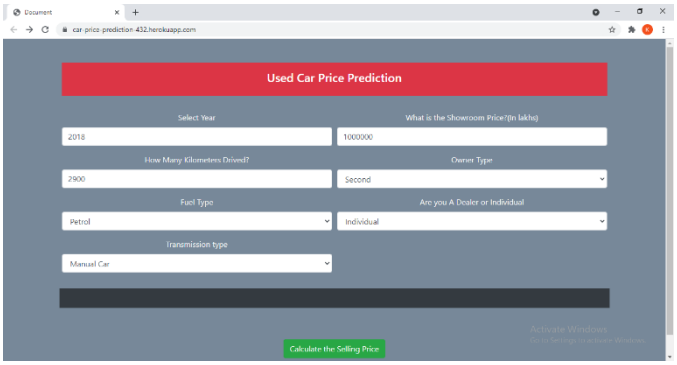
We use sklearn python library to calculate the error rate of our model. The above table shows MAE, MSE, RMSE of our model.

## 6. RESULT

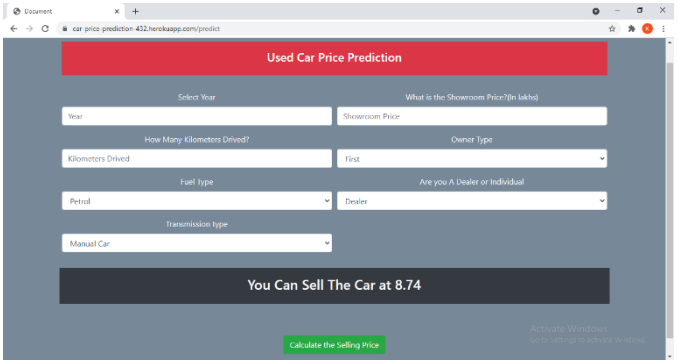
1.Put your information




2.After putting information




3.Final output




## 7. CONCLUSION AND FUTURE WORK

This model is based on the machine learning algorithm. We were trying to predict the selling price of the used cars based on the dataset provided at Kaggle. To predict this dataset we used machine learning algorithms i.e. Random forest. In future we can add login page or we can give authentication to customer, dealer and seller, because of this website will be more secure. The random forest method for predicting car price is efficient.

## ACKNOWLEDGEMENT

We would like to say thanks to our Prof. S. R. Shinge, IT, JJMCOE, for their valuable guidance and suggestions in our project work. We are extremely grateful to him for encourage us to work in right way. He is our biggest motivation in entire duration of project. Also we are thankful to Prof. R. A. Sanadi, Head of Department, IT and Our Departament of Information Technology, for providing required facilities.

Also we express our sincere gratitude to our college, Dr. J. J. magdum College of Engineering, Jasingpur, for giving us various opportunities.

## REFERENCES

[1] Predicting the Price of Used Cars using Machine Learning Techniques. Volume 4, Number 7(2014), pp. 753-764, University of Mauritius.
[2] Car Price Prediction using Machine Learning Techniques, TEM Journal. Volume 8, Issue 1,February 2019. International Burch University.
[3] Used Cars Price Prediction using Supervised Learning Techniques, ASSN: 2249-8958, Volume-9 Issue-1S3, December 2019
[4] Vehicle Price Prediction using SVM Techniques, Volume-9 Issue-8, June 2020.
[5] https://www.kaggle.com/avikasliwal/used-cars-price-prediction.
[6]International Research Journal of Engineering and Technology (IRJET Apr 2021).
[7] https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm
[8]https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.math.mcgill.ca/yyang/resources/doc/randomforest.pdf&ved=2ahUKEwiI1-qVpIXyAhVTyDgGHWidCwMQFjAYegQIEhAC&usg=AOvVaw2xgsSQiKdBpP28HXWE8CzQ&cshid=1627460388114
[9] https://www.irjet.net/archives/V8/i4/IRJET-V8I4278.pdf
[10] https://en.wikipedia.org/wiki/Random_forest
[11] https://www.analyticsvidhya.com/blog/2021/04/steps-to-complete-a-machine-learning-project/
[12] https://www.datatechnotes.com/2019/10/accuracy-check-in-python-mae-mse-rmse-r.html