iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46354

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

Car Price Prediction Using Machine Learning Algorithms

B V Raghurami Reddy[1], Dr. K. Santhi Sree[2]
[1]*Data Science, School of Information Technology, JNTUH*

Abstract: ***Machine learning(ML) is an area of a AI that has been a key component of digitization solutions that have attracted much recognition in the digital arena. ML is used everywhere from automating and do heavy tasks to offering intelligent insights in every industry to benefit from it. The current world already using the devices that are suitable these problems. For example, a wearable fitness tracker like Smart Band or a smart home assistant like Alexa, Google Home. However, there are many more examples of machine learning in use.***
In this project the task is to find out price of a used car. The cars dataset taken from Kaggle, where dataset contains used car details (variables), Our task is to finds out which variables are significant in predicting the price of a used car and how well these variables are important in predicting the price of a car. For this task we were using machine learning algorithms are linear regression, ridge regression, lasso regression, K-Nearest Neighbors (KNN) regressor, random forest regressor, bagging regressor, Adaboost regressor, and XGBoost.
The goal of this project is to build models on above mentioned machine learning algorithms on car dataset. We implement from basic linear regression algorithm to some very good algorithms like Random Forest Regressor and XGBoost Regressor. This project intends to point out the Random Forest and XGBoost Regressor models perform very well in regression problems.
Keywords: XGBoost Regression, Random Forest Regression (RFR), Linear Regression (LR).

I. INTRODUCTION

ML is part of AI that involves data and algorithms to design models, analyze and take decisions by themselves without the need for human activity. It tells how computers work on their own with the help of previous experiences.
The main dissimilarity between regular system software and ML is that a human designer doesn't give codes that instruct the computer how to act in situations, instead, it has to train by a huge amount of data.
ML approaches are divided into Reinforcement Learning, Unsupervised Learning, Supervised Learning and depending on the problem nature. Supervised Learning, there are two types. They are Regression and Classification.

II. PROBLEM STATEMENT

The used car market is a huge and important market for car manufacturers. The second-hand car market is also very likely linked to new car sales. Selling used cars at new car retail and handling lease returns and fleet returns from car rental companies require car manufacturers to be involved in the used car market.
Automakers face several problems in the used market. The deep mess in the world, the general problem of more people, increased competition from other manufacturers and the trend toward electronic cars are just some of the factors that make it difficult to sell used vehicles on the used car market, reducing sales margins. Automakers, therefore, require good decision support systems to maintain the profit of the used car business. A core component of such a system is a predictive model that estimates the selling price based on vehicle attributes and other factors. Although previous studies have explored statistical modelling of resale costs, few studies have attempted to predict resale costs with maximum accuracy to support decision making. As a result, the answers to the following questions are unclear: i) how predictable are resale prices, ii) the relative accuracy of various forecasting methods, and whether some methods are particularly effective. iii) Given those market research agencies specialize in estimating residual values, does it makes sense for automakers to invest in their resale cost prediction models? The purpose of this work is to provide more accurate answers to those questions. The present project comes under the Regression category. This project is all about predicting the used car's prices. In our day to life, everyone wants a car, but budget is the problem, so in this project build a model that will take certain parameters as arguments and result or predict the price of the car based on given parameters. This project's goals are to build a machine learning model which takes car features as input and predicts the cost of the reused car. Compare the most used machine learning regression models which give less error and predict the more accurate value of the price of the car.

III.PROPOSED WORK METHODOLOGY

There are two phases in the build a model:
Training: The model is trained by using the data in the dataset and fits a model based on the model algorithm chosen accordingly.
Testing: The model is provided with the inputs and is tested for its accuracy. Afterwards, the data that is used to train the model or test it, has to be appropriate. The model is built to detect and predict the cost of a used car and good models must be selected.

A. Architecture

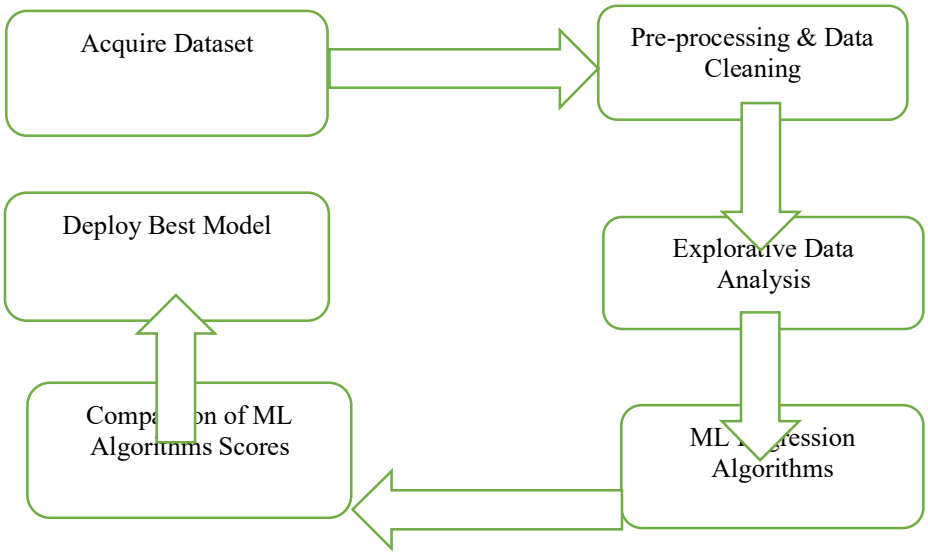


Fig: 1 Architecture of the Proposed System

B. Sample Dataset

The dataset is taken from Kaggle. Take look into sample dataset below.

19	7213839225	bellingham	2004.4	infiniti	g series	fair	5 cylinders	gas	20984	clean	automatic	4wd	full-size	offroad	grey	33.71069	-98.3337	26850
20	7208549803	bellingham	2004.4	infiniti	g series	fair	5 cylinders	gas	20984	clean	automatic	4wd	full-size	offroad	grey	33.71069	-98.3337	11999
21	7213843538	skagit / island / SJI	2004.4	infiniti	g series	fair	5 cylinders	gas	20984	clean	automatic	4wd	full-size	offroad	grey	33.71069	-98.3337	24999
22	7212631321	skagit / island / SJI	2004.4	infiniti	g series	fair	5 cylinders	gas	20984	clean	automatic	4wd	full-size	offroad	grey	33.71069	-98.3337	21850
23	7316814884	auburn	2014	gmc	sierra 150(	good	8 cylinders	gas	57923	clean	other	4wd	full-size	pickup	white	32.59	-85.48	33590
24	7316814758	auburn	2010	chevrolet	silverado 1	good	8 cylinders	gas	71229	clean	other	4wd	full-size	pickup	blue	32.59	-85.48	22590
25	7316814989	auburn	2020	chevrolet	silverado 1	good	8 cylinders	gas	19160	clean	other	4wd	full-size	pickup	red	32.59	-85.48	39590
26	7316743432	auburn	2017	toyota	tundra do(	good	8 cylinders	gas	41124	clean	other	4wd	full-size	pickup	red	32.59	-85.48	30990
27	7316356412	auburn	2013	ford	f-150 xlt	excellent	6 cylinders	gas	128000	clean	automatic	rwd	full-size	truck	black	32.592	-85.5189	15000
28	7316343444	auburn	2012	gmc	sierra 250(	good	8 cylinders	gas	68696	clean	other	4wd	full-size	pickup	black	32.59	-85.48	27990
29	7316304717	auburn	2016	chevrolet	silverado 1	good	6 cylinders	gas	29499	clean	other	4wd	full-size	pickup	silver	32.59	-85.48	34590
30	7316285779	auburn	2019	toyota	tacoma	excellent	6 cylinders	gas	43000	clean	automatic	4wd	full-size	truck	grey	32.6013	-85.444	35000
31	7316257769	auburn	2016	chevrolet	colorado e	good	6 cylinders	gas	17302	clean	other	4wd	full-size	pickup	red	32.59	-85.48	29990
32	7316133914	auburn	2011	chevrolet	corvette g	good	8 cylinders	gas	30237	clean	other	rwd	full-size	other	red	32.59	-85.48	38590

Fig: Sample dataset

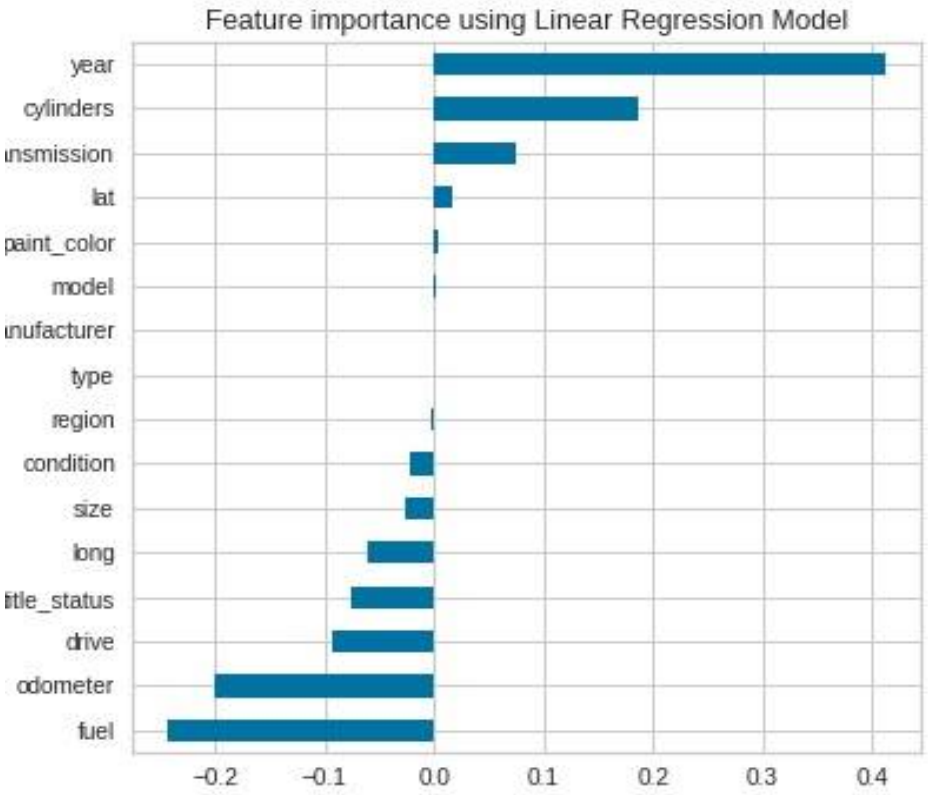
Sample dataset have variables like id, name, year, model, condition, cylinders, fuel type, Odometer, seats, car type, colour, selling price.

IV.IMPLEMENTATION

A. Linear Regression

LR is used to predict the value of a variable based on the value of another feature. The feature you want to predict is called the dependent variable. The label that is used to predict the value of another feature is called the independent variable. The LR equation is of the form A = m + nB, where B is the independent variable, A is the dependent variable, a is the intercept y, and n is the slope of the line.

Linear regression's important features are:



Graph: 1 Linear regression important features

Results of liner regression are:

```
MSLE : 0.0024161658796354 25
R2 Score : 0.625564551513134 3 or 62.5565%
```

B. Ridge Regression

The Ridge regression model is used for analyse any data which faces the multicollinearity problem. This model performs a regularization, particularly L2 regularization. When the problem of multicollinearity comes, the least squares are unbiased and the variances are large, resulting in the predicted outcomes being different from the original outcomes. The cost function for ridge regression:

$$\text{Min } (||Y - X(\text{theta})||^2 + \lambda||\text{theta}||^2)$$

Here the penalty term is Lambda. Lambda is denoted by the symbol λ. So, we control the penalty by changing the alpha values. The more the alpha values, the greater the error and thus the magnitude of the coefficients decreases. It reduces the parameters. Therefore, ridge regression is used to stop multicollinearity and decreases the model complexity by reducing the coefficients.

Results of Ridge Regression are:

```
MSLE : 0.0024161612230360 1
R2 Score : 0.625565295248563 6 or 62.5565%
```

C. Lasso Regression

"LASSO" means least absolute shrinkage and selection, operator. It is a type of linear regression that uses decline. The decline means the data values will decrease towards a central point, such as the mean. It supports simple, sparse models. This kind of regression is useful for algorithms with a more degree of multicollinearity.

Results of Lasso Regression are:

```
MSLE : 0.0024161198691321217
R2 Score : 0.6255752923578952 or 62.5575%
```

D. KNN Regressor

KNN regression is a nonparametric technique that approximates the relationship between an independent variable and a continuous outcome by averaging observations in the same neighbourhood. The size of k should be specified by the analyst. Alternatively, we can choose by cross-validation to choose the size that will minimize the mean squared error.

Results of KNN Regressor are:

```
MSLE : 0.0012246626436424842
R2 Score : 0.8176006701064815 or 81.7601%
```

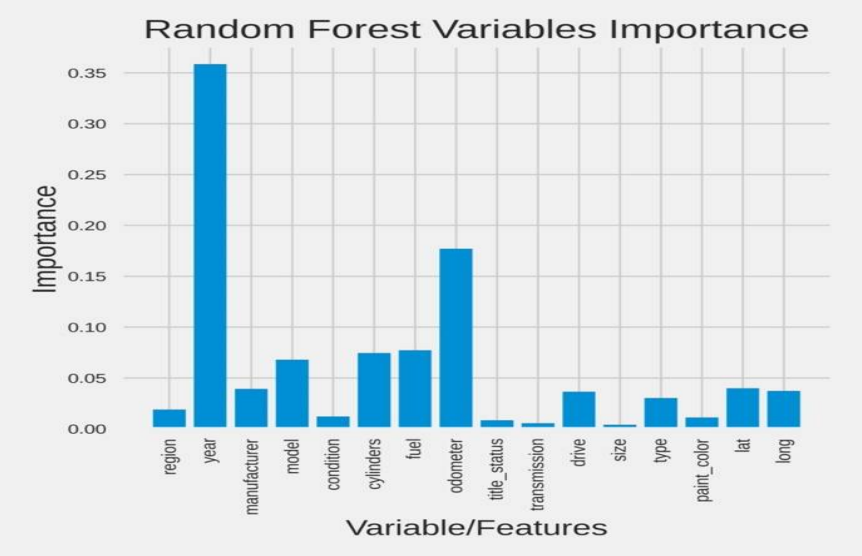
E. Random Forest Regressor

Random Forest Regression comes under a Supervised Machine Learning algorithm which uses an ensemble model for regression. An ensemble learning is a technique that clubs the outcomes of different machine learning models for creating more good predictions than one model.

Results of Random Forest Regressor are:

```
MSLE : 0.0006112110450301187
R2 Score : 0.9118122961528443 or 91.1812%
```

Random forest important variables are:



Graph: 2 Random forest important variables

F. Bagging Regressor

Bagging is short for Bootstrap Aggregating. It uses bootstrap resampling to train multiple models on random variations of the training set. At prediction time, each item's predictions are aggregated to give the final predictions. Bagged decision trees are efficient because each decision tree is suitable for a slightly different training data set, this allows each tree to have subtle differences and make slightly different skill predictions.

Results of Bagging Regressor are:

```
MSLE : 0.0011778460680167315
R2 Score : 0.8267549907211298 or 82.6755%
```

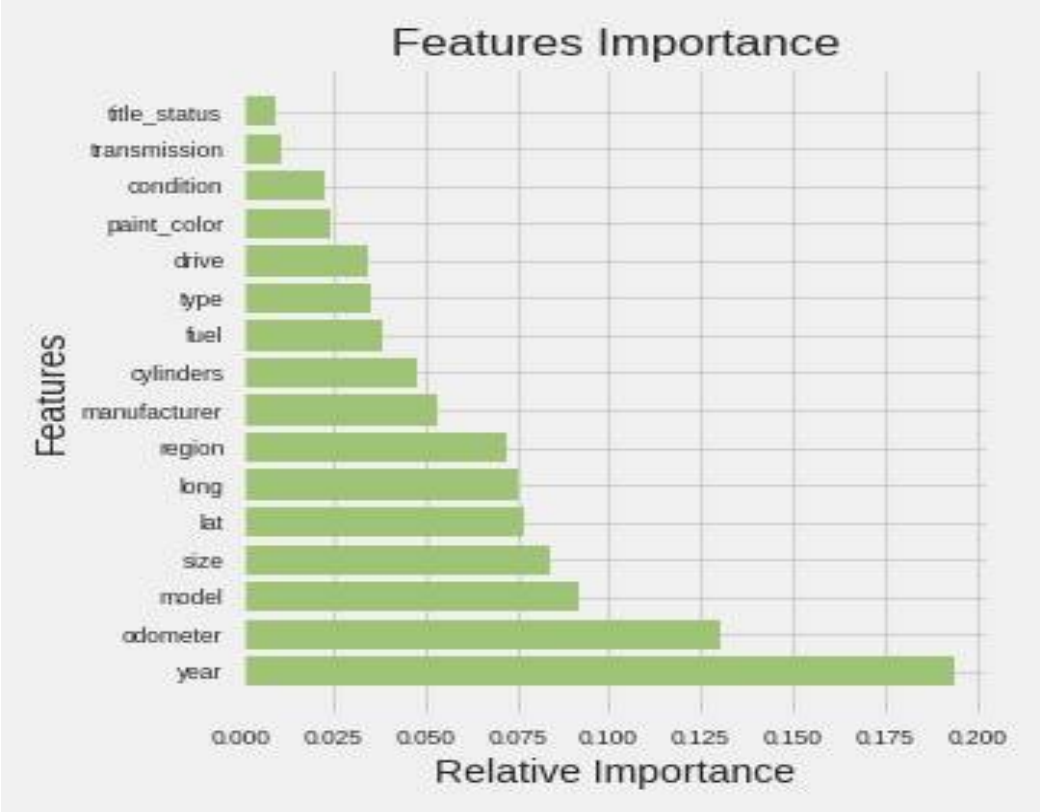
G. Adaboost Regressor

AdaBoost model is a very short single-level decision tree. At first, it models a weak learner and gradually adds it to an ensemble. Each next model will try to modify the predictions or outcomes of the previous models in a series. It is achieved by weighting the train data to focus more on train examples in which old models made prediction errors.

Results of Adaboost regressor are:

```
MSLE : 0.0006373329841582197
R2 Score : 0.9066837500303873 or 90.6684%
```

Adaboost important features are:



Graph: 3 Adaboost important features

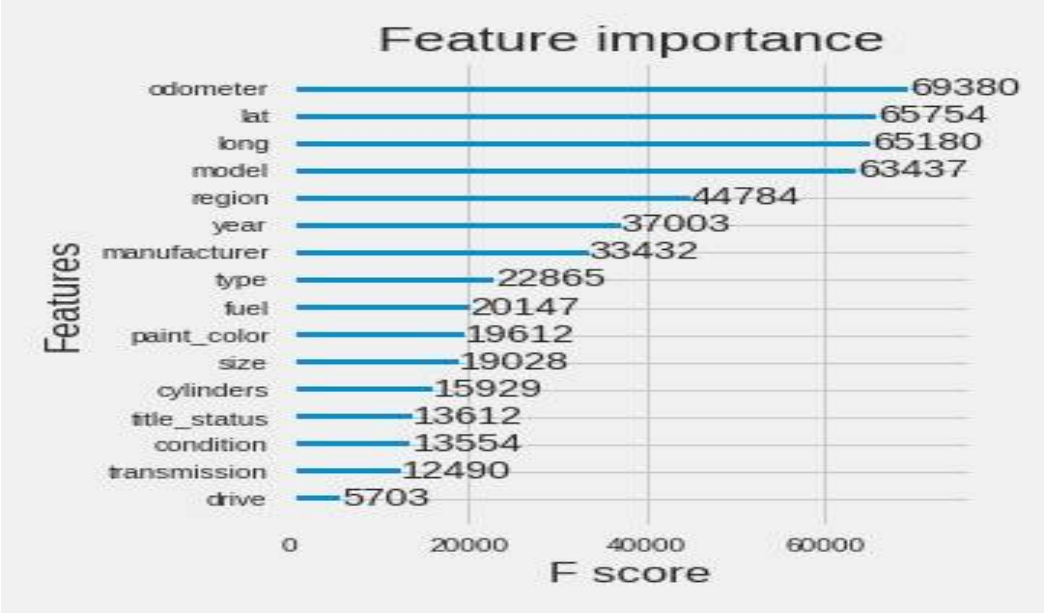
H. XGBoost

XGBoost is a short name for Extreme Gradient Boosting, designed by researchers at Washington University. This library was written in C++. It optimizes gradient boosting training. XGBoost is a family of gradient-boosted decision trees. With this model, the decision tree is built in sequential form. In XGBoost weights play a crucial role. Weights are allocated to all independent variables after that they are fed into a decision tree that predicts outcomes. The weight of the variables that the model predicted wrongly is increased and these variables are then input to the second decision tree model. These individual predictors/trees are then grouped to provide a stronger and more accurate model.

Results of XGBoost Regressor are:

```
MSLE : 0.0005130800524658774
R2 Score : 0.9255950215596382 or 92.5595%
```

XGBoost important features are:



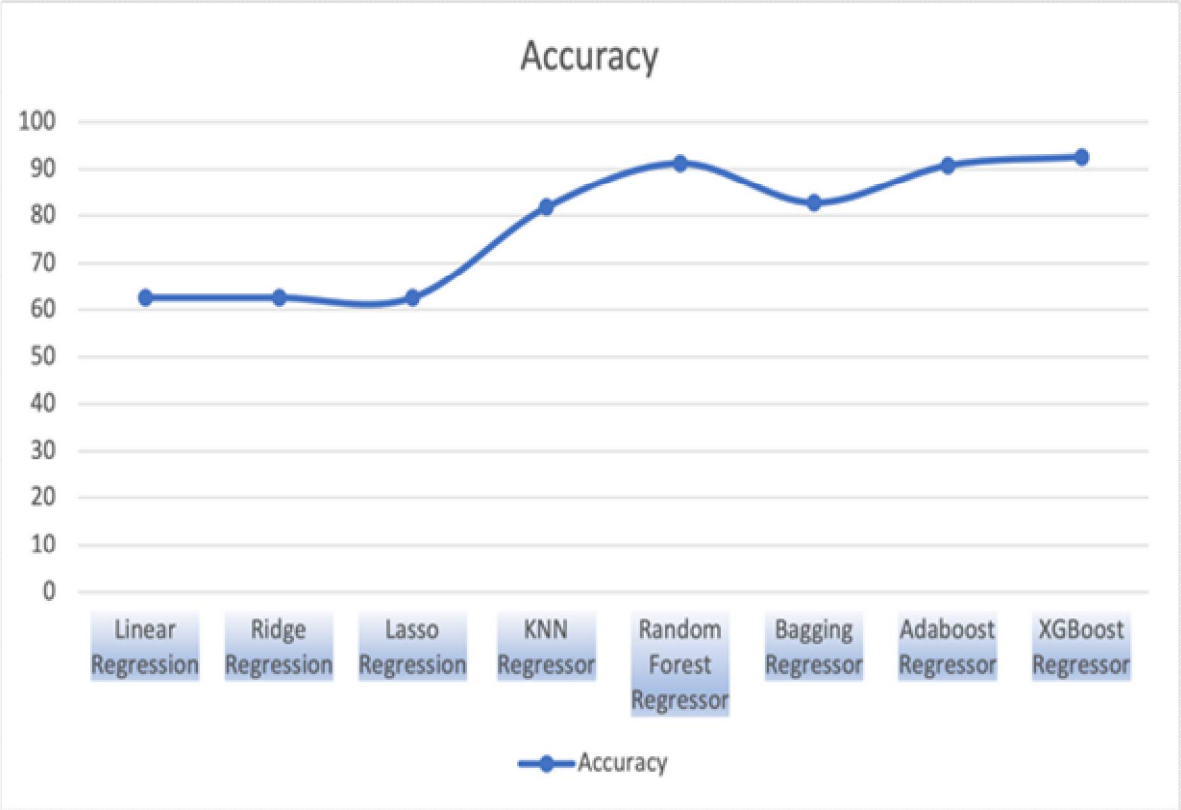
Graph: 4 XGBoost important features

V. RESULTS

By the results above, here XGBoost Regressor gives more accuracy and next is Random Forest Regressor. Therefore, here I can conclude that for regression problems XGBoost and Random Forest Algorithms give more accurate results when compared to Linear, KNN and other Regressions.

Name	MSLE	R2 Score
Linear Regression	0.00241616	0.625564
Ridge Regression	0.00241616	0.625565
Lasso Regression	0.00241611	0.625575
KNN Regressor	0.00122466	0.817600
Random Forest Regressor	0.00061121	0.911812
Bagging Regressor	0.00117784	0.826754
Adaboost Regressor	0.00063733	0.906683
XGBoost Regressor	0.00051308	0.925595

Table: Results



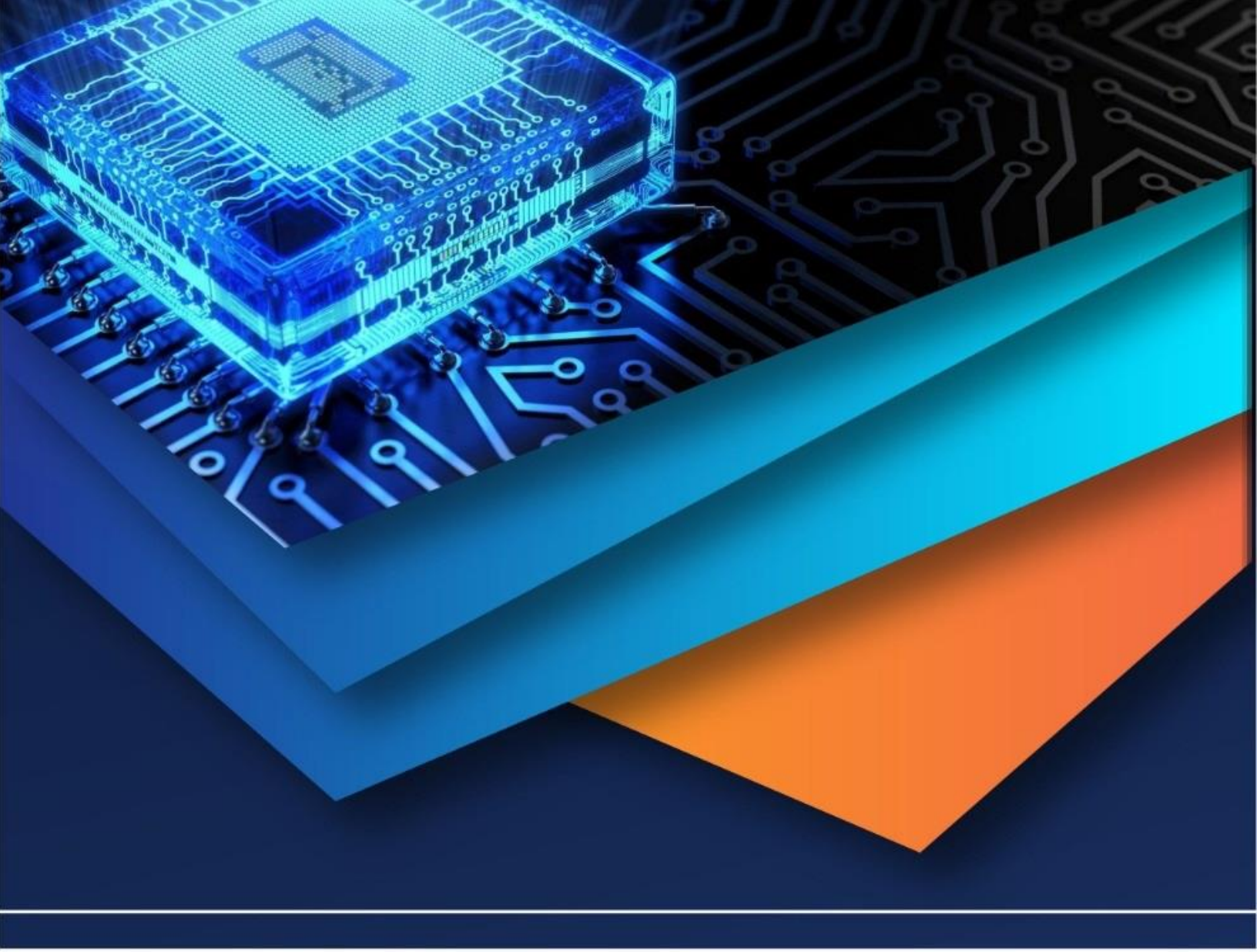
Graph: 5 Comparison of results

VI. CONCLUSION

In this project XGBoost Regressor produces more accuracy and next is Random Forest Regressor. This is because XGBoost take advantage of week learners and gradually learn but Random Forest build different trees without communicating with other learners.

REFERENCES

[1] Advances and Applications in Mathematical Sciences Volume 20, Issue 3, January 2021, Pages 367-375 © 2021 Mili Publications
[2] International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-5S, January 2020
[3] International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1S3, December 2019
[4] Journal of Multidisciplinary Developments. 6(1), 29-43, 2021 e-ISSN: 2564-6095 Predicting Used Car Prices with Heuristic Algorithms and Creating a New Dataset – Bilen
[5] International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764 © International Research Publications House http://www. irphouse.com
[6] Hands-On Machine Learning with Scikit-Learn, Keras, TensorFlow, 2nd Edition, by Aurelien Geron (O'Reilly).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY