iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.49180

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 11 Issue II Feb 2023- Available at www.ijraset.com

Performance Analysis of Regression Algorithms for Used Car Price Prediction: KNIME Analytics Platform

Chitra A R[1], Dr Arjun B C[2]
[1]*M.Tech Student, CSE,* [2]*HOD, ISE, Rajeev Institute Of Technology, Hassan*

Abstract: ***In the recent years people's willingness towards used car has increased. This has reflected in selling and buying of such cars. With the advance in technology online portal for marketing of used cars has come into effect. Many online portals focus to connect available used cars with user needs, present trends and various selection criteria. Using Machine Learning Algorithms such as Linear Regression, Tree Ensemble (Regression), Random forest (Regression), Gradient Boosted Tree(Regression), Simple Regression tree provided by KNIME Analytics Platform used car price predicted is performed. Analysis shows that Gradient Boosted Tree(Regression) prediction is closest to the target.***
Keywords: Machine Learning, KNIME, Regression, Data Analytics, R^2 ***metric***

I. INTRODUCTION

Though sell of new car in market has reduced in recent years second hand car market continues to grow. This is reflected in selling of cars across developing and developed nations such as India, Germany, united states, united kingdom, France and China. Second hand car market has shown scope in business for buyers and sellers. Since used cars are available in second hand market at affordable price people tend to get attracted. These cars do offer resell for profit. The price of used car depends on criteria such as make, type, colour to name a few. Price of used car varies in market, and hence evaluation model to predict the price of used car is necessary.

In the present work used car data set is analysed by five different machine learning regression algorithms provided by KNIME (An Intuitive, OpenSource Platform for Data Science). Machine learning algorithms are trained with data set and evaluated using R^2 metric for predicting the price of used cars.

II. LITERATURE SURVEY

K.Samruddhi et al. [1] used KNN algorithm and got 85% accuracy on prediction of used car price. Praful Rane, et. al. [2] has compared three machine learning algorithms linear regression, Lasso regression and ridge regression. Enis Gegic, et al. [3] focussed on data collection process and preprocessing. PHP scripts are used to normalize the data and avoid noise.

Pattabiraman Venkatasubbu, et al. [4] has conducted studies and concludes that multiple regression and Lasso regression are almost similar. For more accurate results advanced machine learning algorithms such as random forest is suitable.

Linear Regression, Multiple Regression models are used to estimate the cost of second hand cars[7].

Random Forest Regression along with python, flask and HTML are used to examine the price of a used car with the dataset from cardekho.com.[8]

III. PROPOSED METHODOLOGY

Main objective of this project is to decide which category of machine learning algorithm suits our dataset. Since the target is a continuous data we have to go for regression models. In our project, target feature is price which is continuous so we go for regression models. Actual selling price column will be compared with the predicted price from different algorithms. Main idea is to try different Machine Learning Algorithms such as Linear Regression, Tree Ensemble (Regression), Random forest (Regression), Gradient Boosted Tree(Regression), Simple Regression tree to predict the price of a used car, based on dataset downloaded from kaggle as well as perform performance analysis to decide which model gives highest R2 value.

Output of each algorithm is visualised using Scatterplot which is available in KNIME analytics platform. Steps followed in all the machine learning models include

1) New workflow is created for the used car price evaluation model as seen in Figure 1.
2) CSV reader node is fed with the dataset .csv file.
3) CSV reader node is connected to the partitioning node, which is used to separate the given data set into two partitions .This partition is done by choosing the randomly option.
4) Partitioning node one end is connected to the learner node and other end is connected to the predictor node.
5) Learner node output is connected to the predictor node.
6) Predictor node is connected to the numeric scorer node to measure the performance metric score.
7) Predictor node is also connected to the color manager node to give colors for differentiating between actual and predicted values.
8) Finally, color manager node is connected to the scatter plot node to get the data visualization.
9) Nodes are executed at each stage of connection

A. *Evaluating the Model Using Numeric Scorer*

KNIME supports several metrics like

1) R-Squared
2) Mean Absolute Error
3) Mean Squared Error
4) Root Mean Squared Error

In this project R^2 metric is employed using numeric scorer.
Its equation is given by the formula
$R^2 = 1-\Sigma(z_i - n_i)^2/\Sigma(n_i - 1/x*\Sigma n_i)^2$
n_i - numeric column's values and z_i - predicted value
The calculated values is viewed using numeric scorer

B. *Implementation of Different Regression Models*

1) *Linear Regression:* Linear regression which is supervised learning algorithm relates continuous target variable (selling price) with independent variables. Figure-1 discloses linear regression model workflow .Plot of observed vs predicted values is verified for its linearity as shown in the figure -2. Scatter plots is used to check the linearity. A linear regression model tries to fit a regression line to the data points that best represents the relations or correlations.

Nodes used for linear regression in KNIME

a) CSV Reader
b) Partitioning
c) Linear Regression Learner
d) Linear Regression Predictor
e) Numeric Scorer
f) Color Manager
g) Scatter Plot

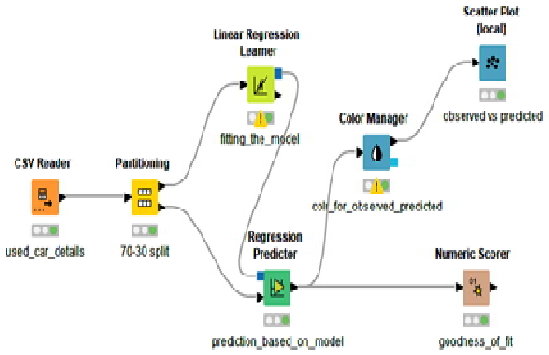


Figure 1. Linear Regression model workflow

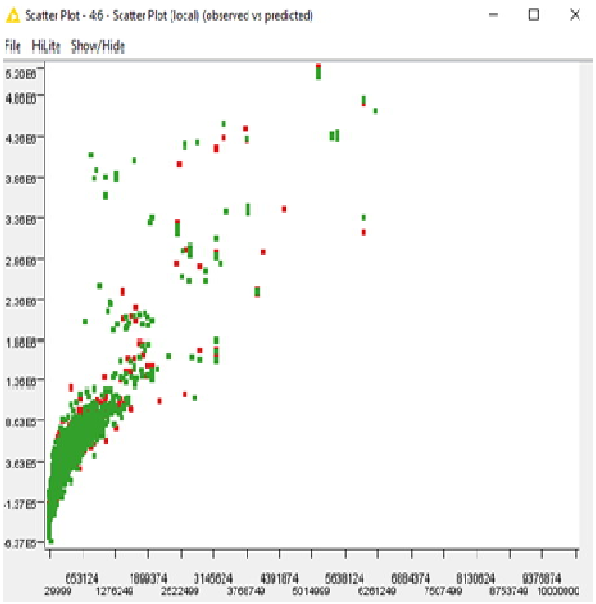


Figure 2. Scatter Plot -Selling Price versus Predicted Selling Price

2) *Tree Ensemble(Regression):* Tree Ensemble is an ensemble of random forest where the predicted value for a leaf node is the mean target value of the records within the leaf .Figure-3 discloses tree ensemble(Regression)model workflow. Plot of observed vs predicted values verified for its linearity as shown in the figure -4.

Nodes used for Tree Ensemble regression in KNIME

a) CSV Reader
b) Partitioning
c) Tree Ensemble Learner(Regression)
d) Tree Ensemble Predictor(Regression)
e) Numeric Scorer
f) Color Manager
g) Scatter Plot

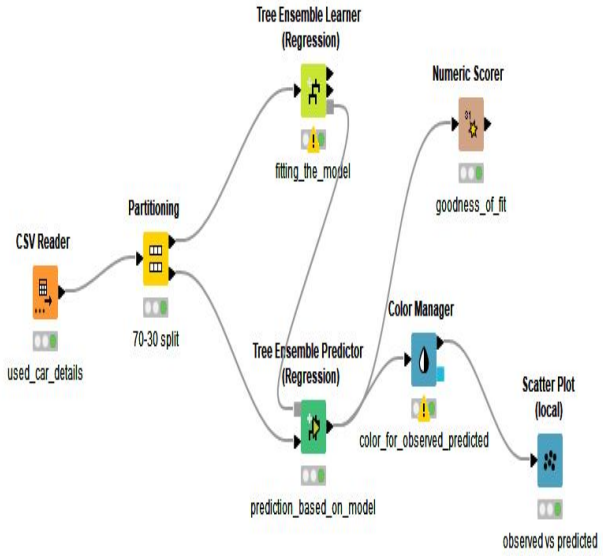


Figure 3. Tree Ensemble(Regression) model workflow

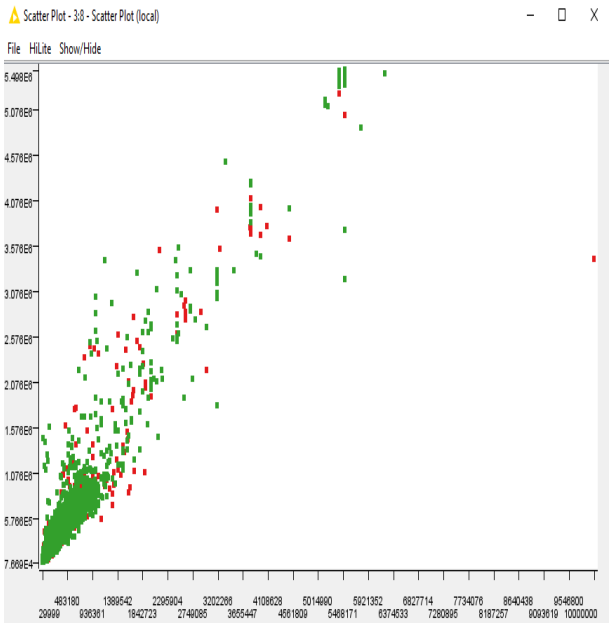


Figure 4. Scatter Plot -Selling Price versus Predicted Selling Price

3) *Random Forest (Regression):* Several decision trees are produced with set of words (row sets) and describing attributes(column). This is carried by bootstrapping that produce new sets of same size as original input. Attributes set necessary for each split of decision tree is selected randomly from learning column. Figure-5 discloses tree ensemble(Regression)model workflow. Plot of observed vs predicted values verified for its linearity as shown in the figure -6.

Nodes used are

a) CSV Reader
b) Partitioning
c) Random Forest Learner(Regression)
d) Random Forest Predictor(Regression)
e) Numeric Scorer
f) Color Manager
g) Scatter Plot

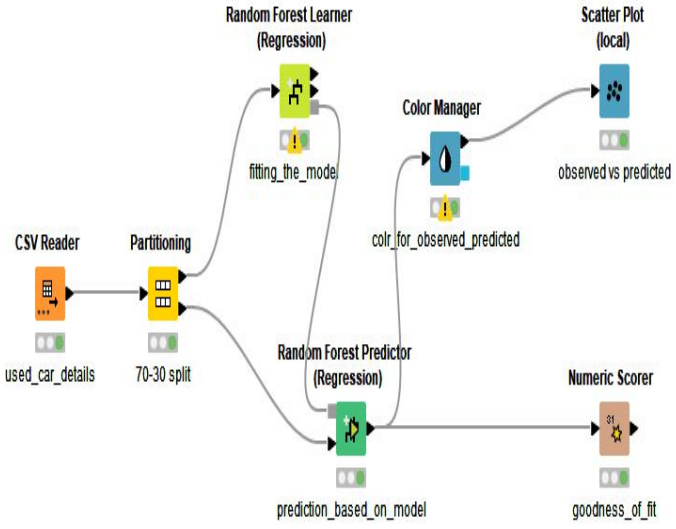


Figure 5. Random forest(Regression) model workflow

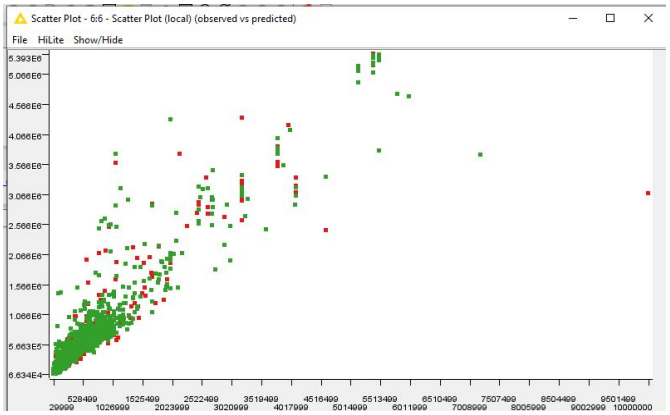


Figure 6. Scatter Plot -Selling Price versus Predicted Selling Price

4) *Gradient Boosted Trees (Regression):*An ensemble of tree is built for the given training set data by using shallow regression trees and special form of boosting. Figure-7 discloses tree ensemble(Regression)model workflow. Plot of observed vs predicted values verified for its linearity as shown in the figure -8

Nodes used are

a) CSV Reader
b) Partitioning
c) Gradient Boosted Trees Learner(Regression)
d) Gradient Boosted Trees Predictor(Regression)
e) Numeric Scorer
f) Color Manager
g) Scatter Plot

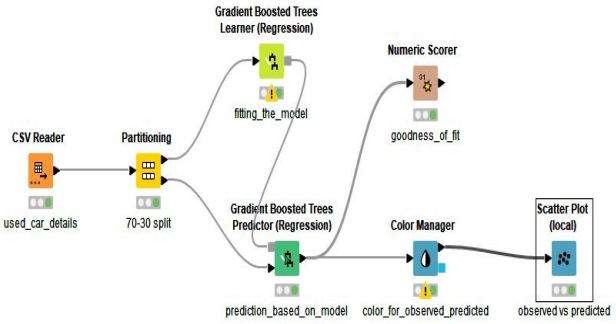


Figure 7. Gradient Boosted Tress (Regression) model workflow

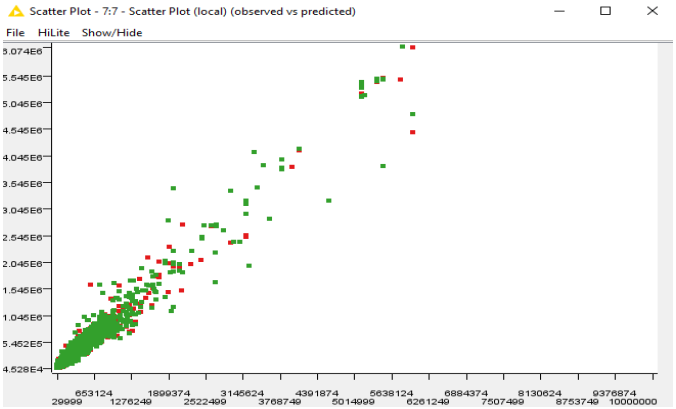


Figure 8. Scatter Plot -Selling Price versus Predicted Selling Price

5) *Simple Regression Tree:* A regression tree model is prepared for the given training set data. It works on the principle of minimizing the variance of target value within a leaf. Sum of squared errors is minimized by splits for respective children. The missing value is verified in each direction of split and the one that provides best result is accepted by the algorithm. Figure-9 discloses tree ensemble(Regression)model workflow. Plot of observed vs predicted values verified for its linearity as shown in the figure -10

Nodes used are

a) CSV Reader
b) Partitioning
c) Simple Regression Tree Learner(Regression)
d) Simple Regression Tree Predictor(Regression)
e) Numeric Scorer
f) Color Manager
g) Scatter Plot

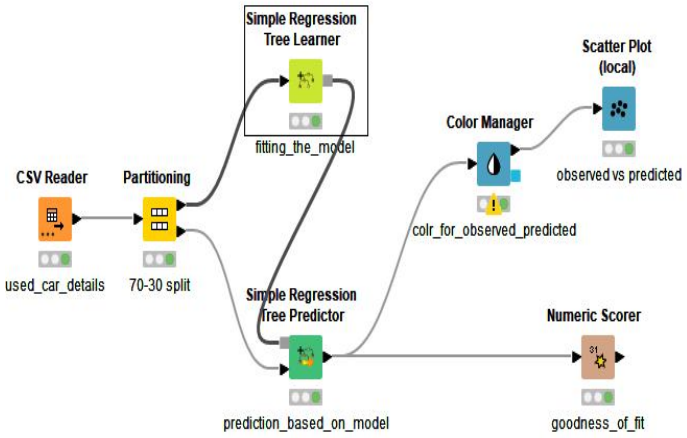


Figure 9. Simple Regression Tree (Regression) model workflow

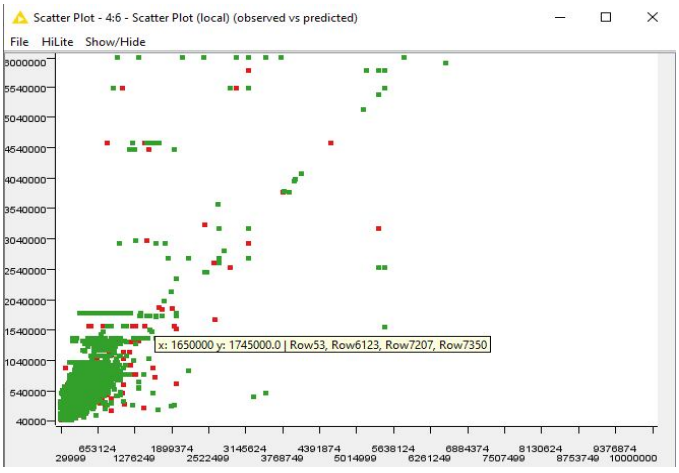


Figure 10. Scatter Plot -Selling Price versus Predicted Selling Price

IV. RESULT ANALYSIS

Used car dataset is taken from Kaggle website. Using this dataset, different regression algorithms are tested to see which model is suitable for better prediction especially the price of the car given the details like km_ driven,region,state province,city,fuel etc. Data used in the project is downloaded from the website [11].

Information (data set) of used cars available in the market has been collected by consulting firm from various market surveys as show in the table -1

Table 1-Dataset attributes

Column Name	Description
Sales_ID	Sales ID
name	Name of the used car
year	Year of the car purchase
selling_price	Current sellling price for used car
km_driven	Total km driven
Region	Region where it is used
State or Province	State or Province where it is used
City	City where it is used
fuel	Fuel type
seller_type	Who is selling the car
transmission	Transmission type of the car
owner	Owner type
mileage	Mileage of the car
engine	engine power
max_power	max power
seats	Number of seats
sold	used car sold or not

One of the most important metrics for evaluating a continuous target is R^2 which is between 0 and 1.
Closer to 1, the better the regression fit. Below mentioned table-2 gives the details of 5 algorithms R^2 value when the whole dataset is partitioned into train and test data set in the ratios 70:30 ,60:40 and 50:50 respectively (provided data samples are chosen in a random fashion).

Table 2:R^2 Values of Regression algorithms

Algorithms	R^2		
Partition(Train/Test datset)	70/30	60/40	50/50
Linear Regression	83.8	85.23	84.08
Tree Ensemble (Regression)	87.6	90.7	86.7
Random Forest (Regression)	87	87.23	85.96
Gradient Boosted Trees(Regression)	96.9	96.1	95.5
Simple Regression tree	60.8	62.6	55

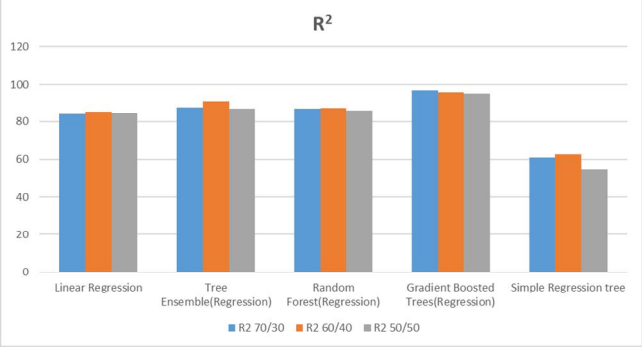


Figure 11. Comparative analysis

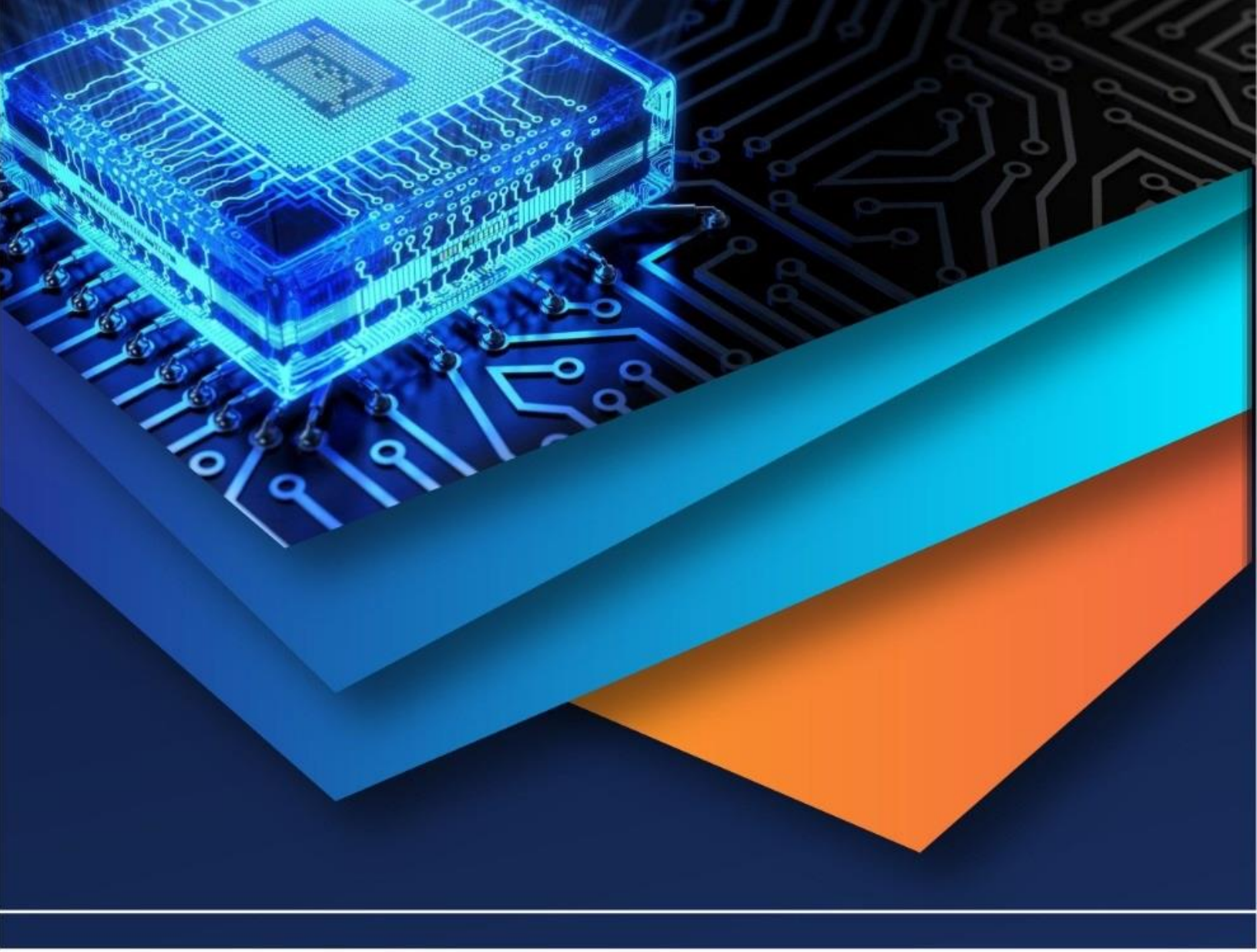
Numeric scorer node is employed to calculate the metric. Statistics of actual values and predicted values is computed by numeric scorer node.By using the data in table-2 comparative analysis is done as shown in the figure-11.

V. CONCLUSION

In this project, model is trained to predict the price of used car for given input features. Linear Regression, Tree Ensemble (Regression), Random forest (Regression), Gradient Boosted Tree(Regression), Simple regression tree algorithms are executed for the same dataset. The experimental analysis conducted using KNIME analytics tool shows that the Gradient Boosted Tree(Regression) model is having the highest R^2 value irrespective of partition(train and test set) percentage as well as linearity is also highest which can be observed from the scatter plot.

REFERENCES

[1] K.Samruddhi and Dr. R.Ashok Kumar Used Car Price Prediction using K-Nearest Neighbor Based Model (IJIRASE 2020)
[2] Praful Rane, Deep Pandya, Dhawal Kotak, USED CAR PRICE PREDICTION ((IRJET Journal 2021)
[3] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric," Car Price Prediction using Machine Learning Techniques"; (TEM Journal 2019)
[4] Enci Liu, Jie Li, Anni Zheng, Haoran Liu and Tao Jiang, Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network) (Article 2022)
[5] Pattabiraman Venkatasubbu, Mukkesh Ganesh Used Cars Price Prediction using Supervised Learning Techniques ((IJEAT 2019)
[6] Ashutosh Datt Sharma ,Vibhor Sharma,Sahil Mittal,Gautam Jain,,Sudha Narang.Predictive Analysis Of Used Car Prices Using Machine Learning
[7] B.Lavanya , Sk.Reshma , N.Nikitha, M.Namitha . Vehicle Resale Price Prediction Using Machine Learning. (UGC Care Group I Listed Journal)
[8] Abhishek Jha, Dr. Ramveer Singh, Manish, Imran Saifi, Shipra Srivastava,Used Car Price prediction(IJARIIT 2021)
[9] The Price Prediction for used Cars using Multiple Linear Regression Model(IJRASET)
[10] https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses
[11] https://www.kaggle.com/code/shubham1kumar/usercardata-pythonfile/data
[12] https://www.knime.com/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY