*Available online at http://www.advancedscientificjournal.com*
*http://www.krishmapublication.com*
*IJMASRI, Vol. 1, issue 10, pp. 246 - 251 Dec. -2021*
**https://doi.org/10.53633/ijmasri.2021.1.10.002**

**INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)**

**ISSN: 2582-9130** **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

**RESEARCH ARTICLE**

# USED CAR PRICE PREDICTION USING MACHINE LEARNING

**[1]Himanshu Dahiya, [1]Chetan Aggarwal, [1]Shubh Goyal and [1]Mini Agarwal**

*[1] Computer Science and Engineering Maharaja Agrasen Institute of Technology, Rohini, Delhi*

## Abstract

Cars are an important asset and their importance has increased exponentially in our life. With the increase in the demand and growing needs, the production of cars has also increased. But due to inflation in the prices of new cars, there are people who still can only afford a used car due to their financial conditions. This whole process has given rise to the used car market, which is outperforming many other industries and is rising every day. The rising market for the used car has also resulted in a great increment in sales of Used Cars. Used Car Sales are on a global increase. But, determining the appropriate listing price of a used car is a challenging task, due to the many factors that drive prices of a used vehicle in the market. And that is why there is an urgent need for a system which can accurately predict the price of a used car. considering all the factors that affect the price of a used car.

**Keywords:** Used Car Price Prediction, Linear Regression, XGBoost, Decision Tree

## Introduction

To overcome this problem we have come up with a model that will be highly effective. Regression Algorithms are used because they provide us with continuously evaluated value as an output and not a categorized value or a value within a range. So,it will help us in predicting the actual price of a car rather than the price range of a car. We will also be providing a user interface which has also been developed which takes input from any user and displays the actual Price of a car according to user's inputs.

## Scope

In future this model can be updated with more recent data which can be used to create a new model which is average of the previous model with new data which would result in more consistency though the accuracy might be affected. Ensembling of models can be used to check if they create a more accurate result and this model may bind with websites which

can be used to provide the price prediction of a used car.

## Literature survey

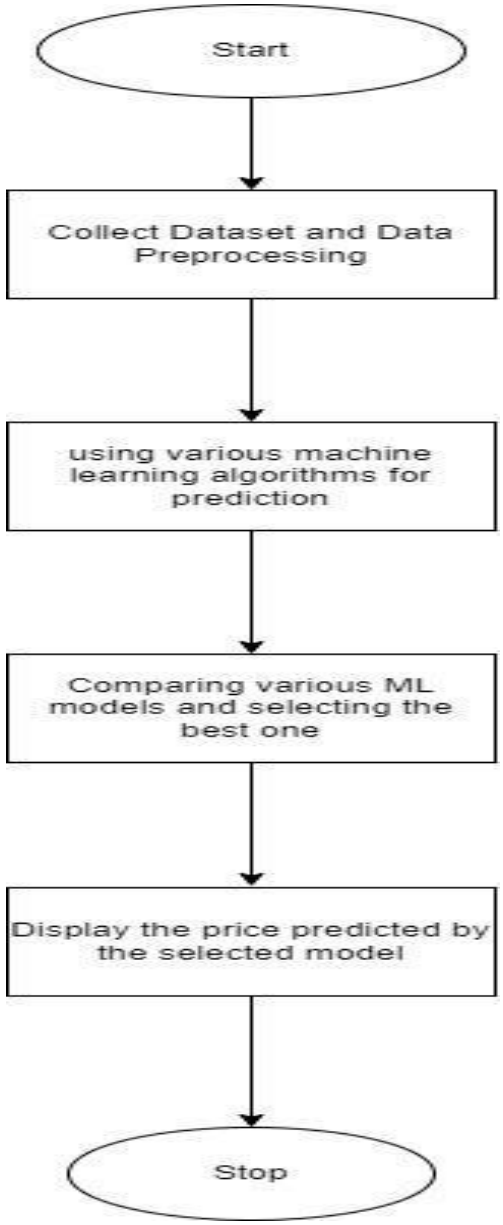
The research paper is Car Price Prediction Using Machine Learning Techniques. For predicting the prices of used cars, the first step is to decide what will be the features for the machine learning model or the attributes which basically decides the price of a used car like fuel type, number of previous owners, number of kilometres run, how many years old, manufacturing year and more[11].

The second paper is Predicting Used Car Prices. The accuracy of a machine learning model depends on the algorithm used for training it. In this paper, they implemented many supervised machine learning algorithms for making the prediction. Techniques such as multiple linear regression analysis, gradient boost, XGBoost and Random Forest have been used to make the predictions.[4]

The third paper is Vehile Price Prediction System using Machine Learning Techniques by Noor and Jan. They used multiple linear regression in their research. They worked on different variables that can influence the price of a vehicle and found out the variables which affect the price of a vehicle most and eliminated the rest of variables[7].

## Proposed architecture

First step includes searching for a relevant dataset for training and testing of the models. Then the data is preprocessed i.e. the cleaning and transformation of data takes place. Later this data is divided into training and testing subset and the training of various models takes place and later testing results are obtained. The comparison is done to find out which model is better among all the developed models, i.e. Multiple Regression, Decision Trees, Ridge Regression, Lasso Regression,




## Methodology

For this project, the dataset from kaggle is used which is based on the online car sales from Car Dekho. Features available in the dataset are name, year, selling price, kms driven, fuel, etc. Here selling price is the dependent variable and all others are dependent variables. Few of the fields were modified according to the needs of our models and then the data was divided into two parts, one for training of model and other for testing of model.

The data was then used to train various models namely Multiple Linear Regression, Decision Tree

Regression, Lasso Regression, Ridge Regression. Among them Decision Tree gave us the best results with least errors so other model which were based on Decision tree ensembling were also used like Random forest Regression, Gradient Boosting Regression and XGBoost Regression

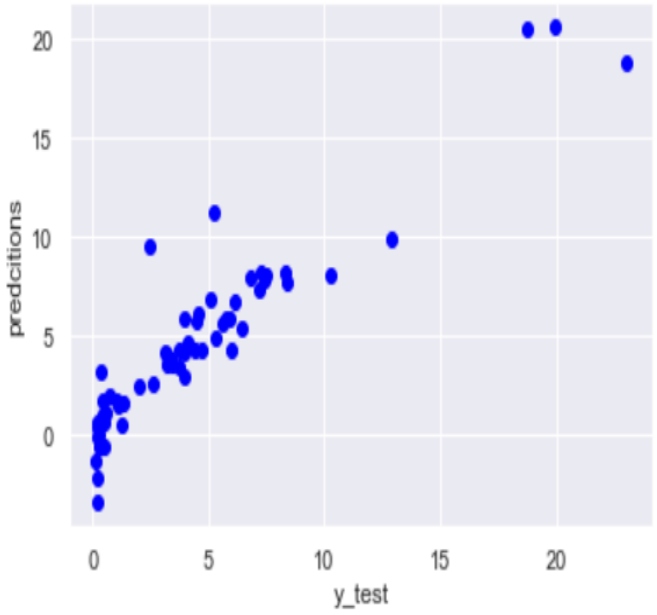
**Multiple Linear Regression**

In multiple linear regressions, multiple independent variables are used to predict the value of the desired variable which is known as dependent variable. The formula used is

$y=m_0 + m_1x_1 + m_2x_2 + \ldots + m_nx_n + c$

Here, y is the dependent variable, that is selling price in this case and $x_1, x_2, \ldots, x_n$ are the dependent variable. $m_1, m_2, \ldots, m_n$ are the weights calculated by the model.

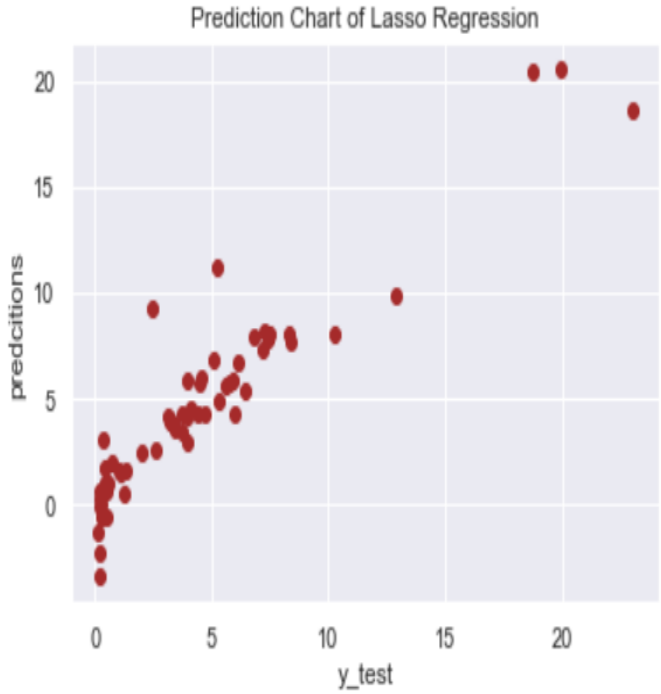
Our result for used car price prediction using multiple linear regressions:




**Lasso Regression**

LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression uses regularization and variable selection techniques to enhance the accuracy of the model. It also uses shrinkage which shrinks the data values towards a central point and the L1 regularization adds a penalty of the absolute value of coefficient's magnitude. It results in making sparse models that have fewer coefficients as some coefficients can become zero and thus gets eliminated from the model[1].

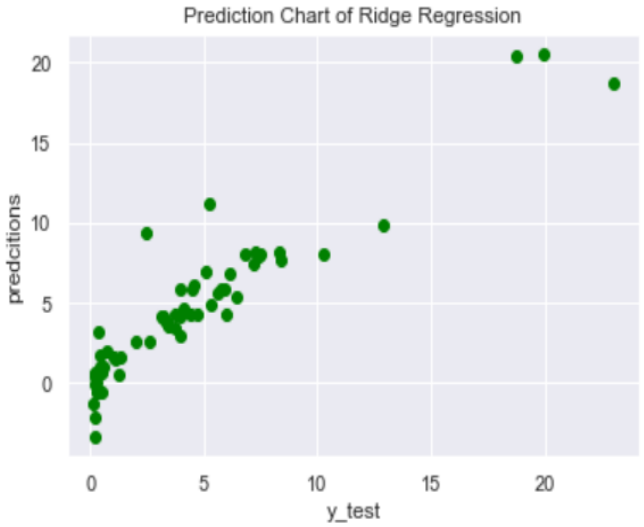
Our results for used car price prediction using Lasso Regression:




**Ridge Regression**

Ridge regression is the method used for multiple regression methods that are having the problem of multicollinearity. In case of multicollinearity, variances are large that lead to predicting the value which is far from the true desired value. It prevents multicollinearity by shrinking the parameters. It also makes use of L2 regularization technique, which adds L2 penalty of value square of the magnitude of coefficients and thus helps in dealing with multicollinearity which happens when independent values are highly correlated.

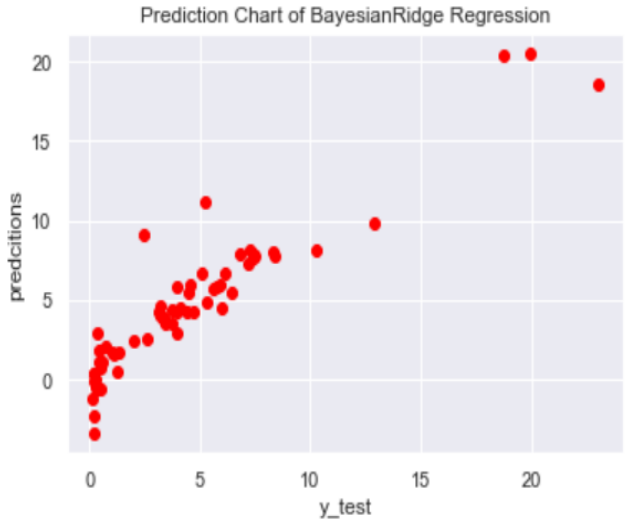
Our result for used car price prediction using ridge regression:




## Bayesian Ridge Regression

Bayesian Regression makes use of regularization parameters for estimation. Bayesian regression is beneficial to use when we have to deal with insufficient or poorly distributed data. Here, the output is calculated from probability distribution in spite of making prediction as a single value. It makes use of ridge regression and its coefficients under the Gaussian distribution to estimate the value of desired variable.

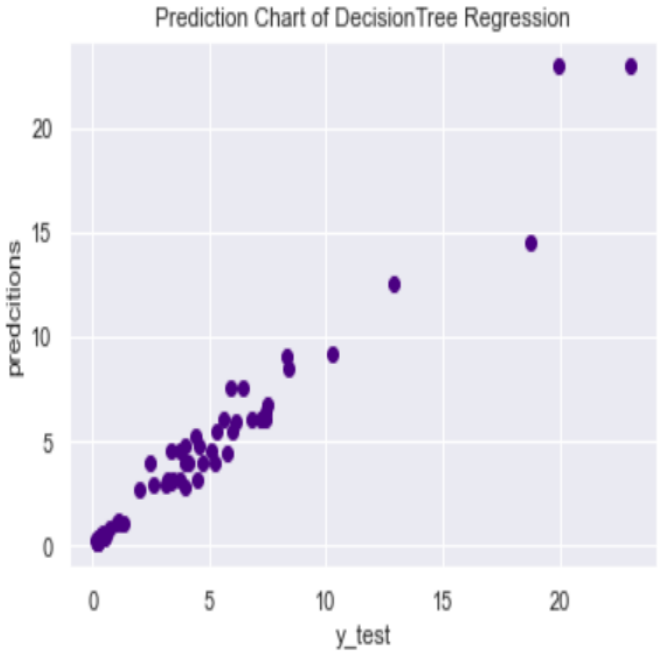
Our result for used car price prediction using Bayesian Ridge regression:




## DecisionTree Regression

In Decision Trees, the regression model is built in the structure like a tree. The dataset is broken into smaller subsets and simultaneously a decision tree is developed incrementally. The Decision tree looks like a flow-chart in which the internal nodes represent the test on an attribute. Branches represent the outcomes of the tests and the leaf nodes represent a class label. It is used to fit a sine curve with additional noisy observations.[8]

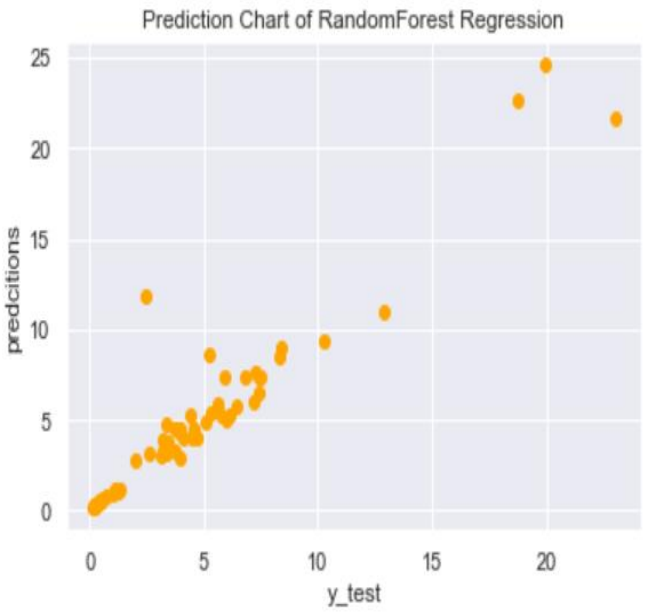
Our result for used car price prediction using Decision Tree regression:




## Random Forest Regression

Random Forest Regression is a type of supervised learning algorithm. It uses the ensemble learning technique for regression. Ensemble Learning technique is a method that combines the result from multiple models to make a more accurate prediction.
A Random Forest works by creating multiple decision trees during training and returning the mean of the predictions of all trees.[9]

Our result for used car price prediction using Random Forest regression:




## Gradient Boosting Regression

Gradient Boosting Regression provides us a model for prediction in the form of an ensemble of weak prediction models. Usually the weak models are decision trees. It usually is better than the random forest[10].

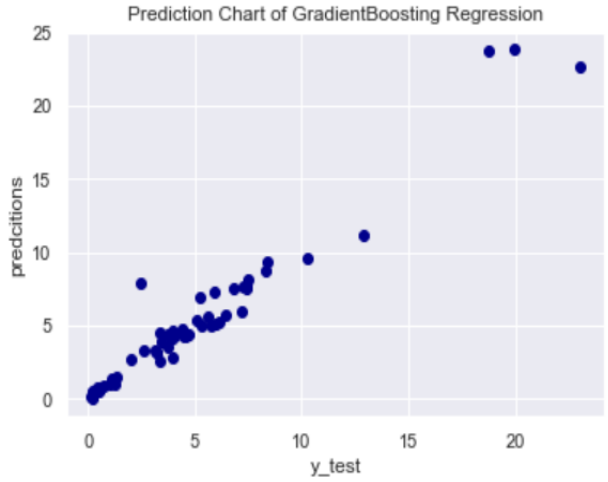
Gradient boosting has three elements:

1. A loss function which needs to be optimised.
2. A weak learning model which will make predictions.
3. A model to add to weak learning models to reduce the loss function.

Our result for used car price prediction using Gradient Boosting Regression:

Gradient Boosting Regression provides us a model for prediction in the form of an ensemble of weak prediction models. Usually the weak models are decision trees. It usually is better than the random forest.[10]

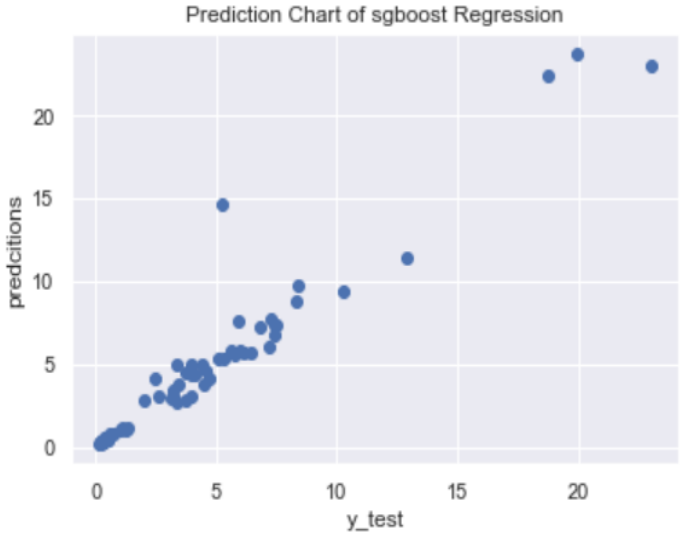
Gradient boosting has three element A loss function which needs to be optimised.




## XGBoost Regression

XGBoost is abv. for "Extreme Gradient Boosting", here "Gradient Boosting" originated from the paper "Greedy Function Approximation: A Gradient Boosting Machine, by Friedman".

XGBoost provides us with an effective and efficient implementation of gradient boosting algorithms and is an open-source library[6]
Our result for used car price prediction using XGBoost Regression




## Result

After training various models, the remaining data was used to test the accuracy of the models and checking if ensembling methods can provide us with better results and the following data was recorded after testing all the models with the test data:

| Errors<br><br>Models | Mean Absolute Error | Mean Squared Error | Root Mean Squared Error |
|---|---|---|---|
| Linear Regression | 1.0998575552990952 | 2.982384861859748 | 1.726958268708236 |
| Lasso Regression | 1.0934873952604163 | 2.907197959149361 | 1.7050507204037542 |
| Ridge Regression Model | 1.108094193398559 | 2.963295353286871 | 1.7214224796042576 |
| BayesianRidge Regression | 1.075017607433412 | 2.8302932475517473 | 1.6823475406561355 |
| Random Forest Regression | 0.7446229508196723 | 2.5644490327868863 | 1.6013897192085649 |
| DecisionTree Regression | 0.6027868852459017 | 0.9108311475409837 | 0.9543747416717312 |
| XGBoost Regression | 0.6822109709899934 | 2.2584028512539818 | 1.502798340182069 |
| GradientBoosting Regression | 0.6474333626420243 | 1.4921474947609719 | 1.2215348929772625 |

## Conclusion

All things which were said to affect the price of a used car were affecting the prices in our model too. We tried ensembling techniques but still the Decision trees Regression gives the best results.

We can see in the model vs errors table that the decision tree is giving us the least error among all the algorithms that we have tried for the model training. Along with that, it can be seen from the graph of all models that the decision tree is the most consistent method to predict the selling prices of used cars.

## Reference


1. https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/
2. https://www.irjet.net/archives/V8/i4/IRJET-V8I4278.pdf
3. https://www.kaggle.com/jpayne/852k-used-car-listings
4. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf
5. http://rstudio-pubs-static.s3.amazonaws.com/342283_57cd964a79004f579d2b0c52876198c9.html
6. https://xgboost.readthedocs.io/en/latest/parameter.html
7. https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf
8. https://www.saedsayad.com/decision_tree_reg.htm.
9. https://builtin.com/data-science/random-forest-algorithm.
10. https://en.wikipedia.org/wiki/Gradient_boosting
11. https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf


*****