CODING & SOLUTIONS

**Function Features**

[[Towards Data Science](https://towardsdatascience.com/?source=post_page-----e3be02d977b2--------------------------------)](https://towardsdatascience.com/?source=post_page-----e3be02d977b2--------------------------------)

Published in

[Towards Data Science](https://towardsdatascience.com/?source=post_page-----e3be02d977b2--------------------------------)

You have **2** free member-only stories left this month.

[Sign up for Medium and get an extra one](https://medium.com/m/signin?operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fused-car-price-prediction-using-machine-learning-e3be02d977b2&source=-----e3be02d977b2---------------------smart_meter-----------)

[[Panwar Abhash Anil](https://abhashpanwar.medium.com/?source=post_page-----e3be02d977b2--------------------------------)](https://abhashpanwar.medium.com/?source=post_page-----e3be02d977b2--------------------------------)

[Panwar Abhash Anil](https://abhashpanwar.medium.com/?source=post_page-----e3be02d977b2--------------------------------)

Aug 3, 2020

·

11 min read

·

Member-only

·

Listen

**Used Car Price Prediction using Machine Learning**

Data Cleaning, Data Preprocessing, 8 Different ML Models and Some Insights from Data

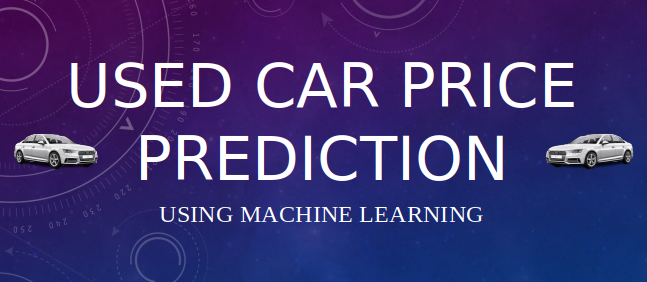


Image By Panwar Abhash Anil

You can reach all Python scripts relative to this on my [*GitHub page*](https://github.com/abhashpanwar/used-car-price-prediction). If you are interested, you can also find the scripts used for data cleaning and data visualization for this study in the same repository.

The project is also deployed using Django on [Heroku](https://abhash-car-price-prediction.herokuapp.com/) and also on Amazon EC2 by [Dockerizing Django App](http://ec2-54-82-16-204.compute-1.amazonaws.com:8080/" \t "_blank)

**Content**

1. Data Cleaning (Identifying null values, filling missing values and removing outliers)
2. Data Preprocessing (Standardization or Normalization)
3. ML Models: Linear Regression, Ridge Regression, Lasso, KNN, Random Forest Regressor, Bagging Regressor, Adaboost Regressor, and XGBoost
4. Comparison of the performance of the models
5. Some insights from data

**Why is price feature scaled by log transformation?**

In the regression model, for any fixed value of X, Y is distributed in this problem data-target value (Price ) not normally distributed, it is right skewed.

To solve this problem, the log transformation on the target variable is applied when it has skewed distribution and we need to apply an inverse function on the predicted values to get the actual predicted target value.

Due to this, for evaluating the model, the [*RMSLE*](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_log_error.html) is calculated to check the error and the [*R2 Score*](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html) is also calculated to evaluate the accuracy of the model.

**Some Key Concepts:**

* **Learning Rate:**Learning rate is a hyper-parameter that controls how much we are adjusting the weights of our network concerning the loss gradient. The lower the value, the slower we travel along the downward slope. While this might be a good idea (using a low learning rate) in terms of making sure that we do not miss any local minima, it could also mean that we’ll be taking a long time to converge — especially if we get stuck on a plateau region.
* **n\_estimators**: This is the number of trees you want to build before taking the maximum voting or averages of predictions. A higher number of trees give you better performance but make your code slower.
* **R² Score:**It is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. 0% indicates that the model explains none of the variability of the response data around its mean.

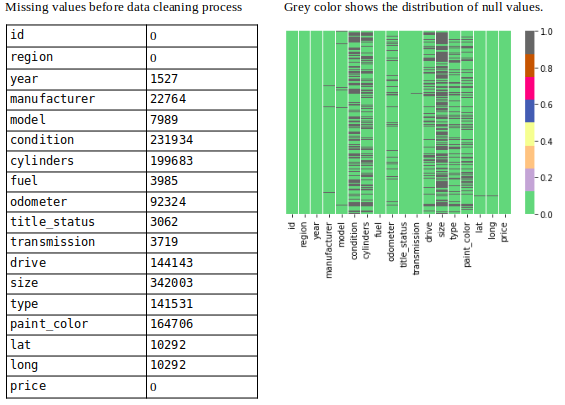
**1. The Data:**

The dataset used in this project was downloaded from [Kaggle](https://www.kaggle.com/austinreese/craigslist-carstrucks-data).

**2. Data Cleaning:**

The first step is to remove irrelevant/useless features like ‘URL’, ’region\_url’, ’vin’, ’image\_url’, ’description’, ’county’, ’state’ from the dataset.

As a next step, check missing values for each feature.

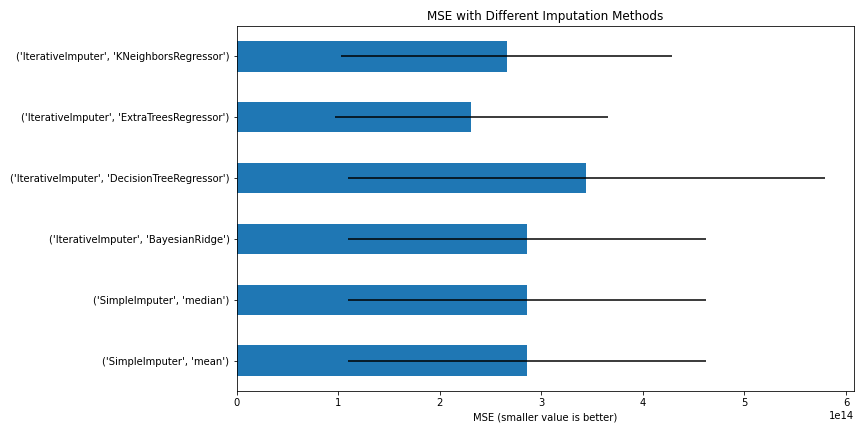


Showing missing values (Image By Panwar Abhash Anil)

Next, now missing values were filled with appropriate values by an appropriate method.

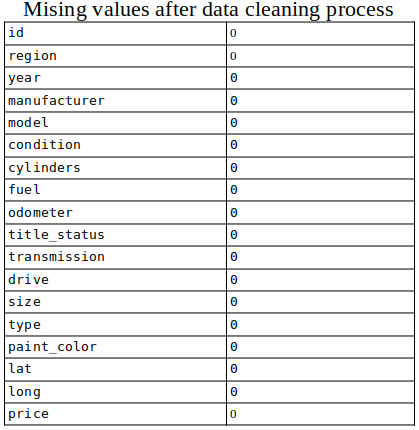
To fill the missing values, *[IterativeImputer](https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html" \t "_blank)* method is used and different estimators are implemented then calculated [*MSE*](https://en.wikipedia.org/wiki/Mean_squared_error) of each estimator using *[cross\_val\_score](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html" \t "_blank)*

1. Mean and Median
2. BayesianRidge Estimator
3. DecisionTreeRegressor Estimator
4. ExtraTreesRegressor Estimator
5. KNeighborsRegressor Estimator



MSE with Different Imputation Methods (Image By Panwar Abhash Anil)

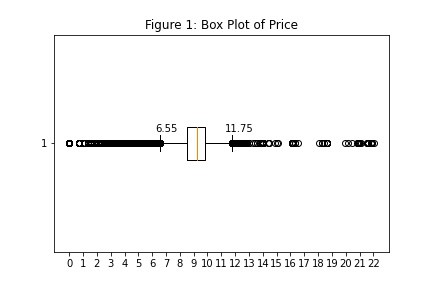
From the above figure, we can conclude that the *ExtraTreesRegressor*estimator will be better for the imputation method to fill the missing value.



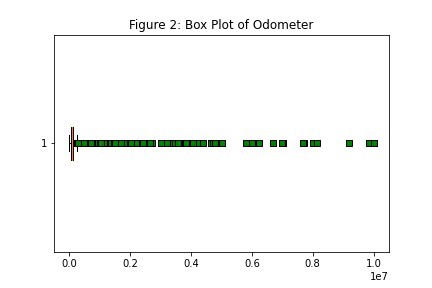
Missing values after filling (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

At last, after dealing with missing values there zero null values.

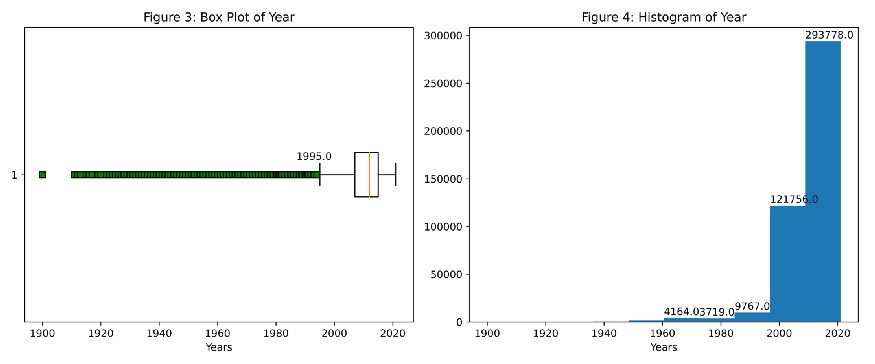
**Outliers:**InterQuartile Range (IQR) method is used to remove the outliers from the data.



Box Plot of price showing outliers (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))



Box Plot of Odometer showing outliers (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))



Box Plot & Histogram of the year (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

* From figure 1, the prices whose log is below 6.55 and above 11.55 are the outliers
* From figure 2, it is impossible to conclude something so IQR is calculated to find outliers i.e. odometer values below 6.55 and above 11.55 are the outliers.
* From figure 3, the year below 1995 and above 2020 are the outliers.

At last, Shape of dataset before process= (435849, 25) and after process= (374136, 18). Total 61713 rows and 7 cols removed.

**3. Data preprocessing:**

**Label Encoder:** In our dataset, 12 features are categorical variables and 4 numerical variables (price column excluded). To apply the ML models, we need to transform these categorical variables into numerical variables. And sklearn library *[LabelEncoder](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html" \t "_blank)* is used to solve this problem.

**Normalization**: The dataset is not normally distributed. All the features have different ranges. Without normalization, the ML model will try to disregard coefficients of features that have low values because their impact will be so small compared to the big value. Hence to normalized, *[sklearn library i.e. MinMaxScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html" \t "_blank)* is used.

**Train the data.** In this process, 90% of the data was split for the train data and 10% of the data was taken as test data.

**4. ML Models:**

In this section, different machine learning algorithms are used to predict price/target-variable.

The dataset is supervised, so the models are applied in a given order:

1. [Linear Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)
2. [Ridge Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ridge_regression.html)
3. [Lasso Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html)
4. [K-Neighbors Regressor](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html)
5. [Random Forest Regressor](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)
6. [Bagging Regressor](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html)
7. [Adaboost Regressor](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html)
8. [XGBoost](https://xgboost.readthedocs.io/en/latest/)

**1) Linear Regression:**

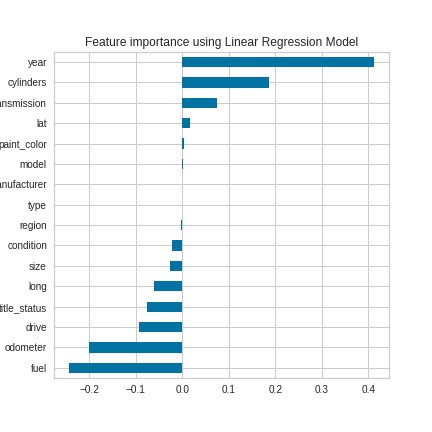
In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. [*More Details*](https://en.wikipedia.org/wiki/Linear_regression)

Coefficients: The sign of each coefficient indicates the direction of the relationship between a predictor variable and the response variable.

* A positive sign indicates that as the predictor variable increases, the response variable also increases.
* A negative sign indicates that as the predictor variable increases, the response variable decreases.

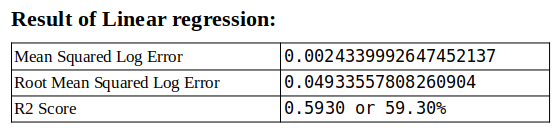


Graph showing performance of Linear regression (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))



Graph showing important feature of the dataset (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

Considering this figure, linear regression suggests that ***year, cylinder, transmission, fuel, and odometer*** these five variables are the most important.

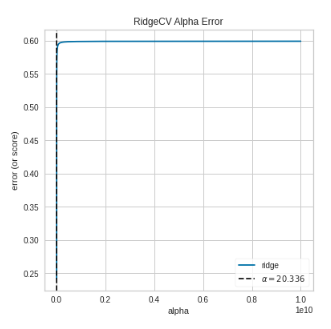


Result of Linear regression (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

**2) Ridge Regression:**

**Ridge Regression** is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.

To find the best alpha value in ridge regression, yellowbrick library *[AlphaSelection](https://www.scikit-yb.org/en/latest/api/regressor/alphas.html" \t "_blank)* was applied.

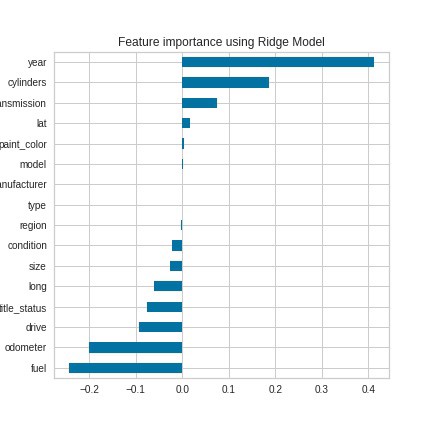


Graph showing best value of Alpha

From the figure, the best value of alpha to fit the dataset is 20.336.

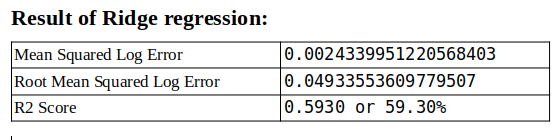
Note: The value of alpha is not constant it varies every time.

Using this value of alpha, Ridgeregressor is implemented.



Graph showing Important Features

Considering this figure, Lasso regression suggests that ***year, cylinder, transmission, fuel, and odometer*** these five variables are the most important.



Final Result of this model (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

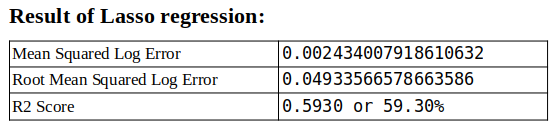
The performance of ridge regression is almost the same as Linear Regression.

**3)Lasso Regression:**

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point as mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

**Why Lasso regression is used?**

The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that cause regression coefficients for some variables to shrink toward zero.



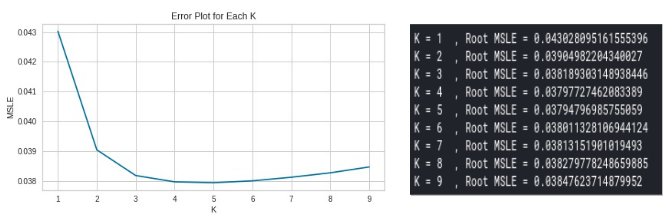
Final Result of this model (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

But for this dataset, there is no need for lasso regression as there no much difference in error.

**4)KNeighbors Regressor: Regression-based on k-nearest neighbors.**

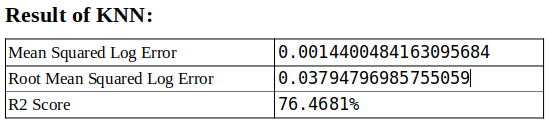
The target is predicted by local interpolation of the targets associated with the nearest neighbours the training set.

*k*-NN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until function evaluation. [***Read More***](https://www.kite.com/python/docs/sklearn.neighbors.KNeighborsRegressor)



Error Plot For Each K range 1–9 (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

From the above figure, for k=5 KNN give the least error. So dataset is trained using n\_neighbors=5 and metric=’euclidean’.



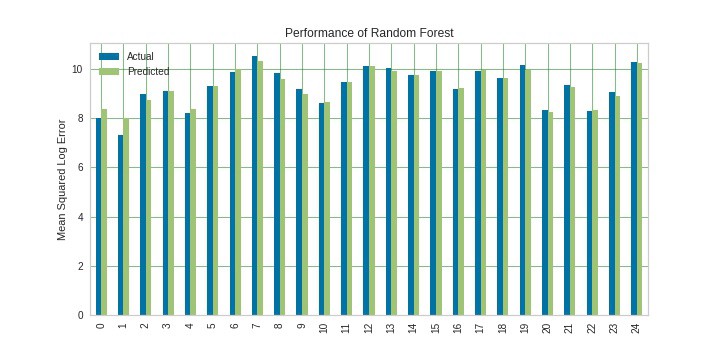
Final result of KNN (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

The performance KNN is better and error is decreasing with increased accuracy.

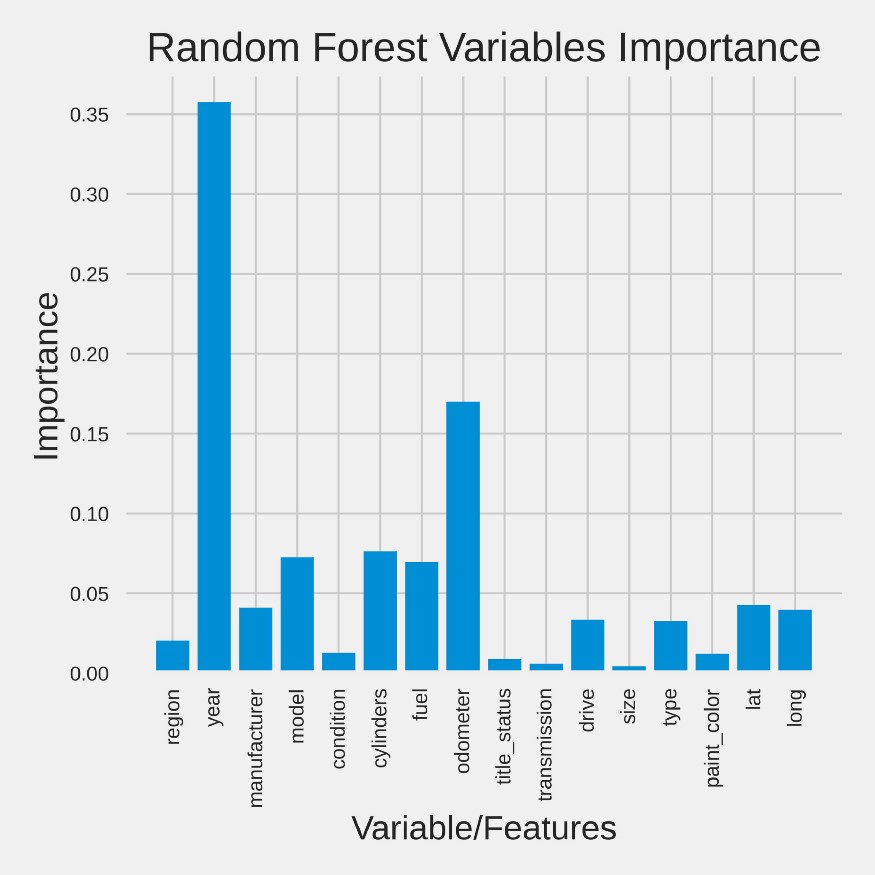
**5) Random Forest:**

The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. [***Read More***](https://en.wikipedia.org/wiki/Random_forest)

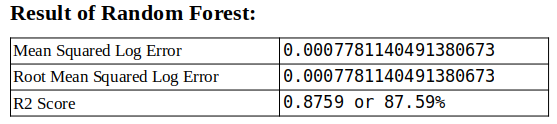
In our model, 180 decisions are created with max\_features 0.5



Performance of Random Forest (True value vs predicted value)



This is the simple bar plot which illustrates that ***year*** is the most important feature of a car and then ***odometer*** variable and then others.



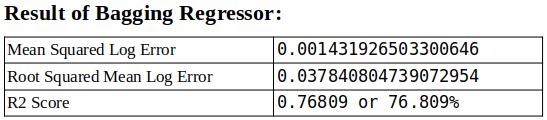
Accuracy Table of Random Forest Model (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

The performance of the Random forest is better and accuracy is increased by approx. 10% which is good. Since the random forest is using bagging when building each tree so next Bagging Regressor will be performed.

**6) Bagging Regressor:**

A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregates their predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it. [***Read More***](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html)

In our model, DecisionTreeRegressor is used as the estimator with max\_depth=20 which creates 50 decision trees and the results show below.



Accuracy Table of Bagging Regressor (Image By [Panwar Abhash Anil](https://www.linkedin.com/in/abhash-panwar-85126976/))

The performance of Random Forest is much better than Bagging regressor.

**The key difference between Random forest and Bagging:**The fundamental difference is that in **Random forests**, only a subset of features are selected at random out of the total **and** the best split feature from the subset is used to split each node in a tree, unlike in **bagging** where all features are considered for splitting a node.

**7) Adaboost regressor:**

AdaBoost can be used to boost the performance of any machine learning algorithm. Adaboost helps you combine multiple “weak classifiers” into a single “strong classifier”. **Library used:**[AdaBoostRegressor](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html" \t "_blank) & [***Read More***](https://en.wikipedia.org/wiki/AdaBoost)

