University of Windsor

Random Forest Regression

Presented By

Darshan Mehta
Lakshesh Vyas
Utsav Patel

Instructor : Yasser Alginahi

6th October, 2023

University of Windsor

Overview

- What is Random Forest Regression ?
- Ensemble Learning
- How Bagging Works?
- Supervised and Unsupervised Learning
- Real Time-Based Application

What is Random Forest Regression?

- Random Forest is an ensemble technique that can handle both regression and classification tasks by combining many decision trees and a technique known as Bootstrap and Aggregation, or bagging.
- The core idea is to use numerous decision trees to determine the final output rather than depending on individual decision trees.
- Random Forest's foundation learning models are numerous decision trees. We randomly select rows and features from the dataset to create sample datasets for each model. This section is known as Bootstrap.

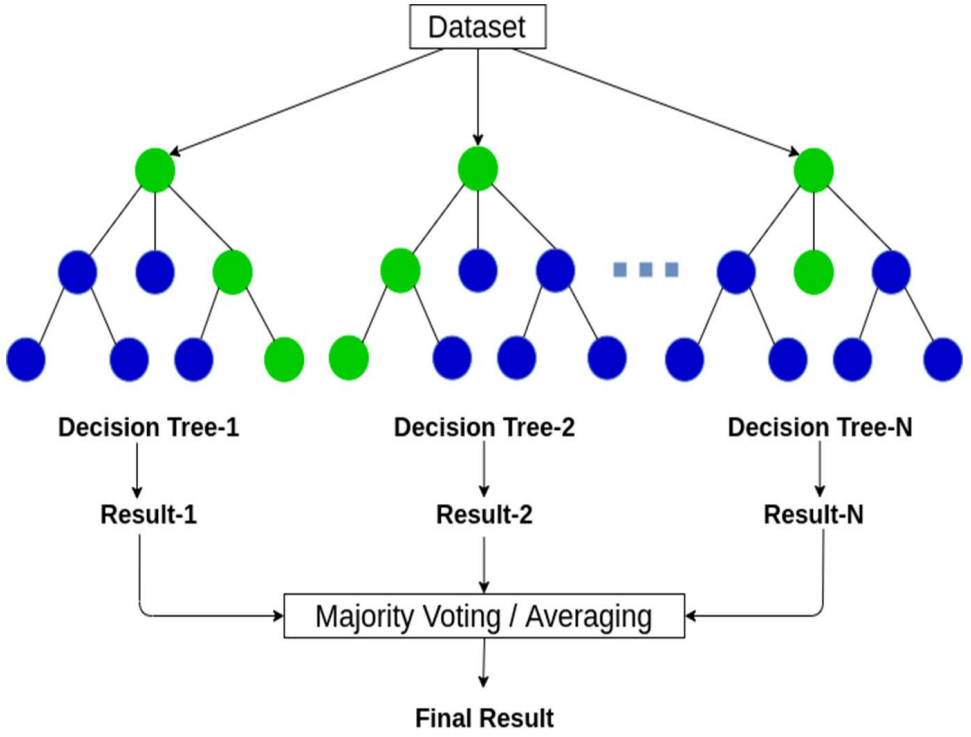

Fig 1. Random Forest Regression [1]

How Bagging works?

- **Bootstrapping:** This resampling technique creates several subsets of the training dataset by randomly and replacement-wise picking data points. This indicates that you have the option to select the same instance more than once each time you choose a data point from the training dataset.

- **Aggregation:** An average or majority of the predictions are taken to compute a more accurate estimate. In the case of regression, an average is taken of all the outputs predicted by the individual classifiers; this is known as soft voting.

What is Bagging ?

Bagging :

Bootstrap Aggregating, a.k.a bagging, is a machine learning ensemble algorithm meant to increase the stability and accuracy of classification and regression machine learning algorithms. It reduces variance and aids in avoiding overfitting.

These weak models are then trained individually or parallelly based on the type of task (regression or classification) after several data samples have been created.

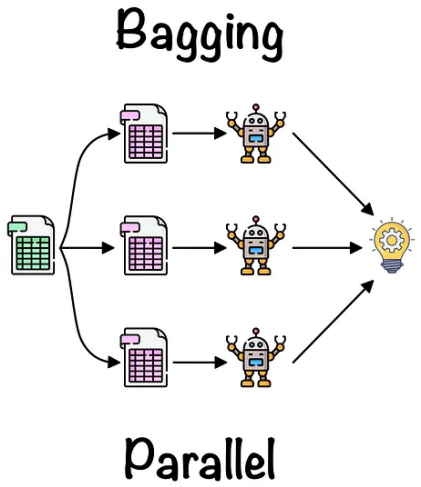

Fig 4. Bagging [3]

What is Boosting?

Boosting :

It is an ensemble modelling technique that generates a strong classifier from a set of weak classifiers. It is accomplished by developing a model in series utilizing weak models. First, a model is constructed using the training data. The second model is then constructed in an attempt to address the faults in the previous model, and it goes on till the last model and the output of the last model is considered as the final output or prediction of the entire model.

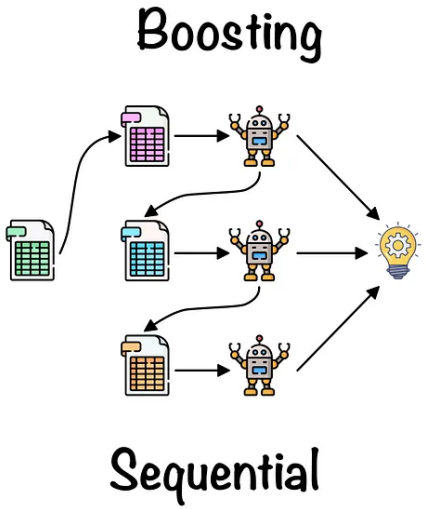

Fig 5. Boosting [3]

Example of Ensemble Learning

Consider how different it would be if one individual were to solve a math problem as opposed to a group of people. If one person answered the question incorrectly, that person would be the only one who could check their work. On the other hand, the group can work together to find a solution, watch out for one another's errors, and decide on the solution by opinion.

“Two heads are better than one”

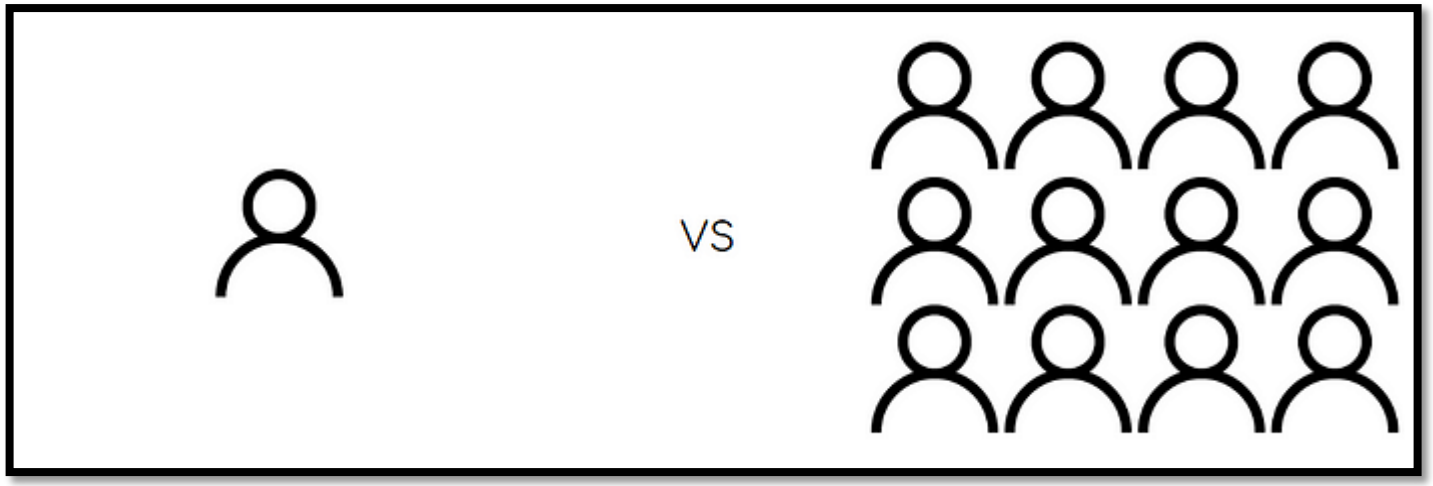

Fig 3. Example of Ensemble Learning

Ensemble Learning

- Ensemble models are a type of machine learning approach that combines several different models in the prediction process. Ensemble models provide a solution to the technical constraints of developing a single estimator.
- An ensemble approach is a methodology that uses numerous independent, comparable or dissimilar models/weak learners to generate an output or make predictions.

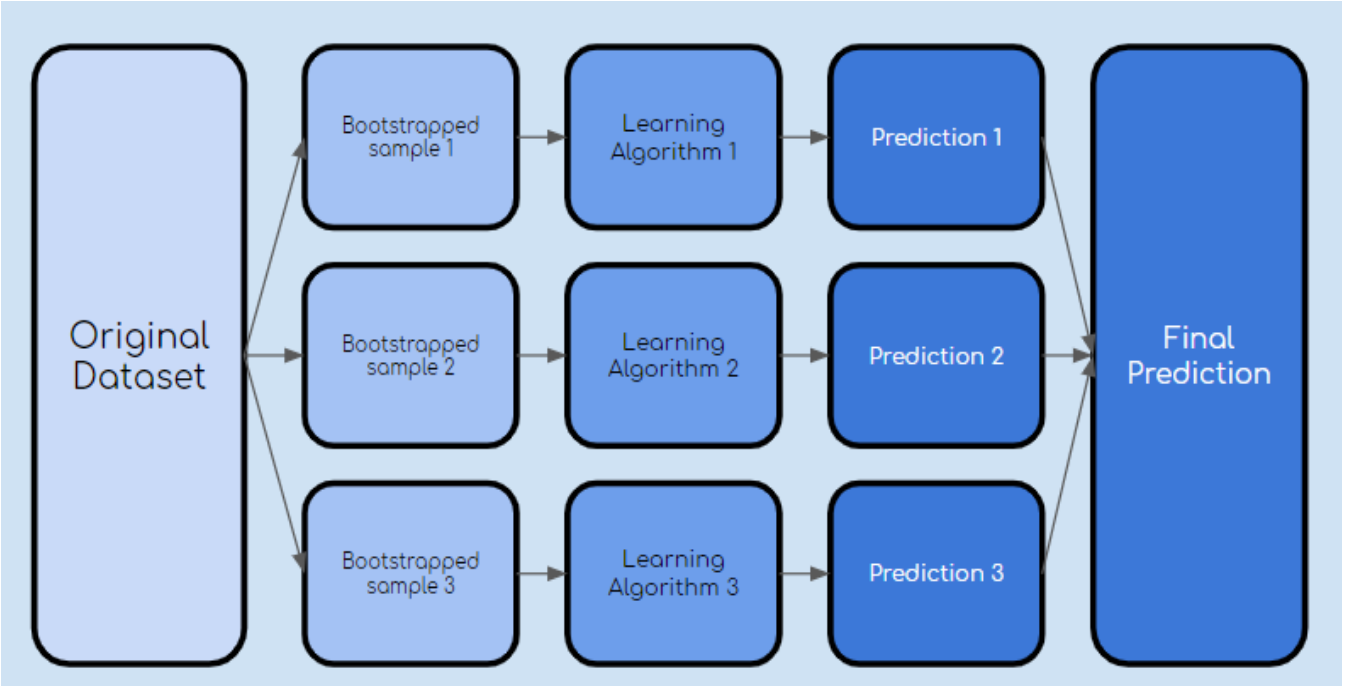

Fig 2. Ensemble Learning [2]

How Ensemble Learning works?

- It will start taking samples from the original dataset using bootstrap method.
- All the weak learners (Decision Trees) will take each sample as input and will start processing.
- Every decision tree will predict the output based on their data samples.
- Finally, aggregation will be performed and the majority or an average will be considered to predict the final output or to form a strong learner.

Why use random forest tree instead of decision tree ?

Decision Tree	Random Forest Tree
Single decision tree.	Ensemble of multiple decision trees
Predicts by averaging leaf node values.	Predicts by averaging predictions from multiple trees.
Prone to overfitting, especially on complex data.	Reduces overfitting through aggregation of multiple trees.
Split data based on feature thresholds, can lead to capturing noise.	Randomly selects subsets of features for each tree to reduce correlations.
Highly interpretable, easy to visualize and follow the decision path.	Less interpretable due to ensemble nature and multiple trees.

Example of Random Forest Regression

Consider a Fruit Box consisting of three fruits Apples, Oranges, and Cherries in training data that n = 3. We are predicting the fruit which is maximum in number in a fruit box. A random forest model using the training data with a number of trees, k = 3.

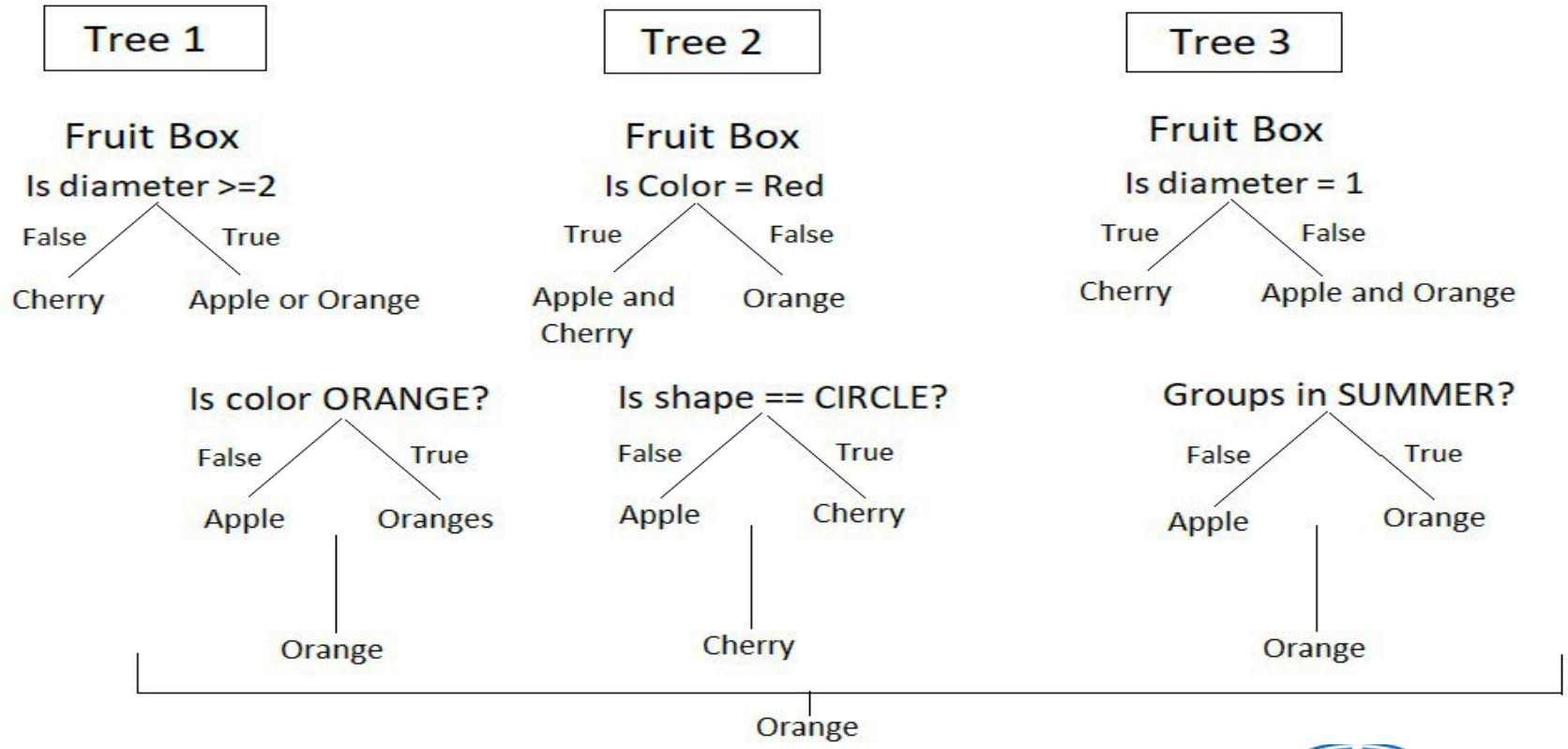

Fig 8. Example of Random Forest Regression

How to implement Random Forest Regression in Python

There are four steps to implement random forest regression in code:

1. Importing Python Libraries and Loading our Data Set into a Data Frame. (such as pandas, NumPy and matplotlib)

2. Splitting our Data Set Into Training Set and Test Set while using sklearn model selection library.

3. Creating a Random Forest Regression Model and Fitting it to the Training Data

4. Visualizing the Random Forest Regression Results

Applications of Random Forest

This algorithm is used to forecast behavior and outcomes in several sectors, including banking and finance, e-commerce, and healthcare. It has been increasingly employed thanks to its ease of application, adaptability, and ability to perform both classification and regression tasks.

Fig 9. Applications [6]

Example: Diabetes Prediction

- **Data collection:-** Get historic stock market data from trading volume , economic indicators and news data.
- **Data Processing:-** This involve handling missing value and creating target variable for prediction.
- **Data splitting:-** Split dataset into training and testing set while using decision tree.
- **Random Forest Regression model:-** Comparing the n number of decision trees to get accurate result instead of using single decision tree.

Fig 10. Stock Market

Code

https://colab.research.google.com/drive/14rqPkI_mhh3TJS1xisxiazlT6cBQM_NM?usp=sharing

Advantages

1. Random forest is not biased
2. Random forest is stable
3. It has a high accuracy rate
4. Works well for larger data sets
5. It can also work well when data has missing values

Disadvantages

1. It is more complex
2. Time Consuming
3. Not ideal for smaller datasets

Youtube Video

This video explains the basic concept of the random forest algorithm and how it works.

The example in the video refers to identify hand written digits recognition with the help of the given data.

https://youtu.be/ok2s1vV9XW0?si=Kn5XDph5culEArA4

References

[1] T. Shin, "Ensemble learning, bagging, and boosting explained in 3 minutes," 24 07 2020. [Online]. [Accessed 20 09 2023].

[2] T. Shin, "Ensemble learning, bagging, and boosting explained in 3 minutes," 24 07 2020. [Online]. Available: https://towardsdatascience.com/ensemble-learning-bagging-and-boosting-explained-in-3-minutes-2e6d2240ae21. [Accessed 1 10 2023].

[3] F. López, "Ensemble Learning: Bagging & boosting," 1 1 2021. [Online]. Available: https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422. [Accessed 2 10 2023].

[4] Labellerr, "Mastering of supervised and unsupervised learning: Know the differences," 02 08 2023. [Online]. Available: https://www.labellerr.com/blog/supervised-vs-unsupervised-learning-whats-the-difference/amp/. [Accessed 2 10 2023].

[5] A. Chakure, "Random forest regression in Python explained," 07 03 2022. [Online]. Available: https://builtin.com/data-science/random-forest-python. [Accessed 4 10 2023].

[6] A. Brital, "Github.io," [Online]. Available: https://anasbrital98.github.io/blog/2021/Random-Forest/. [Accessed 1 10 2023].

Thank You