05 Random Forests

Created Date: 2022-08-29

> Metadata 🗃️
>
> - Title: Decision Trees for Regression (Regression Trees)
> - Author: Andrew Jones
> - Reference: Data Science Infinity

> Links & Tags 📎
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #machinelearning #randomforests

High-Level Overview

Jupyter Notebook: Basic Random Forest Template

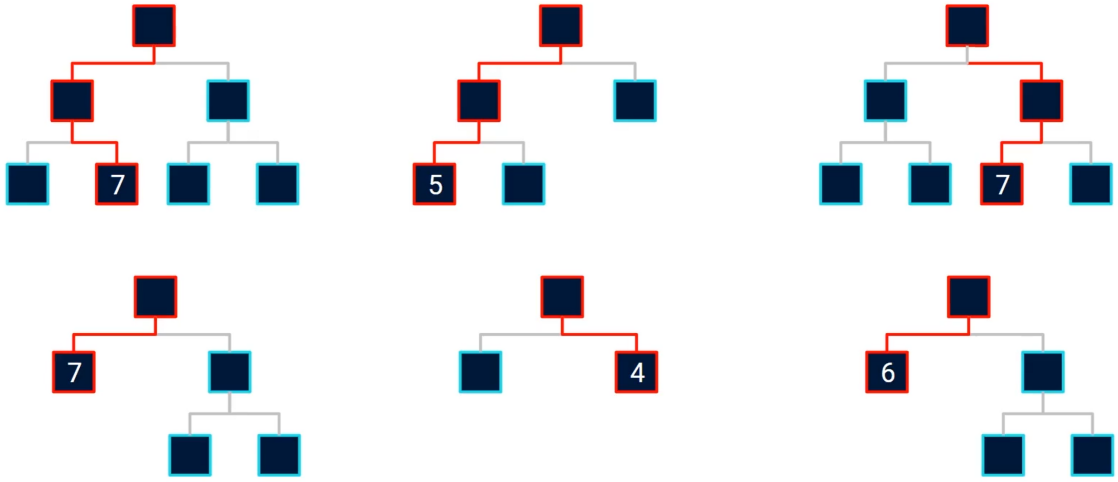
> A Random Forest is an an ensemble model consisting of many Decision Trees working together across different randomly selected subsets of the data, facilitating improved accuracy and stability.

> Bootstrapping is a sampling method that involves iteratively resampling data with replacement, meaning the random rows selected are still available for selection.

- Each decision tree will produce it's own prediction on a random sample of data points and input variables using the bootstrapping sampling

technique
    - At each individual decision tree split, a new random selection of input variables are selected (Default for SciKit Learn is to use all input variables but a parameter can be used to select random input variables)
    - Typically, the number of input variables selected is calculated one of three ways
        - $\sqrt{n}$
        - log^n
        - $n/3$
    - Bootstrapping sampling technique is repeated for each iteration of decision tree
- This intuitively seems like it will hinder the prediction performance but that is the point
    - Each decision tree in the random forest is forced to predict the output variables in different way
- The random forest model will take the average of all these predictions as the final output
- We generally don't have to worry about applying stopping conditions for each decision tree since the sampling of input variables at each split inherently limits the total amount of splits
- We also don't have to remove outliers as each decision tree split is based on data on either side of a line, and distance is not important here
- Lastly, feature selection is not applicable in random forest models since each variable is judged independently

Random Forest for Regression



Advanced Theory

Jupyter Notebook: Advanced Random Forest Template

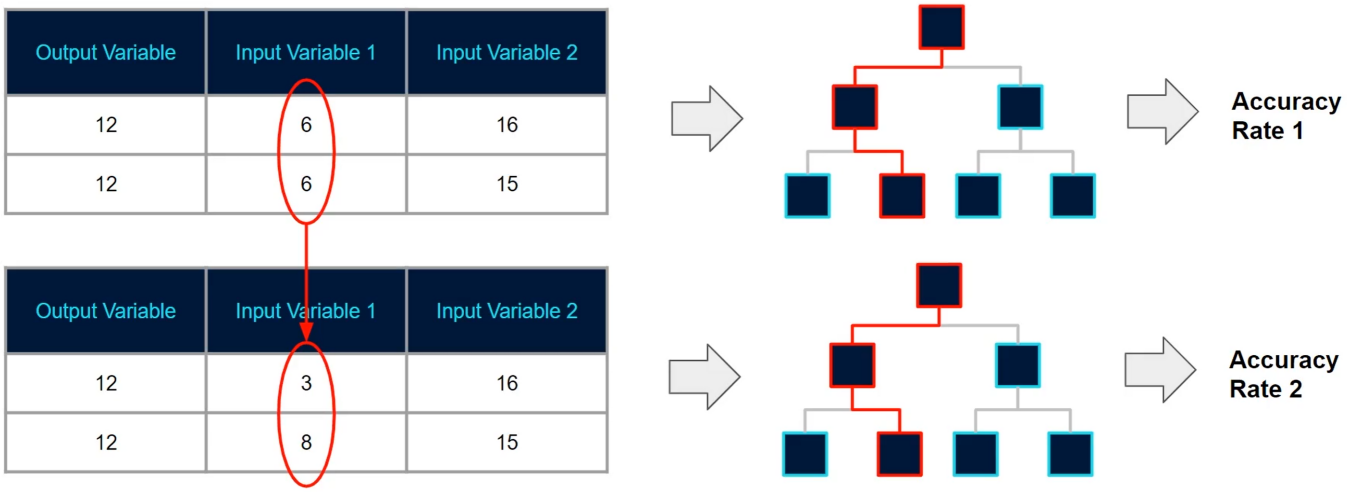
Feature Importance

> How much would accuracy decrease if a specific input variable was removed? If a significant decrease in accuracy is seen when an input variable is removed, that variable is deemed as important.

- There are two common approached for measuring feature importance
  - MSE Approach
    - Identify all the nodes where a particular input variable was used for splitting
    - Compare the MSE before and after the split of those nodes
    - Average the improvements in MSE across the random forest to determine the improvement that input variable causes in the model
    - This can be done for all input variables to compare differences and determine which features increase model performance the most

- Permutation Importance
    - Uses data that was not used during the bootstrapping sampling approach (recall every iteration is random and previously picked rows are available)
        - These rows are called "out of bag" rows and can be used to test the accuracy of a tree
    - These rows are passed through the decision tree to obtain an accuracy score (MSE or R^2)
    - We then randomized the values in one of the input variables to eliminate associations with the output variable, and pass the rows through the decision tree again (obtaining an accuracy score for this tree as well)
    - The difference between the two accuracy scores tells us how important that particular input variable is in determining model performance
    - This can be done for all input variables to determine feature importance
- These techniques are often used a feature selection technique when looking to apply different types of models

Permutation Importance



Evaluating Model Performance

- The most commonly used metric to determine model performance is R Squared
- The R^2 measure (Coefficient of Determination) assesses the accuracy, or goodness of fit, of the decision tree
    - 0 = No Relationship
    - 1 = Perfect Relationship
    - Explains how much of the variation in the output variable (y) is explained by the input variables (x)
    - An $R^2 = 0.70$ is telling us that our input variable(s) are explaining 70% of the variation in the output variable (y)

$$R^2 = \frac{SSR[Mean] - SSR[Model]}{SSE[Mean]}$$

- SSR Sum of Squared Residuals
- SSR Model Sum of Squared Residuals Using the Predicted Values
- SSR Mean Sum of Squared Residuals Using the Mean for Predicted Values
    - Squared residuals are calculated as the difference between the actual output and predicted (or mean) output, squared

$$R^2 = 1 - \frac{RSS}{TSS}$$

- RSS Sum of Squared Residuals
- TSS Total Sum of Squares

Where;

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

- y_i Sample Value

- $f(x_i)$ Predicted Value
- n Number of Observations

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- y_i Sample Value
- $\bar{y}$ Mean Sample Value
- n Number of Observations