08 Random Forests (Classification)

Created Date: 2022-11-02

> Metadata 🗃️
>
> - Title: Random Forests (Classification)
> - Author: Andrew Jones
> - Reference: Data Science Infinity

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

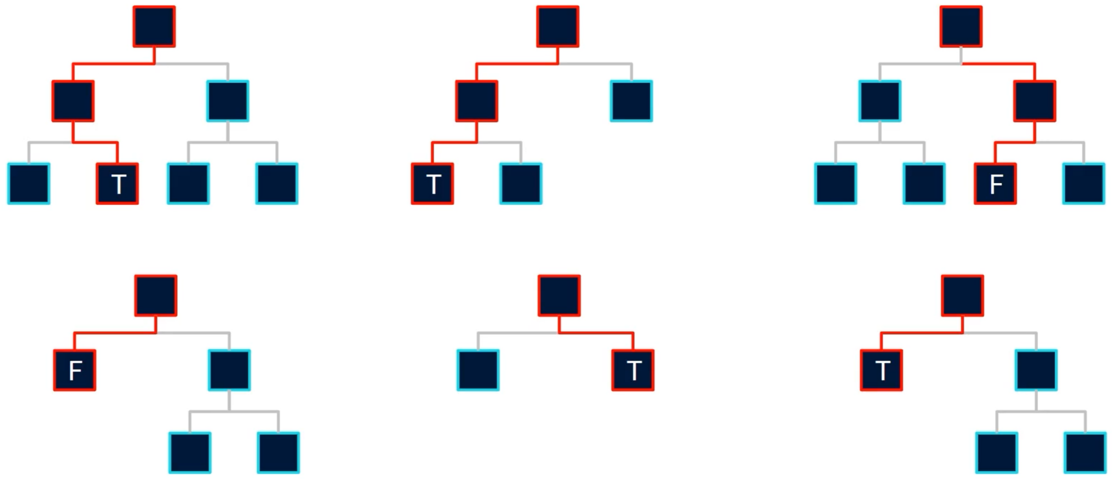
---

High-Level Overview

Jupyter Notebook: Basic Classification Forests Template

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

Random Forest for Classification



- Each decision tree is build using a random sample of data and at each split, the model uses a random sample of input variables
    - Data sampling method is Bootstrapping which involves iteratively resampling data with replacement (data selected is still available for selection)
    - The number of input variables at each split is often $\sqrt{n}$ where n is the number of available input variables
- Each decision tree outputs a classification prediction
- Each decision tree "votes" using its class prediction, and the winning class is the class with the majority of the votes

Advanced Theory

Jupyter Notebook: Advanced Classification Forests Template

Feature Importance

> How much would accuracy (Gini impurity) 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

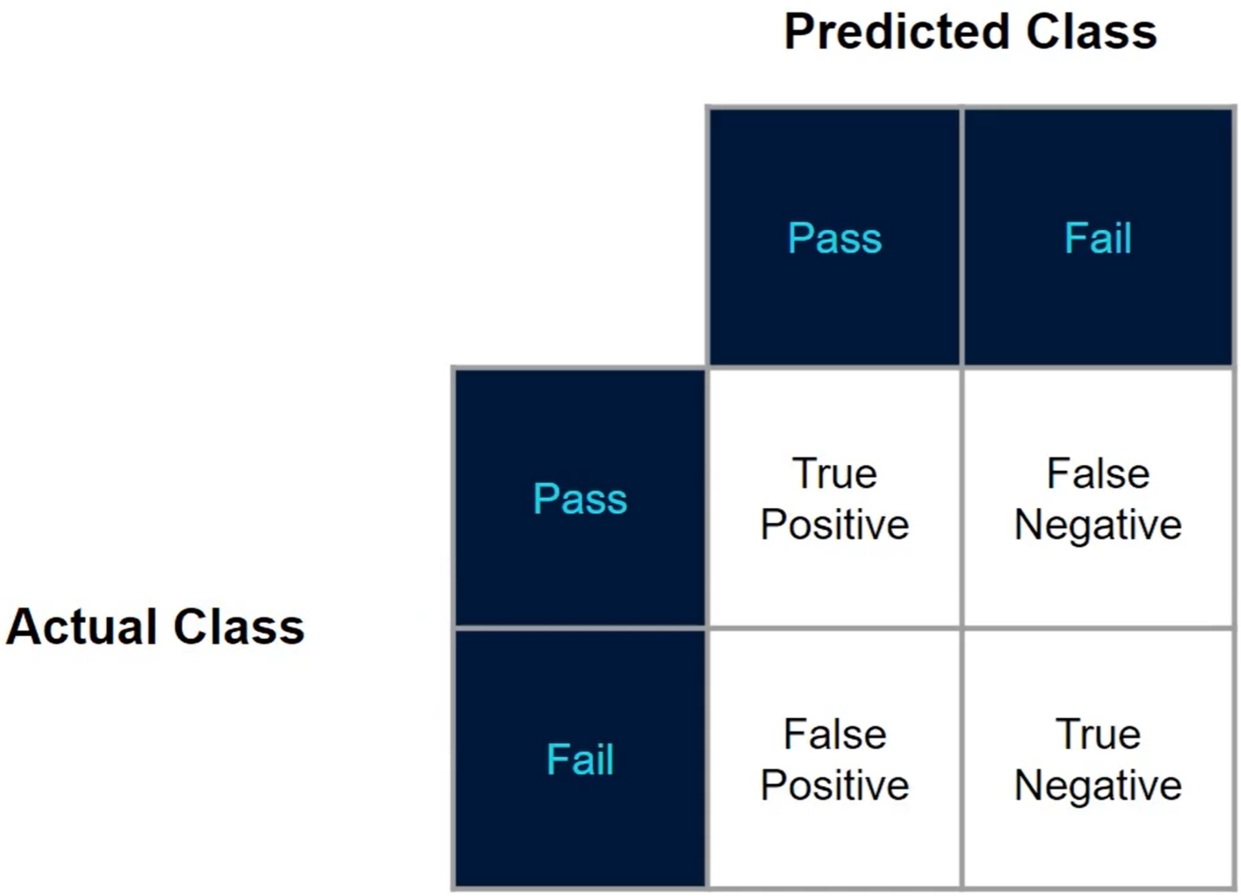
    - Gini Approach
        - Identify all the nodes where a particular input variable was used for splitting
        - Compare the Gini before and after the split of those nodes
        - Average the improvements in Gini across the random forest to determine the improvement that input variable causes in 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 (Gini)
        - 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

Evaluating Classification Accuracy

- $Classification\ Accuracy = \frac{n\ Outcomes\ Classified\ Correctly}{Total\ Outcomes}$

- n Number of Outcomes Classified Correctly

- - Type I Error: False Positive - Type II Error: False Negative - Diagonal (True Positive & True Negative): Outcomes Correctly Classified

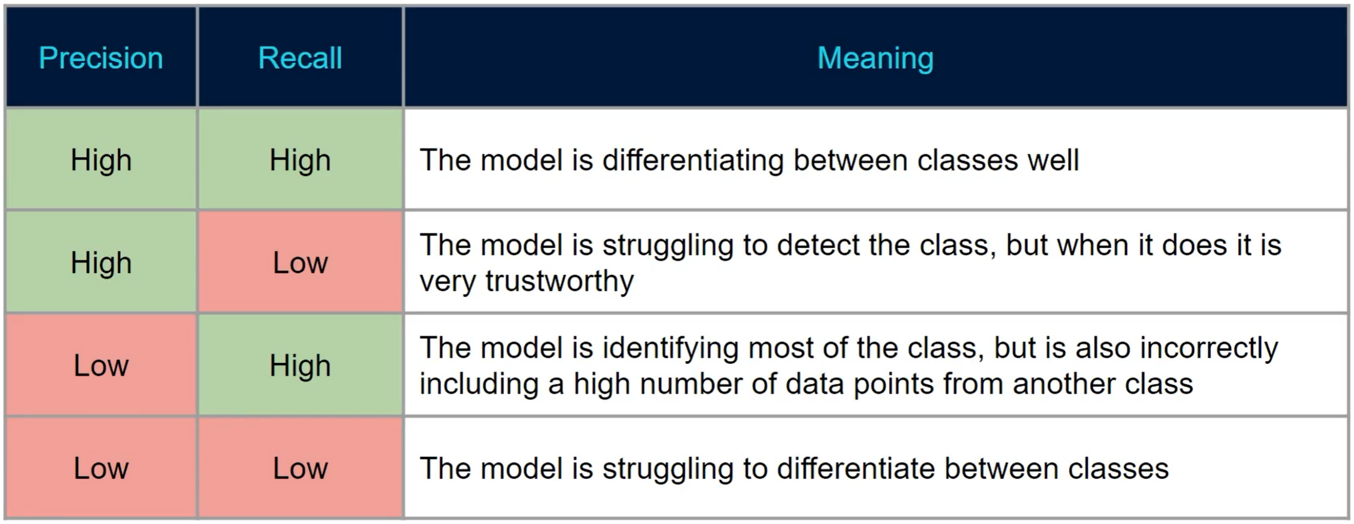


		Predicted Class	
		Pass	Fail
Actual Class	Pass	True Positive	False Negative
	Fail	False Positive	True Negative

Advanced Evaluation Techniques

- When there is a large bias towards one of the classes we have an imbalanced data set
- Advanced techniques help evaluate models when we have imbalanced data
  - Precision evaluates how many observations were predicted as positive who were actually positive

  - Recall (Sensitivity) evaluates how many observations were predicted as positive who were actually positive (also referred to as the True Positive Rate)
  - False Positive Rate evaluates how many observations were predicted as positive who were actually negative
  - F1-Score evaluates the harmonic mean of Precision and Recall
    - A good F1-Score comes when there is a balance between Precision & Recall, rather than a disparity between them
- Precision & Recall can not be optimized together, sometimes it makes sense to adapt a model to optimize one of these metrics
  - As an example, in a disease diagnoses model this would evaluate observations that were not predicted to have a disease who actually have the disease
  - In this example, it may make sense to optimize Recall while still being cognizant that we don't want to misdiagnose people as positive when they are in fact negative



Precision	Recall	Meaning
High	High	The model is differentiating between classes well
High	Low	The model is struggling to detect the class, but when it does it is very trustworthy
Low	High	The model is identifying most of the class, but is also incorrectly including a high number of data points from another class
Low	Low	The model is struggling to differentiate between classes

Advanced Evaluation Metrics

- $Precision = \frac{True\ Positive}{True\ Positive+False\ Positive}$

- $True\ Positive\ Rate\ (Recall\ Sensitivity) = \frac{True\ Positive}{True\ Positive+False\ Negative}$

- $False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive+TrueNegative}$

- $F1\ Score\ (Harmonic\ Mean) = \frac{2 * (Recall * Precision)}{Recall + Precision}$