# Decision Trees and Random Forests

### **Introduction**

I am sure, Steve Jobs, Sundar Pichai, Jeff Bezos and many CEOs constantly need to make decisions to run their big companies.

In fact, all successful people rely on making decision based on the data or information they have

In fact Jeff Bezos on his recent interview states that

**“The great thing about fact-based decisions is that they overrule the hierarchy."**

That is exactly how the concept of Decision Tree and Random Forest works. Generally Machine makes the decision based on the information which it have.

For example:

I am sure you all have played Hangman Game

(The Hangman game is a word-guessing game where the player tries to guess a hidden word by suggesting letters.)

So with every alphabet you predict, you narrow down the correct word from the huge list of word, you generally make the decision to predict the word based on the information you have, That's exactly how the concept of Decision Tree works.

Later I will show you how the concept of decision tree helps to predict the hangman game word accurately.

In short this article will explore the basics of decision trees and the ensemble method, random forests, for improved predictive performance.

### **Decoding Decision Trees**

Decision Tree is machine learning algorithm which is used for regression as well as classification problems.

Lets start with example which illustrate on how Decision Tree works:

Suppose, you are planning a trip for you, Now you have to make Decision on where exactly you would be traveling. So you open makemytrip app and filter your budget, and based on that you chose Ladhakh. And booked the tickets seemlessly.

But here is the trick, Ladakh is closed in winter season because of weather, so it might not be safe for you to travel up there. Therefore based on your decision, you did chose to travel to adverse environment for basic enjoyment.

But lets say, you come up with different strategy for your decision making

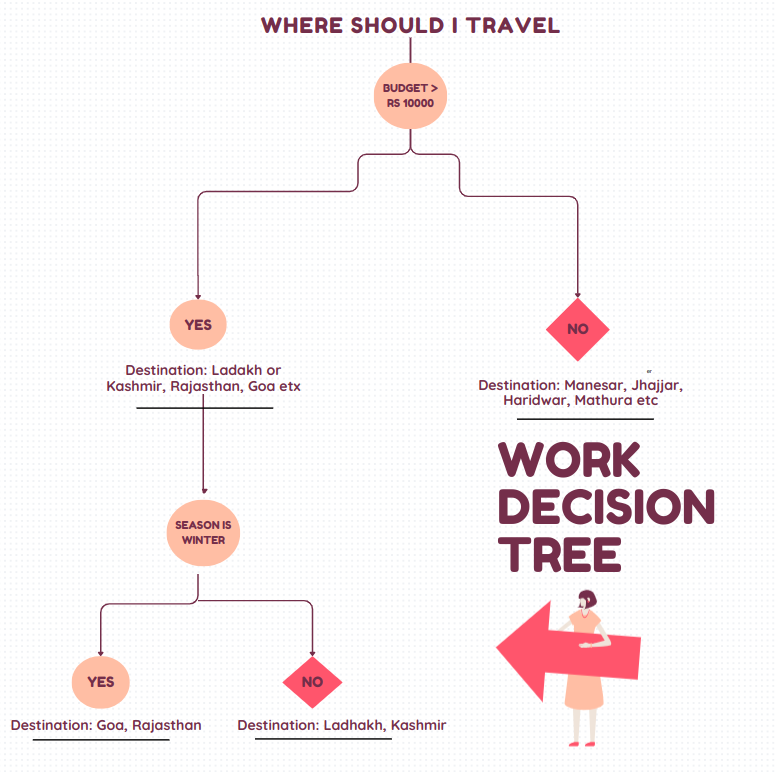
Now you again goes to makemytrip app and again use the budget filter, Now you get many destinations, and suppose you again filter these destination with another filter, which just filter out the places which are available for winter season.

Now based on the second filter, you will be alot more comfortable with your decision making, as you added an extra feature of season along with budget.

Even if this method took more time than the previous one, you will benefit more by relying on this process rather than the previous one

This process of decision making is exactly how concept of decision tree works. So the decision making is primarily based on sequential question and answers and based on the answer or Decision it goes down to certain route, hence named Decision Tree.

That why decision tree is also known as nested if-else classification model



#### **Advantages to using Decision Trees**:

* Simple and least complicated concept
* More Accurate than other complex algorithm
* Because of its simplicity, it is extremely fast

#### **Disadvantages to using Decision Trees**:

* Prone to Overfitting
* Can become highly complex

#### **Why did Decision Trees checks Budget first and not the Season?**

Thats a pretty valid question, Thats where the concept of feature importance emerges. So these features are decided on the basic criteria like Gini Impurity, or Information Gain or Entropy.

By using these criteria, we will be able to to judge the features based on their importance

### **Random Forests - The Ensemble Advantage -**

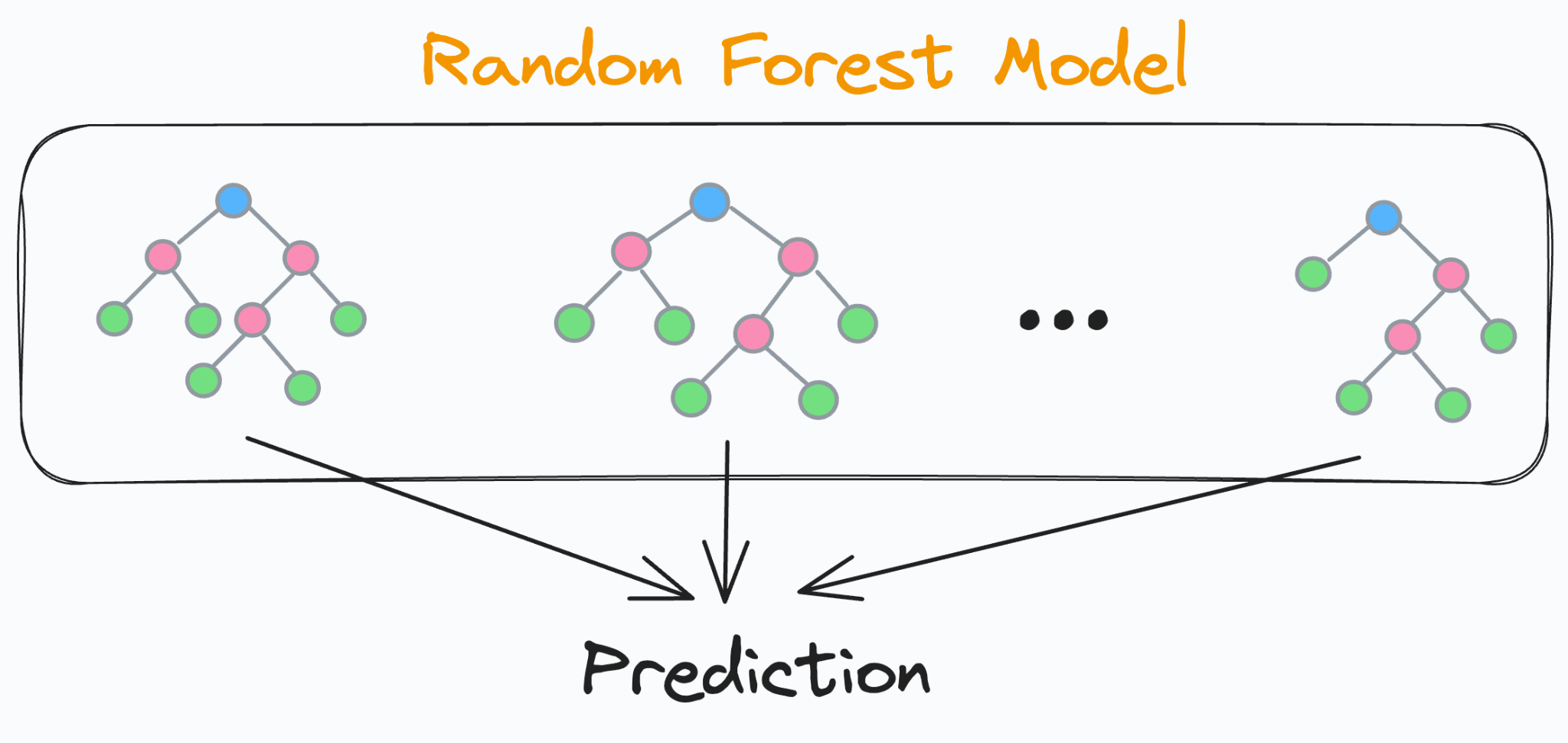
One of the disadvantages of decision tree is that it sometimes becomes so dense and highly complex, which then lead to overfitting.

So to minimize and variance, hence to reduce overfitting, Data Scientist generally use RandomForest.

Random Forest is basically a collection of numerous decision trees. Random forest are collection of decision trees where outcomes are aggregated into one final result.

Hence thats why it is called ‘Forest’ which is a collection of tree

And the word “Random” in “Random Forest” comes because decision trees are randomly created because it is trained on random subset of data



#### **Advantages to using Random Forest**:

* Reduces Overfitting
* It works effectively on large datasets

#### **Disadvantages to using Random Forest**:

* Takes longer time than decision tree on training the model

### **How do Entropy Works?**

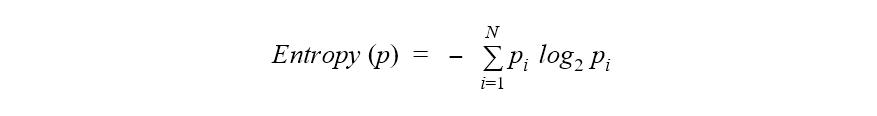
Remember, when I introduce you about the concept of feature importance. Like How do model got to know which feature to pick first and which feature to pick last.

We use criteria such a Gini Impurity, Entrophy based, or Information gain.

The most common concept between all is Entropy

Before solving the problem statement of Decision Tree, the model tends to calculate entropy firstly. Entropy is basically the measure of impurity or disorder in the dataset

Formula of Entropy



h(X) = – [(Pi \* log2 Pi) + (Qi \* log2 Qi)]

Here Pi = Probability of Y = 1 i.e. probability of success of the event

Qi = Probability of Y = 0 i.e. probability of failure of the event

### **Conclusion:**