With increase in computational power, we can now choose algorithms which perform very intensive calculations. One such algorithm is Random Forest, which we will discuss in this article. While the algorithm is very popular in various competitions (e.g. like the ones running on Kaggle), the end output of the model is like a black box and hence should be used judiciously.

Before going any further, here is an example on the importance of choosing the best algorithm.

Importance of choosing the right algorithm

Yesterday, I saw a movie called Edge of tomorrow. I loved the concept and the thought process which went behind the plot of this movie. Let me summarize the plot (without commenting on the climax, of course). Unlike other sci-fi movies, this movie revolves around one single power which is given to both the sides (hero and villain). The power being the ability to reset the day.

Human race is at war with an alien species called Mimics. Mimic is described as a far more evolved civilization of an alien species. Entire Mimic civilization is like a single complete organism. It has a central brain called Omega which commands all other organisms in the civilization. It stays in contact with all other species of the civilization every single second. Alpha is the main warrior species (like the nervous system) of this civilization and takes command from Omega. Omega has the power to reset the day at any point of time.

Now, lets wear the hat of a predictive analyst to analyze this plot. If a system has the ability to reset the day at any point of time, it will use this power, whenever any of its warrior species die. And, hence there will be no single war ,when any of the warrior species (alpha) will actually die, and the brain Omega will repeatedly test the best case scenario to maximize the death of human race and put a constraint on number of deaths of alpha (warrior species) to be zero every single day. You can imagine this as THE BEST predictive algorithm ever made. It is literally impossible to defeat such an algorithm.

Lets now get back to Random Forests using a case study.

Case Study

Following is a distribution of Annual income Gini Coefficients across different countries :

Mexico has the second highest Gini coefficient and hence has a very high segregation in annual income of rich and poor. Our task is to come up with an accurate predictive algorithm to estimate annual income bracket of each individual in Mexico. The brackets of income are as follows :

1. Below $40,000
2. $40,000 150,000
3. More than $150,000

Following are the information available for each individual :

1. Age , 2. Gender, 3. Highest educational qualification, 4. Working in Industry, 5. Residence in Metro/Non-metro

We need to come up with an algorithm to give an accurate prediction for an individual who has following traits:

1. Age : 35 years , 2, Gender : Male , 3. Highest Educational Qualification : Diploma holder, 4. Industry : Manufacturing, 5. Residence : Metro

We will only talk about random forest to make this prediction in this article.

The algorithm of Random Forest

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART model with different sample and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

Back to Case study Disclaimer : The numbers in this article are illustrative

Mexico has a population of 118 MM. Say, the algorithm Random forest picks up 10k observation with only one variable (for simplicity) to build each CART model. In total, we are looking at 5 CART model being built with different variables. In a real life problem, you will have more number of population sample and different combinations of input variables.

Salary bands :

Band 1 : Below $40,000

Band 2: $40,000 150,000

Band 3: More than $150,000

Following are the outputs of the 5 different CART model.

CART 1 : Variable Age

rf1

CART 2 : Variable Gender

rf2

CART 3 : Variable Education

rf3

CART 4 : Variable Residence

rf4

CART 5 : Variable Industry

rf5

Using these 5 CART models, we need to come up with singe set of probability to belong to each of the salary classes. For simplicity, we will just take a mean of probabilities in this case study. Other than simple mean, we also consider vote method to come up with the final prediction. To come up with the final prediction lets locate the following profile in each CART model :

1. Age : 35 years , 2, Gender : Male , 3. Highest Educational Qualification : Diploma holder, 4. Industry : Manufacturing, 5. Residence : Metro

For each of these CART model, following is the distribution across salary bands :

DF

The final probability is simply the average of the probability in the same salary bands in different CART models. As you can see from this analysis, that there is 70% chance of this individual falling in class 1 (less than $40,000) and around 24% chance of the individual falling in class 2.

End Notes

Random forest gives much more accurate predictions when compared to simple CART/CHAID or regression models in many scenarios. These cases generally have high number of predictive variables and huge sample size. This is because it captures the variance of several input variables at the same time and enables high number of observations to participate in the prediction. In some of the coming articles, we will talk more about the algorithm in more detail and talk about how to build a simple random forest on R.

What is Random forest algorithm?

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

If you know the decision tree algorithm. You might be thinking are we creating more number of decision trees and how can we create more number of decision trees. As all the calculation of nodes selection will be same for the same dataset.

Yes. You are true. To model more number of decision trees to create the forest you are not going to use the same apache of constructing the decision with information gain or gini index approach.

If you are not aware of the concepts of decision tree classifier, Please spend some time on the below articles, As you need to know how the decision tree classifier works before you learning the working nature of the random forest algorithm. If you would like to learn the implementation of the decision tree classifier, you can chek it out from the below articles.

How the decision tree works

Implementing the decision tree classifier in Python

Building decision tree classifier in R programming language

How to visualize the modeled decision tree classifier

If you are new to the concept of decision tree. I am giving you a basic overview of the decision tree.

Basic decision tree concept

Decision tree concept is more to the rule based system. Given the training dataset with targets and features, the decision tree algorithm will come up with some set of rules. The same set rules can be used to perform the prediction on the test dataset.

Suppose you would like to predict that your daughter will like the newly released animation movie or not. To model the decision tree you will use the training dataset like the animated cartoon characters your daughter liked in the past movies.

So once you pass the dataset with the target as your daughter will like the movie or not to the decision tree classifier. The decision tree will start building the rules with the characters your daughter like as nodes and the targets like or not as the leaf nodes. By considering the path from the root node to the leaf node. You can get the rules.

The simple rule could be if some x character is playing the leading role then your daughter will like the movie. You can think few more rule based on this example.

Then to predict whether your daughter will like the movie or not. You just need to check the rules which are created by the decision tree to predict whether your daughter will like the newly released movie or not.

In decision tree algorithm calculating these nodes and forming the rules will happen using the information gain and gini index calculations.

In random forest algorithm, Instead of using information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly. Will look about in detail in the coming section.

Next, you are going to learn why random forest algorithm? When we are having other classification algorithms to play with.

Why Random forest algorithm

To address why random forest algorithm. I am giving you the below advantages.

The same random forest algorithm or the random forest classifier can use for both classification and the regression task. Random forest classifier will handle the missing values. When we have more trees in the forest, random forest classifier wont overfit the model. Can model the random forest classifier for categorical values also. Will discuss these advantage in the random forest algorithm advantages section of this article. Until think through the above advantages of random forest algorithm compared to the other classification algorithms.

Random forest algorithm real life example Random Forest Example Random Forest Example

Before you drive into the technical details about the random forest algorithm. Lets look into a real life example to understand the layman type of random forest algorithm.

Suppose Mady somehow got 2 weeks leave from his office. He wants to spend his 2 weeks by traveling to the different place. He also wants to go to the place he may like.

So he decided to ask his best friend about the places he may like. Then his friend started asking about his past trips. Its just like his best friend will ask, You have been visited the X place did you like it?

Based on the answers which are given by Mady, his best start recommending the place Mady may like. Here his best formed the decision tree with the answer given by Mady.

As his best friend may recommend his best place to Mady as a friend. The model will be biased with the closeness of their friendship. So he decided to ask few more friends to recommend the best place he may like.

Now his friends asked some random questions and each one recommended one place to Mady. Now Mady considered the place which is high votes from his friends as the final place to visit.

In the above Mady trip planning, two main interesting algorithms decision tree algorithm and random forest algorithm used. I hope you find it already. Anyhow, I would like to highlight it again.

Decision Tree: To recommend the best place to Mady, his best friend asked some questions. Based on the answers given by mady, he recommended a place. This is decision tree algorithm approach. Will explain why it is a decision tree algorithm approach.

Mady friend used the answers given by mady to create rules. Later he used the created rules to recommend the best place which mady will like. These rules could be, mady like a place with lots of tree or waterfalls ..etc

In the above approach mady best friend is the decision tree. The vote (recommended place) is the leaf of the decision tree (Target class). The target is finalized by a single person, In a technical way of saying, using an only single decision tree.

Random Forest Algorithm:

In the other case when mady asked his friends to recommend the best place to visit. Each friend asked him different questions and come up their recommend a place to visit. Later mady consider all the recommendations and calculated the votes. Votes basically is to pick the popular place from the recommend places from all his friends.

Mady will consider each recommended place and if the same place recommended by some other place he will increase the count. At the end the high count place where mady will go.

In this case, the recommended place (Target Prediction) is considered by many friends. Each friend is the tree and the combined all friends will form the forest. This forest is the random forest. As each friend asked random questions to recommend the best place visit.

Now lets use the above example to understand how the random forest algorithm work.

How Random forest algorithm works

How random forest algorithm works How random forest algorithm works

Lets look at the pseudocode for random forest algorithm and later we can walk through each step in the random forest algorithm.

The pseudocode for random forest algorithm can split into two stages.

Random forest creation pseudocode. Pseudocode to perform prediction from the created random forest classifier. First, lets begin with random forest creation pseudocode

Random Forest pseudocode: Randomly select k features from total m features.

Where k ¡¡ m Among the k features, calculate the node d using the best split point.

Split the node into daughter nodes using the best split.

Repeat 1 to 3 steps until l number of nodes has been reached.

Build forest by repeating steps 1 to 4 for n number times to create n number of trees.

The beginning of random forest algorithm starts with randomly selecting k features out of total m features. In the image, you can observe that we are randomly taking features and observations.

In the next stage, we are using the randomly selected k features to find the root node by using the best split approach.

The next stage, We will be calculating the daughter nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node.

Finally, we repeat 1 to 4 stages to create n randomly created trees. This randomly created trees forms the random forest.

Random forest prediction pseudocode: To perform prediction using the trained random forest algorithm uses the below pseudocode.

Takes the test features and use the rules of each randomly created decision tree to predict the oucome and stores the predicted outcome (target)

Calculate the votes for each predicted target.

Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm we need to pass the test features through the rules of each randomly created trees. Suppose lets say we formed 100 random decision trees to from the random forest.

Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Suppose the 100 random decision trees are prediction some 3 unique targets x, y, z then the votes of x is nothing but out of 100 random decision tree how many trees prediction is x.

Likewise for other 2 targets (y, z). If x is getting high votes. Lets say out of 100 random decision tree 60 trees are predicting the target will be x. Then the final random forest returns the x as the predicted target.

This concept of voting is known as majority voting.

Now lets look into few applications of random forest algorithm.

Random forest algorithm applications

Random Forest Applications

Random Forest Applications

The random algorithm used in wide varieties applications. In this article, we are going address few of them.

Below are some the application where random forest algorithm is widely used.

Banking Medicine Stock Market E-commerce Lets begin with the banking sector.

1.Banking: In the banking sector, random forest algorithm widely used in two main application. These are for finding the loyal customer and finding the fraud customers.

The loyal customer means not the customer who pays well, but also the customer whom can take the huge amount as loan and pays the loan interest properly to the bank. As the growth of the bank purely depends on the loyal customers. The bank customers data highly analyzed to find the pattern for the loyal customer based the customer details.

In the same way, there is need to identify the customer who are not profitable for the bank, like taking the loan and paying the loan interest properly or find the outlier customers. If the bank can identify theses kind of customer before giving the loan the customer. Bank will get a chance to not approve the loan to these kinds of customers. In this case, also random forest algorithm is used to identify the customers who are not profitable for the bank.

2.Medicine In medicine field, random forest algorithm is used identify the correct combination of the components to validate the medicine. Random forest algorithm also helpful for identifying the disease by analyzing the patients medical records.

3.Stock Market In the stock market, random forest algorithm used to identify the stock behavior as well as the expected loss or profit by purchasing the particular stock.

4.E-commerce In e-commerce, the random forest used only in the small segment of the recommendation engine for identifying the likely hood of customer liking the recommend products base on the similar kinds of customers.

Running random forest algorithm on very large dataset requires high-end GPU systems. If you are not having any GPU system. You can always run the machine learning models in cloud hosted desktop. You can use clouddesktoponline platform to run high-end machine learning models from sitting any corner of the world.

Advantages of random forest algorithm Below are the advantages of random forest algorithm compared with other classification algorithms.

The overfitting problem will never come when we use the random forest algorithm in any classification problem. The same random forest algorithm can be used for both classification and regression task. The random forest algorithm can be used for feature engineering. Which means identifying the most important features out of the available features from the training dataset.

Advantages of Random Forest algorithm

Compared with other classification techniques, there are three advantages as the author mentioned.

For applications in classification problems, Random Forest algorithm will avoid the overfitting problem For both classification and regression task, the same random forest algorithm can be used The Random Forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering.