Introduction to Random forest - Simplified

BIG DATA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BIG-DATA/) BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/)

(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/alienvshuman.jpg

With increase in computational power, we can now choose algorithms which pe intensive calculations. One such algorithm is "Random Forest", which we will discuss in While the algorithm is very popular in various competitions (e.g. like the ones running on K 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 algori

Importance of choosing the right algorithm

Yesterday, I saw a movie called " ***Edge of tomorrow***". I loved the concept and the thoug which went behind the plot of this movie. Let me summarize the plot (without commen climax, of course). Unlike other sci-fi movies, this movie revolves around one single pow 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 mo civilization of an alien species. Entire Mimic civilization is like a single complete organis central brain called "Omega" which commands all other organisms in the civilization. contact with all other species of the civilization every single second. "Alpha" is the m species (like the nervous system) of this civilization and takes command from "Omega". "O the power to reset the day at any point of time.

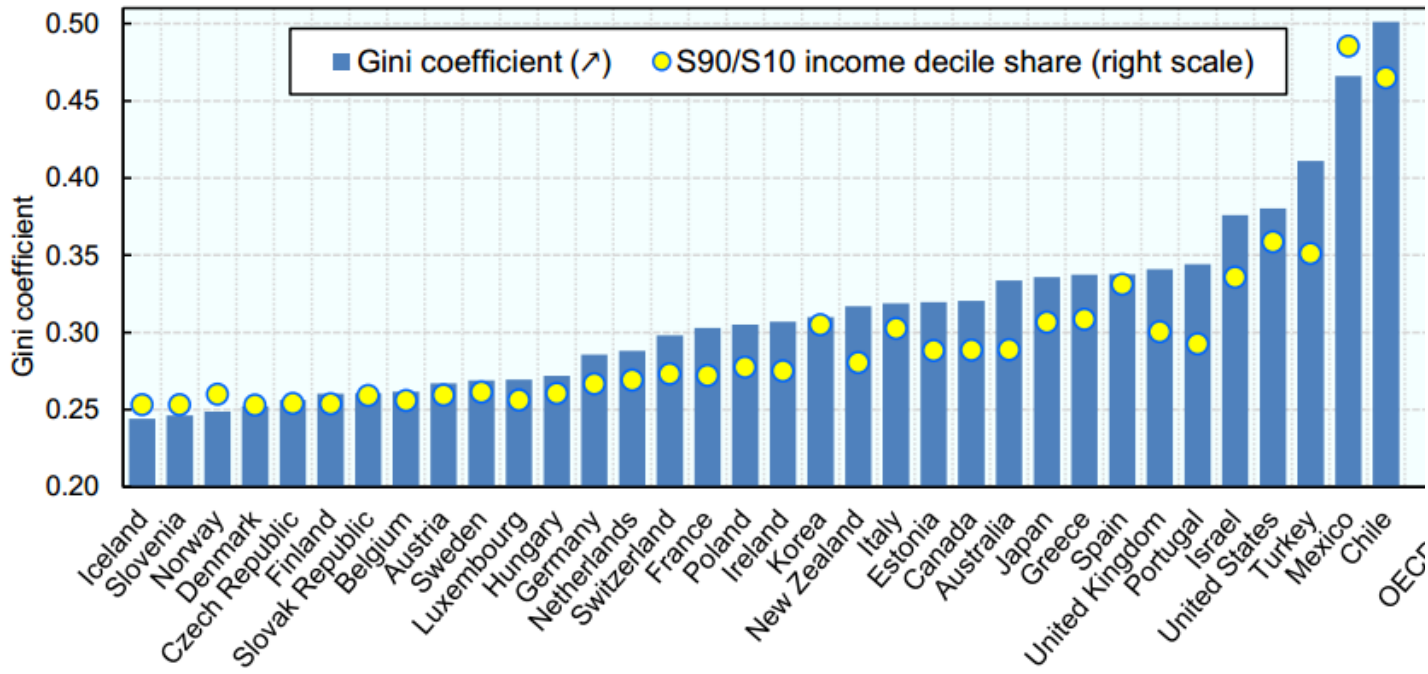
Now, let's wear the hat of a predictive analyst to analyze this plot. If a system has the abi the day at any point of time, it will use this power, whenever any of its warrior species die. A there will be no single war ,when any of the warrior species (alpha) will actually die, an "Omega" will repeatedly test the best case scenario to maximize the death of human race

constraint on number of deaths of alpha (warrior species) to be zero every single day. You c this as "THE BEST" predictive algorithm ever made. It is literally impossible to defe algorithm.

Let's now get back to "Random Forests" using a case study.

Case Study

Following is a distribution of Annual income Gini (http://en.wikipedia.org/wiki/Gini_ Coefficients across different countries :



(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/oecd-income_inequality_2013_2.png)

Mexico has the second highest Gini coefficient and hence has a very high segregatior income of rich and poor. Our task is to come up with an accurate predictive algorithm t 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. Re
Metro/Non-metro

We need to come up with an algorithm to give an accurate prediction for an individu
following traits:

1. Age : 35 years , 2, Gender : Male , 3. Highest Educational Qualification : Diploma holder, 4
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
observation in the complete population with 10 variables. Random forest tries to build mu
model with different sample and different initial variables. For instance, it will take a random
100 observation and 5 randomly chosen initial variables to build a CART model. It will
process (say) 10 times and then make a final prediction on each observation. Final pre
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 obser
only one variable (for simplicity) to build each CART model. In total, we are looking at 5 C
being built with different variables. In a real life problem, you will have more number of
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

	Salary Band	1	2	3
Age	Below 18	90%	10%	0%
	19-27	85%	14%	1%
	28-40	70%	23%	7%
	40-55	60%	35%	5%
	More than 55	70%	25%	5%

(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/rf1.png)

CART 2 : Variable Gender

	Salary Band	1	2	3
Gender	Male	70%	27%	3%
	Female	75%	24%	1%

(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/rf2.png)

CART 3 : Variable Education

	Salary Band	1	2	3
Education	<=High School	85%	10%	5%
	Diploma	80%	14%	6%
	Bachelors	77%	23%	0%
	Master	62%	35%	3%

(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/rf3.png)

CART 4 : Variable Residence

	Salary Band	1	2	3
Residence	Metro	70%	20%	10%
	Non-Metro	65%	20%	15%

(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/rf4.png)

CART 5 : Variable Industry

	Salary Band	1	2	3
Industry	Finance	65%	30%	5%
	Manufacturing	60%	35%	5%
	Others	75%	20%	5%

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

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

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

CART	Band	1	2	3
Age	28-40	70%	23%	7%
Gender	Male	70%	27%	3%
Education	Diploma	80%	14%	6%
Industry	Manufacturing	60%	35%	5%
Residence	Metro	70%	20%	10%
Final probability		70%	24%	6%

(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/DF.png)

The final probability is simply the average of the probability in the same salary bands in diff models. As you can see from this analysis, that there is 70% chance of this individual fallin (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 regression models in many scenarios. These cases generally have high number of predictiv and huge sample size. This is because it captures the variance of several input variables a time and enables high number of observations to participate in the prediction. In some of articles, we will talk more about the algorithm in more detail and talk about how to buil random forest on R.

If you like what you just read & want to continue your learning, subscribe to our (http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya), on twitter (http://twitter.com/analyticsvidhya) or like our facebo (http://facebook.com/analyticsvidhya).

Share this:

(https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/?share=linkedin&nb=1) 102

(https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/?share=facebook&nb=1) 3

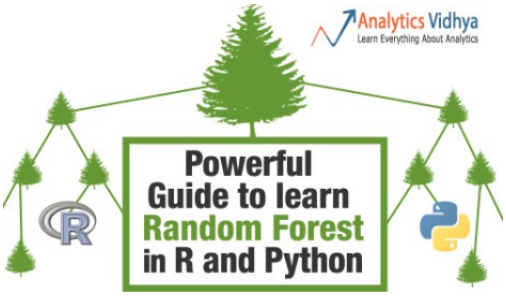
(https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/?share=google-plus-1&nb=1)

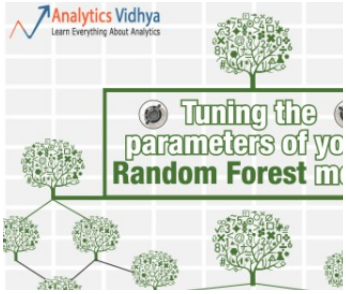
(https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/?share=twitter&nb=1)

(https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/?share=pocket&nb=1)

(https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/?share=reddit&nb=1)

RELATED





(https://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-challenges/)
Powerful Guide to learn Random Forest (with codes in R & Python) (/blog/2015/09/random-forest-algorithm-multiple-challenges/?relatedposts_hit=1&relatedposts_origin=3981&relatedposts_position=0)
September 7, 2015
In "Business Analytics"

(https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/)
How to perform feature selection (i.e. pick important variables) using Boruta Package in R ? (https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/)
March 22, 2016
In "R"

(https://www.analyticsvi
blog/2015/06/tuning-ran
model/)
Tuning the parameters of
Random Forest model
(https://www.analyticsvi
blog/2015/06/tuning-ran
model/)
June 9, 2015
In "Business Analytics"

TAGS: CART (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CART/), CHAID (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CHAID/), DEC
(HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DECISION-TREE/), MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/M/
LEARNING/), R (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/R/), RANDOM FOREST (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RAN

Previous Article
Senior Analyst - VMware, Bangalore (5+ years of experience) (https://www.analyticsvidhya.com/blog/2014/06/senior-analyst-vmware-bangalore-5-years-experience/)

Next Articl
Analyst - Wells Fargo, Chennai (2+ years o experience (https://www.analyticsvidhya.com/blog/2014 wells-fargo-chennai-2-years-experience/


(https://www.analyticsvidhya.com/blog/author/tavish1/)

Author

Tavish Srivastava (https://www.analyticsvidhya.com/blog/author/tavish

I am Tavish Srivastava, a post graduate from IIT Madras in Mechanical Engin
have more than two years of work experience in Analytics. My experience ra

from hands on analytics in a developing country like India to convince banki partners with analytical solution in matured market like US. For last two and years I have contributed to various sales strategies, marketing strategies an Recruitment strategies in both Insurance and Banking industry.

12 COMMENTS

Suman says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
JUNE 10, 2014 AT 6:55 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/#COMMENT-10249)

Enlightening

Anup says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
JUNE 10, 2014 AT 7:05 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/#COMMENT-10250)

Very nice write up. My two cents. A fundamental problem with most empirical research correlated variables? How can analysts be sure that most (if not all) variables that expla process are included? Also what happens if we choose subsets of variables and build p models? Wonder if we can use machine learning tools to deal with this. LASSO models variable selection may be better than Random Forests (maybe) guess but again it is stil the assumption that analysts can provide the entire universe of predictors (which is deb especially in the social sciences. Nevertheless, nice post and look forward to more here

Pankaj says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
JUNE 10, 2014 AT 8:19 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/#COMMENT-10255)

awesome !!! Thanks Man !!!

pradeep says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
JUNE 10, 2014 AT 10:04 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-

SIMPLIFIED/#COMMENT-10263)

This is great. Just one Question why we are looking at only this group

1. Age : 35 years , 2, Gender : Male , 3. Highest Educational Qualification : Diploma holde
: Manufacturing, 5. Residence : Metro ?

Tavish Srivastava says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
JUNE 10, 2014 AT 11:28 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RAN
SIMPLIFIED/#COMMENT-10269)

Pradeep,

That is just a test case on which we are scoring our algorithm.

Tavish

Navdeep says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM
JUNE 21, 2014 AT 5:47 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-
SIMPLIFIED/#COMMENT-11576)

Simple and detailed explanation…really appreciate you work.

Sivakunmar says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
JULY 2, 2014 AT 4:57 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-
SIMPLIFIED/#COMMENT-13030)

Thank u giving such appropriate experience i understand concept of random forest tec

---

olivier says:
REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
OCTOBER 11, 2014 AT 3:22 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FORE
SIMPLIFIED/#COMMENT-28408)

ty.

some include code snippets in R and others don't.

will be very useful to see implementation

ty
O

VJ@DataScience says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
SEPTEMBER 8, 2015 AT 3:33 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOR
SIMPLIFIED/#COMMENT-94494)

Hi Tavish,

really appreciate this and easy to understand the concept of Random Forest.

Question to you:-
In CART model, when we get multiple predictors in a particular model – solution can be
implemented in actual business scenario (e.g. if customer falls in so and so age group &
products in the past and so on…. then probability is 60%)
but in the case of above example, averaging out probabilities of multiple predictors fro
models would leave it as black box – please provide some thoughts around implemen

---

Abdul says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
OCTOBER 20, 2016 AT 1:22 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FORE
SIMPLIFIED/#COMMENT-117341)

How to assign weight to each variable while making random forests and its impact on r

---

Jack Ma says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTION-RANDOM-FOREST-SIMPLIFIED/?REPLYTOCOM=
NOVEMBER 14, 2016 AT 9:44 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/INTRODUCTIO
FOREST-SIMPLIFIED/#COMMENT-118367)

I have the same question. Looking forward to hearing any ideas.

LEAVE A REPLY

Connect with:

(https://www.analyticsvidhya.com/wp-login.php?action=wordpress_social_authenticate&mode=login&provider=Facebook&redirect_to=https%3A%random-forest-simplified%2F)

Your email address will not be published.

Comment

Name (required)

Email (required)

Website

SUBM

ABOUT US

For those of you, who are wondering what is "Analytics Vidhya", "Analytics" can be defined as the science of extracting insights from raw data. The spectrum of analytics starts from capturing data

LATEST POSTS

(https://www.analyticsvidhya.com/blog.to-structuring-customer-complaints/)

and evolves into using insights / trends from this data to make informed decisions. Read More (http://www.analyticsvidhya.com/about-me/)

STAY CONNECTED

8,028
FOLLOWERS
(http://www.twitter.com/analyticsvidhya)

26,148
FOLLOWERS
(http://www.facebook.com/Analyticsvidhya)

1,579
FOLLOWERS
(https://plus.google.com/+Analyticsvidhya)

Email
SUBSCRIBE
(http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya)

Introduction to Structuring (complaints explained with e: (https://www.analyticsvidhy 2017/01/introduction-to-stru customer-complaints/)

YOGESH KULKARNI , JANUARY 27, 20

(https://www.analyticsvidhya.com/blog.

21 Steps to Get Started with Spark using Scala (https://www.analyticsvidhy 2017/01/scala/)

ANKIT GUPTA , JANUARY 25, 2017

(https://www.analyticsvidhya.com/blog. sne-implementation-r-python/)

Comprehensive Guide on t-S with implementation in R & (https://www.analyticsvidhy 2017/01/t-sne-implementati

SAURABH.JAJU2 , JANUARY 22, 2017

(https://www.analyticsvidhya.com/blog. head-to-data-science-hacker/)

MyStory: How I became a Da Hacker from being a Delivery (https://www.analyticsvidhy 2017/01/delivery-head-to-dat hacker/)

GUEST BLOG , JANUARY 21, 2017