(https://www.facebook.com/AnalyticsVidhya) | (https://twitter.com/analyticsvidhya) | (https://plu
(https://www.linkedin.com/groups/Analytics-Vidhya-Learn-everything-about-5057165)
Home (https://www.analyticsvidhya.com/) | Blog (https://www.analyticsvidhya.com/blog/) | Jobs (https:
Trainings (https://www.analyticsvidhya.com/trainings/)
Learning Paths (https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-
DataHack (https://datahack.analyticsvidhya.com) |

Analytics Vidhya
Learn Everything About Analytics (https://www.analyticsvidhya.com)

HOME (HTTPS://WWW.ANALYTICSVIDHYA.COM/) BLOG (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/)
TRAININGS (HTTPS://WWW.ANALYTICSVIDHYA.COM/TRAININGS/) DISCUSS (HTTPS://DISCUSS.ANALYTI
LEARNING PATHS (HTTPS://WWW.ANALYTICSVIDHYA.COM/LEARNING-PATHS-DATA-SCIENCE-BUSINESS-ANAI
DATAHACK (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM) STORIES (HTTPS://WWW.ANALYTICSVIDHYA.C
WRITE FOR US (HTTP://WWW.ANALYTICSVIDHYA.COM/ABOUT-ME/WRITE/) CONTACT US (HTTPS://WWW

Tuning the parameters of your Random Forest m

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

SHARE (http://www.facebook.com/sharer.php?u=https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-

model/&t=Tuning%20the%20parameters%20of%20your%20Random%20Forest%20model) (https://twitter.com/home?status=Tuning%20the%20parameters%20of%20your%20Random%20Forest%20model+https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/) (https://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/) (http://pinterest.com/pin/create/button/?url=https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/&media=https://www.analyticsvidhya.com/wp-content/uploads/2015/06/Tuning-the-parameters-of-your-Random-Forest-model.jpg&description=Tuning%20the%20parameters%20of%20your%20Random%20Forest%20model)

Why to tune Machine Learning Algorithms?

A month back, I participated in a Kaggle c (https://www.analyticsvidhya.com/blog/2015/06/start-journey-kaggle/) called TFI. I start first submission at 50th percentile. Having worked relentlessly on feature engineering for r weeks, I managed to reach 20th percentile. To my surprise, right after tuning the parame machine learning algorithm I was using, I was able to breach top 10th percentile.

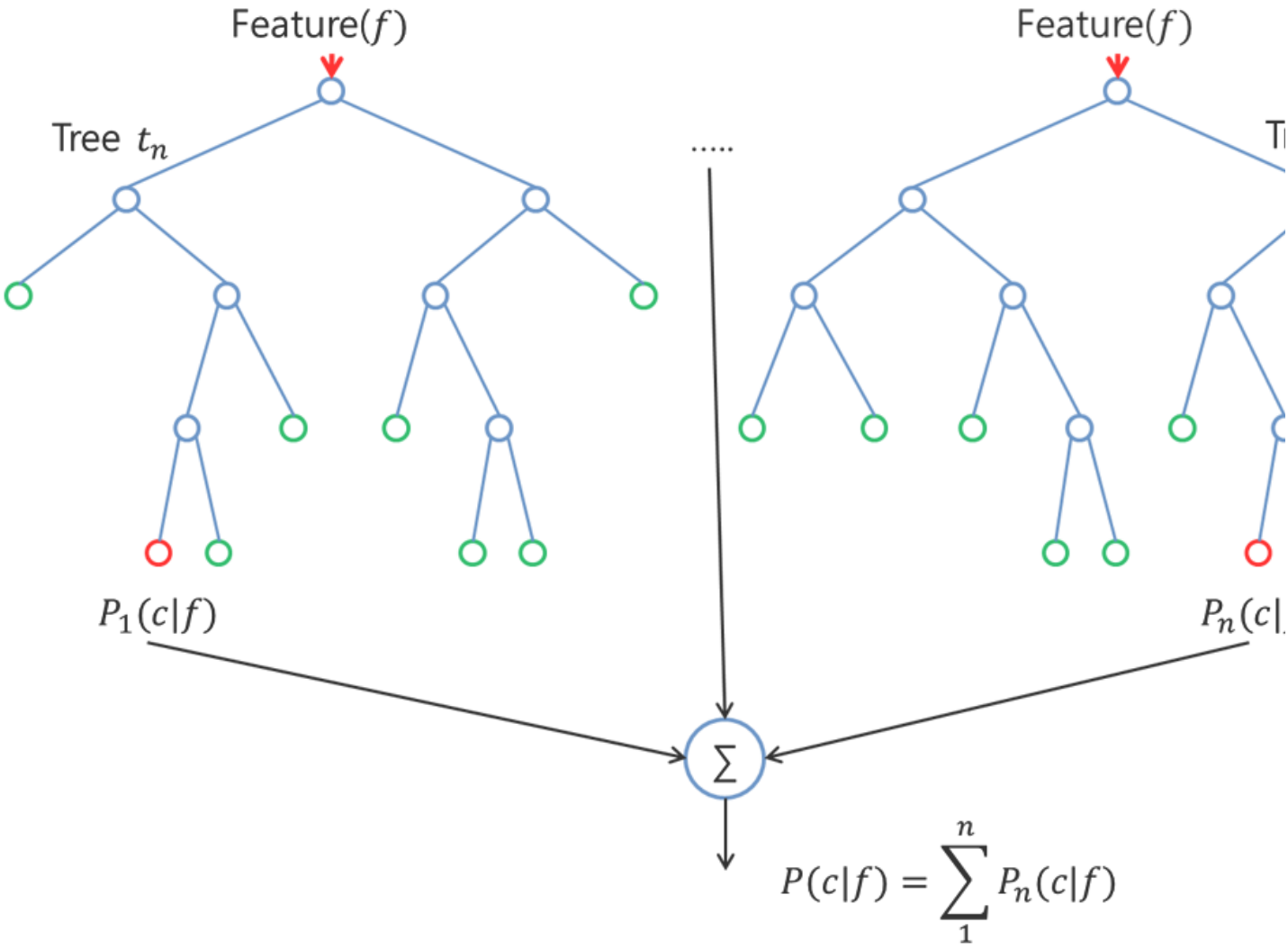
This is how important tuning these machine learning algorithms are. Random Forest is easiest machine learning tool used in the industry. In our previous articles, we have introdu Random Forest (https://www.analyticsvidhya.com/blog/2014/06/introduction-ranc simplified/)and compared it against a CART (https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/). Learning tools are known for their performance.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/06/Tuning-the-parameters Random-Forest-model.jpg)

What is a Random Forest?

Random forest is an ensemble tool which takes a subset of observations and a subset of v build a decision trees. It builds multiple such decision tree and amalgamate them togeth more accurate and stable prediction. This is direct consequence of the fact that by maxin from a panel of independent judges, we get the final prediction better than the best judge.

(https://www.analyticsvidhya.com/wp-content/uploads/2015/06/random-forest4.png)



(https://www.analyticsvidhya.com/wp-content/uploads/2015/06/random-forest7.png)

We generally see a random forest as a black box which takes in input and gives out p without worrying too much about what calculations are going on the back end. This blac have a few levers we can play with. Each of these levers have some effect on either the pe

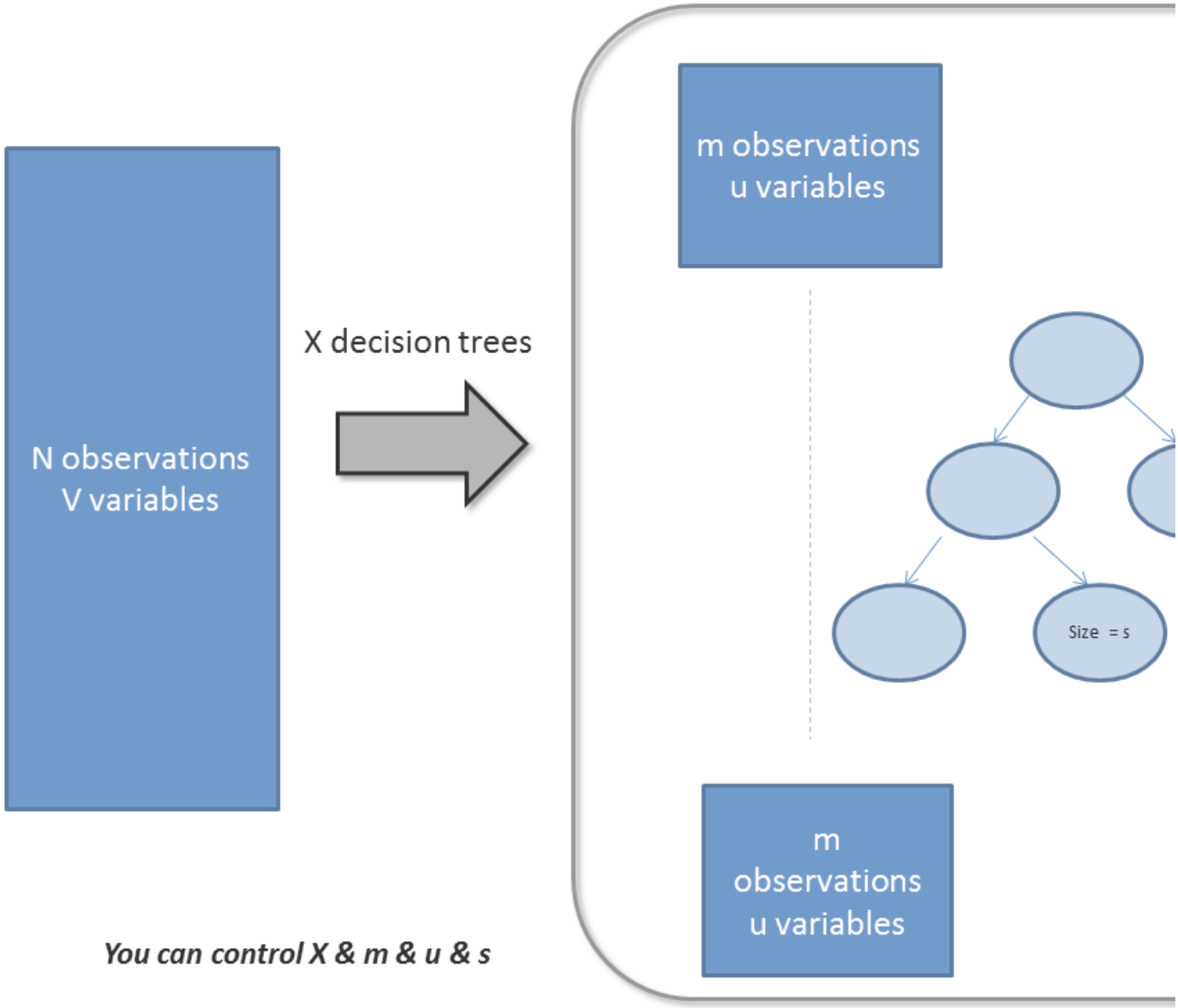
of the model or the resource – time balance. In this article we will talk more about these can tune, while building a random forest model.

Parameters / levers to tune Random Forests

Parameters in random forest are either to increase the predictive power of the model or easier to train the model. Following are the parameters we will be talking about in more d that I am using Python conventional nomenclatures for these parameters) :

(https://www.analyticsvidhya.com/wp-content/uploads/2015/06/knobs.jpg)

1. Features which make predictions of the model better



(https://www.analyticsvidhya.com/wp-content/uploads/2015/06/RF.png)

There are primarily 3 features which can be tuned to improve the predictive power of the m

1.a. max_features:

These are the maximum number of features Random Forest is allowed to try in individual are multiple options available in Python to assign maximum features. Here are a few of ther

1. *Auto/None* : This will simply take all the features which make sense in every tree.Here w not put any restrictions on the individual tree.
2. *sqrt* : This option will take square root of the total number of features in individual run. Fo the total number of variables are 100, we can only take 10 of them in individual tree."log2

similar type of option for max_features.

3. *0.2* : This option allows the random forest to take 20% of variables in individual run. We car value in a format "0.x" where we want x% of features to be considered.

How does "max_features" impact performance and speed?

Increasing max_features generally improves the performance of the model as at each no have a higher number of options to be considered. However, this is not necessarily t decreases the diversity of individual tree which is the USP of random forest. But, fo decrease the speed of algorithm by increasing the max_features. Hence, you need to stril balance and choose the optimal max_features.

1.b. n_estimators :

This is the number of trees you want to build before taking the maximum voting or a predictions. Higher number of trees give you better performance but makes your code s should choose as high value as your processor can handle because this makes your stronger and more stable.

1.c. min_sample_leaf :

If you have built a decision tree before, you can appreciate the importance of minimum s size. Leaf is the end node of a decision tree. A smaller leaf makes the model more prone t noise in train data. Generally I prefer a minimum leaf size of more than 50. However, you multiple leaf sizes to find the most optimum for your use case.

2. Features which will make the model training easier

There are a few attributes which have a direct impact on model training speed. Following parameters which you can tune for model speed :

2.a. n_jobs :

This parameter tells the engine how many processors is it allowed to use. A value of "-1" m is no restriction whereas a value of "1" means it can only use one processor. Here i experiment you can do with Python to check this metric :

```
%timeit
```

```
model = RandomForestRegressor(n_estimator = 100, oob_score = TRUE,n_jobs = 1,random_
```

```
model.fit(X,y)
```

Output ———- 1 loop best of 3 : 1.7 sec per loop

```
%timeit
```

```
model = RandomForestRegressor(n_estimator = 100,oob_score = TRUE,n_jobs = -1,random_
```

```
model.fit(X,y)
```

Output ———- 1 loop best of 3 : 1.1 sec per loop

"%timeit" is an awsum function which runs a function multiple times and gives the faste time. This comes out very handy while scalling up a particular function from prototype to fi

2.b. random_state :

This parameter makes a solution easy to replicate. A definite value of random_state produce same results if given with same parameters and training data. I have personall ensemble with multiple models of different random states and all optimum parameters performs better than individual random state.

2.c. oob_score :

This is a random forest cross validation method. It is very similar to leave one out validation however, this is so much faster. This method simply tags every observation used in differen then it finds out a maximum vote score for every observation based on only trees which this particular observation to train itself.

Here is a single example of using all these parameters in a single function :

```
model = RandomForestRegressor(n_estimator = 100, oob_score = TRUE, n_jobs = -1,rando
50,                                          max_features = "auto", min_samples_leaf
```

```
model.fit(X,y)
```

Learning through a case study

We have referred to Titanic case study in many of our previous articles. Let's try the sam again. The objective of this case here will be to get a feel of random forest parameter tuni getting the right features. Try following code to build a basic model :

```
from sklearn.ensemble import RandomForestRegressor
```

```
from sklearn.metrics import roc_auc_score
```

```
import pandas as pd
```

```
x = pd.read_csv("train.csv")
```

```
y = x.pop("Survived")
```

```
model =  RandomForestRegressor(n_estimator = 100 , oob_score = TRUE, random_state =
```

```
model.fit(x(numeric_variable,y)
```

```
print "AUC - ROC : ", roc_auc_score(y,model.oob_prediction)
```

AUC – ROC : 0.7386

This is a very simplistic model with no parameter tuning. Now let's do some parameter tu have discussed before, we have 6 key parameters to tune. We have some grid search alg in Python, which can tune all parameters automatically.But here let's get our han understand the mechanism better. Following code will help you tune the model for differer

Exercise : Try runing the following code and find the optimal leaf size in the comment bo

```
sample_leaf_options = [1,5,10,50,100,200,500]
```

```
for leaf_size in sample_leaf_options :
```

```
    model = RandomForestRegressor(n_estimator = 200, oob_score = TRUE, n_jobs = -1,
ate =50,                                    max_features = "auto", min_samples_
af_size)
```

```
    model.fit(x(numeric_variable,y)
```

```
    print "AUC - ROC : ", roc_auc_score(y,model.oob_prediction)
```

End Notes

Machine learning tools like random forest, SVM, neural networks etc. are all use performance. They do give high performance, but users generally don't understand how th work. Not knowing the statistical details of the model is not a concern however not knowi model can be tuned well to clone the training data restricts the user to use the algorithr potential. In some of the future articles we will take up tuning of other machine learning alg SVM , GBM and neaural networks.

Have you used random forest before? What parameters did you tune? How did tuning the impact the performance of the model? Did you see any significant benefits by doing the sa us know your thoughts about this guide in the comments section below.

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

Share this:

(https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/?share=linkedin&nb=1) 432

(https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/?share=facebook&nb=1) 174

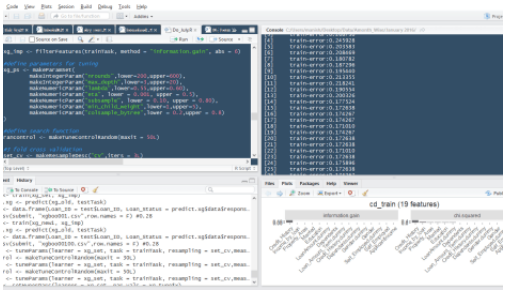
(https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/?share=google-plus-1&nb=1)

(https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/?share=twitter&nb=1)

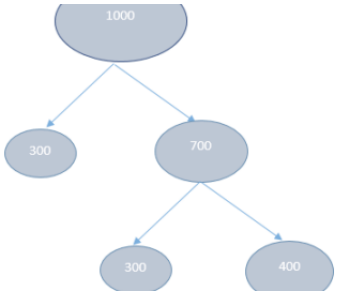
(https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/?share=pocket&nb=1)

(https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/?share=reddit&nb=1)

RELATED



(https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/)
Practicing Machine Learning Techniques in R with MLR Package (https://www.analyticsvidhya.com/blog/2016/08/practicing-machine-learning-techniques-in-r-with-mlr-package/)
August 8, 2016
In "Machine Learning"



(https://www.analyticsvidhya.com/blog/2016/12/detailed-solutions-for-skilltest-tree-based-algorithms/)
45 questions to test Data Scientists on Tree Based Algorithms (Decision tree, Random Forests, XGBoost) (https://www.analyticsvidhya.com/blog/2016/12/detailed-solutions-for-skilltest-tree-based-algorithms/)
December 5, 2016
In "Machine Learning"

(https://www.analyticsvi
blog/2016/05/h2o-data-t
models-large-data-sets/)
Use H2O and data.table t
models on large data set
(https://www.analyticsvi
blog/2016/05/h2o-data-t
models-large-data-sets/)
May 12, 2016
In "Machine Learning"

TAGS: MAX_FEATURES (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MAX_FEATURES/), MIN_SAMPLE_LEAF (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MIN_SAMPLE_LEAF/), N_ESTIMATOR (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/N_ES N_JOBS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/N_JOBS/), PARAMETER TUNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/P TUNING/), PYTHON (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/PYTHON/), RANDOM FOREST (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLO FOREST/)

Previous Article
Cheat Sheet for Exploratory Data Analysis in Python (https://www.analyticsvidhya.com/blog/2015/06/infograph cheat-sheet-data-exploration-python/)

Next Articl
In Conversation with Mr. Stefa Groschupf, Founder and CEO, Datamee (https://www.analyticsvidhya.com/blog/2015 mr-stefan-groschupf-ceo-datameer/

(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.

5 COMMENTS

Aayush Agrawal says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/?REPLYTOCOM=
JUNE 10, 2015 AT 6:50 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/#
88206)

Brilliantly written article. Currently I have used all of these techniques in a Data science was working on and it definitely helps in improving model performance and accuracy. I came across something else also when I was reading some articles on Random Forest, Regularization of Random Forest. The theme was to only split data with some variables splitting is significant enough using Statistical validation, now this is something which c taking Random Forest to next level, as It can help in reducing over-fitting. I tried to use i caret package but I think this technique is computationally expensive so couldn't run it system. I would love to see an article on it to understand it's working and how its perfor be improved.

KARTHI V says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/?REPLYTOCOM=
JUNE 10, 2015 AT 4:00 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/#
88254)

Hi Tavish,

Very useful article.

Ravi says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/?REPLYTOCOM=
SEPTEMBER 11, 2015 AT 6:43 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/#COMMENT-94729)

I love AV and am a fan of your articles. I have heard something like Conditional Inferenc which are similar to Random Forests. Can you share your thoughts on Conditional Infere also? How does it work & its tuning parameters, when does it outcast Random Forests?

Josh says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/?REPLYTOCOM=
SEPTEMBER 17, 2015 AT 2:39 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/#COMMENT-95226)

Great article! I would love to see something similar regarding parameter tuning for the X package.

John S says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MODEL/?REPLYTOCOM=
OCTOBER 12, 2016 AT 12:15 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/06/TUNING-RANDOM-FOREST-MOI
117085)

This was a very nice article. I would still be interested to know if there is a minimum nur trees that can be calculated to reduce computational cost?

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

STAY CONNECTED

LATEST POSTS

(https://www.analyticsvidhya.com/blog.
plan-2017-beginners-data-science/)

Infographic - Learning Plan 2 beginners in data science (https://www.analyticsvidhy 2017/01/learning-plan-2017-b data-science/)

KUNAL JAIN , JANUARY 28, 2017

8,040
FOLLOWERS
(http://www.twitter.com/analyticsvidhya)

26,184
FOLLOWERS
(http://www.facebook.com/Analyticsvidhya)

1,581
FOLLOWERS
(https://plus.google.com/+Analyticsvidhya)

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

(https://www.analyticsvidhya.com/blog/
data-science-plan-for-2017/)

**Infographic - Learning Plan 2
Transitioners in data science
(https://www.analyticsvidhy
2017/01/transitioners-data-s
for-2017/)**

KUNAL JAIN , JANUARY 28, 2017

(https://www.analyticsvidhya.com/blog/
plan-2017-intermediates-data-science/)

**Infographic - Learning Plan 2
Intermediates in data scienc
(https://www.analyticsvidhy
2017/01/learning-plan-2017-
intermediates-data-science/**

KUNAL JAIN , JANUARY 28, 2017

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

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

YOGESH KULKARNI , JANUARY 27, 20

© Copyright 2013-2017 Analytics Vidhya