f (https://www.facebook.com/AnalyticsVidhya) | (https://twitter.com/analyticsvidhya) | (https://plu

in (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)

Comparing a CART model to Random Forest (Par

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

\RE f (http://www.facebook.com/sharer.php?u=https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest- t=Comparing%20a%20CART%20model%20to%20Random%20Forest%20(Part%201)) (https://twitter.com/home? tus=Comparing%20a%20CART%20model%20to%20Random%20Forest%20(Part%201)+https://www.analyticsvidhya.com/blog/2014 t-random-forest-1/) (https://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-f tp://pinterest.com/pin/create/button/?url=https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest- media=https://www.analyticsvidhya.com/wp- tent/uploads/2014/06/tree.png&description=Comparing%20a%20CART%20model%20to%20Random%20Forest%20(Part%201))

I created my first simple regression model with my father in 8th standard (year: 2002) or Obviously, my contribution in that model was minimal, but I really enjoyed the representation of the data. We tried validating all the assumptions etc. for this model. By the exercise, we had 5 sheets of the simple regression model on 700 data points. The ent was complex enough to confuse any person with average IQ level. When I look at my mc which are built on millions of observations and utilize complex statistics behind the scer how machine learning with sophisticated tools (like SAS, SPSS, R) has made our life easy.

Having said that, many people in the industry do not bother about the complex statistics, behind the scene. It becomes very important to realize the predictive power of each tec model is perfect in all scenarios. Hence, we need to understand the data and the surrou system before coming up with a model recommendation.

In this article, we will compare two widely used techniques i.e. CART vs. Random fores Random forest were covered in my last (https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/ take a case study to build a strong foundation of this concept and use R to do the comp dataset used in this article is an inbuilt dataset of R.

As the concept is pretty lengthy, we have broken down this article into two parts

Background on Dataset "Iris"

Data set "iris" gives the measurements in centimeters of the variables : sepal length and petal length and width, respectively, for 50 flowers from each of 3 species of Iris. The data cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, F Species. We intend to predict the Specie based on the 4 flower characteristic variables.

We will first load the dataset into R and then look at some of the key statistics. Yo the following codes to do so.

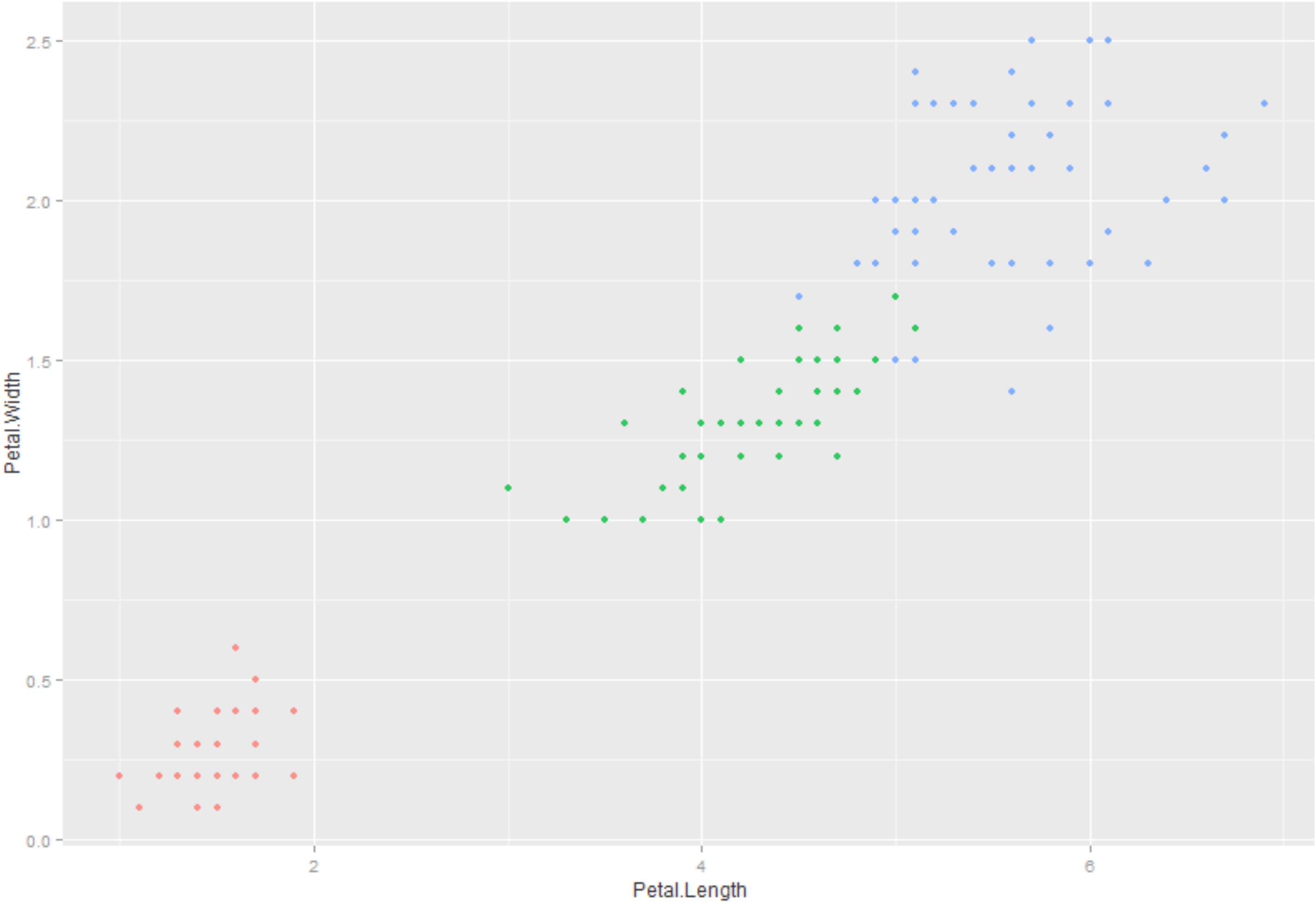
```
data(iris)
```

```
# look at the dataset
```

```
summary(iris)
```

```
# visually look at the dataset
```

```
qplot(Petal.Length,Petal.Width,colour=Species,data=iris)
```



(https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/plot1.png)

The three species seem to be well segregated from each other. The accuracy in pr borderline cases determines the predictive power of the model. In this case, we will install packages for making a CART model.

```
library(rpart)
```

```
library(caret)
```

After loading the library, we will divide the population in two sets: Training and validation. to make sure that we do not overfit the model. In this case, we use a split of 50-50 for t validation. Generally, we keep training heavier to make sure that we capture the key cha You can use the following code to make this split.

```
train.flag <- createDataPartition(y=iris$Species,p=0.5,list=FALSE)
```

```
training <- iris[train.flag,]
```

```
Validation <- iris[-train.flag,]
```

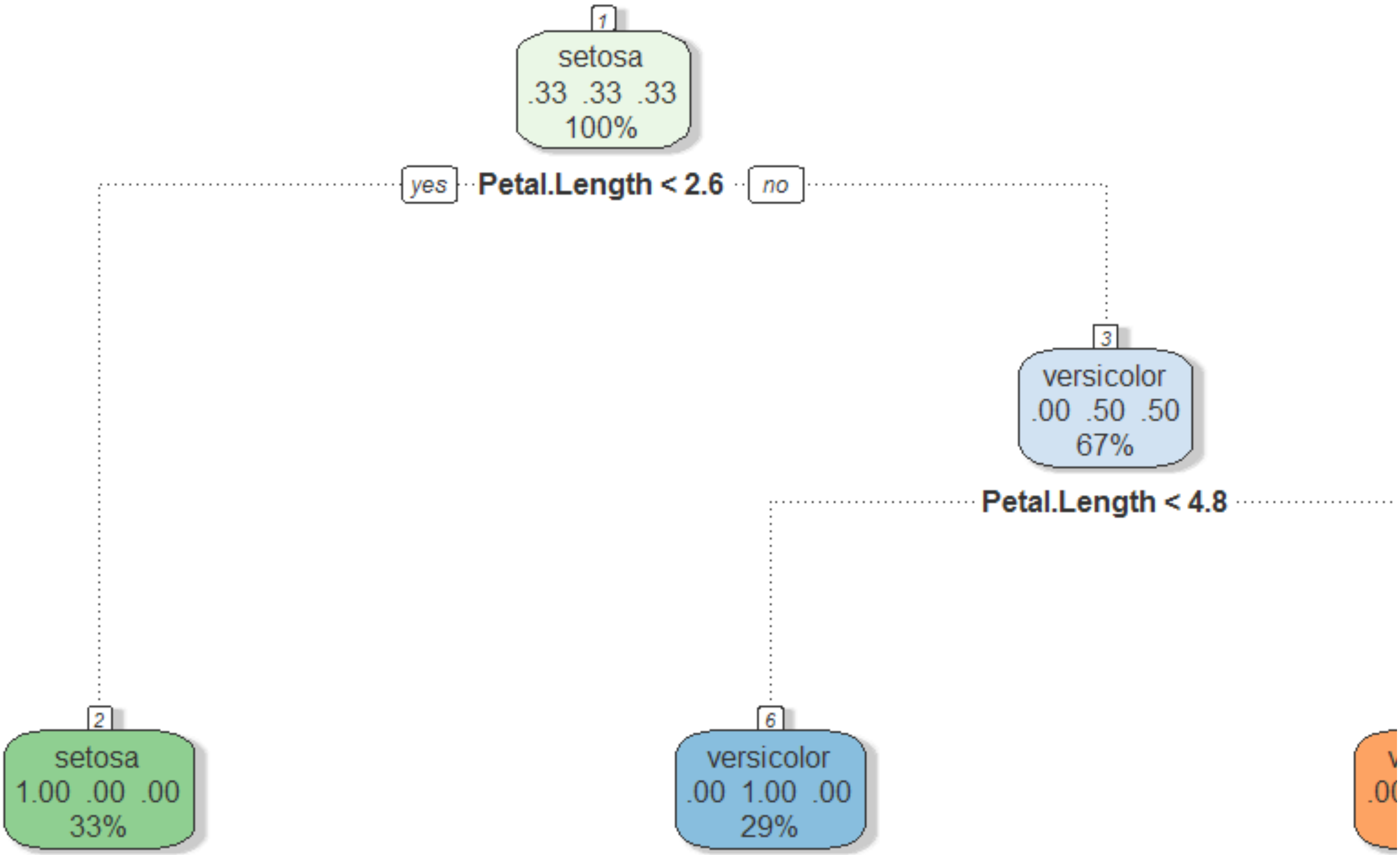
Building a CART model

Once we have the two data sets and have got a basic understanding of data, we now bu model. We have used "caret" and "rpart" package to build this model. However, the representation of the CART model is not graphically appealing on R. Hence, we have used called “rattle” to make this decision tree. “Rattle” builds a more fancy and clean trees, wh easily interpreted. Use the following code to build a tree and graphically check this tree:

```
modfit <- train(Species~.,method="rpart",data=training)
```

```
library(rattle)
```

```
fancyRpartPlot(modfit$finalModel)
```



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

Validating the model

Now, we need to check the predictive power of the CART model, we just built. Here, we are a discordance rate (which is the number of misclassifications in the tree) as the decision use the following code to do the same :

```
train.cart<-predict(modfit,newdata=training)

table(train.cart,training$Species)

train.cart   setosa versicolor virginica


setosa         25          0         0

versicolor     0         22         0

virginica       0          3       25


# Misclassification rate = 3/75
```

Only 3 misclassified observations out of 75, signifies good predictive power. In general, a misclassification rate less than 30% is considered to be a good model. But, the range of a g depends on the industry and the nature of the problem. Once we have built the moc validate the same on a separate data set. This is done to make sure that we are not ove model. In case we do over fit the model, validation will show a sharp decline in the predicti is also recommended to do an out of time validation of the model. This will make sure that is not time dependent. For instance, a model built in festive time, might not hold in regul simplicity, we will only do an in-time validation of the model. We use the following code t time validation:

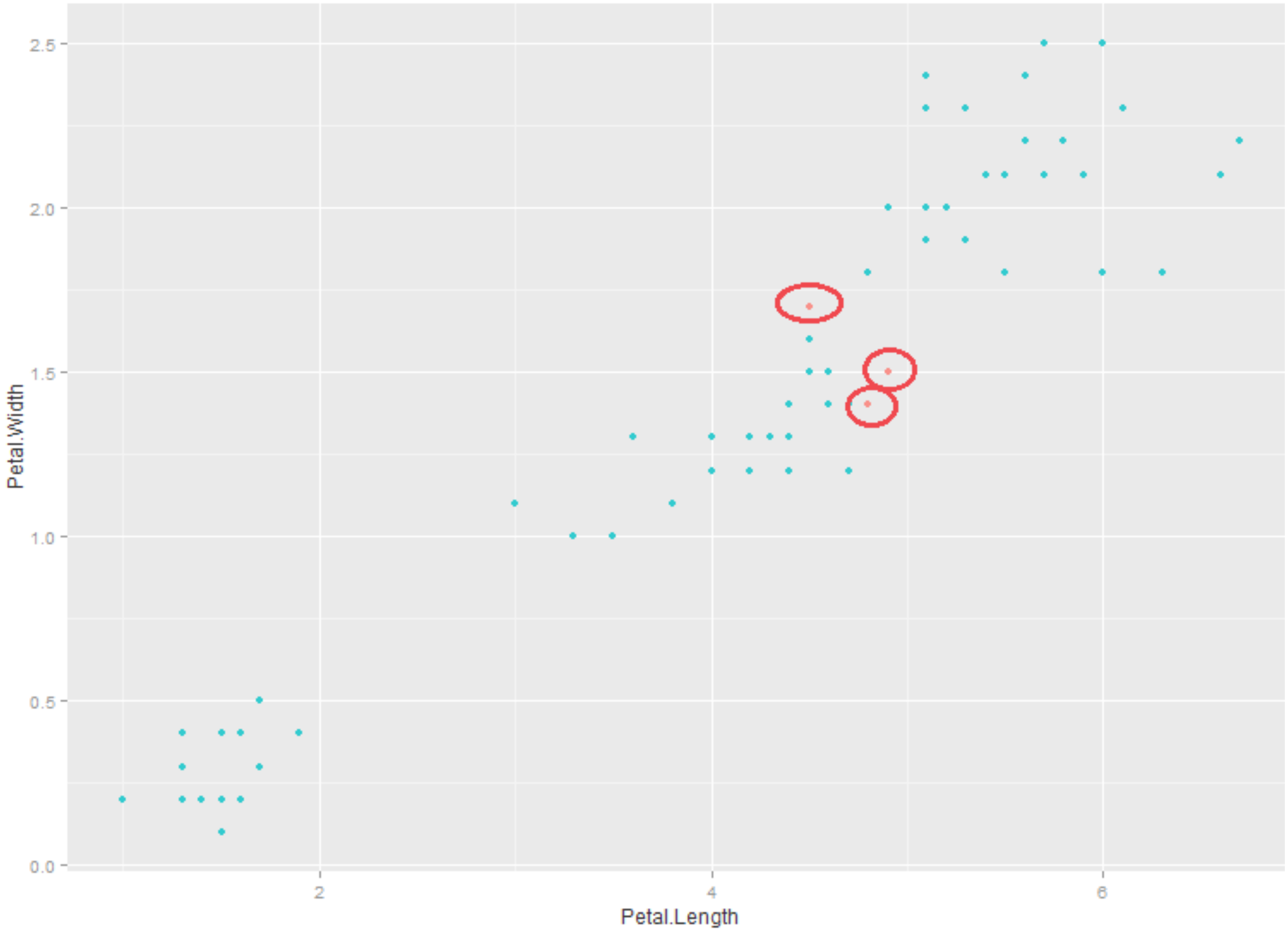
```
pred.cart<-predict(modfit,newdata=Validation)

table(pred.cart,Validation$Species)

pred.cart   setosa versicolor virginica



setosa         25         0         0

versicolor     0         22         1

virginica       0         3       24



# Misclassification rate = 4/75
```

As we see from the above calculations that the predictive power decreased in validation as to training. This is generally true in most cases. The reason being, the model is trained on data set, and just overlaid on validation training set. But, it hardly matters, if the predictiv validation is lesser or better than training. What we need to check is that they are close enc case, we do see the misclassification rate to be really close to each other. Hence, we s CART model in this case study.

Let's now try to visualize the cases for which the prediction went wrong. Following is the c to find the same :

```
correct <- pred.cart == Validation$Species
```

```
qplot(Petal.Length,Petal.Width,colour=correct,data=Validation)
```



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

As you see from the graph, the predictions which went wrong were actually those borde [illegible] before that these are the cases which make or break the con the model. Most of the models will be able to categorize observation far away from ea takes a model to be sharp to distinguish these borderline cases.

Analytics Vidhya Learn Everything About Analytics (https://www.analyticsvidhya.com)
HOME (HTTPS://WWW.ANAL
BLOG (HTTPS://WWW.ANALY
JOBS (HTTPS://WWW.ANALYTICSVIDHYA.COM/JOBS/) TRAININGS (HTTPS://WWW.ANALYTICSVIDHYA.CO
LEARNING PATHS (HTTPS://WWW.ANALYTICSVIDHYA.COM/LEARNING-PATHS-DATA-SCIENCE-BUSINESS-ANA
DATAHACK (HTTPS://DATAHACK.ANALYTICSVIDHYA.COM) STORIES (HTTPS://WWW.ANALYTICSVIDHYA.
WRITE FOR US (HTTP://WWW.ANALYTICSVIDHYA.COM/ABOUT-ME/WRITE/) CONTACT US (HTTPS://WWW

End Notes:

In the next article, we will solve the same problem using a random forest algorithm. We random forest will be able to make even better prediction for these borderline cases. E never generalize the order of predictive power among a CART and a random forest or predictive algorithm. The reason being every model has its own strength. Random fores tends to have a very high accuracy on the training population, because it uses mar characteristics to make a prediction. But, because of the same reason, it sometimes o model on the data. We will see these observations graphically in the next article and ta details on scenarios where random forest or CART comes out to be a better predictive moc

Did you find the article useful? Did this article solve any of your existing dilemmas? compared the two models in any of your projects? If you did, share with us your thoughts o

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), f on twitter (http://twitter.com/analyticsvidhya) or like our facebook pa (http://facebook.com/analyticsvidhya).

Share this:

(https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/?share=linkedin&nb=1) 52

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

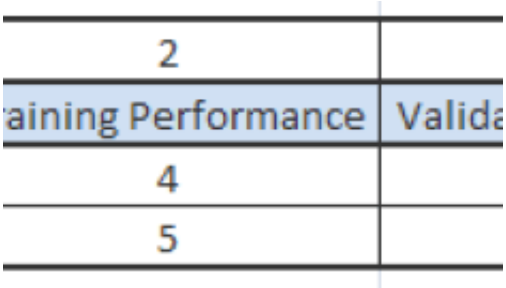
(https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/?share=google-plus-1&nb=1)

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

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

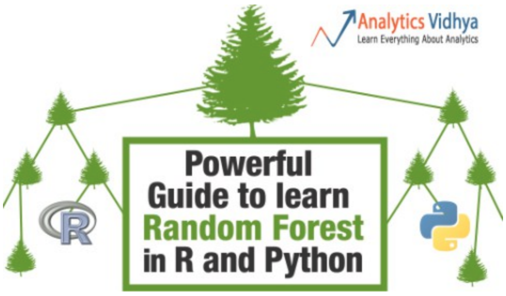
(https://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/?share=reddit&nb=1)

RELATED



2	
aining Performance	Valida
4	
5	

(https://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/) Comparing a Random Forest to a CART model (Part 2) (https://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/)
June 27, 2014
In "Business Analytics"


Analytics Vidhya
Learn Everything About Analytics
Powerful Guide to learn Random Forest in R and Python

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

(https://www.analyticsvi blog/2015/06/hackathon guide-analytics-vidhya/) The Hackathon Practice (Analytics Vidhya (https://www.analyticsvi blog/2015/06/hackathon guide-analytics-vidhya/)
June 5, 2015
In "Analytics Vidhya"

TAGS: ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ANALYTICS/), CART (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CA (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CHAID/), DATA MINING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DATA-MINING/), D (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DECISION-TREE/), RANDOM FOREST (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RAND

Previous Article
Sr. Analyst- MasterCard, Gurgaon (3 - 8 years of experience) (https://www.analyticsvidhya.com/blog/2014/06/sr-analyst-mastercard-gurgaon-3-8-years-experience/)

Next Articl
Leader, Marketing - MasterCard, Gurgao (https://www.analyticsvidhya.com/blog/2014 marketing-mastercard-gurgaon/

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

Author

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

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.

4 COMMENTS

vasudev.D says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-CART-RANDOM-FOREST-1/?REPLYTOCOM=
JULY 12, 2014 AT 3:30 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-CART-RANDOM-FOREST
13994)

its really a great article . may i know the logic or package to use for createDataPartition

vasudev.D says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-CART-RANDOM-FOREST-1/?REPLYTOCOM=
JULY 12, 2014 AT 3:30 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-CART-RANDOM-FOREST
13995)

its really a great article . may i know the logic or package to use for createDataPartition

rahul29 says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-CART-RANDOM-FOREST-1/?REPLYTOCOM=
JUNE 16, 2015 AT 12:31 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-CART-R
1/#COMMENT-88677)

caret package, which has some other functions to split and train the data

A similar ex from caTools package in r -splitting the data according to binary
Dependent var has values as 0 and 1.

split=split(Dep var,SplitRatio=0.7)
train=subset(data,split==TRUE)
test=subset(data=split==FALSE)

LEAVE A REPLY

Connect with:

(https://www.analyticsvidhya.com/wp-login.php?action=wordpress_social_authenticate&mode=login&provider=Facebook&redirect_to=https%3A
cart-random-forest-1%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

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

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

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

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

LATEST POSTS

(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

(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 & I (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