

f (<https://www.facebook.com/AnalyticsVidhya>) | t (<https://twitter.com/analyticsvidhya>)

g+ (<https://plus.google.com/+Analyticsvidhya/posts>)

in (<https://www.linkedin.com/groups/Analytics-Vidhya-Learn-everything-about-5057165>)



(<http://datahack.analyticsvidhya.com/contest/date-your-data>)

Home (<http://www.analyticsvidhya.com/>) > Business Analytics (<http://www.analyticsvidhya.com/blog/category/business...>)

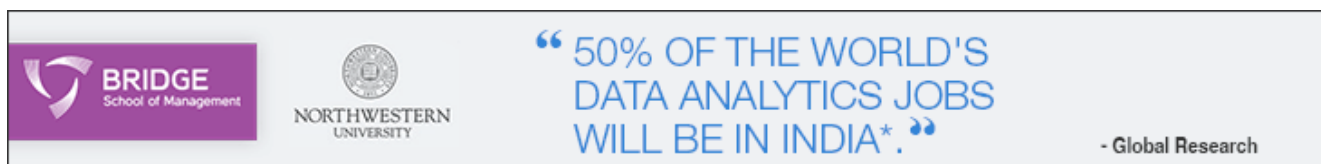
Comparing a Random Forest to a CART model (Part 2)

BUSINESS ANALYTICS ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/](http://www.analyticsvidhya.com/blog/category/business-analytics/)) R

([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/R/](http://www.analyticsvidhya.com/blog/category/r/))

[www.facebook.com/sharer.php?u=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-](http://www.facebook.com/sharer.php?u=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model-20a%20Random%20Forest%20to%20a%20CART%20model%20(Part%202))
[20a%20Random%20Forest%20to%20a%20CART%20model%20\(Part%202\)](https://twitter.com/home?status=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model-20a%20Random%20Forest%20to%20a%20CART%20model%20(Part%202))) t ([https://twitter.com/home?](https://twitter.com/home?status=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model-20a%20Random%20Forest%20to%20a%20CART%20model%20(Part%202))
[g+ \(\[http://pinterest.com/pin/create/button?url=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-\]\(https://plus.google.com/share?url=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-

<a href=\)
\[ription=Comparing%20a%20Random%20Forest%20to%20a%20CART%20model%20\\(Part%202\\)\]\(http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model-20a%20Random%20Forest%20to%20a%20CART%20model%20\(Part%202\)\)\)](https://plus.google.com/share?url=http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model-20a%20Random%20Forest%20to%20a%20CART%20model%20(Part%202))



([http://admissions.bridgesom.com/pba-new/?](http://admissions.bridgesom.com/pba-new/?utm_source=AV&utm_medium=Banner&utm_campaign=AVBanner)
[utm_source=AV&utm_medium=Banner&utm_campaign=AVBanner](http://admissions.bridgesom.com/pba-new/?utm_source=AV&utm_medium=Banner&utm_campaign=AVBanner))

Random forest is one of the most commonly used algorithm in Kaggle competitions. Along with a good predictive power, Random forest model are pretty simple to build. We have previously explained the algorithm of a random forest ([Introduction to Random Forest](http://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/) (<http://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/>)). This article is the second part of the series on comparison of a random forest with a CART model. In the first article, we took an example of an inbuilt R-dataset to predict the classification of an specie. In this article we will build a random forest model on the same dataset to compare the performance with previously built CART model. I did this experiment a week back and found the results very insightful. I recommend the reader to read the first part of this article (Last article (<http://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/>)) before reading this one.

Background on Dataset "Iris"

Data set "iris" gives the measurements in centimeters of the variables : sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of Iris. The dataset has 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, 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. You can use 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)
```

Results using CART Model

The first step we follow in any modeling exercise is to split the data into training and validation. You can use the following code for the split. (We will use the same split for random forest as well)

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

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

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

CART model gave following result in the training and validation :

Misclassification rate in training data = 3/75

Misclassification rate in validation data = 4/75

As you can see, CART model gave decent result in terms of accuracy and stability. We will now model the random forest algorithm on the same training dataset and validate it using same validation dataset.

Building a Random forest model

We have used "caret", "randomForest" and "randomForestSRC" package to build this model. You can use the following code to generate a random forest model on the training dataset.

```
> library(randomForest)
```

```
> library(randomForestSRC)
```

```
> library(caret)
```

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

```
> pred <- predict(modfit,training)
```

```
> table(pred,training$Species)
```

<u>pred</u>	setosa	versicolor	virginica
setosa	25	0	0
versicolor	0	25	0
virginica	0	0	25

Misclassification rate in training data = 0/75

[This is simply awesome!]

Validating the model

Having built such an accurate model, we will like to make sure that we are not over fitting the model on the training data. This is done by validating the same model on an independent data set. We use the following code to do the same :

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

```
> table(train.cart,training$Species)
```

```
> train.cart   setosa versicolor virginica
```

pred	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. However, we see a significant drop in predictive power of this model when we compare it to training misclassification.

Comparison between the two models

Till this point, everything was as per books. Here comes the tricky part. Once you have all performance metrics, you need to select the best model as per your business requirement. We will make this judgement based on 3 criterion in this case apart from business requirements:

1. Stability : The model should have similar performance metrics across both training and validation. This is very essential because business can live with a lower accuracy but not with a lower stability. We will give the highest weight to stability. For this case let's take it as 5.

2. Performance on Training data : This is one of the important metric but nothing conclusive can be said just based on this metric. This is because an over fit model is unacceptable but will get a very high score at this parameter. Hence, we will give a low weight to this parameter (say 2).

3. Performance on Validation data : This metric catch holds of overfit model and hence is an important metric. We will score it higher than performance and lower than stability. For this case let's take it as 3.

Note that the weights and scores entirely depends on the business case. Following is a score table as per my judgement in this case.

weights	5	2	3	
Out of 5	Stability	Training Performance	Validation Performance	Total
CART	5	4	4	45
Random Forest	3	5	5	40

(<http://i1.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2014/06/Comparison.png>)As you can see from the table that however Random forest gives me a better performance, I still will go ahead and use CART model because of the stability factor. Other factor in favor of CART model is the easy business justification. Random forest is very difficult to explain to people working on field. CART models are simple cuts which can be justified by simple business justification/reasons. But the choice of model selection is entirely dependent on business requirement.

End Notes

Every model has its own strength. Random forest, as seen from this case study, has a very high accuracy on the training population, because it uses many different characteristics to make a prediction. But, because of the same reason, it sometimes over fits the model on the data. CART model on the other side is simplistic criterion cut model. This might be over simplification in some case but works pretty well in most business scenarios. However, the choice of model might be business requirement dependent, it is always good to compare performance of different model before taking this call.

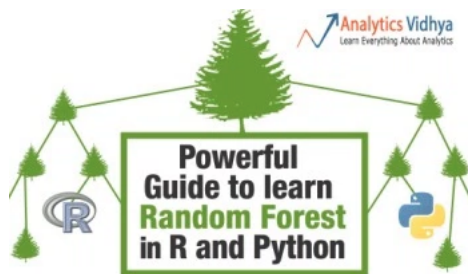
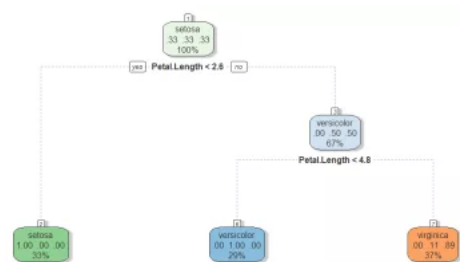
Did you find the article useful? Did this article solve any of your existing dilemmas? Have you compared the two models in any of your projects? If you did, share with us your thoughts on the topic.

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>), follow us on twitter (<http://twitter.com/analyticsvidhya>) or like our facebook page (<http://facebook.com/analyticsvidhya>).

Share this:

-  (<http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/?share=linkedin&nb=1>) 58
-  (<http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/?share=facebook&nb=1>) 16
-  (<http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/?share=google-plus-1&nb=1>)
-  (<http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/?share=twitter&nb=1>)
-  (<http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/?share=pocket&nb=1>)
-  (<http://www.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model/?share=reddit&nb=1>)

RELATED



(<http://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/>)

Comparing a CART model to Random Forest (Part 1)

(<http://www.analyticsvidhya.com/blog/2014/06/comparing-cart-random-forest-1/>)

In "Business Analytics"

(<http://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-challenges/>)

Powerful Guide to learn Random Forest (with codes in R & Python)

(<http://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-challenges/>)

In "Business Analytics"

(<http://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/>)

Introduction to Random forest - Simplified

(<http://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/>)

In "Big data"

TAGS: CART ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CART/](http://www.analyticsvidhya.com/blog/tag/cart/)), CHAID ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CHAID/](http://www.analyticsvidhya.com/blog/tag/chaid/)), COMPARISON ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/COMPARISON/](http://www.analyticsvidhya.com/blog/tag/comparison/)), DECISION TREE ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DECISION-TREE/](http://www.analyticsvidhya.com/blog/tag/decision-tree/)), RANDOM FOREST ([HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RANDOM-FOREST/](http://www.analyticsvidhya.com/blog/tag/random-forest/))

Previous Article

Senior Analyst - AbsolutData, Gurgaon
(2+ years of experience)

(<http://www.analyticsvidhya.com/blog/2014/06/senior-analyst-absolutdata-gurgaon-2-years-experience/>)

Next Article

Director - Bangalore (10 - 15 Years of Experience)

(<http://www.analyticsvidhya.com/blog/2014/06/director-bangalore-10-15-years-experience/>)



(<http://www.analyticsvidhya.com/blog/author/tavish1/>)

Author

Tavish Srivastava

(<http://www.analyticsvidhya.com/blog/author/tavish1/>)

I am Tavish Srivastava, a post graduate from IIT Madras in Mechanical Engineering. I have more than two years of work experience in Analytics. My experience ranges from hands on analytics in a developing country like India to convince banking partners with analytical solution in matured market like US. For last two and a half years I have contributed to various sales strategies, marketing strategies and Recruitment strategies in both Insurance and Banking industry.

3 COMMENTS



Lalit Sachan says:

REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-RANDOM-FOREST-SIMPLE-CART-MODEL/?REPLYTOCOM=12596#RESPOND)
JUNE 28, 2014 AT 8:13 AM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-RANDOM-FOREST-SIMPLE-CART-MODEL/#COMMENT-12596)

I would say this is not fair comparison. Random Forests are not meant for such small datasets essentially. Also try to use CART where $p > n$. Here is a Quote from Kaggle wiki on RFs.

"Unlike single decision trees which are likely to suffer from high Variance or high [Bias] (depending on how they are tuned) Random Forests use averaging to find a natural balance between the two extremes."

With increase in your data in either dimensions, RFs are much more stable in comparison to single trees because of inherent virtue of averaging.



Tavish says:

REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-RANDOM-FOREST-SIMPLE-CART-MODEL/?REPLYTOCOM=12642#RESPOND)
JUNE 28, 2014 AT 4:26 PM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-RANDOM-FOREST-SIMPLE-CART-MODEL/#COMMENT-12642)

Lalit,

Thank you for bringing this up. As mentioned in the article, the choice is always made based on business requirement. The case in hand has low number of observation and has a clean split between different species. This is why CART works so well. For this reason my preference will be CART model, as random forest does not add a significant value in terms of predictive power and stability. But I would say, choosing an algorithm is more of an art than straight forward science. But I appreciate you bringing up a valid point to compare the two models.

Tavish



Chirag says:


REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-RANDOM-FOREST-SIMPLE-CART-MODEL/?REPLYTOCOM=29923#RESPOND)
OCTOBER 15, 2014 AT 7:50 AM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2014/06/COMPARING-RANDOM-FOREST-SIMPLE-CART-MODEL/#COMMENT-29923)

Good explanation Tavish. Thank you for it. Would you mind explaining why you did your validation on the training data.set? Thats the data set you used to make your prediction

model.you also have a validation(test) data set with you,isnt it?I might be wrong as I am a beginner in this.

LEAVE A REPLY

Connect with:

 (http://www.analyticsvidhya.com/wp-login.php?action=wordpress_social_authenticate&mode=login&provider=Facebook&redirect_to=http%3A%2F%2Fwww.analyticsvidhya.com/blog/2014/06/comparing-random-forest-simple-cart-model%2F)

Your email address will not be published.

Comment

Name (required)

Email (required)






Website

☐ Notify me of follow-up comments by email.


☐ Notify me of new posts by email.

SUBMIT COMMENT

TOP USERS

Rank	Name	Points
1	 Nalin Pasricha http://datahack.analyticsvidhya.com/user/profile/Nalin	3478
2	 SRK (http://datahack.analyticsvidhya.com/user/profile/SRK)	3364
3	 Aayushmnit http://datahack.analyticsvidhya.com/user/profile/aayushmnit	3075
4	 binga (http://datahack.analyticsvidhya.com/user/profile/binga)	2623
5	 vikash (http://datahack.analyticsvidhya.com/user/profile/vikash)	2190


More Rankings (<http://datahack.analyticsvidhya.com/users>)



TRANSFORM
YOUR CAREER

Enrol to India's No.1 Analytics Course

- Chosen by **500 +** professionals
- 1,75,000+** learning hours delivered
- Designed for **working professionals**.
LEARN while you WORK



http://pgpba.greatlakes.edu.in/?utm_source=AVM&utm_medium=Banner&utm_campaign=Pgpba_decjan

POPULAR POSTS

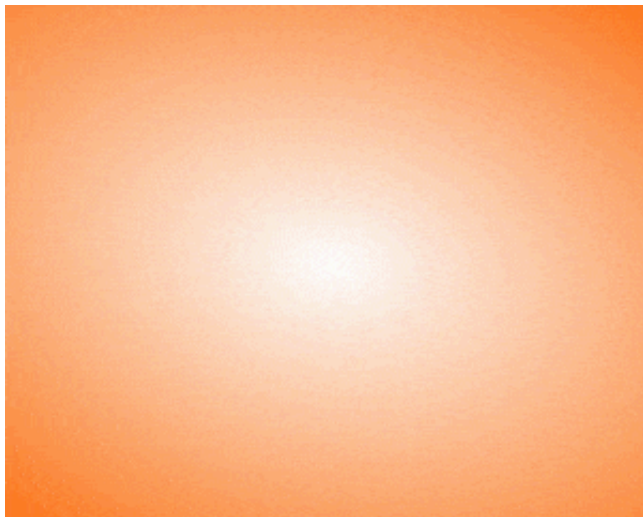
- Free Must Read Books on Statistics & Mathematics for Data Science
(<http://www.analyticsvidhya.com/blog/2016/02/free-read-books-statistics-mathematics-data-science/>)
- A Complete Tutorial to Learn Data Science with Python from Scratch
(<http://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/>)
- Essentials of Machine Learning Algorithms (with Python and R Codes)
(<http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/>)
- A Complete Tutorial on Time Series Modeling in R
(<http://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/>)
- Complete guide to create a Time Series Forecast (with Codes in Python)
(<http://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/>)
- 4 tricky SAS questions commonly asked in interview
(<http://www.analyticsvidhya.com/blog/2013/11/4-sas-tricky-analytics-interview/>)
- SAS vs. R (vs. Python) – which tool should I learn?
(<http://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/>)
- 7 Important Model Evaluation Error Metrics Everyone should know
(<http://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/>)



Online Course from MIT

Learn programming from MIT faculty.
Starts march 2. Enroll now.

www.edx.org/MITx



([http://imarticus.org/programs/business-analytics-](http://imarticus.org/programs/business-analytics-professional/)

professional/)

RECENT POSTS



([http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-](http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)

science-scratch/)

A Complete Tutorial to learn Data Science in R from Scratch

(<http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)

MANISH SARASWAT , FEBRUARY 28, 2016



([http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-](http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)

segmentation/)

Guide to Build Better Predictive Models using Segmentation

(<http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/>)

GUEST BLOG , FEBRUARY 26, 2016



([http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-](http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

salary-report-2016/)

Quick Insights: India Analytics and Big Data Salary Report 2016

(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 24, 2016



(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)

India Exclusive: Analytics and Big Data Salary Report 2016

(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 22, 2016



([http://www.edvancer.in/certified-business-analytics?](http://www.edvancer.in/certified-business-analytics?utm_source=AV&utm_medium=AVads&utm_campaign=AVads1&utm_content=cbapavad)

[utm_source=AV&utm_medium=AVads&utm_campaign=AVads1&utm_content=cbapavad](http://www.edvancer.in/certified-business-analytics?utm_source=AV&utm_medium=AVads&utm_campaign=AVads1&utm_content=cbapavad))

GET CONNECTED



4,162

FOLLOWERS

(<http://www.twitter.com/analyticsvidhya>)



915

FOLLOWERS

(<https://plus.google.com/+Analyticsvidhya>)



11,954

FOLLOWERS

(<http://www.facebook.com/Analyticsvidhya>)



Email

SUBSCRIBE

(<http://feedburner.google.com/fb/a/mailverify?>



uri=analyticsvidhya)

([http://www.analyticsvidhya.com/blog/2016/02/quick-](http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

STAY CONNECTED



4,162

FOLLOWERS

(<http://www.twitter.com/analyticsvidhya>)



915

FOLLOWERS

(<https://plus.google.com/+Analyticsvidhya>)



11,954

FOLLOWERS

(<http://www.facebook.com/Analyticsvidhya>)



Email

SUBSCRIBE

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

LATEST POSTS



([http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-](http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)

science-scratch/)

A Complete Tutorial to learn Data Science in R from Scratch

(<http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)

MANISH SARASWAT , FEBRUARY 28, 2016



([http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-](http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)

segmentation/)

Guide to Build Better Predictive Models using Segmentation

(<http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/>)

GUEST BLOG , FEBRUARY 26, 2016



(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

Quick Insights: India Analytics and Big Data Salary Report 2016

(<http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 24, 2016



(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)

India Exclusive: Analytics and Big Data Salary Report 2016

(<http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/>)

KUNAL JAIN , FEBRUARY 22, 2016

QUICK LINKS

Home (<http://www.analyticsvidhya.com/>)

About Us (<http://www.analyticsvidhya.com/about-me/>)

Our team (<http://www.analyticsvidhya.com/about-me/team/>)

Privacy Policy

(<http://www.analyticsvidhya.com/privacy-policy/>)

Refund Policy

(<http://www.analyticsvidhya.com/refund-policy/>)

Terms of Use

(<http://www.analyticsvidhya.com/terms/>)

TOP REVIEWS

© Copyright 2015 Analytics Vidhya