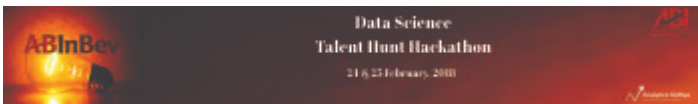


f (<https://www.facebook.com/AnalyticsVidhya>)

t (<https://twitter.com/analyticsvidhya>)

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

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



(https://datahack.analyticsvidhya.com/contest/data-science-talent-hunt-hackathon/?utm_source=AVhome_top)

Home (<https://www.analyticsvidhya.com/>) > R (<https://www.analyticsvidhya.com/blog/category/r/>) > How to perform feature sele...

How to perform feature selection (i.e. pick important variables) using Boruta Package in R ?

R (<https://www.analyticsvidhya.com/blog/category/r/>)

u=[https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-%20selection%20\(i.e.%20pick%20important%20variables\)%20using%20Boruta%20Package%20in%20R%20?%20election%20\(i.e.%20pick%20important%20variables\)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-%20selection%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?%20election%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/))

t ([https://twitter.com/home?selection%20\(i.e.%20pick%20important%20variables\)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/](https://twitter.com/home?selection%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/))

g+ ([https://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-%20selection%20\(i.e.%20pick%20important%20variables\)%20using%20Boruta%20Package%20in%20R%20?%20election%20\(i.e.%20pick%20important%20variables\)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/](https://plus.google.com/share?url=https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-%20selection%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?%20election%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/))

com/pin/create/button/?url=https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-%20selection%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?%20election%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/

m/wp-

ion=How%20to%20perform%20feature%20selection%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?%20election%20(i.e.%20pick%20important%20variables)%20using%20Boruta%20Package%20in%20R%20?/select-important-variables-boruta-package/

Acadgild

REGISTER NOW

BATCHES STARTING SOON

([https://acadgild.com/big-data/data-science-training-certification?](https://acadgild.com/big-data/data-science-training-certification?aff_id=6014&source=AV&account=paid&utm_source=paid&utm_medium=cpc&utm_campaign=Feb_AV_Banner)

[aff_id=6014&source=AV&account=paid&utm_source=paid&utm_medium=cpc&utm_campaign=Feb_AV_Banner](https://acadgild.com/big-data/data-science-training-certification?aff_id=6014&source=AV&account=paid&utm_source=paid&utm_medium=cpc&utm_campaign=Feb_AV_Banner))

Introduction

Variable selection is an important aspect of model building which every analyst must learn. After all, it helps in building predictive models free from correlated variables, biases and unwanted noise.

A lot of novice analysts assume that keeping all (or more) variables will result in the best model as you are not losing any information. Sadly, that is not true!

How many times has it happened that removing a variable from model has increased your model accuracy?

At least, it has happened to me. Such variables are often found to be correlated and hinder achieving higher model accuracy. Today, we'll learn one of the ways of how to get rid of such variables in R. I must say, R has an incredible CRAN repository. Out of all packages, one such available package for variable selection is Boruta Package.

In this article, we'll focus on understanding the theory and practical aspects of using Boruta Package. I've followed a step wise approach to help you understand better.

I've also drawn a comparison of boruta with other traditional feature selection algorithms. Using this, you can arrive at a more meaningful set of features which can pave the way for a robust prediction model. The terms "features", "variables" and "attributes" have been used interchangeably, so don't get confused!



What is Boruta algorithm and why such a strange name ?

Boruta is a feature selection algorithm. Precisely, it works as a wrapper algorithm around Random Forest. This package derive its name from a demon in Slavic mythology who dwelled in pine forests.

We know that feature selection is a crucial step in predictive modeling. This technique achieves supreme importance when a data set comprised of several variables is given for model building.

Boruta can be your algorithm of choice to deal with such data sets. Particularly when one is interested in understanding the mechanisms related to the variable of interest, rather than just building a black box predictive model with good prediction accuracy.

How does it work?

Below is the step wise working of boruta algorithm:

1. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
2. Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
3. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs.

What makes it different from traditional feature selection algorithms?

Boruta follows an all-relevant feature selection method where it captures all features which are in some circumstances relevant to the outcome variable. In contrast, most of the traditional feature selection algorithms follow a minimal optimal method where they rely on a small subset of features which yields a minimal error on a chosen classifier.

While fitting a random forest model on a data set, you can recursively get rid of features in each iteration which didn't perform well in the process. This will eventually lead to a minimal optimal subset of features as the method minimizes the error of random forest model. This happens by selecting an over-pruned version of the input data set, which in turn, throws away some relevant features.

On the other hand, boruta find all features which are either strongly or weakly relevant to the decision variable. This makes it well suited for biomedical applications where one might be interested to determine which human genes (features) are connected in some way to a particular medical condition (target variable).

Boruta in Action in R (Practical)

Till here, we have understood the theoretical aspects of Boruta Package. But, that isn't enough. The real challenge starts now. Let's learn to implement this package in R.

First things first. Let's install and call this package for use.

```
> install.packages("Boruta")
> library(Boruta)
```

Now, we'll load the data set. For this tutorial I've taken the data set from Practice Problem Loan Prediction (<http://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction>)

```
> setwd("../Data/Loan_Prediction")
> traindata <- read.csv("train.csv", header = T, stringsAsFactors = F)
```

Let's have a look at the data.

```
> str(traindata)
> names(traindata) <- gsub("_", "", names(traindata))
```

gsub() function is used to replace an expression with other one. In this case, I've replaced the underscore(_) with blank("").

Let's check if this data set has missing values.

```
> summary(traindata)
```

We find that many variables have missing values. It's important to treat missing values prior to implementing boruta package. Moreover, this data set also has blank values. Let's clean this data set.

Now we'll replace blank cells with NA. This will help me treat all NA's at once.

```
> traindata[traindata == ""] <- NA
```

Here, I'm following the simplest method of missing value treatment i.e. list wise deletion. More sophisticated methods & packages of missing value imputation can be found [here](https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/) (https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/).

```
> traindata <- traindata[complete.cases(traindata),]
```

Let's convert the categorical variables into factor data type.

```
> convert <- c(2:6, 11:13)
> traindata[,convert] <- data.frame(apply(traindata[convert], 2, as.factor))
```

Now is the time to implement and check the performance of boruta package. The syntax of boruta is almost similar to regression (lm) method.

```
> set.seed(123)
> boruta.train <- Boruta(LoanStatus~.-LoanID, data = traindata, doTrace = 2)
> print(boruta.train)
```

Boruta performed 99 iterations in 18.80749 secs.

5 attributes confirmed important: ApplicantIncome, CoapplicantIncome, CreditHistory, LoanAmount, LoanAmountTerm.

4 attributes confirmed unimportant: Dependents, Education, Gender, SelfEmployed.

2 tentative attributes left: Married, PropertyArea.

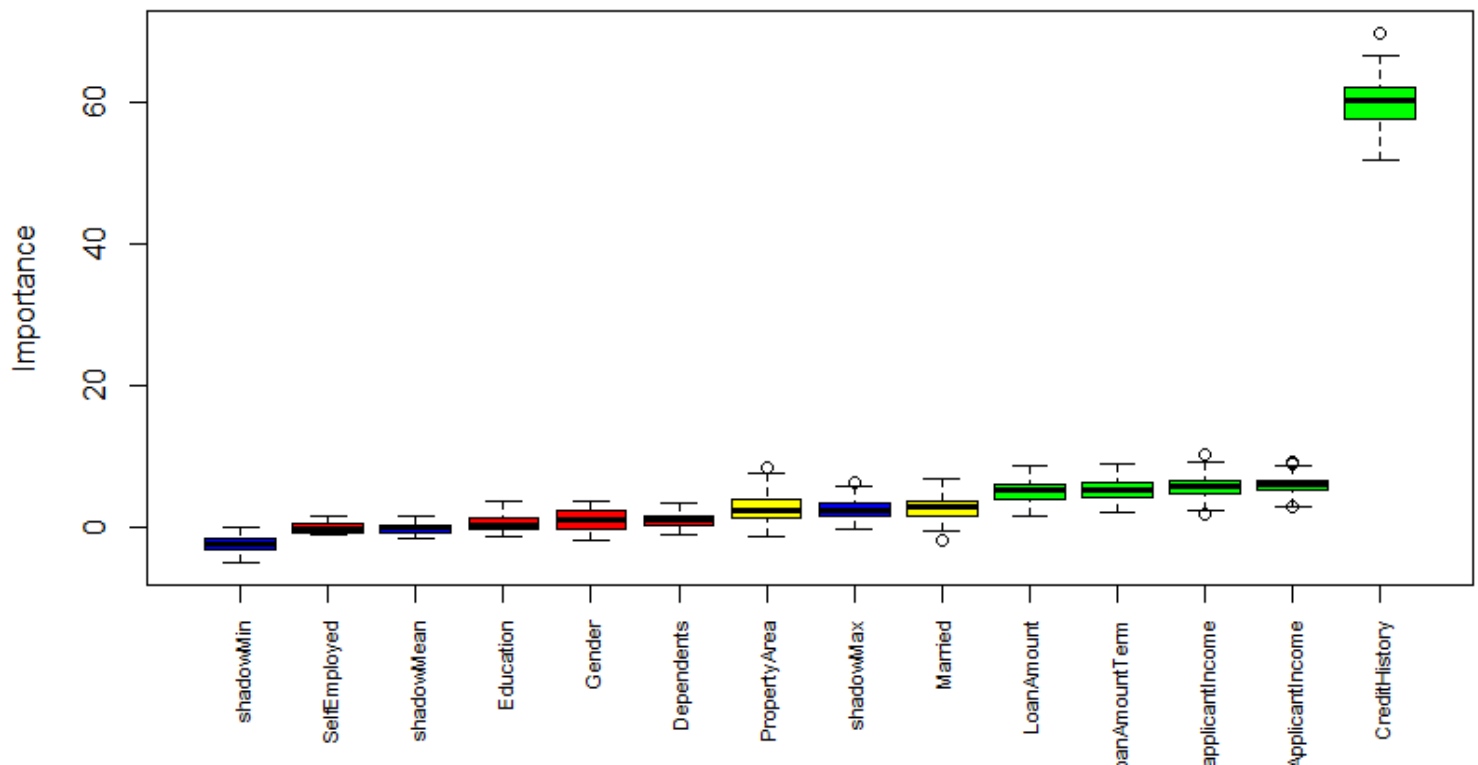
Boruta gives a crystal clear call on the significance of variables in a data set. In this case, out of 11 attributes, 4 of them are rejected and 5 are confirmed. 2 attributes are designated as tentative. Tentative attributes have importance so close to their best shadow attributes that Boruta is not able to make a decision with the desired confidence in default number of random forest runs.

Now, we'll plot the boruta variable importance chart.

By default, plot function in Boruta adds the attribute values to the x-axis horizontally where all the attribute values are not displayed due to lack of space.

Here I'm adding the attributes to the x-axis vertically.

```
> plot(boruta.train, xlab = "", xaxt = "n")
> lz<-lapply(1:ncol(boruta.train$ImpHistory),function(i)
boruta.train$ImpHistory[is.finite(boruta.train$ImpHistory[,i]),i])
> names(lz) <- colnames(boruta.train$ImpHistory)
> Labels <- sort(sapply(lz,median))
> axis(side = 1,las=2,labels = names(Labels),
at = 1:ncol(boruta.train$ImpHistory), cex.axis = 0.7)
```



Blue boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Red, yellow and green boxplots represent Z scores of rejected, tentative and confirmed attributes respectively.

Now is the time to take decision on tentative attributes. The tentative attributes will be classified as confirmed or rejected by comparing the median Z score of the attributes with the median Z score of the best shadow attribute. Let's do it.

```
> final.boruta <- TentativeRoughFix(boruta.train)
> print(final.boruta)
```

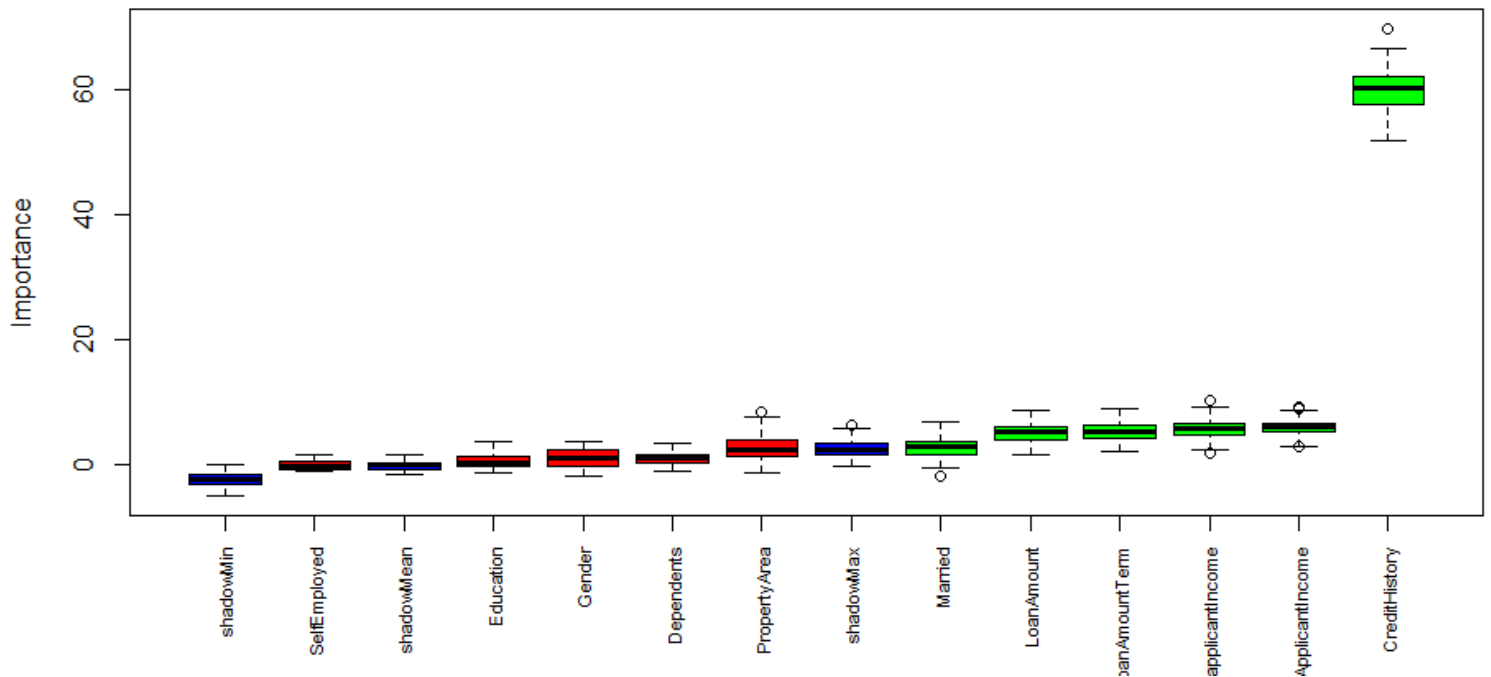
Boruta performed 99 iterations in 18.399 secs.

Tentatives roughfixed over the last 99 iterations.

6 attributes confirmed important: ApplicantIncome, CoapplicantIncome, CreditHistory, LoanAmount, LoanAmountTerm and 1 more.

5 attributes confirmed unimportant: Dependents, Education, Gender, PropertyArea, SelfEmployed.

Boruta result plot after the classification of tentative attributes



It's time for results now. Let's obtain the list of confirmed attributes

```
> getSelectedAttributes(final.boruta, withTentative = F)
[1] "Married"          "ApplicantIncome"  "CoapplicantIncome" "LoanAmount"
[5] "LoanAmountTerm"   "CreditHistory"
```

We'll create a data frame of the final result derived from Boruta.

```
> boruta.df <- attStats(final.boruta)
> class(boruta.df)
[1] "data.frame"
> print(boruta.df)
```

	meanImp	medianImp	minImp	maxImp	normHits	decision
Gender	1.04104738	0.9181620	-1.9472672	3.767040	0.01010101	Rejected
Married	2.76873080	2.7843600	-1.5971215	6.685000	0.56565657	Confirmed
Dependents	1.15900910	1.0383850	-0.7643617	3.399701	0.01010101	Rejected

Education	0.64114702	0.4747312	-1.0773928	3.745441	0.03030303	Rejected
SelfEmployed	-0.02442418	-0.1511711	-0.9536783	1.495992	0.00000000	Rejected
ApplicantIncome	6.05487791	6.0311639	2.9801751	9.197305	0.94949495	Confirmed
CoapplicantIncome	5.76704389	5.7920332	1.9322989	10.184245	0.97979798	Confirmed
LoanAmount	5.19167613	5.3606935	1.7489061	8.855464	0.88888889	Confirmed
LoanAmountTerm	5.50553498	5.3938036	2.0361781	9.025020	0.90909091	Confirmed
CreditHistory	59.57931404	60.2352549	51.7297906	69.721650	1.00000000	Confirmed
PropertyArea	2.77155525	2.4715892	-1.2486696	8.719109	0.54545455	Rejected

Let's understand the parameters used in Boruta as follows:

- *maxRuns*: maximal number of random forest runs. You can consider increasing this parameter if tentative attributes are left. Default is 100.
- *doTrace*: It refers to verbosity level. 0 means no tracing. 1 means reporting attribute decision as soon as it is cleared. 2 means all of 1 plus additionally reporting each iteration. Default is 0.
- *holdHistory*: The full history of importance runs is stored if set to TRUE (Default). Gives a plot of Classifier run vs. Importance when the *plotImpHistory* function is called.

For more complex parameters, please refer to the package documentation (<https://cran.r-project.org/web/packages/Boruta/Boruta.pdf>) of Boruta.

Boruta vs Traditional Feature Selection Algorithm

Till here, we have learnt about the concept and steps to implement boruta package in R.

What if we used a traditional feature selection algorithm such as recursive feature elimination on the same data set. Do we end up with the same set of important features? Let us find out.

Now, we'll learn the steps used to implement recursive feature elimination (RFE). In R, RFE algorithm can be implemented using caret package.

Let's start by defining a control function to be used with RFE algorithm. We'll load the required libraries:

```
> library(caret)
> library(randomForest)
> set.seed(123)
> control <- rfeControl(functions=rfFuncs, method="cv", number=10)
```

Here we have specified a random forest selection function through *rfFuncs* option (which is also the underlying algorithm in Boruta)

Let's implement the RFE algorithm now.

```
> rfe.train <- rfe(traindata[,2:12], traindata[,13], sizes=1:12, rfeControl=control)
```

I'm sure this is self explanatory. `traindata[,2:12]` refers to selecting all independent variables except the ID variable. `traindata[,13]` selects only the dependent variable. It might take some time to run.

We can also check the outcome of this algorithm.

```
> rfe.train
```

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
1	0.8083	0.4702	0.03810	0.1157	*
2	0.8041	0.4612	0.03575	0.1099	
3	0.8021	0.4569	0.04201	0.1240	
4	0.7896	0.4378	0.03991	0.1249	
5	0.7978	0.4577	0.04557	0.1348	
6	0.7957	0.4471	0.04422	0.1315	

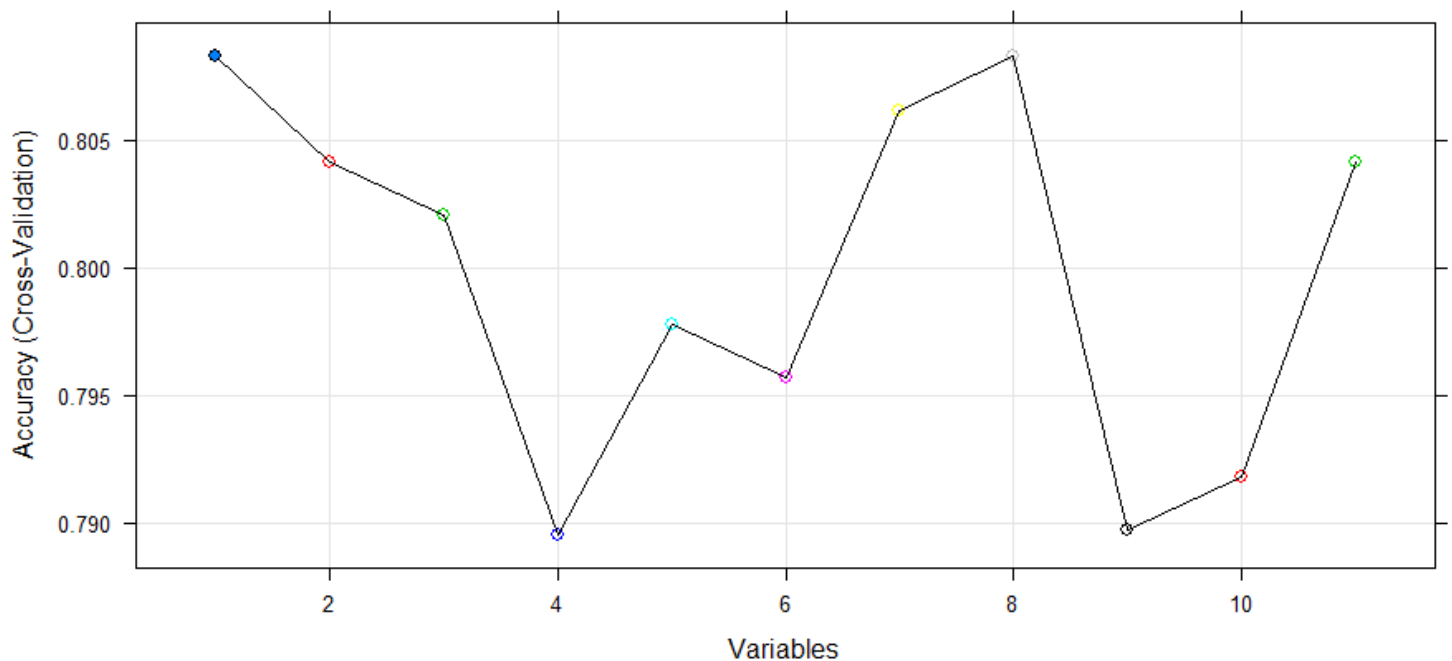
7	0.8061	0.4754	0.04230	0.1297
8	0.8083	0.4767	0.04055	0.1203
9	0.7897	0.4362	0.05044	0.1464
10	0.7918	0.4453	0.05549	0.1564
11	0.8041	0.4751	0.04419	0.1336

The top 1 variables (out of 1):

CreditHistory

This algorithm gives highest weightage to Credit History. Now, we'll plot the result of RFE algorithm and obtain a variable importance chart.

```
> plot(rfe.train, type=c("g", "o"), cex = 1.0, col = 1:11)
```



Let's extract the chosen features. I am confident it would result in Credit History.

```
> predictors(rfe.train)
```

```
[1] "CreditHistory"
```

Hence, we see that recursive feature elimination algorithm has selected "CreditHistory" as the only important feature among the 11 features in the dataset.

As compared to this traditional feature selection algorithm, boruta returned a much better result of variable importance which was easy to interpret as well ! I find it awesome to work on R where one has access to so many amazing packages. I'm sure there would be many other packages for feature selection. I'd love to read about them.

End notes

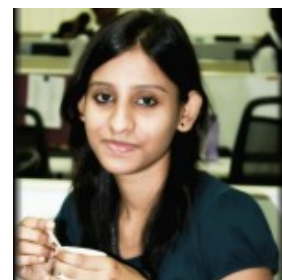
Boruta is an easy to use package as there aren't many parameters to tune / remember. You shouldn't use a data set with missing values to check important variables using Boruta. It'll blatantly throw errors. You can use this algorithm on any classification / regression problem in hand to come up with a subset of meaningful features.

In this article, I've used a quick method to impute missing value because the scope of this article was to understand boruta (theory & practical). I'd suggest you to use advanced methods of missing value imputation. After all, information available in data is all we look for ! Keep going.

Did you like reading this article ? What other methods of variable selection do you use? Do share your suggestions / opinions in the comments section below.

About the Author

Debarati Dutta (<https://ca.linkedin.com/in/debaratidutta8>) is MA Econometrics graduate from University of Madras. She has more than 3 years of experience in data analytics and predictive modeling across multiple domains. She has worked in companies such as Amazon, Antuit, Netlink. Currently, she's based out of Montreal, Canada.



Debarati is the first winner of Blogathon (<http://datahack.analyticsvidhya.com/contest/blogathon>). She won amazon voucher worth INR 5000.

You can test your skills and knowledge. Check out Live Competitions (<http://datahack.analyticsvidhya.com/contest/all>) and compete with best Data Scientists from all over the world.

Share this:

 (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?share=linkedin&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?share=facebook&nb=1>)

96

 (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?share=google-plus-1&nb=1>)

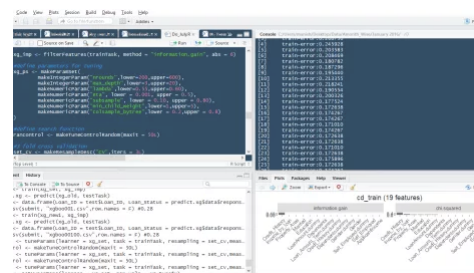
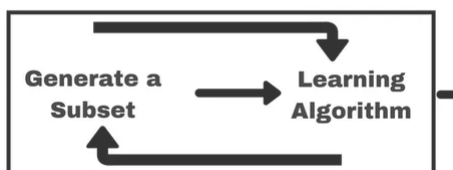
 (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?share=twitter&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?share=pocket&nb=1>)

 (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?share=reddit&nb=1>)

RELATED

Selecting the Best Subset



(<https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>)
Introduction to Feature Selection methods with an example (or how to select the right variables?)
(<https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>)
December 1, 2016

(<https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)
A Complete Tutorial to learn Data Science in R from Scratch
(<https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>)
February 28, 2016
In "Business Analytics"

(<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"

In "Machine Learning"

TAGS: BORUTA ALGORITHM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BORUTA-ALGORITHM/](https://www.analyticsvidhya.com/blog/tag/boruta-algorithm/)), BORUTA PACKAGE ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BORUTA-PACKAGE/](https://www.analyticsvidhya.com/blog/tag/boruta-package/)), CARET PACKAGE ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CARET-PACKAGE/](https://www.analyticsvidhya.com/blog/tag/caret-package/)), MACHINE LEARNING ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING/](https://www.analyticsvidhya.com/blog/tag/machine-learning/)), MISSING VALUES ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MISSING-VALUES/](https://www.analyticsvidhya.com/blog/tag/missing-values/)), MISSING VALUES IMPUTATION IN R ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MISSING-VALUES-IMPUTATION-IN-R/](https://www.analyticsvidhya.com/blog/tag/missing-values-imputation-in-r/)), RANDOM FOREST ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RANDOM-FOREST/](https://www.analyticsvidhya.com/blog/tag/random-forest/)), RECURSIVE FEATURE SELECTION ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/RECURSIVE-FEATURE-SELECTION/](https://www.analyticsvidhya.com/blog/tag/recursive-feature-selection/))

Next Article

Exploring Recommendation System (with an implementation model in R)

(<https://www.analyticsvidhya.com/blog/2016/03/exploring-building-banks-recommendation-system/>)

Previous Article

Course Review – Big data and Hadoop Developer Certification Course by Simplilearn

(<https://www.analyticsvidhya.com/blog/2016/03/review-big-data-hadoop-developer-certification-simplilearn/>)



(<https://www.analyticsvidhya.com/blog/author/guest-blog/>)

Author

Guest Blog (<https://www.analyticsvidhya.com/blog/author/guest-blog/>)

This article is quite old now and you might not get a prompt response from the author. We would request you to post this comment on Analytics Vidhya **Discussion portal** (<https://discuss.analyticsvidhya.com/>) to get your queries resolved.

34 COMMENTS



Sreenivas Malahasthi says

(<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108048#respond>)
MARCH 23, 2016 AT 4:18 AM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108048>)

Gr8 Share Debarati. I was looking for the same kind of info. Keep sharing. Thanks.



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108112#RESPOND)

MARCH 23, 2016 AT 8:07 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108112)

Thanks Sreenivas. Glad that it helped.



Dr.D.K.Samuel says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108051#RESPOND)

MARCH 23, 2016 AT 5:11 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108051)

Beautiful, meaningful info, thanks a lot



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108113#RESPOND)

MARCH 23, 2016 AT 8:08 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108113)

Thank you so much Dr. Samuel.



Vlad says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108066#RESPOND)

MARCH 23, 2016 AT 8:05 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108066)

Hi,

This is not the best package for the determination of the importance of predictors. See this article.<https://www.mql5.com/en/articles/2029> (<https://www.mql5.com/en/articles/2029>)



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108133#RESPOND)

MARCH 23, 2016 AT 10:24 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108133)

Hi Vlad,

Thanks for the interesting article. Well, I would say "best" is more likely a relative term which depends a lot on the problem we have in hand as well as our needs. As mentioned in another comment, if prediction accuracy is your only concern, it might / might not be the best method for

feature selection. But, if you are also interested in understanding the relationships in your data, it would do a much better job.

Hence, application of machine learning techniques involve a lot of trial and error to arrive at the "best" method.



geneseo2000 says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108071#RESPOND)
MARCH 23, 2016 AT 9:22 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108071)

Great tutorial! Thanks!



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108114#RESPOND)
MARCH 23, 2016 AT 8:09 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108114)

Thanks @geneseo2000.



Mathew says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108076#RESPOND)
MARCH 23, 2016 AT 10:08 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108076)

Hi Debrati,
Has this model of feature selection helped in improving the predictions.



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108117#RESPOND)
MARCH 23, 2016 AT 8:39 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108117)

Hi Mathew,

Good question. Well, the answer wouldn't be yes in all cases. In case you only care about good prediction accuracy, it might / might not be the best method for feature selection. But, if you are also interested in inference, it will help you come up with a subset of features, both strongly and weakly relevant to the outcome variable. In this case, although the feature sets obtained from the two methods are quite different, but still it would have very negligible difference in prediction

accuracy. It is due to the fact that the extra variables confirmed by Boruta are weakly relevant to the outcome variable (as evident from the Boruta plot) and hence, might not be playing a major role in prediction accuracy. This might differ from case to case. Hope this helps.



Hunaidkhan Pathan (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108080#respond>)
MARCH 23, 2016 AT 11:01 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108080](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108080))

Really useful package and thanks Debarati really helpful article .

only one Comment for the readers please do change Loan_status to LoanStatus and Loan_ID to LoanID , otherwise it will throw an error



Debarati says: ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108119#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108119#respond))
MARCH 23, 2016 AT 8:46 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108119](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108119))

Hi Hunaidkhan,

Glad that it helped. And thank you so much for pointing it out. Cleaning the variable names using gsub is an optional step and can be avoided.



joo says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108081#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108081#respond))
MARCH 23, 2016 AT 11:06 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108081](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108081))

thanks for this info , please can you send me the traindata
thanks



Debarati says: ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108120#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108120#respond))
MARCH 23, 2016 AT 8:48 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108120](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108120))

Hi @joo,

Glad that you liked it. Can you please send me your e-mail ID?



joo says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108262#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108262#respond))
MARCH 25, 2016 AT 3:32 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108262](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108262))

thanks for reply ,

this is my email address : selmiyoussef50@yahoo.fr (mailto:selmiyoussef50@yahoo.fr)

thanks in advance



joo says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108263#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108263#respond))
MARCH 25, 2016 AT 3:35 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108263](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108263))

thanks for reply ,this is my email address : selmiyoussef50@yahoo.fr
(mailto:selmiyoussef50@yahoo.fr)



michael.saelzer@gmail.com says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108086#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108086#respond))
MARCH 23, 2016 AT 12:40 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108086](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108086))

It would be nice to obtain a decent print of this and other articles.



Debarati says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108121#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108121#respond))
MARCH 23, 2016 AT 8:50 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108121](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108121))

Hi Michael,

Glad that you liked it.



Pallavi says: REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108093#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108093#respond))
MARCH 23, 2016 AT 3:53 PM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108093](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108093))

Hi,

Nice article! This seems to be computationally very expensive. I have a dataset which has 80K rows and 150 columns and Boruta feature selection is taking over 3 hours... Is there any way to optimize the calculations?



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108130#RESPOND)
MARCH 23, 2016 AT 10:01 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108130)

Hi Pallavi,

Glad that it helped. Well, it is computationally expensive as it is a permutation-based feature selection method. You can try out a couple of things to make it a little faster. Try specifying `holdHistory = F` while implementing boruta to prevent it from saving the full history of variable importance runs.

```
boruta.train <- Boruta(Loan_Status~.-Loan_ID, data = traindata, doTrace = 2, holdHistory = F)
```

Alternatively, you can also try decreasing the value of `maxRuns` parameter in case you are not getting any tentative attributes with the default number of random forest runs.

By default, Boruta uses Random forest mean decrease accuracy as the variable importance measure. You can try using a faster Random ferns based variable importance measure. Random ferns is a simplified variation of random forest algorithm.

```
install.packages("rFerns")  
library(rFerns)  
set.seed(123)  
boruta.train <- Boruta(LoanStatus~.-LoanID, data=traindata, doTrace = 2, getImp=getImpFerns,  
holdHistory = F)
```

In this case, you might obtain a different subset of features as the underlying algorithm is different. Hope this helps.



Michael says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108175#RESPOND)
MARCH 24, 2016 AT 7:38 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108175)

Really nice article thanks. Donyou know if the algorithm can be implemented in Python?



Debarati says: (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108225#RESPOND)
MARCH 25, 2016 AT 6:31 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108225)

Hi Michael,

Glad that you liked it. The Python implementation of Boruta can be found in this github account.

https://github.com/danielhomola/boruta_py (https://github.com/danielhomola/boruta_py)



james saye REPLY ([HTTPS://WWW.ANALYTCSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108203#RESPOND](https://www.analytcsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108203#respond))
MARCH 25, 2016 AT 1:35 AM ([HTTPS://WWW.ANALYTCSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108203](https://www.analytcsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108203))

I've DATE data in my csv file. Boruta can differentiate them, but RFE cannot take in DATE format ?

it seems RFE can take in only numerical variables ?

Eg mydate is "3/10/2016"



Debarati says: ([HTTPS://WWW.ANALYTCSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108229#RESPOND](https://www.analytcsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108229#respond))
MARCH 25, 2016 AT 7:11 AM ([HTTPS://WWW.ANALYTCSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108229](https://www.analytcsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108229))

Hi james,

I presume your mydate variable is of class "character" until you convert it to R date format.

Well, boruta can handle predictor variables from classes numeric, factor and character but RFE is only able to handle variables of classes numeric and factor. The reason behind this is although both the algorithms function as wrappers around random forest, they implement random forest algorithm using two different R packages.

Boruta runs random forest from ranger package which allows automatic coercion of character variables to factor labels whereas RFE runs the algorithm from randomForest package which doesn't automatically convert the character variables to factors labels. So, in case of RFE, the character variables in the input dataset are coerced to numeric and hence, NAs will be introduced by coercion. Hope this helps.



james saye REPLY ([HTTPS://WWW.ANALYTCSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=108258#RESPOND](https://www.analytcsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108258#respond))
MARCH 25, 2016 AT 1:49 PM ([HTTPS://WWW.ANALYTCSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-108258](https://www.analytcsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108258))

Hi

How do I pick up Boruta variables that are confirmed, rejected and tentative. ? Or can I get the column numerical so that I can input to RFE for reselection since RFE only select numerical and factors ?



Debarati says: (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108374#respond>)
MARCH 27, 2016 AT 2:11 AM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108374>)

Hi James,

You can use "getSelectedAttributes" function on boruta.train to obtain the list of confirmed variables.

```
confirmed.var <- getSelectedAttributes(boruta.train, withTentative = F)
```

Alternatively, you can also subset the result of "attStats" function to obtain the list of confirmed variables.

```
boruta.df <- attStats(boruta.train)
```

```
# The "decision" variable in boruta.df has to be converted to class character.  
boruta.df$decision <- as.character(boruta.df$decision)
```

```
# Subsetting boruta.df by confirmed, tentative and rejected variables  
confirmed.var <- subset(rownames(boruta.df), boruta.df$decision == "Confirmed")  
tentative.var <- subset(rownames(boruta.df), boruta.df$decision == "Tentative")  
rejected.var <- subset(rownames(boruta.df), boruta.df$decision == "Rejected")
```

To input the list of variables confirmed by Boruta to RFE algorithm

```
rfe.train <- rfe(traindata[,confirmed.var], traindata[,13], sizes=1:5, rfeControl=control)
```

Hope this helps.



venugopal rao says: (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108461#respond>)
MARCH 28, 2016 AT 8:05 AM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108461>)

Good one



karthika (<https://www.thinkittraining.in/hadoop/says-select-important-variables-boruta-package/?replytocom=108662#respond>)
MARCH 30, 2016 AT 10:56 AM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108662>)

thanks for sharing this nice information.wonderful explanation.your way of explanation is good.,it was more impressive to read ,which helps to design more in effective ways



Anu says: REPLY (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108892#respond>)
APRIL 3, 2016 AT 5:04 PM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108892>)

Really good article.

when I tried the same approach on my data, I received a bit different result.

I found 4 variables through Boruta and 5 through the variables, even the 4 variables are not subset of the 5 variables. I am wondering what could be the reason of this.



Debarshi says: (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=108936#respond>)
APRIL 4, 2016 AT 2:14 PM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-108936>)

Thanks Debarati for sharing this. Though I have not tested this, but have few questions:

1. If my ultimate aim is to use a GLM (Generalized Linear Model), do you think it's useful to use a random forest based feature analysis, or should I try something else which are based on GLM methods?
2. What's the 'importance' criterion in this model based on (e.g. Goodness of fit, information criterion,...)
3. Do you think it's useful for regression analysis where the predicted value is a continuous variable?



astrude says: (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=109166#respond>)
APRIL 8, 2016 AT 5:45 AM (<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-109166>)

Have you tried Gradient Boosting based Feature Importance? It's a very powerful technique. It gave superior results to Random Forest results on all my datasets



Sivaji says REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=109304#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=109304#respond))
APRIL 12, 2016 AT 1:38 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-109304](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-109304))

Hi,

Thanks for sharing the information.

Although you explained it clearly the logic behind the Boruta package, I still surprised that most of the features shown as confirmed are not significant in simple glm. Here my analysis is not to improve the accuracy of the model but towards understanding the relationships within the data. I also tried with simple t-test/chi square on the confirmed features and found all numeric features are not significant. As you pointed Boruta is a wrapper algorithm around random forest and looks like it is biased towards numeric features.

My analysis suggests cforest from party package is providing reasonable features which are aligned with my EDA like t-test/chi square test .Even for finding strong relationship with target variable, not sure this package is doing justice to it.



Amit says REPLY ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=112846#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=112846#respond))
JUNE 30, 2016 AT 12:56 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-112846](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-112846))

Great Share



KISHORE says ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/?REPLYTOCOM=119429#RESPOND](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/?replytocom=119429#respond))
DECEMBER 8, 2016 AT 10:46 AM ([HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2016/03/SELECT-IMPORTANT-VARIABLES-BORUTA-PACKAGE/#COMMENT-119429](https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#comment-119429))

Hello #Debarathi Dutta , This really helped to improve my model.
Thank you very much for sharing .

LEAVE A REPLY

Your email address will not be published.

Comment






Name (required)

Email (required)

Website

SUBMIT COMMENT

TOP ANALYTICS VIDHYA USERS

Rank	Name		Points
1		vopani (https://datahack.analyticsvidhya.com/user/profile/Rohan Rao)	8204
2		SRK (https://datahack.analyticsvidhya.com/user/profile/SRK)	7707
3		binga (https://datahack.analyticsvidhya.com/user/profile/binga)	5269
4		Aayushmnit (https://datahack.analyticsvidhya.com/user/profile/aayushmnit)	5258
5		Mark Landry (https://datahack.analyticsvidhya.com/user/profile/mark12)	5243

[More Rankings \(http://datahack.analyticsvidhya.com/users\)](http://datahack.analyticsvidhya.com/users)



(<http://www.greatlearning.education/analytics/?>

utm_source=avm&utm_medium=avmbanner&utm_campaign=pgpba+bda)



(<https://upgrad.com/data-science?>

utm_source=AV&utm_medium=Display&utm_campaign=DS_AV_Banner&utm_term=DS_AV_Banner&utm_conte

POPULAR POSTS

- Essentials of Machine Learning Algorithms (with Python and R Codes)
(<https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>)
- A Complete Tutorial to Learn Data Science with Python from Scratch
(<https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/>)
- 7 Types of Regression Techniques you should know!
(<https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>)

- Understanding Support Vector Machine algorithm from examples (along with code)
(<https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>)
- 6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)
(<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>)
- A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)
(<https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/>)
- The Ultimate Learning Path to Becoming a Data Scientist in 2018
(<https://www.analyticsvidhya.com/blog/2018/01/ultimate-learning-path-becoming-data-scientist-2018/>)
- 17 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely)
(<https://www.analyticsvidhya.com/blog/2016/10/17-ultimate-data-science-projects-to-boost-your-knowledge-and-skills/>)

RECENT POSTS



(<https://www.analyticsvidhya.com/blog/2018/02/pytorch-tutorial/>)

**An Introduction to PyTorch –
A Simple yet Powerful Deep
Learning Library**

(<https://www.analyticsvidhya.com/blog/2018/02/pytorch-tutorial/>)

FAIZAN SHAIKH , FEBRUARY 22, 2018



(<https://www.analyticsvidhya.com/blog/2018/02/top-5-github-repositories-january-2018/>)

5 Data Science & Machine Learning Repositories on GitHub in Jan 2018

(<https://www.analyticsvidhya.com/blog/2018/02/top-5-github-repositories-january-2018/>)

PRANAV DAR , FEBRUARY 19, 2018



books/)

(<https://www.analyticsvidhya.com/blog/2018/02/10-free-must-read-machine-learning-e-books/>)

10 Free Must-Read Machine Learning E-Books For Data Scientists & AI Engineers

(<https://www.analyticsvidhya.com/blog/2018/02/10-free-must-read-machine-learning-e-books/>)

T
o
p

PRANAV DAR , FEBRUARY 16, 2018



(<https://www.analyticsvidhya.com/blog/2018/02/audio-beat-tracking-for-music-information-retrieval/>)

Learn Audio Beat Tracking for Music Information Retrieval (with Python codes)

(<https://www.analyticsvidhya.com/blog/2018/02/audio-beat-tracking-for-music-information-retrieval/>)

FAIZAN SHAIKH , FEBRUARY 14, 2018



([http://www.edvancer.in/certified-data-scientist-with-python-](http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad)

[course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad](http://www.edvancer.in/certified-data-scientist-with-python-course?utm_source=AV&utm_medium=AVads&utm_campaign=AVadsnonfc&utm_content=pythonavad))

GET CONNECTED



14,454

FOLLOWERS

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



2,574

FOLLOWERS

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



43,299

FOLLOWERS

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



Email

SUBSCRIBE

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



engineering-talent-hunt-hackathon/

talent-hunt-hackathon/?utm_source=AV/home_top)



talent-hunt-hackathon/?utm_source=AV/home_top)

© Copyright 2013-2018 Analytics Vidhya.

Privacy Policy (<https://www.analyticsvidhya.com/privacy-policy/>)

Terms of Use (<https://www.analyticsvidhya.com/terms/>)

Refund Policy (<https://www.analyticsvidhya.com/refund-policy/>)

(<https://datahack.analyticsvidhya.com/contest/data-engineering-talent-hunt-hackathon/>)

DATA SCIENTISTSCOMPANIES

JOIN OUR COMMUNITY :

Blog

Post Jobs

(<https://www.analyticsvidhya.com/blog/>)

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

Hackathon

Trainings

(<https://datahack.analyticsvidhya.com/>)

Our Team

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

Career

(<https://www.analyticsvidhya.com/career-analytics-vidhya/>)

Contact Us

(<https://www.analyticsvidhya.com/contact/>)

Write for us

(<https://www.analyticsvidhya.com/about-me/write/>)

Discussions

(<https://discuss.analyticsvidhya.com/>)

Apply Jobs

(<https://www.analyticsvidhya.com/jobs/>)

Leaderboard

(<https://datahack.analyticsvidhya.com/users/>)

Real-time

(<https://www.analyticsvidhya.com/real-time/>)

(<https://datahack.analyticsvidhya.com/contest/data-science-talent-hunt-hackathon/>)

Followers

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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