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

✓ (https://twitter.com/analyticsvidhya)

8+ (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.analyticsvidhyarppynhytp://wwwianapyhoswytha.competior/zors.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhytp://wwwianapyhoswytha.competior/zors.analyticsvidhyarppynhytp://wwwianapyhoswytha.competior/zors.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhytp://wwwianapyhoswytha.competior/zors.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhytp://wwwianapyhoswytha.competior/zors.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhytp://wwwianapyhoswytha.competior/zors.analyticsvidhyarppynhytp://www.analyticsvidhyarppynhy

Decision Tree - Simplified!

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

SHARE **f** (http://www.facebook.com/sharer.php?u=http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/&t=Decision%20Tree%20-%20Simplified!) **f** (https://twitter.com/home?status=Decision%20Tree%20-%20Simplified!+http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/) **g** (https://plus.google.com/share?url=http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/) **p** (http://pinterest.com/pin/create/button/?url=http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/&media=http://www.analyticsvidhya.com/wp-content/uploads/2015/01/29.jpg&description=Decision%20Tree%20-%20Simplified!)



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

I started working as a business analyst in my previous organisation. I transitioned from a Business Intelligence (BI) Analyst to become a Business Analyst. During the initial days of tenure as a business analyst, I had a bias towards using a classification technique – **DECISION TREE**. This was because of its inherent simplicity and many advantages. We will discuss these in more details later in this article.

Later on, I figured out that decision trees are one of the most commonly used technique among all business analysts. It can not only help us with prediction and classification, but also is a very effective tool to understand the behavior of various variables. In this article, we will discuss this algorithm in detail.

What is a Decision Tree?

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

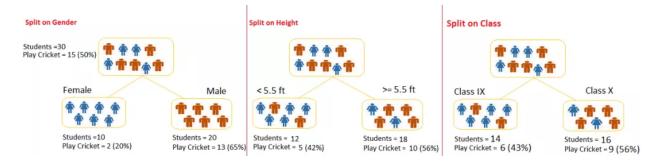


(http://i1.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Decision_Tree_3.png)

Example:-

Let's say we have a sample of 30 students with three variables Gender (Boy/ Girl), Class(IX/ X) and Height (5 to 6 ft). 15 out of these 30 play cricket in leisure time. Now, I want to create a model to predict who will play cricket during leisure period? In this problem, we need to segregate students who play cricket in their leisure time based on highly significant input variable among all three.

This is where decision tree helps, it will segregate the students based on all values of three variable and identify the variable, which creates the best homogeneous sets of students (which are heterogeneous to each other). In the snapshot below, you can see that variable Gender is able to identify best homogeneous sets compared to the other two variables.



(http://io.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Test.png)

As mentioned above, decision tree identifies the most significant variable and it's value that gives best homogeneous sets of population. Now the question which arises is, how does it identify the variable and the split? To do this, decision tree uses various algorithms, which we will discuss in next article.

Types of Decision Tree

Types of decision tree is based on the type of target variable we have. It can be of two types:

- 1. **Binary Variable Decision Tree:** Decision Tree which has binary target variable then it called as Binary Variable Decision Tree. Example:- In above scenario of student problem, where the target variable was "Student will play cricket or not" i.e. YES or NO.
- 2. **Continuous Variable Decision Tree:** Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Example:- Let's say we have a problem to predict whether a customer will pay his renewal premium with an insurance company (yes/no). Here we know that income of customer is a significant variable but insurance company does not have income details for all customers. Now, as we know this is an important variable, then we can build a decision tree to predict customer income based on occupation, product and various other variables. In this case, we are predicting values for continuous variable.

Terminology related to Decision Trees:

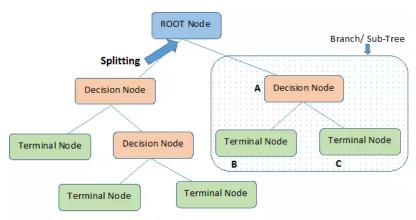
Let's look at the basic terminology used with Decision trees:

ROOT Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.

SPLITTING: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.

Leaf/Terminal Node: Nodes do not split is called Leaf or Terminal node.



Note:- A is parent node of B and C.

(http://i2.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/Decision_Tree_2.png)

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.

Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree

Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

These are the terms commonly used for decision trees. As we know that every algorithm has advantages and disadvantages, below I am discussing some of these for decision trees.

Advantages:

- 1. **Easy to Understand**: Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
- 2. **Useful in Data exploration**: Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. You can refer article (Trick to enhance power of regression model (http://www.analyticsvidhya.com/blog/2013/10/trick-enhance-power-regression-model-2/)) for one such trick. It can also be used in data exploration stage. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
- 3. **Less data cleaning required:** It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
- 4. Data type is not a constraint: It can handle both numerical and categorical variables.
- 5. **Non Parametric Method**: Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Disadvantages:

- 1. **Overfit:** Over fitting is one of the most practical difficulty for decision tree models. This problem gets solved by use of random forests, which we will discuss some other day.
- 2. **Not fit for continuous variables**: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories.
- 3. Not Sensitive to Skewed distributions: Since this is a non-parametric method, this method is not sensitive with skewed distribution.

End Notes:

In this article, we looked at one of the most famous and commonly used technique for predictive models / exploratory analysis. This method is very effective for rapid prototyping of models as well. In our next article, we will look at algorithms behind decision tree, How it splits the population or sample and identify the best split.

- P.S. Have you joined Analytics Vidhya Discuss (http://discuss.analyticsvidhya.com/) yet? If not, you are missing out on awesome data science discussions. Here are some of the discussions happening on modeling techniques:
- 1. Decision tree with continuous variables (http://discuss.analyticsvidhya.com/t/decision-tree-with-continuous-variables/201)
- 2. Datasets to practice modeling techniques (http://discuss.analyticsvidhya.com/t/from-where-i-can-get-data-sets-to-practice-modelling-techniques/155)

If you like what you just read & want to continue your analytics learning, subscribe to our emails

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

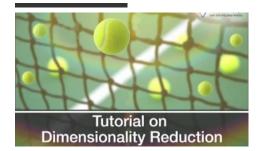
uri=analyticsvidhya), follow us on twitter

(http://twitter.com/analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

Share this:

- in (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?share=linkedin&nb=1) 105
- f (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?share=facebook&nb=1) 56
- G+ (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?share=google-plus-1&nb=1)
- (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?share=twitter&nb=1)
- (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?share=pocket&nb=1)
- (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?share=reddit&nb=1)

RELATED



(http://www.analyticsvidhya.com /blog/2015/07/dimensionreduction-methods/) **Beginners Guide To Learn Dimension Reduction Techniques** (http://www.analyticsvidhya.com /blog/2015/07/dimensionreduction-methods/)

In "Business Analytics"



(http://www.analyticsvidhya.com /blog/2015/09/random-forestalgorithm-multiple-challenges/) Powerful Guide to learn Random Forest (with codes in R & Python) (http://www.analyticsvidhya.com /blog/2015/09/random-forestalgorithm-multiple-challenges/) In "Business Analytics"



(http://www.analyticsvidhya.com /blog/2014/06/introductionrandom-forest-simplified/) Introduction to Random forest -**Simplified** (http://www.analyticsvidhya.com /blog/2014/06/introductionrandom-forest-simplified/) In "Big data"

TAGS: ALGORITHM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ALGORITHM/). CART MODEL (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CART-MODEL/), CHAID MODEL (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CHAID-MODEL/), DECISION TREE (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/DECISION-TREE/), MACHINE LEARNING (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/MACHINE-LEARNING/), ORANGE PYTHON (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ORANGE-PYTHON/)

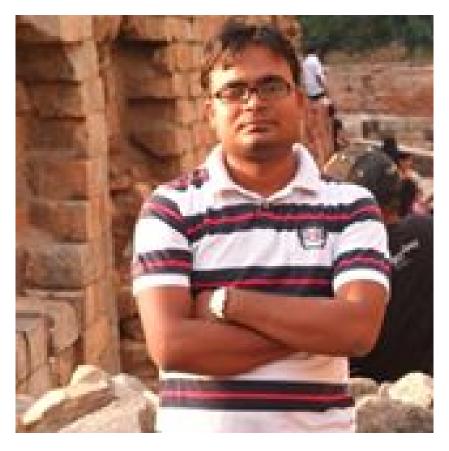
Previous Article

bangalorechennai-7-years-experience/)

Olikview (DAS) - Technical Lead -Mindtree - Bangalore/Chennai (7+ years of experience) (http://www.analyticsvidhya.com/blog/2015/01/qlikviewdelivery-leads-analyttica-bangalore-2-10das-technical-lead-mindtree-

Analytical Delivery Leads - Analyttica -Bangalore (2-10 years of experience) (http://www.analyticsvidhya.com/blog/2015/01/analytica years-experience/)

Next Article



(http://www.analyticsvidhya.com/blog/author/sunil-ray/) **Author**

Sunil Ray (http://www.analyticsvidhya.com/blog/author/sunil-ray/)

I am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years.

4 COMMENTS



Diptesh says: REPLY (HTTP://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?replytocom=68096#respond)
JANUARY 16, 2015 AT 6:41 AM (HTTP://www.analyticsvidhya.com/blog/2015/01/decision-treeSIMPLIFIED/#COMMENT-68096)

Great article.



Samarth says: REPLY (HTTP://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?replytocom=68425#respond)
JANUARY 16, 2015 AT 5:39 PM (HTTP://www.analyticsvidhya.com/blog/2015/01/decision-treeSIMPLIFIED/#COMMENT-68425)

While Decision Trees are prone to overfitting, the way to solve that is by using reducederror post pruning (or by setting a depth / number of nodes limit).

Random Decision Forests belong to an altogether different class of algorithms, ensemble methods to be precise and to say overfitting is solved by quoting another algorithm is, imho, incorrect.

Also, Decision Trees are very very prone to outliers and noise, unless pruning is performed.

Just wanted to point that out. Great article...

Cheers.



C.S. Ganit says:REPLY (HTTP://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/?replytocom=68669#respond)
JANUARY 17, 2015 AT 4:15 AM (HTTP://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/#comment-68669)

An excellent starting point , very clear and in simple temple explained the concepts rather intuitively .

Decision Tree Algorithms - Simplified (http://www.analyticsvidhya.com/blog/2015/01/decision-tree-algorithms-simplified/) says:

JANUARY 22, 2015 AT 5:59 PM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/01/DECISION-TREE-SIMPLIFIED/#COMMENT-71842) [...] last article, we looked at the basics of Decision tree and how it helps in classifications. We also looked at advantages and disadvantages of using [...]

LEAVE A REPLY

Connect with:



(http://www.analyticsvidhya.com/wp-login.php?

 $action=wordpress_social_authenticate\&mode=login\&provider=Facebook\&redirect_to=http\%3A\%2F\%2F\ tree-simplified\%2F)$

Your email address will not be published.

Comment	
Name (required)	
Email (required)	
Website	
Notify me of follow-up comments by email.	SUBMIT COMMENT
Notify me of new posts by email.	

TOP USERS

Rank Name Points

Nalin Pasricha

1	(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)



(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-datascience/)
- A Complete Tutorial to Learn Data Science with Python from Scratch

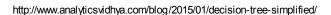
(http://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)

- 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/)

Decision Tree Forests

DTREG creates forests of decision trees to model and predict data







(http://imarticus.org/programs/business-analytics-

professional/)

RECENT POSTS



(http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-

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-

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-

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?

utm_source=AV&utm_medium=AVads&utm_campaign=AVads1&utm_content=cbapavad)

GET CONNECTED



4,159

FOLLOWERS

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

g+

915

FOLLOWERS

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

f

11,950

FOLLOWERS

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

3

Email

SI IRSCDIRE

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



uri=analyticsvidhya)

(http://www.analyticsvidhya.com/blog/2016/02/quick-

FULL TO BE SECOND POUL 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. insights-analytics-big-data-salary-report-2016/)

STAY CONNECTED



4,159

FOLLOWERS

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



915

FOLLOWERS

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

LATEST POSTS



FOLLOWERS

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



SUBSCRIBE

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



(http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-

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-

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-

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	REV I	IEWS
-----	--------------	-------------

© Copyright 2015 Analytics Vidhya