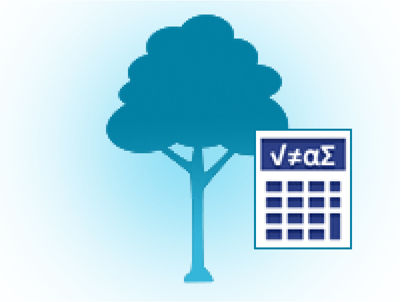
Random Forest

Model for classification and regression based on a forest of trees using random inputs.



Product: IBM® SPSS® Modeler

Extension type: Model

Table of Contents

Description

This is an SPSS® Modeler 'model' node for classification and regression based on a forest of trees using random inputs, utilizing conditional inference trees as base learners. Simply install the node, choose the target and predictors and specify additional settings.

Requirements

- SPSS Modeler v16.0 or later
- SPSS Modeler 'R Essentials' plugin
- R v2.15.x

Installation

Close SPSS Modeler. Save the *.cfe* file in the CDB folder, located by default on Windows in "*C:\ProgramData\IBM\SPSS\Modeler\16\CDB*" or under your modeler 16 installation directory.

Restart SPSS Modeler: the node will now appear in the Model palette.

R Packages used

The R packages will be installed the first time the node is used as long as an Internet connection is available.

- 'party' by Torsten Hothorn [aut, cre], Kurt Hornik [aut], Carolin Strobl [aut], Achim Zeileis [aut]

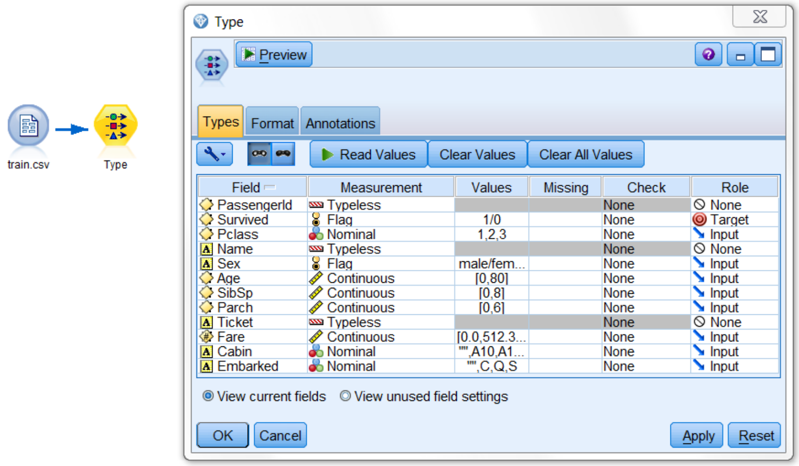
  http://cran.r-project.org/web/packages/party/

Result example

For this tutorial, we will predict the survival probability of the Titanic passengers; the data is available on the popular analytics competitions website http://kaggle.com/

1. Create a ‘Var. File’ source node. Select the *‘train.csv’* file that contains training data for the model to learn. Be careful: Select ‘comma’ and ‘newline’ as field delimeters. Select ‘Pair and discard’ for double quotes. Click on preview to verify the data fills the columns normally.

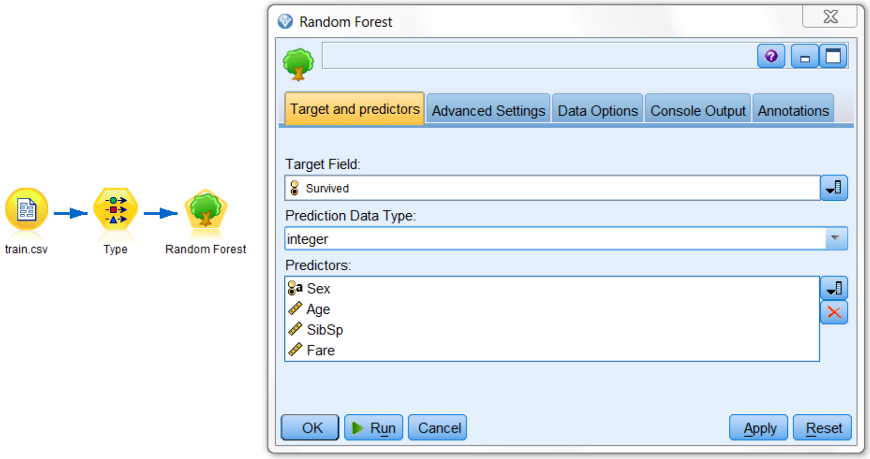
Add a type node and select the type as follow:



Field	Measurement	Values	Missing	Check	Role
PassengerId	Typeless			None	None
Survived	Flag	1/0		None	Target
Pclass	Nominal	1,2,3		None	Input
Name	Typeless			None	None
Sex	Flag	male/fem...		None	Input
Age	Continuous	[0,80]		None	Input
SibSp	Continuous	[0,8]		None	Input
Parch	Continuous	[0,6]		None	Input
Ticket	Typeless			None	None
Fare	Continuous	[0.0,512.3...		None	Input
Cabin	Nominal	"",A10,A1...		None	Input
Embarked	Nominal	"",C,Q,S		None	Input

● View current fields ○ View unused field settings

2. Add a ‘Random Forest’ node from the model palette. Then double-click on it:



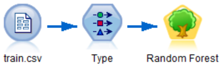
Target Field:
Survived

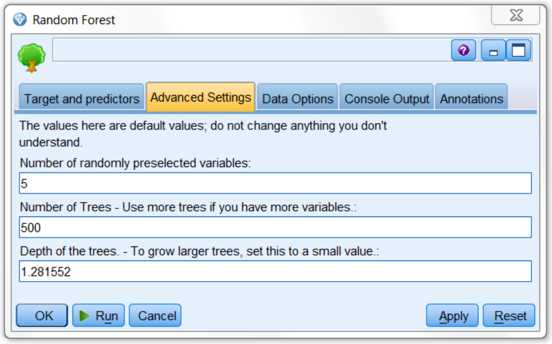
Prediction Data Type:
integer

Predictors:
Sex
Age
SibSp
Fare

Choose a Target, the supposed type of the prediction and the predictors. Here you want to predict if the passenger survived or not, so the target is ‘Survived’. The predictors should be for example: Sex, Age, SbSp (number of sibillings and spouses), Fare ($).

Then you can specify additionnal settings:





The values here are default values; do not change anything you don't understand.

Number of randomly preselected variables: 5

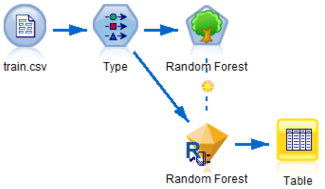
Number of Trees - Use more trees if you have more variables.: 500

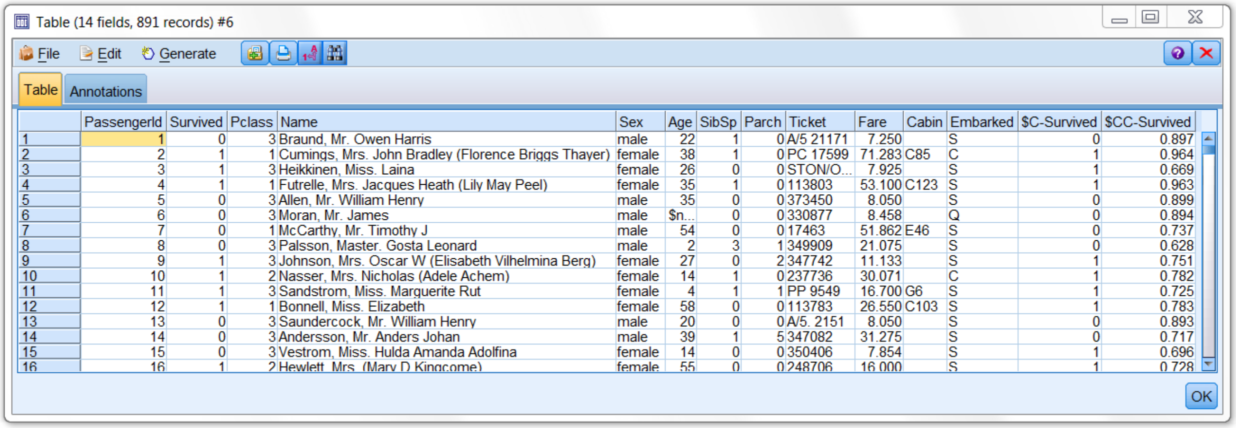
Depth of the trees. - To grow larger trees, set this to a small value.: 1.281552

- Number of randomly preselected variables. Default: 5
- Number of trees. The more variables, the more trees you should use. Default: 500
- Depth of the trees. To grow larger trees, set this to a small value. Default: 1.281552

3. Select the Random Forest node and run the stream. A golden nugget should appear now.

Select it and add a table node from the output palette. Then run again the stream.

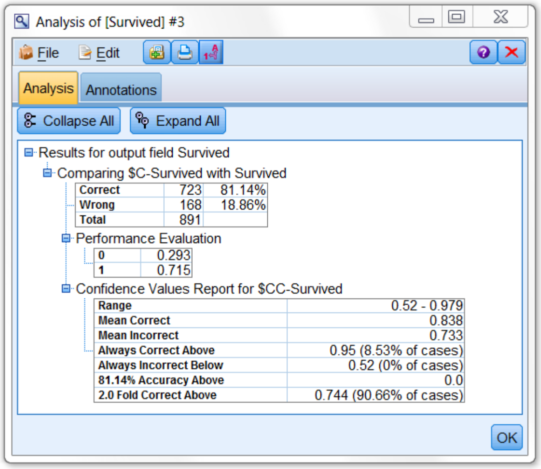




	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	$C-Survived	$CC-Survived
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.250		S	0	0.897
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.283	C85	C	1	0.964
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O...	7.925		S	1	0.669
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.100	C123	S	1	0.963
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.050		S	0	0.899
6	6	0	3	Moran, Mr. James	male	$n...	0	0	330877	8.458		Q	0	0.894
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.862	E46	S	0	0.737
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S	0	0.628
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.133		S	1	0.751
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.071		C	1	0.782
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.700	G6	S	1	0.725
12	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.550	C103	S	1	0.783
13	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.050		S	0	0.893
14	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S	0	0.717
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.854		S	1	0.696
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16.000		S	1	0.728

Two new columns appear: $C-Survived (the prediction); $CC-Survived (the probability it’s true).

You can evaluate the accuracy of the model by adding an ‘Analysis’ node after the nugget and running the stream again:



Analysis of [Survived] #3

Results for output field Survived

Comparing $C-Survived with Survived

Correct	723	81.14%
Wrong	168	18.86%
Total	891	

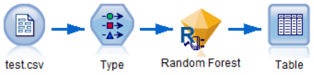
Performance Evaluation

0	0.293
1	0.715

Confidence Values Report for $CC-Survived

Range	0.52 - 0.979
Mean Correct	0.838
Mean Incorrect	0.733
Always Correct Above	0.95 (8.53% of cases)
Always Incorrect Below	0.52 (0% of cases)
81.14% Accuracy Above	0.0
2.0 Fold Correct Above	0.744 (90.66% of cases)

4. Now you can use your model on other data to predict if the passenger survived. Re-do the step one, this time with *‘test.csv’*. Then add the golden ‘Random Forest’ nugget from the ‘Models’ tab on the top-right of the screen. Add a table; then run the stream.



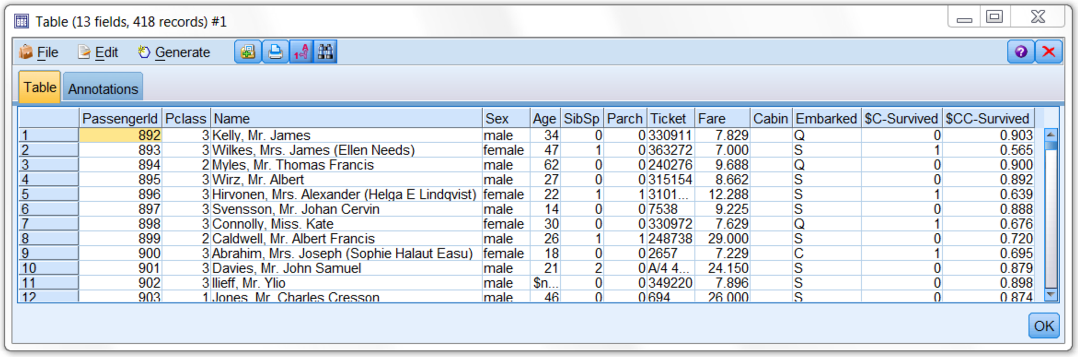


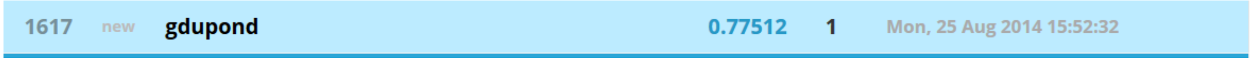
Table (13 fields, 418 records) #1

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	$C-Survived	$CC-Survived
1	892	3	Kelly, Mr. James	male	34	0	0	330911	7.829		Q	0	0.903
2	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7.000		S	1	0.565
3	894	2	Myles, Mr. Thomas Francis	male	62	0	0	240276	9.688		Q	0	0.900
4	895	3	Wirz, Mr. Albert	male	27	0	0	315154	8.662		S	0	0.892
5	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22	1	1	3101...	12.288		S	1	0.639
6	897	3	Svensson, Mr. Johan Cervin	male	14	0	0	7538	9.225		S	0	0.888
7	898	3	Connolly, Miss. Kate	female	30	0	0	330972	7.629		Q	1	0.676
8	899	2	Caldwell, Mr. Albert Francis	male	26	1	1	248738	29.000		S	0	0.720
9	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18	0	0	2657	7.229		C	1	0.695
10	901	3	Davies, Mr. John Samuel	male	21	2	0	A/4 4...	24.150		S	0	0.879
11	902	3	Ilieff, Mr. Ylio	male	$n...	0	0	349220	7.896		S	0	0.898
12	903	1	Jones, Mr. Charles Cresson	male	46	0	0	694	26.000		S	0	0.874

Thanks to your model you predicted the survival of the other passengers.

5 – bonus:

You can now send your results to Kaggle and see if the predictions were true. We use a filler node from the field ops palette to do this; and a Flat File export node to get the data as a csv.



1617	new	gdupond	0.77512	1	Mon, 25 Aug 2014 15:52:32

You see how easily you can get good results with only a few nodes. With further data preparation you can expect to get a greater rank.

Important links

Learn

- Learn more about SPSS software.
- Visit developerWorks Business analytics for more technical analytics resources for developers.
- The Comprehensive R Archive Network is the main site for the R project and each R package. The help pages and manuals that are associated with `optimx`, `nlmrt`, and `Rcgmin` are detailed. Numerous references are provided.
- Read "Do I need to learn R?" (Catherine Dalzell, developerWorks, September 2013) to learn why R is a valuable tool for data analytics that was expressly designed to reflect the way that statisticians think and work.
- "Calling R from SPSS" describes how to use R code inside IBM SPSS Modeler 16.
- Read "Using Google maps API" to discover how to use Google Maps API with R.
- Read “Create new nodes for IBM SPSS Modeler 16 using R” to learn how to create new extensions easily.

Get products and technologies

- Download the R plug-in for SPSS plugin.
- Download the R 2.15.2 for Windows package.

Discuss

- Visit the IBM SPSS DevCentral developerWorks community to share tips and experiences with other IBM SPSS developers.
- Follow developerWorks on Twitter to be among the first to hear about new resources.