SAS AND OPEN SOURCE – A MATCH MADE IN ENTERPRISE MINER

TORONTO DATA MINING USER GROUP

AGENDA | SAS AND OPEN SOURCE

- Open Source analytics in Business
- Open Source Integration Node
- Output modes
- Workflow examples to incorporate R models
- Careful considerations

OPEN SOURCE INTEGRATION

THIS IS ACHIEVED WITH SAS ANALYTICS IN ACTION

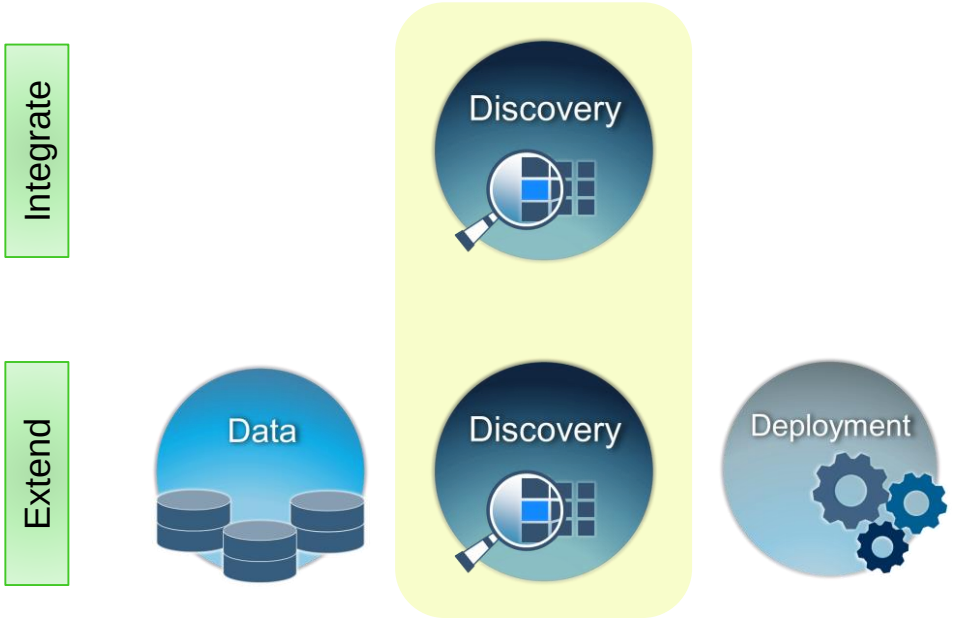
SAS ANALYTICS IN ACTION =

- Data is about gathering data from the different data sources and locations, unifying it and making it ready for modeling
- Discovery is about having the flexibility to prototype analytical models to uncover business value
- Deployment is about engineering enterprise level solutions from those prototypes with governance measures to ensure quality

OPEN SOURCE INTEGRATION

SAS DOES IT BY “INTEGRATING” AND “EXTENDING” IT

Where do we integrate? Where do we extend?



THE OPEN SOURCE INTEGRATION NODE

- Enables the **execution of R code** within an Enterprise Miner workflow.
- Transfers data, metadata, and results automatically between Enterprise Miner and R

THE OPEN SOURCE INTEGRATION NODE

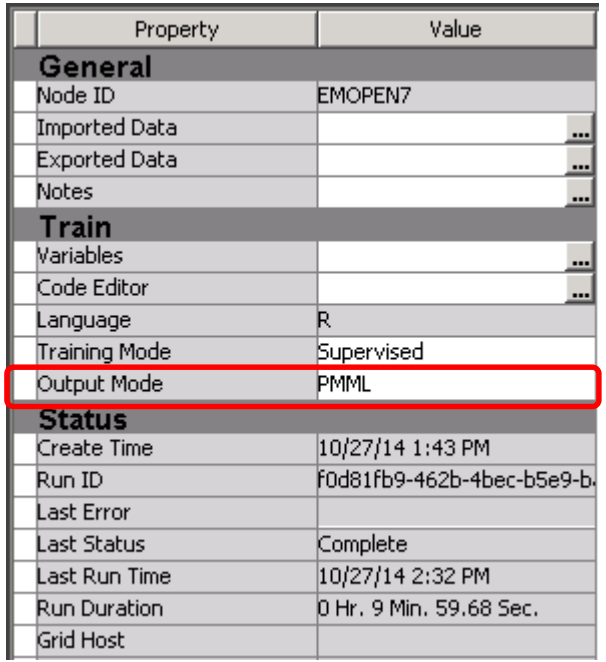
- Facilitates **multitasking** in R
- Generates **text and graphical output** from R
- Integrates both **supervised and unsupervised** learning tasks

PMML OUTPUT

Predictive modeling markup language (PMML) is an open standard enabling certain R models to be **translated into SAS DATA step code**

Currently supported R models include:

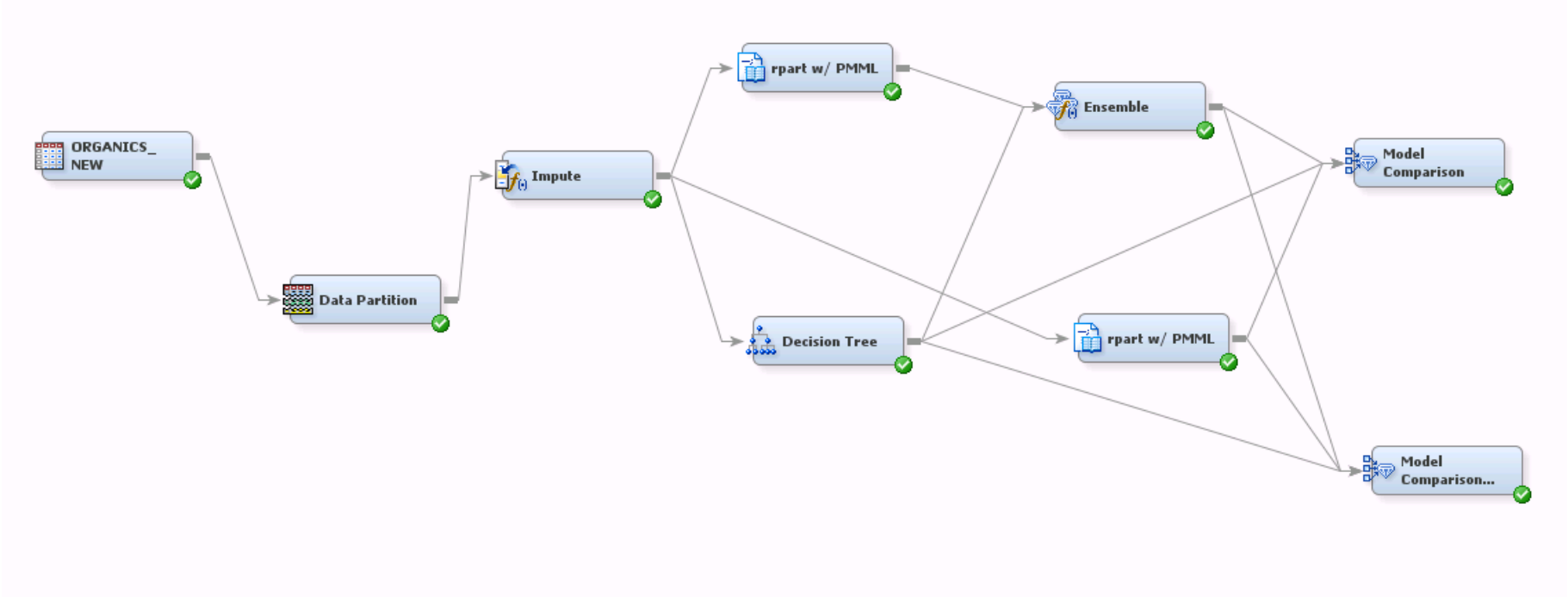
- Linear Models (lm)
- Multinomial Log-Linear Models (multinom (nnet))
- Generalized Linear Models (glm (stats))
- Decision Trees (rpart)
- Neural Networks (nnet)
- k-means Clustering (kmeans (stats))



Property	Value
General	
Node ID	EMOPEN7
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Code Editor	...
Language	R
Training Mode	Supervised
Output Mode	PMML
Status	
Create Time	10/27/14 1:43 PM
Run ID	f0d81fb9-462b-4bec-b5e9-b...
Last Error	
Last Status	Complete
Last Run Time	10/27/14 2:32 PM
Run Duration	0 Hr. 9 Min. 59.68 Sec.
Grid Host	

USING R IN SAS ENTERPRISE MINER

PMML MODE



Enterprise Miner - EM_R_Demo

File Edit View Actions Options Window Help

EM_R_Demo
- Data Sources
- Diagrams
  - Organics_R_Merge
  - Organics_R_PMML
  - Organics_SAS_R
- Model Packages

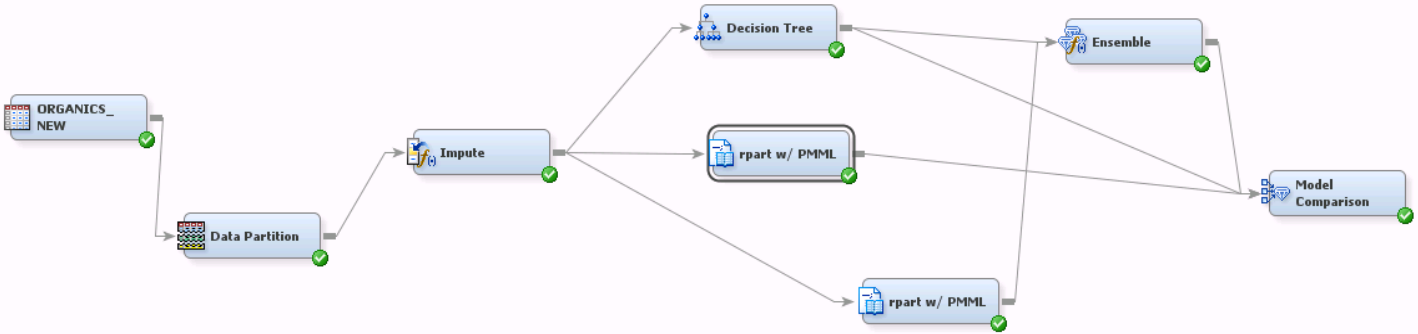
Property	Value
General	
Node ID	EMOPEN2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Code Editor	
Language	R
Training Mode	Supervised
Output Mode	PMML
Status	
Create Time	5/2/16 10:01 AM
Run ID	625ff050-7099-4efd-acee-ae
Last Error	
Last Status	Complete
Last Run Time	5/2/16 10:49 AM
Run Duration	0 Hr. 8 Min. 58.71 Sec.
Grid Host	
User-Added Node	No

General

General Properties

Sample | Explore | Modify | Model | Assess | Utility | Credit Scoring | HPDM | Applications | Text Mining | Time Series

Organics_R_PMML



Code Editor

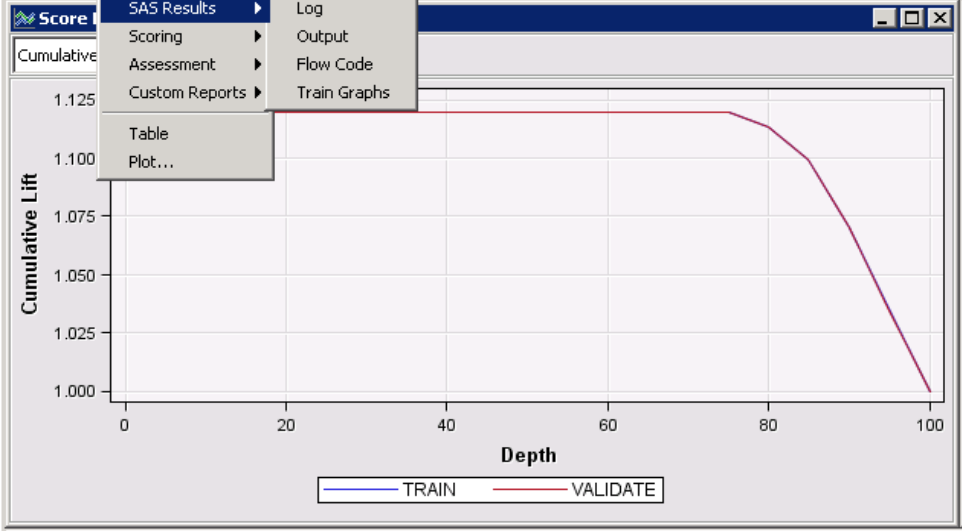
```
library(rpart)
library(rpart.plot)

&EMR_MODEL <- rpart(&EMR_CLASS_TARGET ~ &EMR_CLASS_INPUT + &EMR_NUM_INPUT, data= &EMR_IMPORT_DATA, method= "class")

printcp(&EMR_MODEL) # display the results
plotcp(&EMR_MODEL) # visualize cross-validation results
summary(&EMR_MODEL) # detailed summary of splits



png("Classification_Tree.png")
split.fun <- function(x, labs, digits, varlen, faclen)
{
    # replace commas with spaces (needed for strwrap)
    labs <- gsub(",", " ", labs)
    for(i in 1:length(labs)) {
        # split labs[i] into multiple lines
        labs[i] <- paste(strwrap(labs[i], width=20), collapse="\n")
    }
    labs
}

prp(&EMR_MODEL, type=2, extra=100, tweak=2.0, faclen=2, compress=TRUE, ycompress=TRUE, split.fun=split.fun)

#plot(&EMR_MODEL, uniform=TRUE, compress=TRUE, main="Classification_Tree")#
#text(&EMR_MODEL, use.n=TRUE)#
dev.off()
```

Results - Node: rpart w/ PMML Diagram: Organics_R_PMML

File Edit View Window

Properties
SAS Results
Scoring
Assessment
Custom Reports
Table
Plot...

Log
Output
Flow Code
Train Graphs



Output

```
1   *------------------------------------------------------------*
2   User:                  sasdemo
3   Date:                  May 02, 2016
4   Time:                  10:58:55
5   *------------------------------------------------------------*
6   * Training Output
7   *------------------------------------------------------------*
8
9
10
11
12  Variable Summary
13
14                Measurement    Frequency
15  Role            Level         Count
16
17  ID             INTERVAL           1
18  ID             NOMINAL            1
```

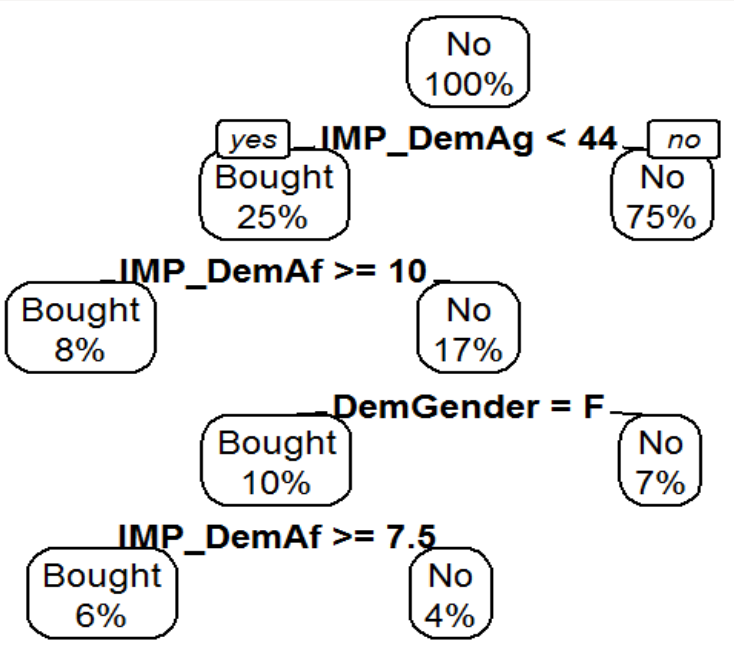
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetBuy	Organics Purchase	_ASE_	Average Squared Error	0.150916	0.151178	.
TargetBuy	Organics Purchase	_DIV_	Divisor for ASE	2026734	1351162	.
TargetBuy	Organics Purchase	_MAX_	Maximum Absolute Error	0.842275	0.842275	.
TargetBuy	Organics Purchase	_NOBS_	Sum of Frequencies	1013367	675581	.
TargetBuy	Organics Purchase	_RASE_	Root Average Squared Error	0.388479	0.388816	.
TargetBuy	Organics Purchase	_SSE_	Sum of Squared Errors	305867.1	204265.6	.
TargetBuy	Organics Purchase	_DISF_	Frequency of Classified Cases	1013367	675581	.
TargetBuy	Organics Purchase	_MISC_	Misclassification Rate	0.193598	0.194011	.
TargetBuy	Organics Purchase	_WRONG_	Number of Wrong Classificati...	196186	131070	.

Results - Node: rpart w/ PMML Diagram: Organics_R_PMML

File Edit View Window

Train Graphs



sasdemo
May 02, 2016
10:58:55

g Output

Summary

Measurement Level	Frequency Count
INTERVAL	1
NOMINAL	1

	Validation	Test
	0.150916	0.151178
	2026734	1351162
	0.842275	0.842275
	1013367	675581
	0.388479	0.388816
	305867.1	204265.6
	1013367	675581
	0.193598	0.194011
	196186	131070

MERGE OUTPUT MODE

Merge output mode enables integration with thousands of R packages that are not supported in PMML output mode.

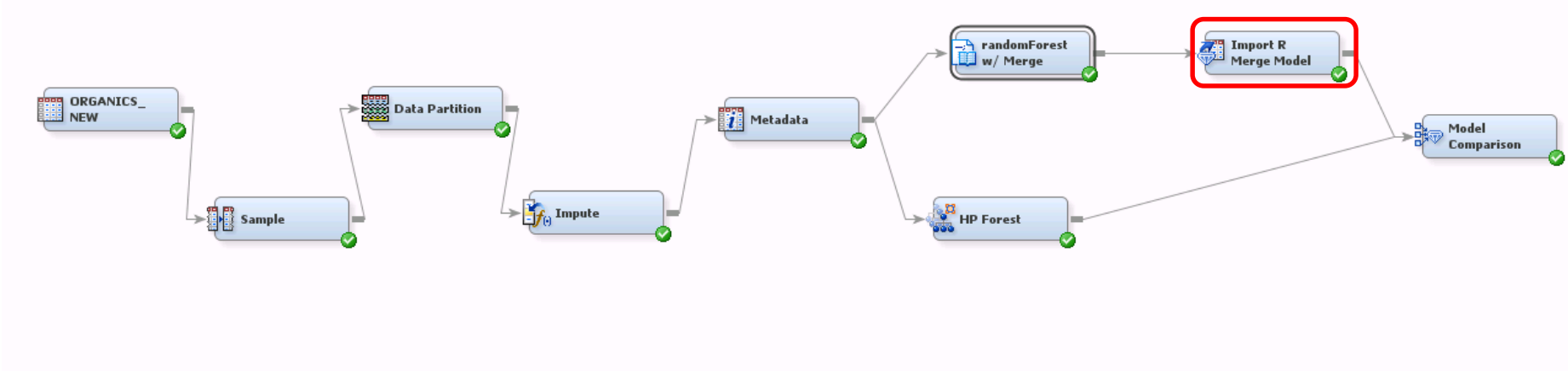
Variables created in R are merged with SAS Enterprise Miner data sources **by the user**.

SAS DATA step code is not created.

Property	Value
General	
Node ID	EMOPEN
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Code Editor	...
Language	R
Training Mode	Supervised
Output Mode	Merge
Status	
Create Time	4/29/16 11:43 AM
Run ID	4818815a-9824-4244-894a-a
Last Error	
Last Status	Complete
Last Run Time	4/29/16 11:52 AM
Run Duration	0 Hr. 1 Min. 16.02 Sec.
Grid Host	
User-Added Node	No

USING R IN SAS ENTERPRISE MINER

MERGE MODE



Enterprise Miner - EM_R_Demo

File Edit View Actions Options Window Help

EM_R_Demo
- Data Sources
- Diagrams
  - Organics_R_Merge
  - Organics_R_PMML
  - Organics_SAS_R
- Model Packages

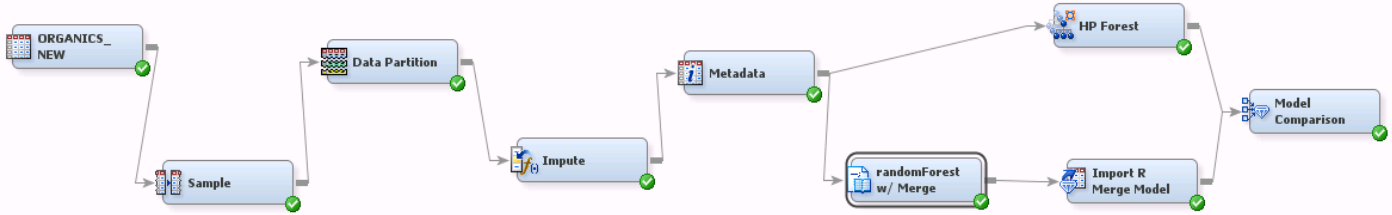
Property	Value
General	
Node ID	EMOPEN
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Code Editor	
Language	R
Training Mode	Supervised
Output Mode	Merge
Status	
Create Time	5/2/16 9:46 AM
Run ID	d1680455-37b0-40f5-9b1a-c
Last Error	
Last Status	Complete
Last Run Time	5/2/16 10:03 AM
Run Duration	0 Hr. 0 Min. 51.93 Sec.
Grid Host	
User-Added Node	No

General

General Properties

Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series

Organics_R_Merge



Code Editor

```
library(randomForest)

&EMR_MODEL <- randomForest(&EMR_CLASS_TARGET ~  &EMR_CLASS_INPUT + &EMR_NUM_INPUT, ntree= 500, mtry= 5, data= &EMR_IMPORT_DATA, importance= TRUE)

&EMR_EXPORT_TRAIN <- predict(&EMR_MODEL, &EMR_IMPORT_DATA, type="prob")
&EMR_EXPORT_VALIDATE <- predict(&EMR_MODEL, &EMR_IMPORT_VALIDATE, type="prob")

&EMR_EXPORT_TRAIN[1:10,]

png("EMR_forestMsePlot.png")
plot(&EMR_MODEL, main= 'randomForest MSE Plot')
dev.off()

write.table(round(importance(&EMR_MODEL),2), file = "EMR_forestImportance.csv", sep=",", row.names = TRUE, col.names = TRUE)
```

SOME PRECAUTIONS

Some items to consider when running R models in Open Source note:

- Missing Values may be an issue
- Ensure Categorical Variables are not high in cardinality
- Memory issues

ADDITIONAL RESOURCES

SAS Webinar: How SAS Adds Value to Open Source

Video: SAS Enterprise Miner and R

Whitepaper: The Use of Open Source is Growing. So Why Do Organizations Still Turn to SAS?

Whitepaper: SAS Analytics and Open Source

QUESTIONS