SAS Grid:
Above and Beyond Open-Source

Mark Ezzo

Data Scientist/

VINCI SAS Administrator

April 11, 2019

Poll Question

- What is your primary usage of the SAS Grid?
- A – ETL
- B – Data Preparation (SQL & SAS)
- C – Analysis (including Modeling)
- D - Other

Poll Question

- What is your primary usage of Open-Source?
- A – ETL
- B – Data Preparation
- C – Analysis (including Modeling)
- D - Other

Enterprise vs. Open-Source

- **Open source software** (OSS) refers to the software which uses the code freely available on the Internet. The code can be copied, modified or deleted by other users and organizations. As the software is open to the public, the result is that it constantly updates, improves and expands as more people can work on its improvement.
- **Closed source software** (CSS) is opposite to OSS and means the software which uses the proprietary and closely guarded code. Only the original authors of software can access, copy, and alter that software. In case with closed source software, you are not purchasing the software, but only pay to use it.

Comparison

- Here is a comparison of five basic aspects: pricing, security, support, source availability, and usability.
- **#1 Price Policy**
- Open source often referred as a free of cost software. It can, however, have costs for extras like assistance, additional services or added functionality. Thus, you may still pay for a service with OSS.
- Closed source software is usually a paid software. The costs can vary depending on the complexity of the software. While the price can be higher, what you get is a better product, full support, functionality and innovation. However, most companies provide free trials to convince the purchaser that their software is the right fit.

Comparison (continued)

- **#2 Security**
- The question of security is very controversial as each software has two sides of the coin. The code of open source software can be viewed, shared and modified by community, which means anyone can fix, upgrade and test the broken code. The bugs are fixed quickly, and the code is checked thoroughly after each release. However, because of availability, the source code is open for hackers to practice on.
- On the contrary, closed source software can be fixed only by a vendor, this is rare. If something goes wrong with the software, you send a request and wait for the answer from the support team. Solving the problem can take much longer than compared to OSC.
- When it comes to choosing the most secure software, the answer is that each of them has its pros and cons. Thus, it is often a challenge for firms which work in particular industry.

Comparison (continued)

- **#3 Quality of Support**
- Comparing open source and closed source software support, it is obvious that CSS is predominant in this case. The costs for it include an option to contact support and get it in one business day in most cases. The response is well organized and documented.
- For open source software such option is not provided. The only support options are forums, useful articles and hired expert. However, it is not surprising that using such kind of service you will not receive a high level of response.
- **#4 Source Code Availability**
- Open source software provides an ability to change the source code without any restrictions. Individual users can develop what they want and get benefits from innovation developed by others within the user community. As the source code is easily accessible, it enables the software developers to improve the already existing programs.
- Closed source software is more restricted than open source software because the source code cannot be changed or viewed. However, such limitation is what may contribute to CSS security and reliability.

Comparison (continued)

- **#5 Usability**
- Usability is a painful subject of open source software. User guides are written for developers rather than to layperson users. Also, these manuals are failing to conform to the standards and structure.
- For closed source software usability is one of the merits. Documentation is usually well-written and contains detailed instructions.

Enterprise Advantages

- **Companies can meet their engineering needs within minutes, and Development teams never need to follow convoluted procurement cycles.**
- **Enterprise software improves efficiency and offers strong security features that many industries consider essential to their operations.**
- **Flexible licensing systems combine multiple licensing arrangements under single umbrella contracts.**
- **Finding, testing, verifying and hardening open source software entails many unanticipated costs for maintenance, internal development and generating user-friendly support processes. Enterprise software solutions eliminate these worries and expenses, and you have a support team ready to take your call and provide answers or fixes when needed.**
- **To qualify as truly enterprise-grade, databases need to ensure ACID transactions. The ACID acronym refers to atomicity, consistency, isolation and durability, and transaction processing needs to satisfy its guidelines.**
  - **Atomicity**: Data changes need to function as single transactions, causing simultaneous updates in related databases.
  - **Consistency**: Zero-sum totals must balance after inputting data updates.
  - **Isolation**: Transaction processes remain invisible until posted simultaneously in all related databases.
  - **Durability**: Changes in databases persist even after catastrophic system failures.

Poll Question

- Which venue is your preference?
- A – Enterprise (Closed-Source)
- B – Open-Source
- C - Combination
- D – No Preference
- E – No Opinion

Poll Question

- If you chose Enterprise (Closed-Source), please pick the primary reason?
- A – Pricing
- B – Security
- C – Support
- D – Source Code Availability
- E – Usability (Documentation)
- F – Other

Poll Question

- If you chose Open-Source, please pick the primary reason?
- A – Pricing
- B – Security
- C – Support
- D – Source Code Availability
- E – Usability (Documentation)
- F – Other

Poll Question

- If you chose Combination, please pick the primary reason?
- A – Pricing
- B – Security
- C – Support
- D – Source Code Availability
- E – Usability (Documentation)
- F – Other

Public Code Repositories

- Both SAS and Open-Source have these code repositories for about anything that you can desire, here are some examples:
- SAS: https://github.com/trending/sas
- R: https://r-forge.r-project.org
- Python: https://pythontips.com/2013/07/30/20-python-libraries-you-cant-live-without/
- Et Al…….

Our Current SAS Grid

Capability	Why it Matters
Centralized/Shared Workload Management	Effectively manage jobs and users, prioritization, compliance/auditing
High Availability	Avoid user or service disruption, rolling maintenance
Distributed Processing	Improved performance, meet changing demands, grow incrementally
Leverage Commodity Hardware	Reduce Costs

Call R Packages from PROC IML

- You do not need to do anything special to call an R package. Provided that an R package is installed, you can call library(package) from inside a SUBMIT block to load the package. You can then call the functions in the package.
- The example in this section calls an R package and imports the results into a SAS data set.
  - Define the data and transfer the data to R.
  - Call R functions to analyze the data.
  - Transfer the results of the analysis into SAS/IML vectors.

http://support.sas.com/documentation/cdl/en/imlug/65547/HTML/default/viewer.htm#imlug_r_sect012.htm

Call R Packages from PROC IML – Example A

1. Define the data in the SAS/IML vector q and then transfer the data to R by using the ExportMatrixToR subroutine. In R, the data are stored in a vector named rq.

```
proc iml;
q = {3.7, 7.1, 2, 4.2, 5.3, 6.4, 8, 5.7, 3.1, 6.1, 4.4, 5.4, 9.5, 11.2};
RVar = "rq";
call ExportMatrixToR( q, RVar );
```

2. Load the KernSmooth package. Because the functions in the KernSmooth package do not handle missing values, the nonmissing values in q must be copied to a matrix p. (There are no missing values in this example.) The Sheather-Jones plug-in bandwidth is computed by calling the dpik function in the KernSmooth package. This bandwidth is used in the bkde function (in the same package) to compute a kernel density estimate.

```
submit RVar / R;
   library(KernSmooth)
   idx <-which(!is.na(&RVar))        # must exclude missing values (NA)
   p <- &RVar[idx]                   #    from KernSmooth functions
   h = dpik(p)                       # Sheather-Jones plug-in bandwidth
   est <- bkde(p, bandwidth=h)       # est has 2 columns
endsubmit;
```

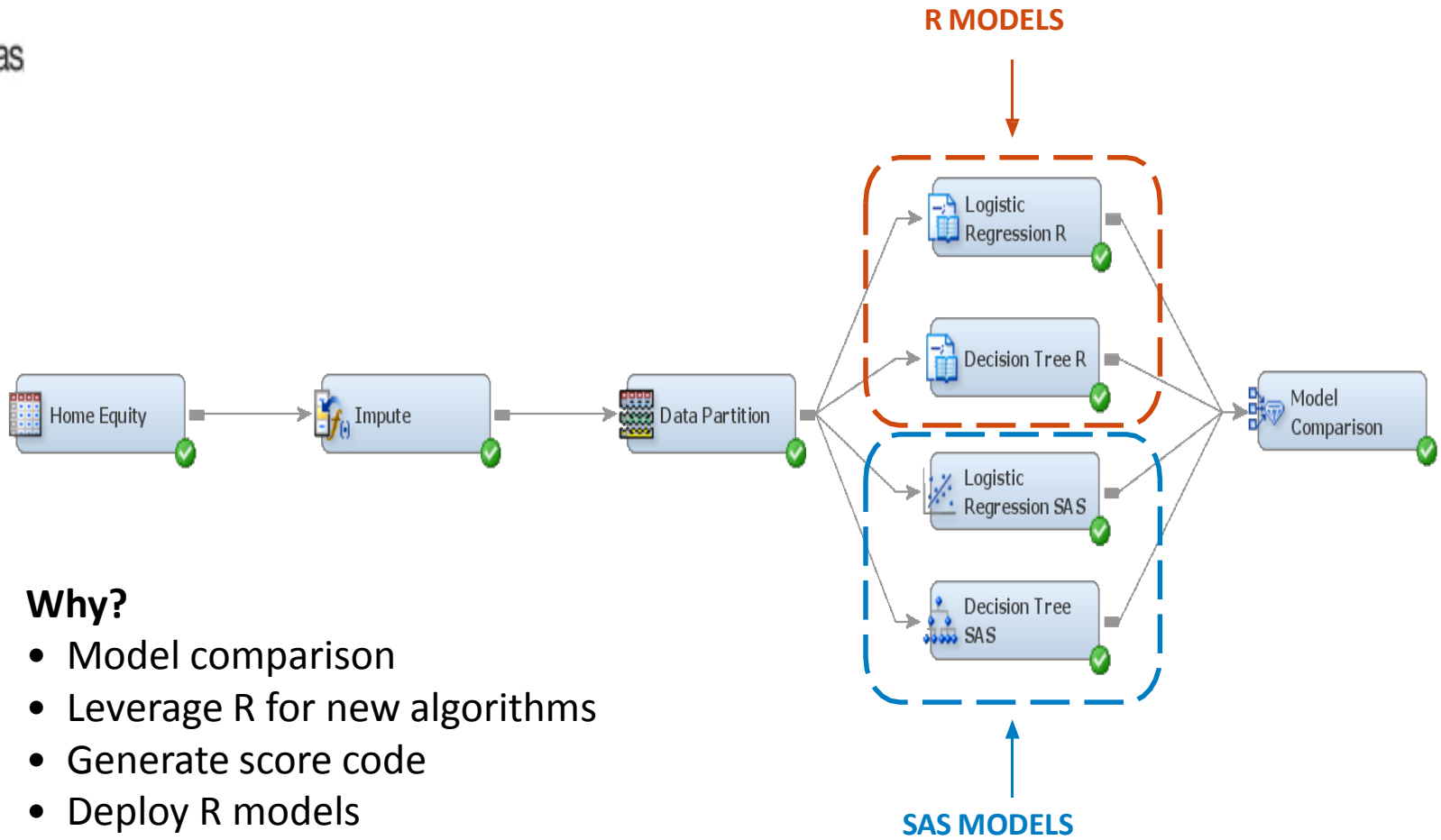
3. Copy the results into a SAS data set or a SAS/IML matrix, and perform additional computations. For example, the following statements use the trapezoidal rule to numerically estimate the density that is contained in the tail of the density estimate of the data:

```
call ImportMatrixFromR( m, "est" );
/* estimate the density for q >= 8 */
x = m[,1];                /* x values for density */
idx = loc( x>=8 );        /* find values x >= 8 */
y = m[idx, 2];            /* extract corresponding density values */

/* Use the trapezoidal rule to estimate the area under the density curve.
   The area of a trapezoid with base w and heights h1 and h2 is
   w*(h1+h2)/2. */
w = m[2,1] - m[1,1];
h1 = y[1:nrow(y)-1];
h2 = y[2:nrow(y)];
Area = w * sum(h1+h2) / 2;
print Area;
```

http://support.sas.com/documentation/cdl/en/imlug/65547/HTML/default/viewer.htm#imlug_r_sect012.htm

USE SAS TO INTEGRATE R



Why?

- Model comparison
- Leverage R for new algorithms
- Generate score code
- Deploy R models

SAS FROM R

RStudio

File Edit Code View Plots Session Build Debug Tools Help

RcallingSAS.R*

```
library("RCurl")
tempDir <- tempfile()
dir.create(tempDir)

myhtml <- getURL(url="http://joeloonix-sasbiws2.na.sas.com:7980/SASBIWS/rest/storedProcesses/CABdemo/RandomForest/da
                 httpheader=c(Accept="application/xml,text/html",
                              'Content-Type' = "application/xml", Authorization="Basic c2FzZGVtbzpwYXNzd29yZA==="),
                 postfields="<RandomForest><parameters><dataset>shoptrain</dataset><numtrees>100</numtrees><varstotr
                 verbose = TRUE)

write(myhtml, file.path(tempDir, "sasout.html"))
rstudio::viewer(file.path(tempDir, "sasout.html"))
```

Console

```
Content-Length: 131

* upload completely sent off: 131 out of 131 bytes
< HTTP/1.1 200 OK
< Date: Tue, 26 May 2015 13:05:29 GMT
< Server: Apache-Coyote/1.1
< Content-Type: text/html;charset=utf-8
< Transfer-Encoding: chunked
<
* Connection #0 to host joeloonix-sasbiws2.na.sas.com left intact
> source('~/.active-rstudio-document')
* Hostname was NOT found in DNS cache
*   Trying 10.12.38.157...
* Connected to joeloonix-sasbiws2.na.sas.com (10.12.38.157) port 7980 (#0)
> POST /SASBIWS/rest/storedProcesses/CABdemo/RandomForest/dataTargets/_WEBOUT HTTP/1.1
Host: joeloonix-sasbiws2.na.sas.com:7980
Accept: application/xml,text/html
Content-Type: application/xml
Authorization: Basic c2FzZGVtbzpwYXNzd29yZA==
Content-Length: 131

* upload completely sent off: 131 out of 131 bytes
< HTTP/1.1 200 OK
< Date: Tue, 26 May 2015 14:29:25 GMT
< Server: Apache-Coyote/1.1
< Content-Type: text/html;charset=utf-8
< Transfer-Encoding: chunked
<
* Connection #0 to host joeloonix-sasbiws2.na.sas.com left intact
>
```

Environment History

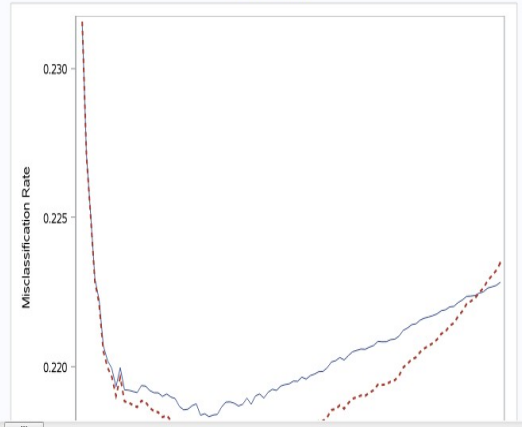
Files Plots Packages Help Viewer

USER_FACTOR2	1271	0.000548	-0.00055	0.001095	-0.00001
USER_FACTOR1	1425	0.000653	-0.00056	0.001305	0.00008
USER_FACTOR16	1398	0.000584	-0.00057	0.001168	0.00003
USER_FACTOR10	1492	0.000644	-0.00057	0.001288	0.00007
USER_FACTOR20	1547	0.000594	-0.00057	0.001187	0.00002
USER_FACTOR11	1447	0.000623	-0.00058	0.001245	0.00005
USER_FACTOR9	1412	0.000615	-0.00059	0.001230	0.00006
USER_FACTOR17	1557	0.000604	-0.00060	0.001208	0.00001
USER_FACTOR19	1610	0.000658	-0.00062	0.001316	0.00005
USER_FACTOR18	1793	0.000763	-0.00068	0.001527	0.00011

Procedure Task Timing

Task	Seconds	Percent
Reading Data	2.18	22.14%
Training Forest	7.64	77.79%
Saving Model	0.01	0.07%

OOB vs Training



Live Demos

- Proc IML with R on the Grid
- Enterprise Miner with Open Source

Possible Future

- How can you open your analytics program to all types of programming languages and all levels of users?
- How can you ensure consistency across your models?

SAS® Viya®

- Architecture that is enabled by wide range of capabilities
  - End-to-end data cleansing/data mining and machine-learning process with a comprehensive, visual (and programming) interface

- Extension of the success of SAS GRID @ VINCI
  - Today the VINCI Grid utilizes SAS via user interfaces such as Base SAS, SAS Enterprise Guide, and batch submit

9.4 & Viya™

- It's an enhances current strategy (Viya extends 9.4)
- Each is designed to solve different use cases
- Can co-exist
- Data, models and code can be accessed directly
- In 9.4M5, it's just another library engine

SAS Grid will continue to be needed to provide workload management of existing/new customer SAS programs. It continues to be the cornerstone for modernizing existing SAS environments.

SAS Viya provides, among other things, the CAS server – an in-memory , distributed run-time engine – to enable visualization, deep learning, machine learning, etc. against very big data.

Using SAS Viya to extend a SAS Grid environment becomes seamless with 9.4m5.

SAS 9.4M5

Will have CAS client connectivity within SAS 9.4M5

Do everything from SAS 9.4M5 client

Run ALL existing SAS Grid jobs WITHOUT modifications

Run CAS engine, CAS statement, CAS enabled procs by adding them to existing SAS Grid jobs or create new jobs

SAS Viya and SAS 9 is an and strategy. Each is designed to solve different use cases. With the latest releases, you can access SAS Viya processing from any SAS 9 coding client (like SAS Enterprise Guide, SAS Display Manager, and all other clients with a coding interface). No longer is SAS/CONNECT bridges needed – many of the SAS 9.4 products have direct coding connections within workflows as well, like SAS Enterprise Miner, SAS DI Studio.

Open: Multiple Interfaces, Single Code Base

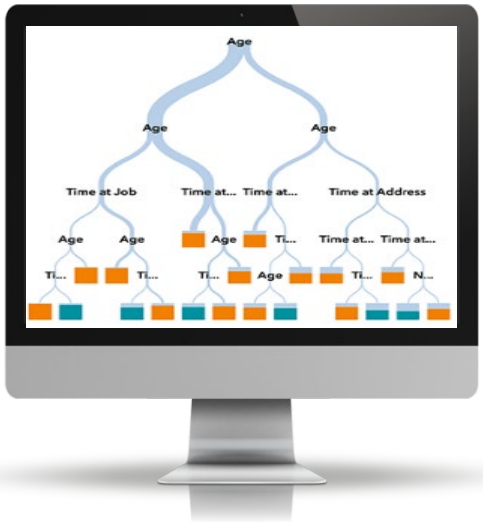
Visual Interfaces

Programming Interfaces

python sas Java R

API Interfaces

REST API



THE ANALYTIC LIFECYCLE

EXTEND

Discovery & *Development of Analytics*

Deployment & *Execution of Analytics*

Wish List

- Best possible analytics
- Flexibility of tools
- Productivity
- Greater insights = models
- Trusted models

How SAS Extends…

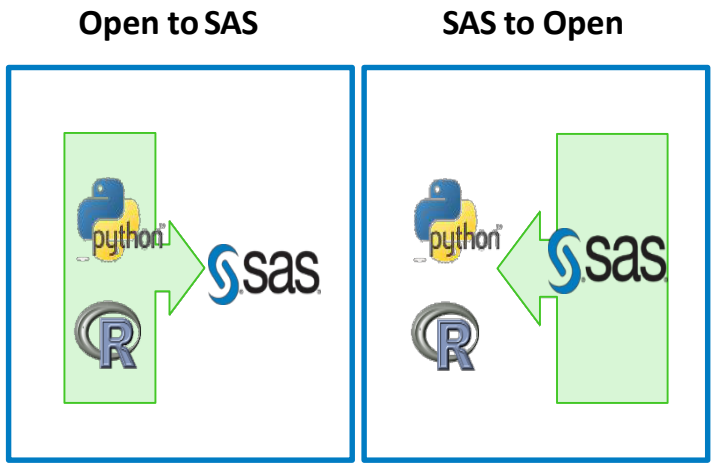
- A variety of options to develop models
- Allows data scientist to code in language of choice
- Ability to scale to any data volume
- Handle complex graphics

SAS AS AN ENHANCEMENT

AND

SAS can augment open source

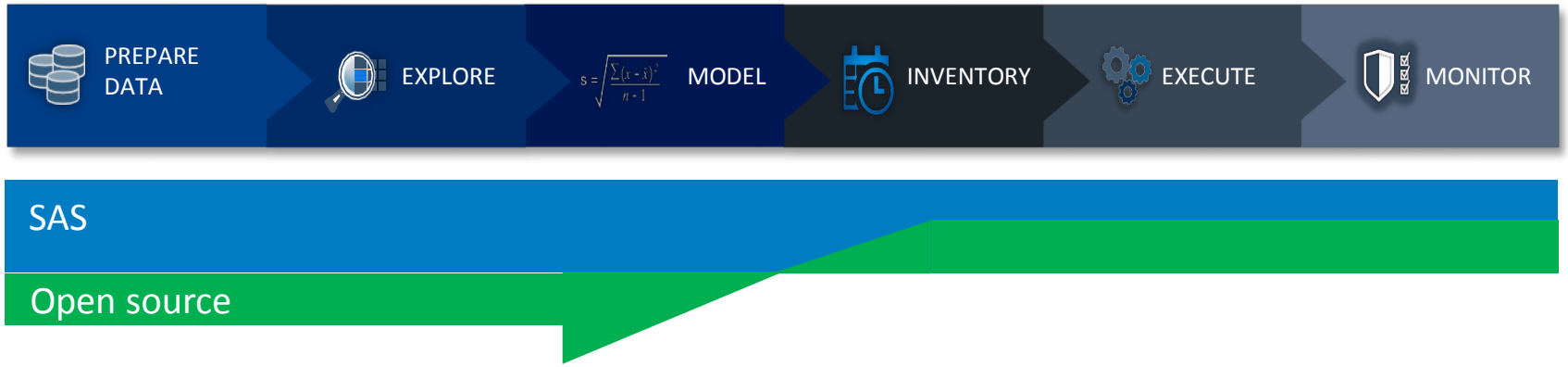
- Increase productivity
- Leverage your assets, people and platforms
- Bring the power of SAS to open source
- Create deployable analytics
- Goal is to ‘embrace’ and ‘extend’



THE ANALYTIC LIFECYCLE: SAS AND OPEN SOURCE

Discovery & *Development of Analytics*

Deployment & *Execution of Analytics*



- SAS embraces open source for Data Prep
- Open source and SAS work well for Discovery and Development
- SAS can extend open source
  - inventory, register and manage models
  - deploy and execute models in Hadoop and in database
  - enhance models and provide monitoring and reporting

EMBRACE + ***EXTEND***

The Need to Unify Analytics Assets

- Analytics Governance
- Precision Results You Can Trust
- User Appropriate Interfaces
- Analytics For The Masses
- Streamlined Deployment
- Effective Processing
- Scalability For Large Complex Or Time-sensitive Problems

Open:
Consistent Results Across Different Interfaces

Python in Jupyter Notebook

Fit Statistics			
Rowid	Description	Training	Validation
M2LL	-2 Log Likelihood	26293.763289	11351.559792
AIC	AIC (smaller is better)	26397.763289	11455.559792
AICC	AICC (smaller is better)	26397.880956	11455.83476
SBC	SBC (smaller is better)	26853.060156	11866.79791
ASE	Average Square Error	0.082181822	0.0829827159
M2LLNULL	-2 Log L (Intercept-only)	36367.490885	15585.606217
RSQUARE	R-Square	0.1933015701	0.1899502038
ADJRSQ	Max-rescaled R-Square	0.3582874055	0.3520858071
MCFADDEN	McFadden's R-Square	0.2769981473	0.2716638908
MISCLASS	Misclassification Rate	0.1050600252	0.1086123688
DIFFMEAN	Difference of Means	0.277629269	0.2702324535

SAS Studio

Fit Statistics		
Description	Training	Validation
-2 Log Likelihood	26294	11352
AIC (smaller is better)	26398	11456
AICC (smaller is better)	26398	11456
SBC (smaller is better)	26853	11867
Average Square Error	0.08218	0.08298
-2 Log L (Intercept-only)	36367	15586
R-Square	0.19330	0.18995
Max-rescaled R-Square	0.35829	0.35209
McFadden's R-Square	0.27700	0.27166
Misclassification Rate	0.10506	0.10861
Difference of Means	0.27763	0.27023

Model Studio

Fit Statistics

Data Role ▲	Sum of Frequencies	Average Squared Error	Divisor for ASE	Root Average Squared Error	Misclassification Rate	Multi-Class Log Loss	KS (Youden)
TRAIN	46,897	0.0822	46,897	0.2867	0.1051	0.2803	0.5150
VALIDATE	20,099	0.0830	20,099	0.2881	0.1086	0.2824	0.5081

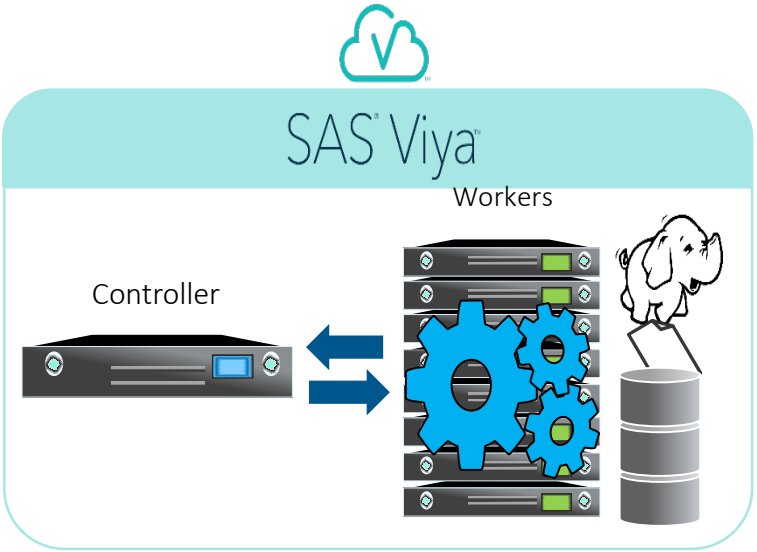
Open: Single Code Base Example

API
s

```
proc print data = hmeq (obs = 10);
run;
```

```
df = s.CASTable('hmeq')
df.head(10)
```

```
df <- defCasTable(s, 'hmeq')
head(df, 10)
```



Translated Action

```
[table.fetch]
    table.name = "hmeq"
    from = 1 to = 10
```

SAS AND OPEN SOURCE

SAS 9.4

SAS® Viya™

EMBRACE

open source by including it and leveraging it where we can

+

EXTEND

open source by improving its interoperability and utility for the enterprise

FOR MORE INFORMATION

Empowering the SAS/IML user with the functionality of R

Documentation: *IML User's Guide - Calling Functions in the R Language*
http://support.sas.com/documentation/cdl/en/imlug/66845/HTML/default/viewer.htm#imlug_r_toc.htm

Video: *Calling R Procedures from SAS/IML® Software*
https://www.youtube.com/watch?v=rUaTTre24kI

Video: *SAS/IML and R: Using Them Together*
https://www.youtube.com/watch?v=nmRQ3MtkG6A

Blogs: *The DO Loop – R tags*
http://blogs.sas.com/content/iml/tag/r/

Paper (p 14-17): *Rediscovering SAS/IML® Software: Modern Data Analysis for the Practicing Statistician*
http://support.sas.com/resources/papers/proceedings10/329-2010.pdf

Article: *Versions of R that are supported by SAS/IML*
http://blogs.sas.com/content/iml/2013/09/16/what-versions-of-r-are-supported-by-sas.html

FOR MORE INFORMATION - EXTENDING R

Video: *Using R in SAS Enterprise Miner*
https://www.youtube.com/watch?v=TbXo0xQCqDw

Blogs: *Spectral Clustering in SAS® Enterprise Miner™ Using Open Source Integration Node*
https://communities.sas.com/docs/DOC-8011

Blogs: *How to execute a Python script in SAS® Enterprise Miner™*
https://communities.sas.com/docs/DOC-10832

Blogs: *Open Source Integration Using the Base SAS Java Object*
https://communities.sas.com/docs/DOC-10746

Article: *The Open Source Integration node installation cheat sheet*
https://communities.sas.com/docs/DOC-9988

Usage Notes:
http://support.sas.com/dsearch?Find=Search&ct=&qt=open+source&col=suppprd&nh=25&qp=&qc=suppsas&ws=1&qm=1&st=1&lk=1&rf=0&oq=&rq=0

FOR MORE INFORMATION MATERIALS ON GITHUB

This organization Search

sassoftware

Sas integration and sample code
Integration with R, Python
https://github.com/sassoftware/enlighten-integration

Integration with Jupyter Notebook and Python
https://github.com/sassoftware/sas_kernel
https://github.com/sassoftware/saspy

Sample codes of SAS Machine Learning methods
https://github.com/sassoftware/enlighten-apply

SAS Enterprise Miner process flow diagrams
https://github.com/sassoftware/dm-flow

enlighten-integration Java 23 20

Example code and materials that illustrate techniques for integrating SAS with popular open source analytics technologies like Python and R.

Updated a day ago

sas_kernel Jupyter Notebook 18 6

A Jupyter kernel for SAS. This opens up all the data manipulation and analytics capabilities of your SAS system within a notebook interface. Use the Jupyter Notebook interface to execute SAS code and view results inline.

Updated 2 days ago

saspy Python 8 5

An interface module to the SAS System. It works with Linux SAS, and is currently intended as a support module for the sas_kernel project which provides a Jupyter Notebook kernel which surfaces the SAS Language and SAS ODS Output to Jupyter Notebooks. Additionally, provides magics which allow SAS code to be submitted for notebooks with other kern...

Updated 4 days ago

enlighten-apply SAS 40 31

Example code and materials that illustrate applications of SAS machine learning techniques.

Updated 8 days ago

dm-flow 9 6

Library of SAS Enterprise Miner process flow diagrams to help you learn by example about specific data mining topics.

Updated 21 days ago

Resources Offered

- **The VINCI SAS Admins offer online training for you or your group.**
- **You can contact VINCISASAdmins@va.gov for SAS Specific questions or VINCI@va.gov for access, network, data or other questions.**
- ***VINCI Central site: http://vhacdwweb02.vha.med.va.gov/vincicentral/projectsites/SASGrid/Shared%20Documents/Forms/AllItems.aspx***

Good Gridding!

Thank you for attending.

Please contact VINCI SAS Administrators:

VINCI SAS Admins
VINCISASAdmins@va.gov
with any questions or comments.