WHITE PAPER

SAS® in the Open Ecosystem

How a unifying platform can bring together diverse data and analytics to drive measurable value for your organization

Contents

The Need to Unify Analytics Assets

Envision this situation at a growing bank. Its competitive landscape demands an agile response to evolving customer needs. Fortunately, analytically minded professionals in different divisions are seeing results that positively affect the bottom line.

- A data scientist in the business development team analyzes data to create customized experiences for premium customers.
- A digital marketer tracks and influences the customer journey for prospective mortgage customers.
- A risk analyst builds risk models for the bank's loan portfolios.
- A data analyst examines data about local customers.
- A technical architect defines a new system to protect bank data from internal and external cyberthreats.
- An application developer builds a new mobile app for online customer portfolio management.

Between them, these employees might be using more than a dozen packages for analytics and data management.

This mix could include open source technologies, commercial software solutions, enterprise-hosted applications and cloud deployments. When data and analytical professionals can use the programming languages and tools of their choice, they are happier and more efficient. And when technical skills are in short supply, flexibility around skills helps organizations find the talent they need and make the most of what they already have.

How can the IT/analytics department at this bank make sure these projects are using trustworthy data, the best models and a rigorous process that will guarantee compliant, useful and repeatable results? And whose responsibility is it to stitch together all of the disparate code bases and business scenarios to monitor the outcomes or find other opportunities for business improvements?

Organizations need an environment that unifies their data and analytical silos. One that can integrate existing and emerging technologies to bring together analytics endeavors and provide shareable access to analytical assets across divisions – with whatever code and tools are in the mix. This gives data and analytical professionals the freedom to create, experiment, test and rapidly deploy different methods easily, using whatever skills they have. It allows IT leaders to easily manage the entire analytics life cycle, from development to discovery to deployment, for all analytical assets from centralized controls.

What can help? A comprehensive, unifying analytics platform that ensures governed analytical assets and produces tangible results across the enterprise.

How Open Source Is Used in Modern Computing Environments

Open source refers to a computing program or infrastructure in which the source code is publicly available for use and modification by a community of users. By definition, open source is collaborative, where programmers improve upon the source code and share their changes with a broader community. Programmers using open source programs have direct access to online communities of like-minded individuals and often an array of prebuilt algorithms that others have contributed. Functions, projects and applications are built on published open source code, which can speed application development and reduce initial costs for analytics software.

Open source technologies gained popularity in the last decade for several reasons. They solved problems other programs didn't, they're flexible and students (who later enter the workforce) learn to use them because the software is free and easily available.

At the same time, with the expanding demand for analytics and data management, many organizations found themselves lacking appropriately skilled employees. Open source technologies provide the flexibility to address data and analytics needs through an array of software programming languages. Giving data and analytics professionals the ability to work in the programming language and environment of their choice extends the pool of available analytics talent.

Do open source technologies meet enterprise analytics needs?

Open source can help bring together a community of in-house contributors, each building on top of code developed by others. There's no need to reinvent the wheel. But the use of open source technologies for enterprise analytics can fall short for organizations seeking to streamline enterprise operational efficiency. Here are some possible issues:

Analytical silos and IT complexity. When each team of practitioners uses its preferred tools, the result is a set of analytical silos. Each silo has its own data, algorithms, methods, code and versioning. Each silo needs its own management and maintenance. This adds to the complexity of deploying analytics into production, which detracts from enterprise-level consistency and control.

Recoding models can produce inconsistent results. In real-world analytical problem-solving, multiple methods are needed for the best results. When models are written in different languages, additional coding is required to convey the outputs of one result as inputs into another. It's not unusual for data scientists to recode models using different languages or different versions of the same language, which takes time and can lead to mismatched results.

Code can be difficult for others to interpret. Because developers are focused on solving a particular problem at a certain time, their code may be difficult for others to reuse or interpret. Lack of documentation about the code, its purpose and other helpful information can be problematic.

Data management and scalability issues. Data manipulation in open source can be very time-consuming. And some open source solutions don't offer the scalability needed to handle large data sets.

Lack of governance. It can be a serious struggle to corral independently developed and lightly documented analytical assets in different areas of an organization. How can you consistently and efficiently deploy analytics built in different languages? How do you provide transparency and traceability into different development and data processes? What will happen when your current data scientists leave?

Problematic deployment. IT departments need to verify that software meets their standards and protocols before they deploy it. With open source technologies, the validation to satisfy enterprise deployment is often lacking. The costs of evaluating each new open source package and ensuring version consistency between model building and operational deployment could quickly negate the savings from open source technologies in production environments. And because there's no assurance of backward compatibility, the software put in production today might not work tomorrow if the underlying open source packages are updated to a newer version.

The Advantages of Integrating SAS® and Open Source Technologies

The addition of the SAS Platform to an open analytics ecosystem brings numerous benefits.

Analytics governance. The SAS Platform brings the governance necessary to unify an enterprise-level analytics infrastructure so an organization can connect all the parts of a disparate analytics ecosystem. It helps data scientists manage models coded in different languages and helps IT trace and audit analytics for an effective compliance strategy. It consolidates information about model versioning, authorization, model lineages and source data location. With one governed inventory for all analytical assets, organizations get trusted and traceable insights with speed and agility – and can easily manage their entire analytics portfolio.

SAS provides the broadest, most operationally vetted range of analytics capabilities available.

Precision results you can trust. When it comes to business-critical functions, especially those with regulatory ramifications such as risk, fraud or cybersecurity, "close enough" isn't an option. Modern organizations need precision. SAS provides the broadest and most operationally vetted range of analytics capabilities available. From the mission critical to the experimental, SAS offers analytics for even the most complex tasks.

User-appropriate interfaces. SAS provides both code-based and highly visual user experiences. Analysts can get started with data preparation and analytics in a robust visual interface, with or without predefined process workflows. Data practitioners can write code their way (such as in Python, R, Lua, etc.) while benefiting from productivity improvements driven by the SAS Platform, including automated analytical model comparison, scalable performance, automated deployment code and model decay analysis.

Analytics for the masses. From programming interfaces to self-service, visual exploration tools, the SAS Platform empowers a wide variety of organizational stakeholders. Everyone from data scientists, citizen data scientists and business analysts to upper management can turn unstructured and structured data into trusted insights.

Streamlined deployment. The SAS Platform lets you inventory, execute and monitor all analytical assets – SAS and otherwise. Organizations can deploy and operationalize analytical models once and reuse them throughout the organization. These models can be exported easily and efficiently from development to production in a consistent, reliable and repeatable fashion. And when your deployment environment changes – no worries. Your code stays the same because SAS is portable.

Effective processing. From public/private clouds to on-site deployments to the edge in the Internet of Things, SAS can access and analyze data wherever it lives. When users of open source languages move from a single-threaded environment to the multi-threaded, distributed architecture of SAS, they can see processing times shrink from hours to minutes. And SAS functions can run natively inside Hadoop and Teradata (popular storage and processing frameworks) as well as other big data stores so you don't have to move data to use it. Being able to run more iterations and use all your data – not just a sample – increases model accuracy. And, it's important to note that modern machine learning and artificial intelligence algorithms are very data hungry. Even with huge volumes of data, SAS processing produces fast results.

Scalability for large, complex or time-sensitive problems. High-performance analytics right out of the box can tackle any problem or data size. SAS functions run the same way whether you're processing hundreds – or hundreds of millions – of rows, without changing the function itself. So the same process can be used on any data set within the organization and can scale to handle changes when your organization grows.

From SAS programmers to those who code in third-party languages to those who just want point-and-click insights, SAS lets users choose how they want to interact with their data.

How the SAS® Platform Can Help

The SAS Platform helps you get maximum value from your data and analytical assets by addressing all aspects of analytics life cycle, from data to discovery to deployment. By combining the power of the SAS Platform with open source technologies, you can unify disparate toolsets and analytics assets into a streamlined, governed and collaborative environment that improves productivity, fosters business agility and delivers tangible results.

From simple Excel spreadsheets to big data in Hadoop, the SAS Platform scales analytics to match the volume, velocity and variety of your data.

	SAS® augments open source with …	**Results**
Prepare data	Native access to all data, including in database, batch and in stream. Ability to embed key analytics functions to reduce data movement or adjust erroneous elements.	Work with more data faster, identify new patterns and anomalies, and uncover new insights. Minimize movement of data to accelerate performance. Provide trusted, high-quality data for all.
Explore data	A visual interface that enables business users and junior analysts to start exploring data. Embedded data preparation and transformations. Data quality and governance features.	Give more people access to data stored in all systems. Improve governance by working with data inside big data sources. Governance that is supported by business processes.
Build models	Ability to program in language of choice to facilitate better integration from disparate environments. Collaborative, interactive and highly visual environment designed for multiple stakeholders. Robust, comprehensive suite of algorithms that scale to all data with automatic identification of champion models.	Democratize analytics to all types of professional skills. Free up data science resources and solve more complex business problems by reducing model build time. Increase model accuracy by using all your data – not just a sample – and running more iterations more often. Reduce latency and improve time to value by analyzing data closer to the source.
Manage model inventory	Comprehensive, simultaneous model management for enterprise model inventory, including SAS and open source models. Collaborative environment for monitoring model health and audit performance to identify model decay. Documentation, versioning and model lineage.	Manage analytics as an enterprise asset. Run your organization on fact-based decisions. Create trusted models with data-to-deployment traceability. Manage risk and compliance. Embed analytics into production systems, data in motion or data at rest – all with the same consistent code. Flexible deployment options for private or public clouds – or no cloud at all.
Execute models	Portable code, deployable anywhere. Automated execution processes.	
Monitor model performance	Robust analytics to assess model performance, including retraining. Champion/challenger facility to identify the best-performing models.	

How SAS® Works in the Open Ecosystem

Augmenting open source technologies with SAS increases the efficiencies needed for enterprise analytics. The SAS Platform delivers capabilities that integrate with and take advantage of many open source technologies, including:

- Programming languages such as Python, R, Lua, Scala and Java.
- Data frameworks such as Hadoop.
- Changing physical and virtual hardware environments.

With SAS, you can:

- Access powerful SAS Analytics from both SAS and other coding interfaces, including native programming access from SAS, Python, Java, R, Scala and Lua and through REST application programming interfaces (APIs).
- Process operations written in any analytical coding language, faster and more efficiently. A multithreaded, in-memory, massively parallel processing engine reduces processing times of complex analytics from days to minutes.
- Optimize all analytics capabilities for popular IT environments with a cloud-ready platform that can run within any public cloud, private cloud or on-site infrastructure – or a combination of environments.
- Take SAS into an open source technology environment or bring open source technology elements into the SAS environment.

Let's take a closer look at options for using SAS and open source technologies together.

Taking SAS® to Open Source Environments

Data scientists can code in their language and interface of choice, while gaining the advantages of SAS, such as automatic delivery of model performance metrics and the ability to scale to any data volume without editing code.

Using SAS in open source applications can ease the transition for users who aren't experienced with SAS. For example, calling SAS via stored processes or REST APIs from other programming interfaces is a simple way for open source programmers to access SAS.

In addition, SAS provides an expanding set of open source projects, including:

- A SAS scripting wrapper for analytics transfer (SWAT) package that allows users to execute SAS actions and process results from open source languages. It provides a familiar environment for Python and R programmers, translating their code into SAS production-ready code. Versions are available for Python and R.
- A Python interface module to SAS (saspy) lets you start a SAS session and run analytics from Python.
- The Jupyter kernel for SAS opens up all the SAS data management and analytics capabilities within a Jupyter Notebook interface. Python coders can call SAS procedures from the Python kernel in Jupyter Notebook, enabling them to take advantage of the power of both languages from within a single interface. Use the interface to execute SAS code and view results inline.

Improve model lift and performance by creating ensemble models that combine the best of SAS and open source technology.

- The SAS deep learning Python package (python-dlpy) contains the high-level Python APIs to SAS deep learning algorithms. It allows users to build deep learning models using friendly Keras-like APIs.
- The SAS pipefitter package provides a Python API for developing pipelines for data transformation and model fitting as stages of a repeatable machine learning workflow in SAS.
- Sasopty is a Python package that provides a modeling interface for SAS Optimization solvers.

SAS maintains a GitHub repository to allow the community to consume and enhance integrations for SAS. These and other open source projects led by SAS can be found at https://github.com/sassoftware.

SAS maintains a GitHub repository to allow the community to consume and enhance integrations for SAS. Open source projects led by SAS can be found at https://github.com/sassoftware.

Bringing Open Source Technology Interactions to SAS®

Integration capabilities built into SAS software enable it to take advantage of open source technology. For example, Base SAS software offers a Java class object that incorporates a variety of languages, including Python.

You can also easily integrate R and Python code inside of a SAS Visual Data Mining and Machine Learning pipeline or a SAS® Enterprise Miner™ process flow diagram. This enables you to perform data transformation and exploration, as well as train and score supervised and unsupervised models in other programming languages. The value of this integration is the ability to create ensemble models using open source and SAS. A blended model combines the best of both SAS and open source technology to achieve the greatest improvement overall performance.

Open source models can also be compared with SAS models. Business-oriented, user-friendly assessment reports are created automatically. Experimental models that are not ready for production can be used for benchmarking. Model flows automatically generate code, including the score code operational deployment – all from within an easy-to-use, drag-and-drop interface. SAS automatically generates documentation capturing best practices, which promotes collaboration and preserves continuity when analytical professionals move on.

Using and Contributing to the Open Ecosystem

At SAS, we build our software using open source technology, on open source operating systems, using open standards. We're productive using open source browsers and open source data formats, integrated with open source cloud services. SAS has always been extensible in that it provides integration points into third-party programming tools, data types and operating systems. There are plenty of commercial tools and services in the mix too. It's all a part of the modern technology ecosystem.

SAS also has formal commitments to open data initiatives. For example, as a member of the Open Data Platform initiative (ODPi), SAS remains committed to ensuring that our applications work with, and exploit, the Hadoop distribution of our customers' choice, while adding the stability and quality needed in demanding business environments.

SAS is a collaborating partner in the Data Governance Initiative (DGI), which was chartered to bring a common, metadata-powered approach to data governance into the open source community, and to establish a framework with the flexibility to be applied across industries.

Unifying Analytics Deployment

The best data scientists don't just develop one model to solve a business problem. They develop a set of competing models and use different techniques to address nuances in the data and the problem. At any given time, there are likely hundreds of models in various stages of development for different business purposes.

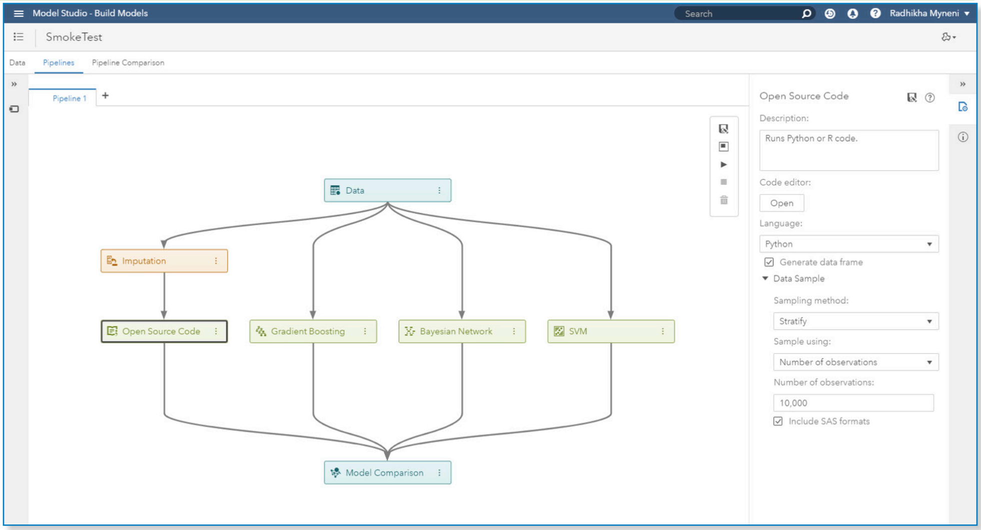
Managing all of these models is no easy feat. The process of deploying analytical models – embedding them into production systems, business processes and applications – can take weeks or months when models are in different languages and deployment isn't centrally controlled. Managing the status and performance of hundreds or thousands of models across their lifetimes can be a complex task.

SAS streamlines deployment of analytical models, bridging the gaps between model developers and IT with a common, centralized and managed deployment process. This reduces time to deployment and hastens the value driven by analytics to the organization.

SAS can inventory, publish, score, monitor and retrain both SAS and open source models. With automation, workflow support, centralized business rule logic and a unified experience, SAS helps organizations transition their models to the production ecosystem more easily and efficiently.

Open source integration

- Execute R and Python code and models from within SAS.
- After execution, results can be displayed in the user interface.
- Produce model assessment and compare with other models (SAS and open source) to pick a champion.



Open Source Code

Description: Runs Python or R code.

Code editor: Open

Language: Python

☑ Generate data frame

Data Sample

Sampling method: Stratify

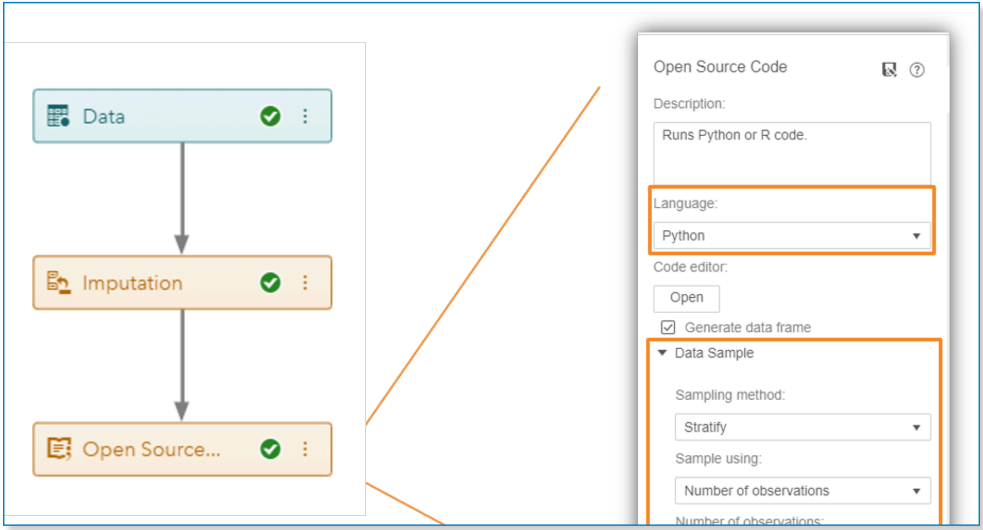
Sample using: Number of observations

Number of observations: 10,000

☑ Include SAS formats

Properties

- Select sampling strategies.
- Code snippets are generated behind the scenes and appended to user code for convenience.



Open Source Code

Description: Runs Python or R code.

Language: Python

Code editor: Open

☑ Generate data frame

Data Sample

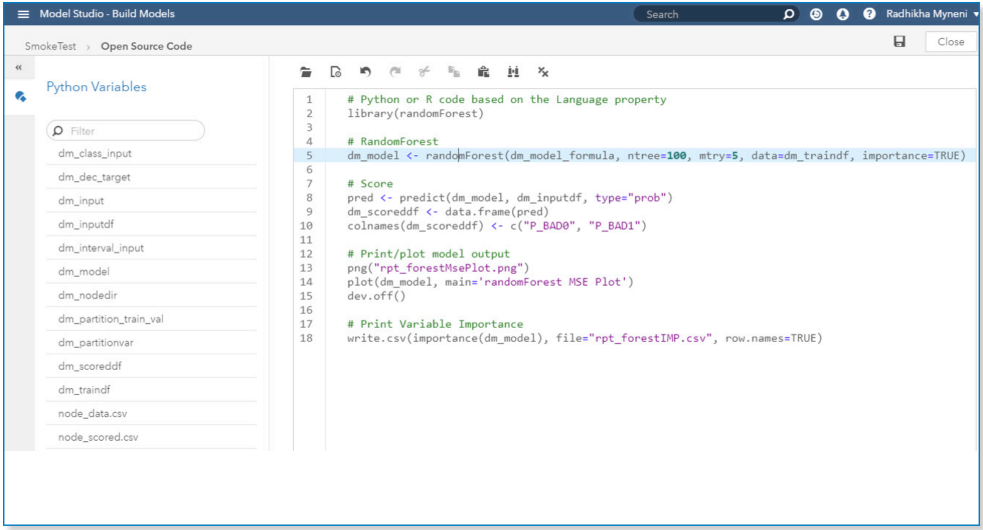
Sampling method: Stratify

Sample using: Number of observations

Number of observations:

Code editor

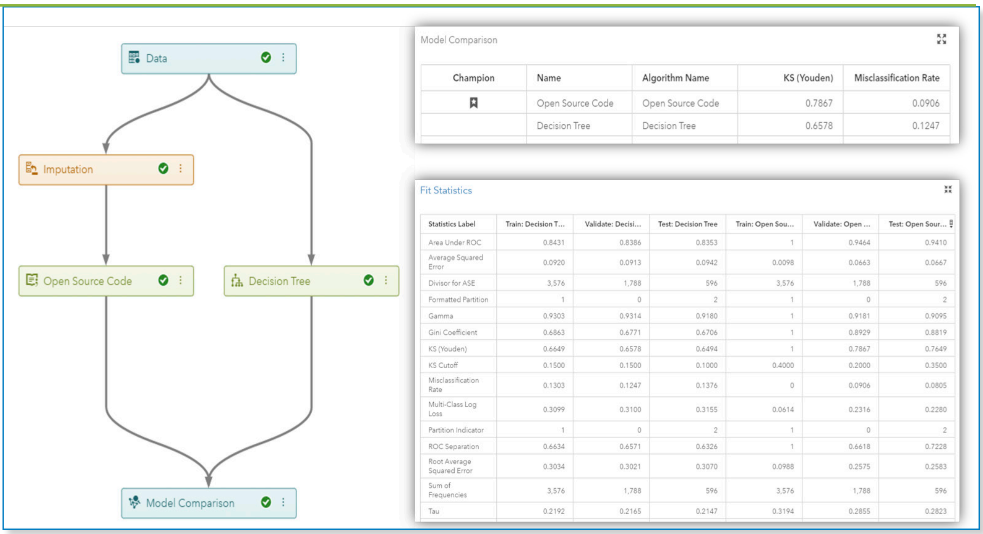
- Flexibility to code in Python or R.
- Execute R or Python code from SAS.
- Use prebuilt variables (shown on left) for faster data manipulation and model building.



```
# Python or R code based on the Language property
library(randomForest)

# RandomForest
dm_model <- randomForest(dm_model_formula, ntree=100, mtry=5, data=dm_traindf, importance=TRUE)

# Score
pred <- predict(dm_model, dm_inputdf, type="prob")
dm_scoreddf <- data.frame(pred)
colnames(dm_scoreddf) <- c("P_BAD0", "P_BAD1")

# Print/plot model output
png("rpt_forestMsePlot.png")
plot(dm_model, main='randomForest MSE Plot')
dev.off()

# Print Variable Importance
write.csv(importance(dm_model), file="rpt_forestIMP.csv", row.names=TRUE)
```

Model assessment

- Quickly compare fit statistics of open source models along with other models in the visual pipeline.
- Choose a champion model based on a list of selection criteria.



Model Comparison

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
🔖	Open Source Code	Open Source Code	0.7867	0.0906
	Decision Tree	Decision Tree	0.6578	0.1247

Fit Statistics

Statistics Label	Train: Decision T...	Validate: Decisi...	Test: Decision Tree	Train: Open Sou...	Validate: Open ...	Test: Open Sour...
Area Under ROC	0.8431	0.8386	0.8353	1	0.9464	0.9410
Average Squared Error	0.0920	0.0913	0.0942	0.0098	0.0663	0.0667
Divisor for ASE	3,576	1,788	596	3,576	1,788	596
Formatted Partition	1	0	2	1	0	2
Gamma	0.9303	0.9314	0.9180	1	0.9181	0.9095
Gini Coefficient	0.6863	0.6771	0.6706	1	0.8929	0.8819
KS (Youden)	0.6649	0.6578	0.6494	1	0.7867	0.7649
KS Cutoff	0.1500	0.1500	0.1000	0.4000	0.2000	0.3500
Misclassification Rate	0.1303	0.1247	0.1376	0	0.0906	0.0805
Multi-Class Log Loss	0.3099	0.3100	0.3155	0.0614	0.2316	0.2280
Partition Indicator	1	0	2	1	0	2
ROC Separation	0.6634	0.6571	0.6326	1	0.6618	0.7228
Root Average Squared Error	0.3034	0.3021	0.3070	0.0988	0.2575	0.2583
Sum of Frequencies	3,576	1,788	596	3,576	1,788	596
Tau	0.2192	0.2165	0.2147	0.3194	0.2855	0.2823

Closing Thoughts

SAS helps deliver on the promise of enterprise analytics by providing a flexible and adaptable technology landscape that embraces many open technologies to ensure interoperability. It can be used to centralize analytics across various groups and helps organizations consolidate analytics assets with a unifying, governed platform.

Only SAS enables organizations to maximize the value from analytics assets with:

- An environment that unifies diverse business silos, providing trusted insights developed by analytics professionals using SAS and other programming languages.
- An analytics platform that consistently, reliably and repeatedly enables experimentation, analytical model building and streamlined deployment of SAS and other analytics assets regardless of scale.
- Centralized governance, specialized for analytics workload processing, driving efficiencies and optimizing investments made in analytics.

Organizations can make more accurate, analytics-driven decisions and improve performance by taking advantage of the best of both worlds – SAS and open source technologies – to solve problems in innovative ways.

To deliver business value, models must be in production systems and processes. SAS helps transition models – both SAS and open source – from development to production more easily and efficiently.

Learn More

Read this recommended article for more details about the advantages of having the right platform to get the best of enterprise software in a governed, collaborative way without giving up the features of open source software: “Keeping an open mind about open analytics.”

To contact your local SAS office, please visit: sas.com/offices