PharmaSUG 2018

SAS® Visual Interface for Data Exploration and Modeling

Alex Ford, SAS, Cary, NC
Andrea Coombs, SAS, East Lansing, MI

ABSTRACT

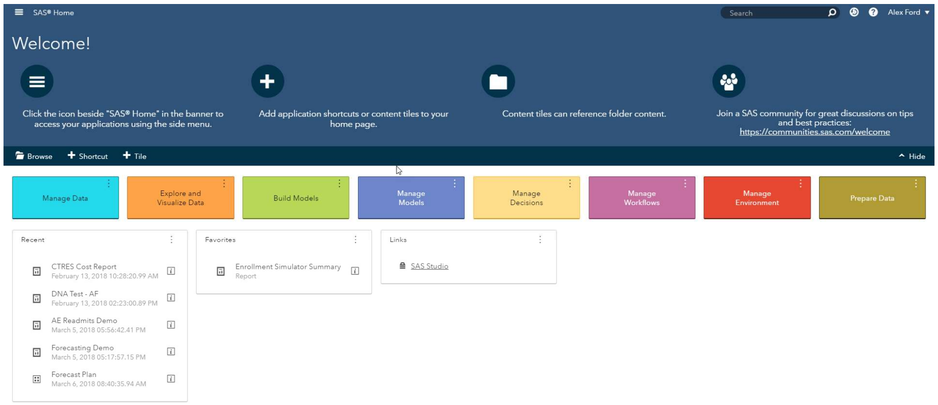
Across the globe, health and life science organizations need the ability to access, explore, and model data. Study start-up units view enrollment and screen fail rates across sites; monitors investigate serious adverse events; biostatisticians test interactions between predictive variables and health outcomes; business analysts forecast the budgetary spend necessary to complete a trial. In this presentation we will discuss how SAS® Visual Analytics, Visual Statistics, and Visual Data Mining and Machine Learning can help answer these questions and many more facing the health and life science industry. This suite of visualization tools provides a modern, integrated environment for governed data discovery that can be leveraged by many users across the organization including those without advanced analytical or programming skills. We will demonstrate data exploration, descriptive and predictive modeling, and machine-learning techniques for analyzing structured and unstructured data.

INTRODUCTION

A visual presentation of data is one of the most straightforward, concise, and reader-friendly tools that can be leveraged to derive and distribute insights across an organization, especially for clinical trials. Unfortunately, creating nice-looking and informative graphs using standard coding techniques can become cumbersome and is a task that is often easier said than done. To help all levels of an organization get the answers they need quickly, SAS has layered an analytic-ready user interface on top of the SAS® Viya in-memory compute engine to allow users to quickly explore data and generate repeatable analyses at the click of a button in one unified environment. In this hands-on workshop, we will discuss options for loading data into SAS® Viya, as well as basic methods for exploring data. Next, we will see an overview of building, comparing, and putting models into production along with some various options for sharing reports. Join us as we discover the ease of usability in the latest offerings from SAS Institute and navigate through the platform to build appealing high-powered analytic reports.

SAS® HOME

The SAS® Viya landing screen – SAS® Home - is the starting point from which actions are performed and reports can be created. The screen contains tiles or shortcuts which correspond to different actions that a user can make when performing an analysis. For example, the first tile is to “Manage Data”, in other words to view, access, load, and profile data for use in generating reports. The home screen can be customized for other user actions such as browsing recently generated reports, saving favorite reports, and embedding links.



Welcome!

Click the icon beside "SAS® Home" in the banner to access your applications using the side menu.

Add application shortcuts or content tiles to your home page.

Content tiles can reference folder content.

Join a SAS community for great discussions on tips and best practices: https://communities.sas.com/welcome

CTRES Cost Report
February 13, 2018 10:28:20.99 AM

DNA Test - AF
February 13, 2018 02:23:00.89 PM

AE Readmits Demo
March 5, 2018 05:56:42.41 PM

Forecasting Demo
March 5, 2018 05:17:57.15 PM

Forecast Plan
March 6, 2018 08:40:35.94 AM

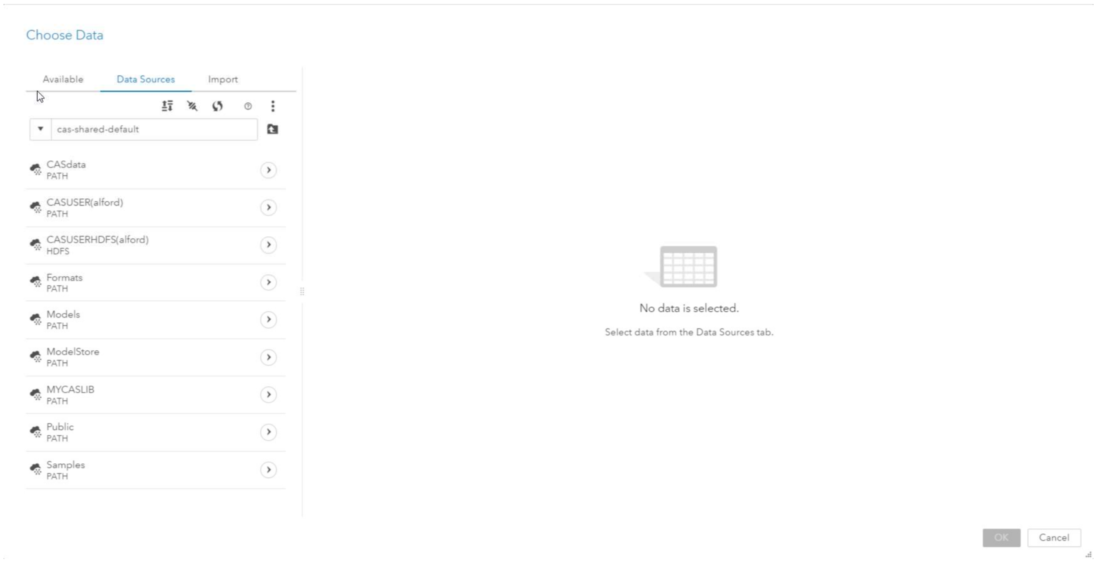
Enrollment Simulator Summary
Report

SAS Studio

Figure 1. SAS® Home

DATA

Users can set up access to data and files and load data into memory using the “Manage Data” tile. The Available tab lists data currently loaded into memory. The Data Sources tab allows users to establish connections to data sources and lists tables with access authorization. The Import tab allows users to load local flat files or access and pull social media feeds for analysis.



Choose Data

Available | Data Sources | Import

cas-shared-default

CASdata
PATH

CASUSER(alford)
PATH

CASUSERHDFS(alford)
HDFS

Formats
PATH

Models
PATH

ModelStore
PATH

MYCASLIB
PATH

Public
PATH

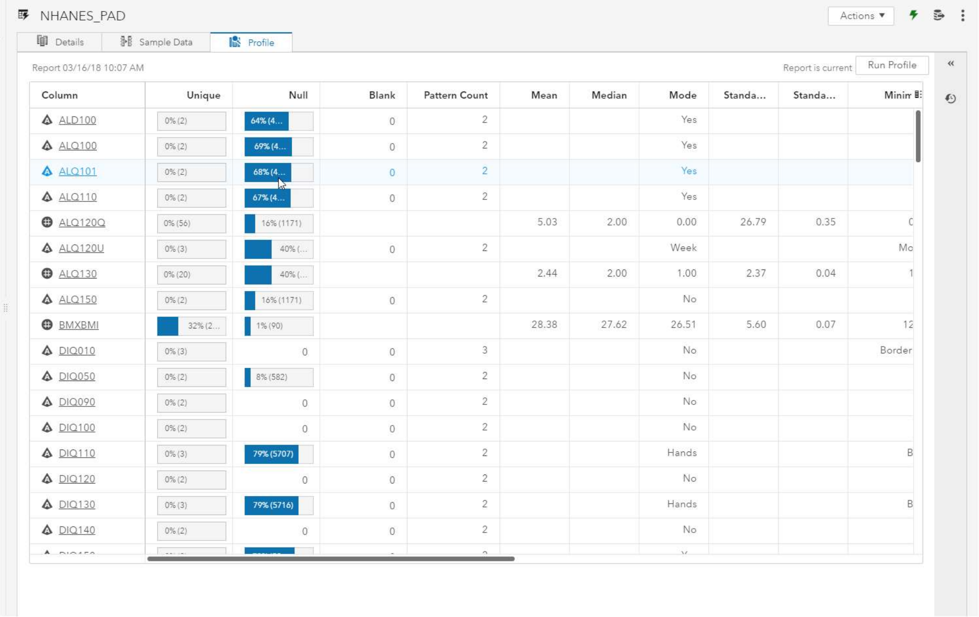
Samples
PATH

No data is selected.

Select data from the Data Sources tab.

Figure 2. Data Access

Once a dataset has been loaded into memory, it is available for use on the platform. The first step in building analyses off a dataset is to figure out what is in the dataset. There are three tabs that appear on the right half of the screen after selecting a single dataset within the “Manage Data” tile. The first tab, “Details”, provides high level details about the dataset – the number of variables and rows, dataset size, when the dataset was created and last modified, and the type of each variable. The “Sample Data” tab is a print of the dataset. A user can select the number of rows that are printed for viewing. Lastly, the “Profile” provides summary statistics for each of the variables in the dataset such as count of unique levels, percentage of the variable that is null, and basic summary statistics.



NHANES_PAD

Report 03/16/18 10:07 AM

Report is current

Column	Unique	Null	Blank	Pattern Count	Mean	Median	Mode	Standa...	Standa...	Minim
ALD100	0% (2)	64% (4...	0	2			Yes			
ALQ100	0% (2)	69% (4...	0	2			Yes			
ALQ101	0% (2)	68% (4...	0	2			Yes			
ALQ110	0% (2)	67% (4...	0	2			Yes			
ALQ120Q	0% (56)	16% (1171)			5.03	2.00	0.00	26.79	0.35	C
ALQ120U	0% (3)	40% (...	0	2			Week			Mo
ALQ130	0% (20)	40% (...			2.44	2.00	1.00	2.37	0.04	1
ALQ150	0% (2)	16% (1171)	0	2			No			
BMXBMI	32% (2...	1% (90)			28.38	27.62	26.51	5.60	0.07	12
DIQ010	0% (3)	0	0	3			No			Border
DIQ050	0% (2)	8% (582)	0	2			No			
DIQ090	0% (2)	0	0	2			No			
DIQ100	0% (2)	0	0	2			No			
DIQ110	0% (3)	79% (5707)	0	2			Hands			B
DIQ120	0% (2)	0	0	2			No			
DIQ130	0% (3)	79% (5716)	0	2			Hands			B
DIQ140	0% (2)	0	0	2			No			

Figure 3. Profile Data

EXPLORE DATA

Once comfortable with the contents of the dataset, we can use the “Actions” dropdown at the top right of Figure 3 and begin exploring the data using the “Explore and Visualize Data” tile. On the screen that appears, there is an “Objects” section on the left side, which contains the tables, listings, graphs, forecasts, and models that a user can select to create in a report. There is also a pop out bar descending vertically along the right side of the screen in which a user defines the roles of the variables used, options for the layout and appearance, and additional choices for variable actions, rules, and filters for each object. To create an object, drag and drop the name out of the left panel onto the graphing space.

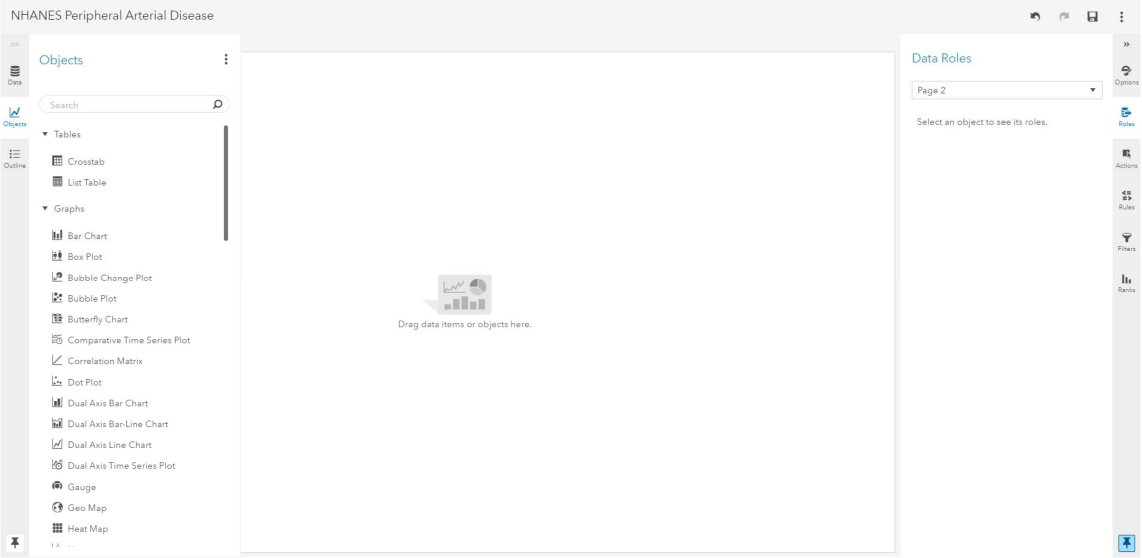


Figure 4. Explore and Visualize Data Homescreen

To perform exploratory analysis of a dataset, objects are created based upon variables of interest and customized to gather insights. In Figure 5, objects are created based on the response variable PAD (Peripheral Arterial Disease). A donut chart shows frequency of PAD, a boxplot graphs BMI by diabetes status, a dual-axis bar line chart illustrates time since quit smoking cigarettes by marital status segregated by PAD status, and a button bar controls view based on country of origin.

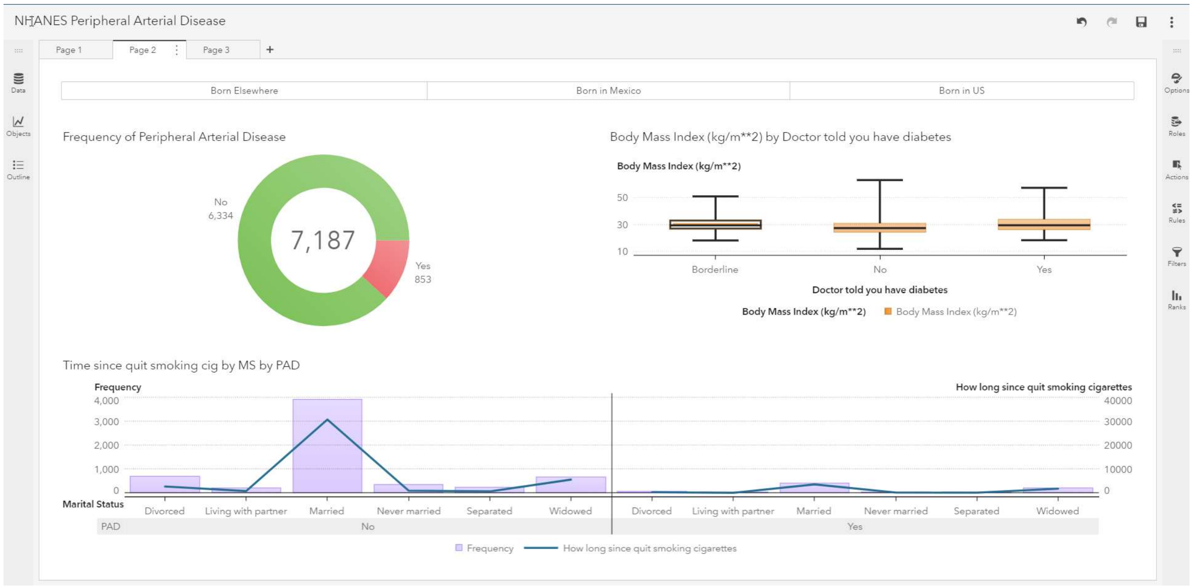
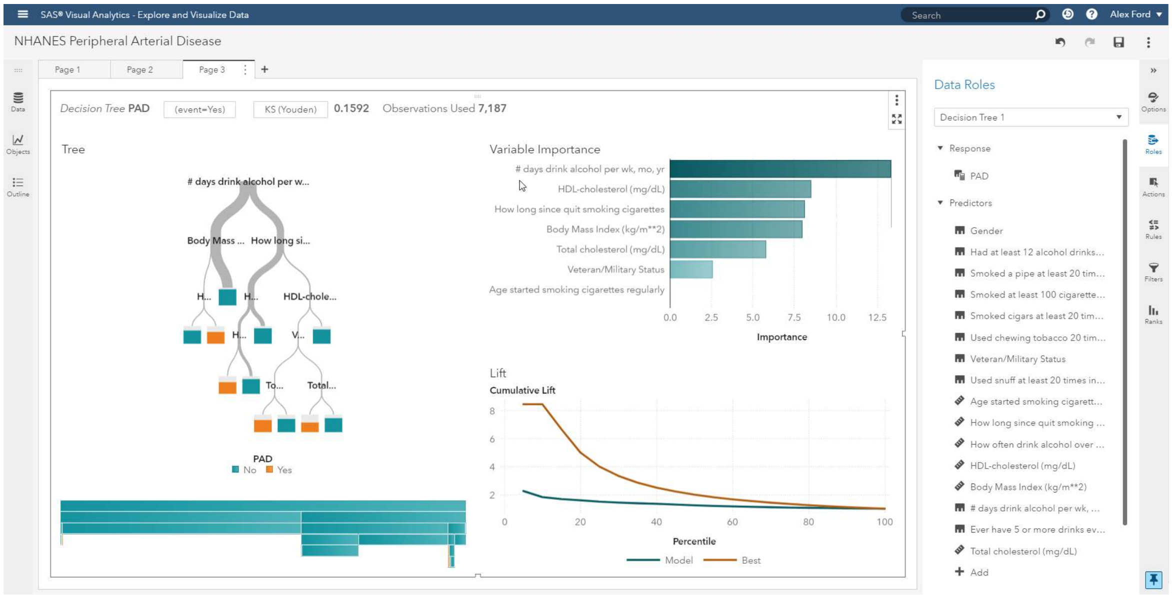


Figure 5. SAS® Visual Analytics Plots in Explore and Visualize Data

BUILDING MODELS

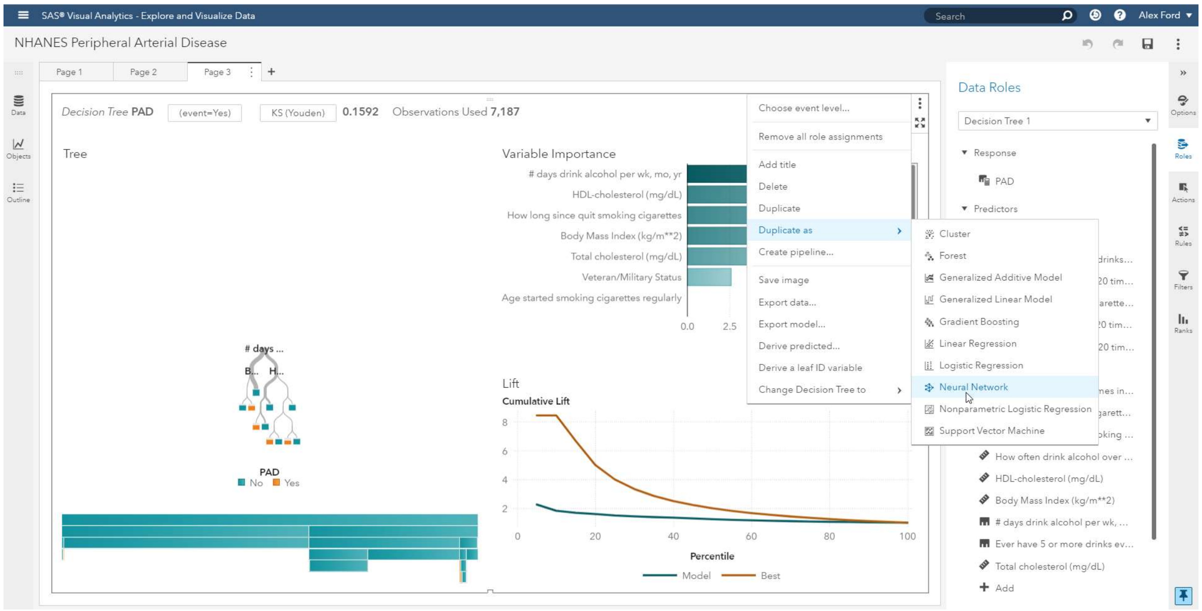
To begin creating models in the report, scroll the objects tab to the SAS® Visual Statistics section and drag and drop a decision tree into the graphics space. Choosing PAD as the response variable and a list of various predictor variables, generates a decision tree showing that the number of days drinking alcohol per week is the variable of highest importance from those selected. Note, that variables can easily be added or removed from the model by clicking the "+ Add" button at the bottom of the predictors list. Users can also choose the event level, statistic of interest, and customize additional selections in the options tab.



SAS® Visual Analytics - Explore and Visualize Data
Search
Alex Ford
NHANES Peripheral Arterial Disease
Page 1
Page 2
Page 3
Data
Objects
Outline
Decision Tree PAD (event=Yes) KS (Youden) 0.1592 Observations Used 7,187
Tree
# days drink alcohol per w...
Body Mass ...
How long si...
H...
H...
HDL-chole...
H...
V...
To...
Total...
PAD
No
Yes
Variable Importance
# days drink alcohol per wk, mo, yr
HDL-cholesterol (mg/dL)
How long since quit smoking cigarettes
Body Mass Index (kg/m**2)
Total cholesterol (mg/dL)
Veteran/Military Status
Age started smoking cigarettes regularly
0.0
2.5
5.0
7.5
10.0
12.5
Importance
Lift
Cumulative Lift
8
6
4
2
0
20
40
60
80
100
Percentile
Model
Best
Data Roles
Decision Tree 1
Response
PAD
Predictors
Gender
Had at least 12 alcohol drinks...
Smoked a pipe at least 20 tim...
Smoked at least 100 cigarette...
Smoked cigars at least 20 tim...
Used chewing tobacco 20 tim...
Veteran/Military Status
Used snuff at least 20 times in...
Age started smoking cigarett...
How long since quit smoking ...
How often drink alcohol over ...
HDL-cholesterol (mg/dL)
Body Mass Index (kg/m**2)
# days drink alcohol per wk, ...
Ever have 5 or more drinks ev...
Total cholesterol (mg/dL)
Add
Options
Roles
Actions
Rules
Filters
Ranks

Figure 6. Decision Tree from SAS® Visual Statistics on Viya

Once satisfied with the decision tree, select the three vertical dots in the upper right-hand corner of this object and choose "Duplicate As" and then click the "Neural Network" option from the SAS® Visual Data Mining and Machine Learning (VDMML) package. This selection will automatically duplicate the decision tree from Figure 6 into a Neural Network model. The "Duplicate As" option will show the entire list of available models available for exploring.



SAS® Visual Analytics - Explore and Visualize Data
Search
Alex Ford
NHANES Peripheral Arterial Disease
Page 1
Page 2
Page 3
Decision Tree PAD (event=Yes) KS (Youden) 0.1592 Observations Used 7,187
Tree
Variable Importance
# days drink alcohol per wk, mo, yr
HDL-cholesterol (mg/dL)
How long since quit smoking cigarettes
Body Mass Index (kg/m**2)
Total cholesterol (mg/dL)
Veteran/Military Status
Age started smoking cigarettes regularly
0.0
2.5
Choose event level...
Remove all role assignments
Add title
Delete
Duplicate
Duplicate as
Create pipeline...
Save image
Export data...
Export model...
Derive predicted...
Derive a leaf ID variable
Change Decision Tree to
Cluster
Forest
Generalized Additive Model
Generalized Linear Model
Gradient Boosting
Linear Regression
Logistic Regression
Neural Network
Nonparametric Logistic Regression
Support Vector Machine
# days ...
B...
H...
PAD
No
Yes
Lift
Cumulative Lift
8
6
4
2
0
20
40
60
80
100
Percentile
Model
Best
Data Roles
Decision Tree 1
Response
PAD
Predictors
How often drink alcohol over ...
HDL-cholesterol (mg/dL)
Body Mass Index (kg/m**2)
# days drink alcohol per wk, ...
Ever have 5 or more drinks ev...
Total cholesterol (mg/dL)
Add

Figure 7. Duplicate Figure 6 as a Neural Network

MODEL COMPARISON

After creating a set of models, add a new page by clicking the plus sign that is next to the page tabs. In the objects pane, drag and drop a “Model Comparison” object onto the graphing screen. Upon doing so, a text box will appear prompting the user to choose the parameters and select from the previously created models for comparison. In this example, only the Decision Tree and Neural Network are available, however, additional models can be created or replicated with different parameters to be included in the model comparison. In Figure 8, we see that Decision Tree 1 is the selected champion model using a KS fit statistic. Note in the options that the fit statistic used for comparison can be changed.

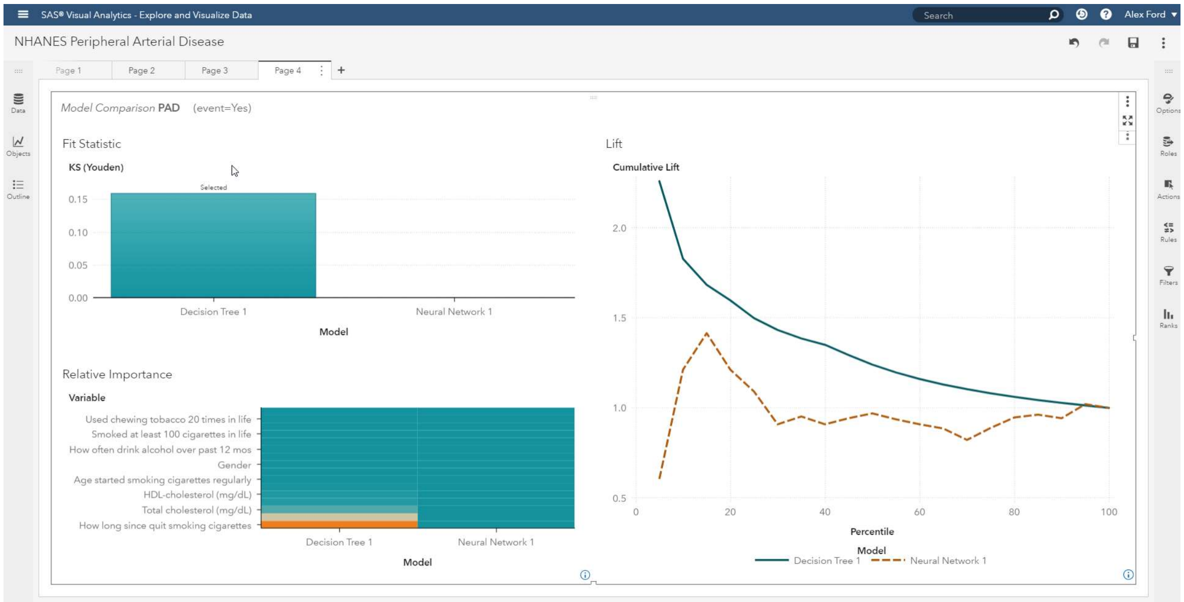
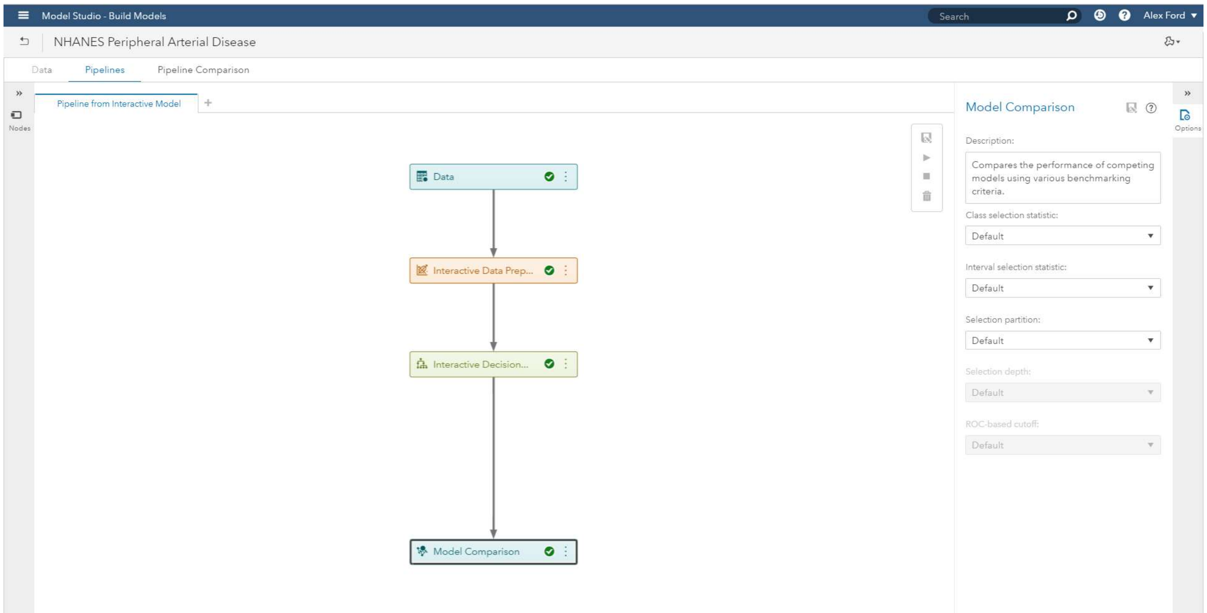


Figure 8. Model Comparison

GENERATING PIPELINES

We will generate a pipeline based upon the process that we have followed up until now. To do so, right-click within the model comparison object and select “Create Pipeline…”. This will open a project in SAS® Model Studio that contains the steps which were taken to create the model comparison from figure 8. By right-clicking on nodes within the pipeline, a user can add additional steps into the pipeline which will be accounted for when running using the play button. Pipelines have additional criteria that can be defined, including statistics for class selection, interval selection, and partition and depth selections.

In addition to adding new nodes, users can also create additional pipelines which follow a different process and compare pipelines. For example, the Decision Tree champion from Figure 8, may not be the selected model when compared against a new pipeline.



Model Comparison

Description:
Compares the performance of competing models using various benchmarking criteria.

Class selection statistic: Default

Interval selection statistic: Default

Selection partition: Default

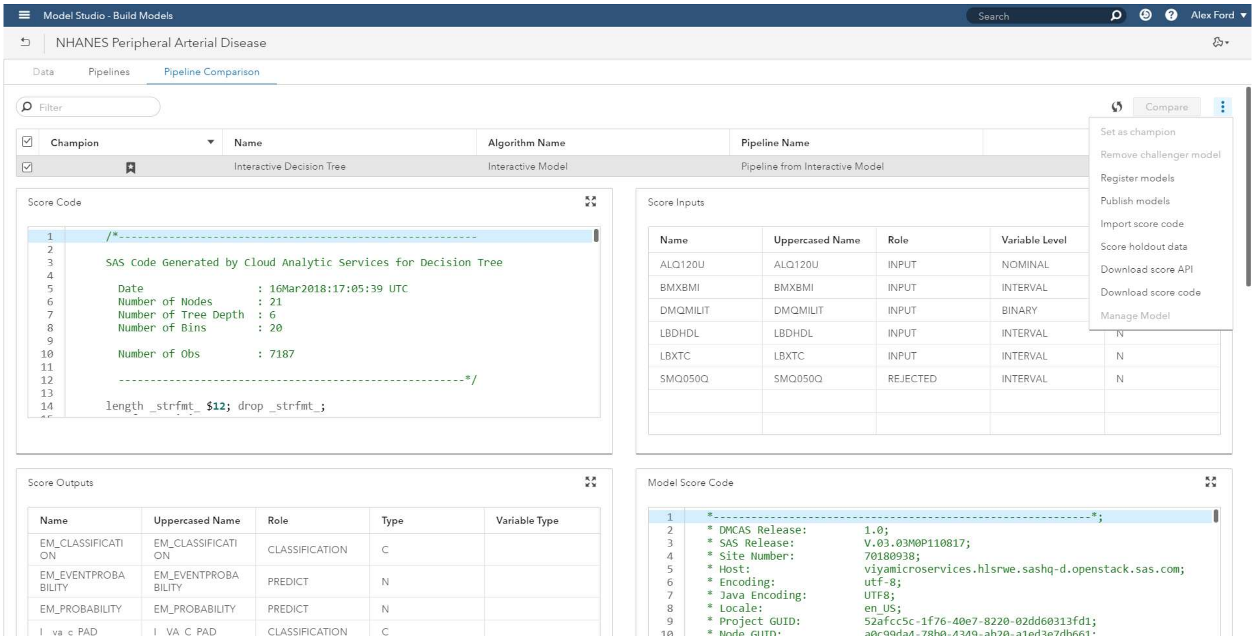
Selection depth: Default

ROC-based cutoff: Default

Figure 9. Pipeline

Once the pipeline(s) has successfully run, select the “Pipeline Comparison” tab at the top of the screen. This tab provides various diagnostic attributes from the pipeline that was just created including the score code, ROC Reports, Lift Reports, and Output.

Using the gear in the top right, users can choose to download the underlying batch code in either SAS, Python, or REST code. Additionally, using the three vertical dots in the top right, models can be published into a specific folder or database for use, making it extremely easy to publish models once they are complete.



☑	Champion	Name	Algorithm Name	Pipeline Name
☑		Interactive Decision Tree	Interactive Model	Pipeline from Interactive Model

Score Code

```
        /*------------------------------------------------------

        SAS Code Generated by Cloud Analytic Services for Decision Tree

          Date                   : 16Mar2018:17:05:39 UTC
          Number of Nodes        : 21
          Number of Tree Depth   : 6
          Number of Bins         : 20

          Number of Obs          : 7187

          ------------------------------------------------------*/

        length _strfmt_ $12; drop _strfmt_;
```

Score Inputs

Name	Uppercased Name	Role	Variable Level	
ALQ120U	ALQ120U	INPUT	NOMINAL	
BMXBMI	BMXBMI	INPUT	INTERVAL	
DMQMILIT	DMQMILIT	INPUT	BINARY	
LBDHDL	LBDHDL	INPUT	INTERVAL	N
LBXTC	LBXTC	INPUT	INTERVAL	N
SMQ050Q	SMQ050Q	REJECTED	INTERVAL	N

Score Outputs

Name	Uppercased Name	Role	Type	Variable Type
EM_CLASSIFICATION	EM_CLASSIFICATION	CLASSIFICATION	C	
EM_EVENTPROBABILITY	EM_EVENTPROBABILITY	PREDICT	N	
EM_PROBABILITY	EM_PROBABILITY	PREDICT	N	
I va c PAD	I VA C PAD	CLASSIFICATION	C	

Model Score Code

```
*------------------------------------------------------------*;
* DMCAS Release:          1.0;
* SAS Release:            V.03.03M0P110817;
* Site Number:            70180938;
* Host:                   viyamicroservices.hlsrwe.sashq-d.openstack.sas.com;
* Encoding:               utf-8;
* Java Encoding:          UTF8;
* Locale:                 en_US;
* Project GUID:           52afcc5c-1f76-40e7-8220-02dd60313fd1;
* Node GUID:              a0c99da4-78b0-4349-ab20-a1ed3e7db661;
```

Figure 10. Pipeline Comparison

CONCLUSION

The suite of visualization tools running on SAS® Viya are specifically designed to make generating high quality analytic reports and validated predictive models easy and fast. With these tools you can access your data wherever it is housed, build and share reports across your organization, and decrease time to insight generation. This workshop barely scratches the surface of what is available on the platform and is meant to serve exclusively as an extremely high-level overview. Once you familiarize yourself with the platform and all that is available within, you will wonder how your organization ever derived insights without it.

RECOMMENDED READING

Box, Jim. 2018. "Build Models without Code with the new SAS® Viya™ Visual Interface". *Proceedings of the PharmaSUG 2018 Conference.*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Alex Ford
SAS Institute
100 SAS Campus Drive
Cary, NC, 27513
Alex.Ford@sas.com

Andrea Coombs
SAS Institute
100 SAS Campus Drive
Cary, NC, 27513
Andrea.Coombs@sas.com