THE POWER TO KNOW.

Getting Started with SAS® Factory Miner 14.1

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *Getting Started with SAS® Factory Miner 14.1*. Cary, NC: SAS Institute Inc.

Getting Started with SAS® Factory Miner 14.1

Copyright © 2016, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

February 2016

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

Contents

Getting Started with SAS® Factory Miner 14.1 . . . . . *v*

Chapter 1 • Getting Started with SAS Factory Miner **1**
About SAS Factory Miner 1
Segmentation Modeling 1
SAS Factory Miner Workflow 2

Chapter 2 • Preparing to Create a SAS Factory Miner Project **3**
Import the Data Source 3
Define New Factory Metadata 3
Create the Model Templates 4
Specify Default Model Templates 6

Chapter 3 • Creating and Retraining a SAS Factory Miner Project **7**
Create the Project 7
Build the Profile 7
Review the Model Templates 7
Run the Model and Review the Results 8
Retrain the Models 8

Chapter 4 • Quick Reference **9**
Open Model Logs 9
Download Score Code 9
Download Model Reports 10
Where to Find Additional Documentation 10

Getting Started with SAS® Factory Miner 14.1

Audience

The audience for this book includes business analysts, data scientists, or model builders that are interested in learning how to use SAS Factory Miner.

Chapter 1

Getting Started with SAS Factory Miner

About SAS Factory Miner 1
Segmentation Modeling 1
SAS Factory Miner Workflow 2

About SAS Factory Miner

SAS Factory Miner is a web-based data mining application that is especially designed for segmentation modeling. With SAS Factory Miner, you identify segmentation variables, a target variable, and your desired models. SAS Factory Miner uses this information to build a collection of models within each segment and chooses a champion model based on a pre-specified criterion. You can easily import new data as it becomes available and retrain your models.

SAS Factory Miner is designed to build and retrain many models over a large number of data segments. Therefore, it includes two important tools for expediting the model building process. First, you can define factory metadata on a per variable basis. For example, if you are dealing with home loan applications and your target variable is a binary variable named BAD, you can define this information in the factory metadata. For every new data set imported into SAS Factory Miner that contains a variable named BAD, SAS Factory Miner will assign it the role of target and type of binary. Second, you can create a collection of model templates, and each template can contain your favorite setting for that model type. For example, suppose that your business rules permit only those decision trees that contain two splits per branch and a maximum depth of five levels. You can define a decision tree model template that contains these features and set it as a default model. Every new SAS Factory Miner project will contain that model.

Segmentation Modeling

When working with a large data set, you often find several heterogeneous subgroups. In this case, it is beneficial to segment the data into these subgroups and create separate models for each group. Segmentation modeling assumes that the observations in each subgroup are more similar to each other than they are to the data set as a whole. Because of this similarity, the model created for each segment should be more powerful within that segment than a single model created on the entire data set.

In SAS Factory Miner, you use the measurement levels of up to three categorical variables to define the model segments. Each unique combination of measurement levels defines a segment. For example, suppose your data set contains categorical variables for customers' ZIP codes that contains 10 levels, and income brackets that contains three levels. You can specify both the ZIP code and income bracket variables as segment variables. SAS Factory Miner creates a segment for each unique combination of ZIP code and income bracket. This means that your project will contain 30 segments.

SAS Factory Miner Workflow

The SAS Factory Miner workflow contains eight distinct steps:

1. Import the Data Source — Identify your data sets of interest and import them into SAS Factory Miner.
2. Define New Factory Metadata — Factory metadata constitutes knowledge about specific variables. This knowledge is automatically applied when a project is created.
3. Create the Model Templates — A model template defines a specific statistical model and any applicable performance parameters that can be used in any SAS Factory Miner project.
4. Specify Default Model Templates — Indicate which model templates you want to include in all new projects.
5. Create the Project — Select a specific data set to investigate. Applicable factory metadata is applied when the project is created.
6. Build the Profile — Creating a project profile prepares the input data for segmentation modeling.
7. Review the Model Templates — Make any final adjustments to the models used in your project.
8. Retrain the Models — As new data becomes available, you import that data into your project and retrain the models.

This document contains a complete example that demonstrates each of the steps in the SAS Factory Miner workflow. The data used in this document is vehicle fuel economy and emissions information published by the United States Environmental Protection Agency (EPA). You will use the data from 2010 through 2013 to create a project , and the data from 2014 to retrain the project. This data is provided in a ZIP file

Chapter 2

Preparing to Create a SAS Factory Miner Project

Import the Data Source 3

Define New Factory Metadata 3

Create the Model Templates 4
Create a Decision Tree Model Template 4
Create a Second Decision Tree Model Template 4
Create a Bayesian Network Model Template 5
Create a Neural Network Model Template 5
Create a Regression Model Template 6

Specify Default Model Templates 6

Import the Data Source

Before you can import a data source, your administrator must save that data source in a library that SAS Factory Miner can use to access and register the data source. An example of these steps is provided in *SAS Factory Miner: Administration and Configuration*. The example in this book assumes that you or your administrator has followed those steps and that the EPA data is available in SAS Factory Miner.

To import the EPA data sets:

1. Click ☰ and select **Data Sources**.
2. In the upper right corner of the screen, click ⊕ to add a new data source.
3. In the Open window, navigate to the **EPA10_13** data set. Select **EPA10_13** and click **Open**.
4. Repeat steps 2 and 3 for the **EPA14** data set.

Define New Factory Metadata

1. On the Data Sources page, click **Select** in the upper right corner of the page. Select the **EPA10_13** data set.
2. Click ▤ to open the Define Data window.

3. Select **Vehicle_Type** and click **Segment**.
4. Select **THC_g_mi** and click **Target**.
5. Click **Save**.

The only way to delete factory metadata is from the Factory Metadata page.

Create the Model Templates

A model template contains all relevant data preparation and model configuration settings necessary to build a statistical model. Before creating a project, you should create a model template for each statistical model that you want to include in the project.

Create a Decision Tree Model Template

1. In the upper right corner of the Model Template page, click to add a new model template.
2. Enter `Decision Tree GS` in the **Name** field.
3. Select **Decision Tree** in the **Model algorithm** field.
4. Click **Save**. The Decision Tree GS window appears.
5. Drag **Impute** from the **Ignore** area to the **Use** area. This example uses the default imputation settings.
6. Make the following changes to the **Decision Tree** settings:
   - Set **Nominal target splitting criterion** to **Information Gain Ratio**.
   - Set **Maximum Depth** to **20**.
   - Set **Nominal target assessment measure** to **Misclassification**.
7. Click **Save**.

Create a Second Decision Tree Model Template

1. In the upper right corner of the Model Templates page, click to add a new model template.
2. Enter `Decision Tree GS 2` in the **Name** field.
3. Select **Decision Tree** in the **Model algorithm** field.
4. Click **Save**. The Decision Tree GS 2 window appears.
5. Drag **Variable selection** from the **Ignore** area to the **Use** area.

   Make the following changes to the **Variable selection** properties:
   - Set **Selection method** to **LAR**.
   - Set **Stop criterion** to **AICC**.
6. Drag **Filter** from the **Ignore** area to the **Use** area. **Filter** should appear above **Variable selection**. Use the default filtering settings.
7. Click **Decision Tree** and make the following changes:

- Set **Nominal target splitting criterion** to **Information Gain Ratio**.
- Set **Maximum Depth** to **20**.
- Set **Nominal target assessment measure** to **Misclassification**.

8. Click **Save**.

Create a Bayesian Network Model Template

1. In the upper right corner of the Model Templates page, click to add a new model template.
2. Enter `Bayesian Network GS` in the **Name** field.
3. Select **Bayesian Network** in the **Model algorithm** field.
4. Click **Save**. The Bayesian Network GS window appears.
5. Drag **Impute** from the **Ignore** area to the **Use** area. This example uses the default imputation settings.
6. Click **Bayesian Network Classifier** to edit the Bayesian network settings. Set **Network model** to **BAN**.
7. Click **Save**.

Create a Neural Network Model Template

1. In the upper right corner of the Model Templates page, click to add a new model template.
2. Enter `Neural Network GS` in the **Name** field.
3. Select **Neural Network** in the **Model algorithm** field.
4. Select **Mark as default**.
5. Click **Save**. The Neural Network GS window appears.
6. Drag **Variable selection** from the **Ignore** area to the **Use** area.

   Make the following changes to the **Variable selection** properties:

   - Set **Target model** to **Unsupervised selection**.
   - Set **Unsupervised maximum steps** to `500`.
   - Set **Selection method** to **LAR**.
   - Set **Stop criterion** to **AICC**.

7. Click **Neural Network** and make the following changes:

   - Set **Architecture** to **two layers with direct architecture**.
   - Set **Number of hidden neurons** to **250**.
   - Set **Number of tries** to **5**.
   - Select **Use missing as level**.

8. Click **Save**.

Create a Regression Model Template

1. In the upper right corner of the Model Templates page, click to add a new model template.
2. Enter `Regression GS` in the **Name** field.
3. Select **Regression** in the **Model algorithm** field.
4. Click **Save**. The Regression GS window appears.
5. Drag **Variable selection** from the **Ignore** area to the **Use** area. This example uses the default variable selection settings.
6. Click **Regression** and make the following changes:
   - Select **Two-factory interactions**.
   - Select **Use missing as level**.
   - Set **Selection method** to **Stepwise**.
   - Set **Stop criterion** to **Significance Level**.
7. Click **Save**.

Specify Default Model Templates

1. On the Model Templates page, click **Select**.
2. Select the following models:
   - **Bayesian Network GS**
   - **Bayesian Network template**
   - **Decision Tree GS**
   - **Decision Tree GS 2**
   - **Decision Tree template**
   - **Neural Network GS**
   - **Neural Network template**
   - **Regression GS**
   - **Regression template**
3. Click ✓ to set these model templates as defaults.

   Note: If any of your selected model templates is already a default model template, then ✓ is unavailable. Deselect those models before trying again.
4. Click **Select**. Select any model template not listed above that is labeled **Default**. Click ✕ to remove these items from the list of default models.

Chapter 3

Creating and Retraining a SAS Factory Miner Project

Create the Project 7
Build the Profile 7
Review the Model Templates 7
Run the Model and Review the Results 8
Retrain the Models 8

Create the Project

1. On the Projects page, click to create a new project.
2. Enter `Getting Started` in the **Name** field.
3. Select **EPA10_13** in the **Data source** field.
4. Click **Save**. The Data page appears.

   From the Data page, you can reassign variable type, role, and level. You can also perform some basic data exploration tasks. Because you assigned factory metadata for **THC_g_mi** and **Vehicle_Type**, no further work is necessary. You are encouraged to repeat this example with different target and segment variables.

Build the Profile

1. On the Data page, click **Build Profile** in the upper right corner of the page.
2. Review the project profile. The Profile page enables you to specify whether a segment is modeled based on the number of observations in the training data set.

Review the Model Templates

1. On the Profile page, click **Model Templates** in the upper left corner of the page.

2. Ensure that the model templates that you created in the previous chapter are included here. If they are not included, click ➕ in the upper right corner of the page and select the model templates that you want to include.

Run the Model and Review the Results

This example contains two models per segment that are not successfully trained and includes instructions for retrieving model logs.

1. On the Model Templates page, click **Run** in the upper right corner of the page.

   The Results page opens and is initially unpopulated. As the models are completed, the results table and segment summary graphs are updated. Wait until all models are completed and the segment summary graphs appear.

2. In the summary table, notice that two models failed in each segment. Click **Segment 2** to investigate this segment.

3. On the Results page for Segment 2, you can see that both of the Bayesian Network models failed. To investigate, open the log for that model.

   a. Select **Bayesian Network GS**.

   b. Click **Actions** ⇨ **Log**.

   c. Open the log file with an appropriate application.

   d. Use the model log to determine why the model failed to complete and make the appropriate adjustments in the model template.

      In this case, because your target variable is numeric, Bayesian networks are inappropriate models.

   e. Close the log file and return to SAS Factory Miner.

4. Click **Decision Tree GS 2** because it was chosen as the champion model. Review the plots and graphs for this model.

5. Review the plots and graphs for the other models as well.

6. Repeat steps 3–6 for the other two segments in your project.

Retrain the Models

1. On the Results page, click to import a new data set.

2. Select **EPA2014**.

3. Click **Save**.

4. You are taken back to the Profile page, and all segments are designated to run. Click **Run**.

5. Review the segments as you did in the previous section.

Chapter 4

Quick Reference

Open Model Logs 9
Download Score Code 9
Download Model Reports 10
Where to Find Additional Documentation 10

Open Model Logs

1. On the Results page, click the segment that contains the model of interest.
2. On the Segment page, select the model of interest.
3. In the upper right corner of the screen, select **Actions** ⇨ **Log**.
4. Select an appropriate program to open the log file. The log file is a plain text file.

Download Score Code

To download score code for multiple segments:

1. From the Results page, click in the upper right corner of the screen.
2. In the Download Score Code window, specify whether you want to download the score code for all of the models or just the champion models.

To download score for a single mode:

1. From the Results page, click the segment that contains the model of interest.
2. On the Segment page, select the model of interest.
3. In the upper right corner of the screen, select **Actions** ⇨ **Download Score Code**.
4. Save the score code file or select an appropriate program to open the score code file. The score code file is a ZIP file.

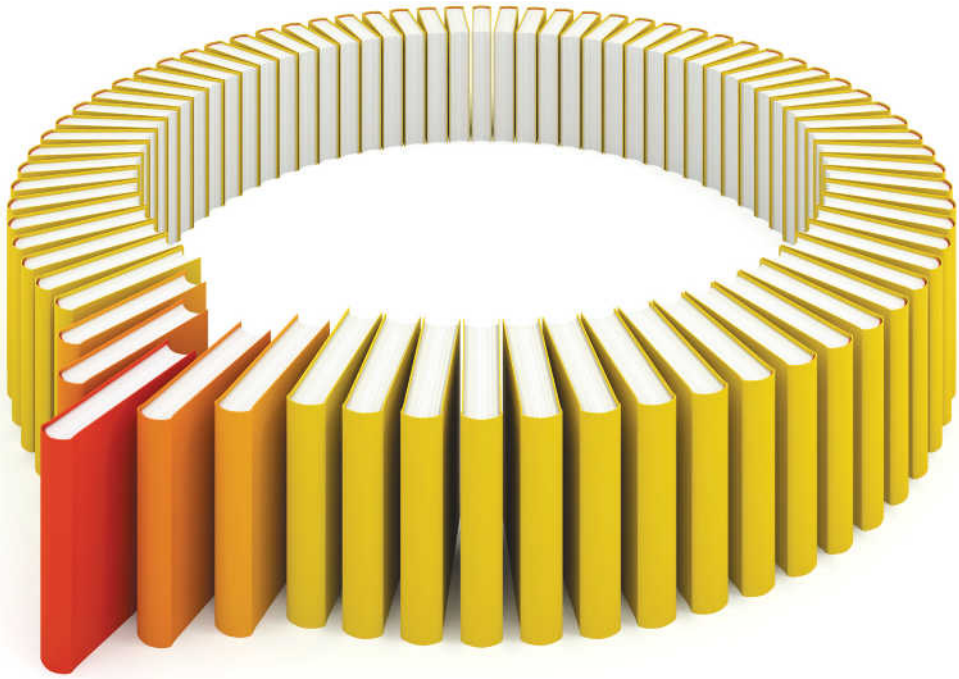
Download Model Reports

1. From the Results page, click the segment that contains the model of interest.
2. On the Segment page, select the models of interest.
3. In the upper right corner of the screen, select **Actions** ⇨ **Report**.
4. In the Report window, specify whether you want to include model comparisons and the file format. Click **Generate**.
5. After the report is generated, click **Yes** in the Report window.
6. Save the report file or select an appropriate program to open the score code file.

Where to Find Additional Documentation

The most current technical resources for SAS Factory Miner are available on the SAS Factory Miner page on the SAS support site.

Your experience with SAS software should be as smooth as possible. Please submit your feedback.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2013 SAS Institute Inc. All rights reserved. S107969US.0613