



Insights Hub

Predictive Learning

System Manual
03/2024

Welcome to Predictive Learning Help 1

Predictive Learning Help 2

Open Source Software 3

Legal information

Warning notice system

This manual contains notices you have to observe in order to ensure your personal safety, as well as to prevent damage to property. The notices referring to your personal safety are highlighted in the manual by a safety alert symbol, notices referring only to property damage have no safety alert symbol. These notices shown below are graded according to the degree of danger.

DANGER

indicates that death or severe personal injury **will** result if proper precautions are not taken.

WARNING

indicates that death or severe personal injury **may** result if proper precautions are not taken.

CAUTION

indicates that minor personal injury can result if proper precautions are not taken.

NOTICE

indicates that property damage can result if proper precautions are not taken.

If more than one degree of danger is present, the warning notice representing the highest degree of danger will be used. A notice warning of injury to persons with a safety alert symbol may also include a warning relating to property damage.

Qualified Personnel

The product/system described in this documentation may be operated only by **personnel qualified** for the specific task in accordance with the relevant documentation, in particular its warning notices and safety instructions. Qualified personnel are those who, based on their training and experience, are capable of identifying risks and avoiding potential hazards when working with these products/systems.

Proper use of Siemens products

Note the following:

WARNING

Siemens products may only be used for the applications described in the catalog and in the relevant technical documentation. If products and components from other manufacturers are used, these must be recommended or approved by Siemens. Proper transport, storage, installation, assembly, commissioning, operation and maintenance are required to ensure that the products operate safely and without any problems. The permissible ambient conditions must be complied with. The information in the relevant documentation must be observed.

Trademarks

All names identified by ® are registered trademarks of Siemens AG. The remaining trademarks in this publication may be trademarks whose use by third parties for their own purposes could violate the rights of the owner.

Disclaimer of Liability

We have reviewed the contents of this publication to ensure consistency with the hardware and software described. Since variance cannot be precluded entirely, we cannot guarantee full consistency. However, the information in this publication is reviewed regularly and any necessary corrections are included in subsequent editions.

Table of contents

1. Welcome to Predictive Learning Help	5
1.1. Welcome to Predictive Learning Help!.....	5
2. Predictive Learning Help	6
2.1. Introduction to Predictive Learning.....	6
2.2. Navigating Predictive Learning.....	7
2.3. Predictive Learning Workflow.....	8
2.4. About External Data.....	9
2.5. Connecting to IoT Data.....	9
2.6. Managing Data Imports.....	10
2.7. Managing Files and Folders.....	11
2.8. Managing Datasets.....	14
2.9. Managing Spark Pipeline Models.....	17
2.10. Managing Analytics Workspaces.....	18
2.11. Navigating to the Manage Analytics Workspaces Page.....	18
2.12. Using the Advanced Configuration Dialog.....	19
2.13. Specifying the S3 Bucket Location.....	20
2.14. Defining Cluster Auto Shutdown Time.....	20
2.15. Starting a Cluster.....	22
2.16. Creating a Workspace.....	25
2.17. Opening an Existing Workspace.....	27
2.18. Adding an Exploration Panel.....	27
2.19. Viewing Data Distributions.....	28
2.20. Adding an Exploration Panel.....	29
2.21. Linear Regression Analysis.....	30
2.22. Logistic Regression Analysis.....	34
2.23. Adding a Transformation Panel.....	38

2.24. Selecting Columns to Carry Over.....	40
2.25. Transforming Data Using Joins.....	40
2.26. Filtering a Dataset.....	42
2.27. Filtering Workspace Data.....	43
2.28. Replacing Missing Values.....	46
2.29. Principal Component Analysis (PCA).....	47
2.30. Normalizing Data With StandardScaler.....	49
2.31. Running an Analysis.....	51
2.32. Launching a Service.....	52
2.33. Making API Calls from Zeppelin Notebook.....	53
2.34. Using Jupyter Notebook.....	58
2.35. Using GPU.....	67
2.36. Viewing Usage Metrics.....	67
2.37. Manage Environment Configurations.....	68
2.38. Manage Environments.....	72
2.39. Managing Analytical Models (External).....	74
2.40. Managing Analytical Models Details.....	79
2.41. Managing Docker Models.....	81
2.42. Running and Managing Jobs.....	87
2.43. Running Docker Containers as Jobs.....	90
2.44. Running Scheduled Jobs.....	91
2.45. Managing Sources.....	94
2.46. Predictive Learning (PrL) API.....	97
3. Open Source Software.....	100
3.1. Open Source Software.....	100

Welcome to Predictive Learning Help

1.1 Welcome to Predictive Learning Help!

Predictive Learning combines analytics, statistics, and machine learning algorithms to provide unmatched insight into trends in your data. Combining predictive technology with IoT, service, field, and other data streams allows you to generate a deeper impact on the customer experience. Leveraging as-used data to identify patterns and sequences of events, companies can engage with customers and resolve potential issues before problems arise.

This data insight allows you to see into the future, refine your systems to produce the highest level of product performance and quality. The Predictive Learning interface gives data scientists, business analysts, and application developers the tools they need to increase customer satisfaction and improve net promoter scores.

Predictive Learning allows you to:

- Easily bring external data into a Predictive Learning workspace.
- Rapidly build and execute predictive models using both statistical and machine learning algorithms.
- Pinpoint where in the supply chain an issue originates.
- Intervene ahead of negative manufacturing events or product performance issues.

With your feedback, future releases will continue to enhance Predictive Learning through integrating more algorithms and statistical analysis capabilities.

2.1 Introduction to Predictive Learning

Predictive Learning provides a workspace into which users can bring data in structured, unstructured, and semi-structured formats, and multiple sources. Once your data is in the workspace, you can run analyses that help you anticipate coming events and predict when issues might arise. This will allow you to boost product quality and proactively:

- Intervene with preventive maintenance
- Disrupt fatal sequences of manufacturing events
- Reduce the number of field failures
- Enhance product performance
- Enhance customer experience

Predictive Learning employs machine learning (ML) algorithms, statistics functions, transformations, and filtering to bring you the most comprehensive and flexible means to access and work with your data.

Features of Predictive Learning

Predictive Learning helps companies attain unprecedented product quality. Forecasting field performance reveals not only potential failure sequences, but also optimal sequences for higher product performance. By correlating field performance data with product design and manufacturing process data, Predictive Learning enables you to improve and optimize product quality – before the product leaves the factory. Tremendous value accrues as these higher quality products translate into lower service and warranty costs. Predictive Learning allows you to benefit from:

- Simple data collection and preparation from external sources
- Rapidly build and execute predictive models
- Automatically analyze numerous sequences through predictive algorithms
- Discover the myriad sequences that create product field failures and preemptively disrupt fatal sequences
- Proactively push fixes (parts, software, hardware, or firmware) to resolve predicted points of failure or degraded performance

- Forecast field performance to improve product quality

You can also schedule preventive maintenance before failures occur, and:

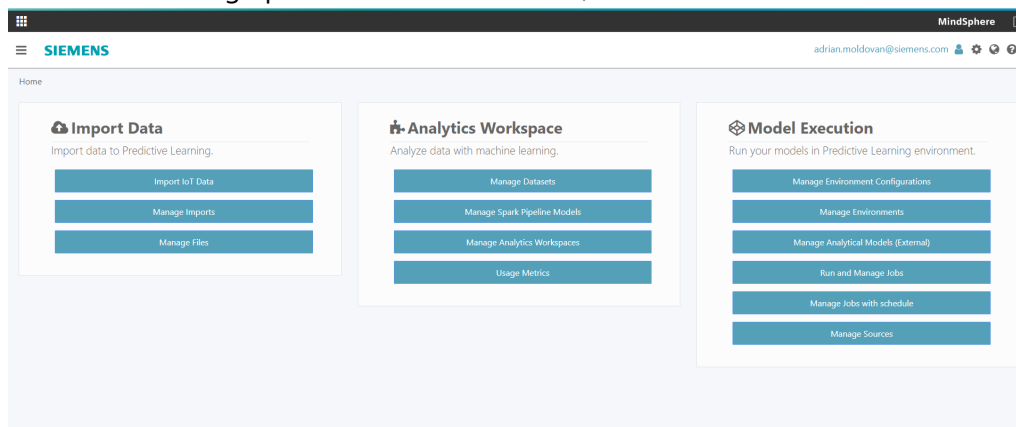
- Identify patterns within normal operations that may otherwise go unnoticed
- Identify precursors, such as sudden spikes within a normal operating range and specific test results
- Correlate the precursors with machine maintenance records to predict which machines may suddenly require unscheduled maintenance or downtime
- Schedule preventative maintenance to avoid unscheduled machine downtime

Roles and Permissions in Predictive Learning

Roles and permissions in Predictive Learning are created and edited in the Settings application on the Launchpad. All Predictive Learning administrators and some users can create and edit roles and permissions via the Settings application. If the Settings application does not appear on the Launchpad application, please contact your Tenant administrator.

2.2 Navigating Predictive Learning

Predictive Learning opens from the main menu, which looks like this:



Here is a brief description of Predictive Learning functionality:

Application Tool Bar: click the icon in the top right corner to log out of Predictive Learning.

Predictive Learning Tool Bar: click the user icon to access your user profile. Click the question mark to access Predictive Learning Help or Support.

Import IoT Data: use this page to import your IoT data from various connected devices. An import job is created that you can access anywhere datasets are available.

Manage Imports: shows all IoT import jobs you created and their status. Also allows you to delete any import jobs you no longer need.

Manage Files: use this page to upload your data files from your local machine to an S3 bucket, and then download the files from the S3 bucket to your local machine. You can also view a list of

files and folders and perform various actions on those files and folders.

Manage Datasets: once you create a dataset, it resides in a table that displays statistics about all of the datasets you have created or have access to. On the Manage Datasets page, you can open, rename, share, or delete datasets.

Manage Spark Pipeline Models: provides a list of all Spark models you have created in the Predictive Learning app, and provides details about each.

Manage Analytics Workspaces: on this page, you can view and open all the existing workspaces, create new workspaces, or delete those you no longer need. You can also start a cluster, open a workspace, add exploration and transformation panels, and launch external services.

Usage Metrics: lets you see how many Predictive Learning compute hours remain for your organization, and lists your individual transactions and the number of compute hours you have used.

Manage Environment Configurations: allows administrators to create, update, delete, or view a list of environment configurations based on an available configuration template.

Manage Environments: provides a list of environment configurations created and saved by your administrator and allows you to select and start or stop an environment to run your model on.

Manage Analytical Models (External): use this page to upload models developed outside of Predictive Learning and store them in an external S3 bucket.

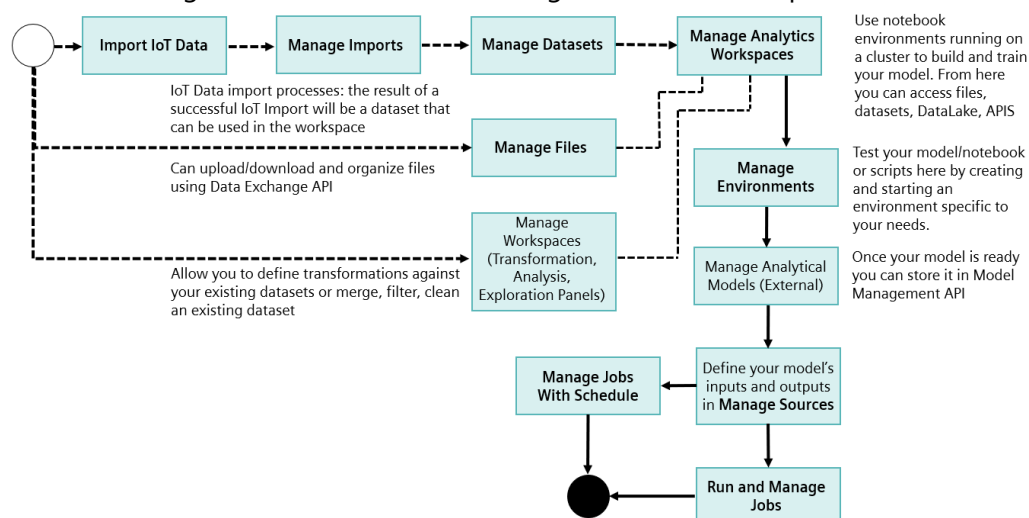
Run and Manage Jobs: use this page to select a model and execute the job. You can also view a list of jobs you have run and the details about each.

Manage Jobs with schedule: here you can define jobs that are managed by schedules.

Manage Sources: in this page you can define various data sources for your jobs, based on how you intend to use them, that is, as sources for input or output.

2.3 Predictive Learning Workflow

This diagram illustrates and describes the Predictive Learning workflow, from the import of data sources, through transformations and management, to final outputs.



2.4 About External Data

External data in Predictive Learning is stored and accessed through Amazon S3 buckets. External data is accessible from the Zeppelin environment. You can load data from your external S3 bucket (e.g., with Spark APIs), and then save the dataset in Predictive Learning.

The data buckets are set up at the tenant level with user folders. The following are supported for external buckets and Predictive Learning:

- One external bucket per user can be configured using the Advanced Configuration dialog box on the Manage Analytics Workspace page.
- The external bucket must exist in the same AWS region where the Predictive Learning application is hosted (currently eu-central-1 only)

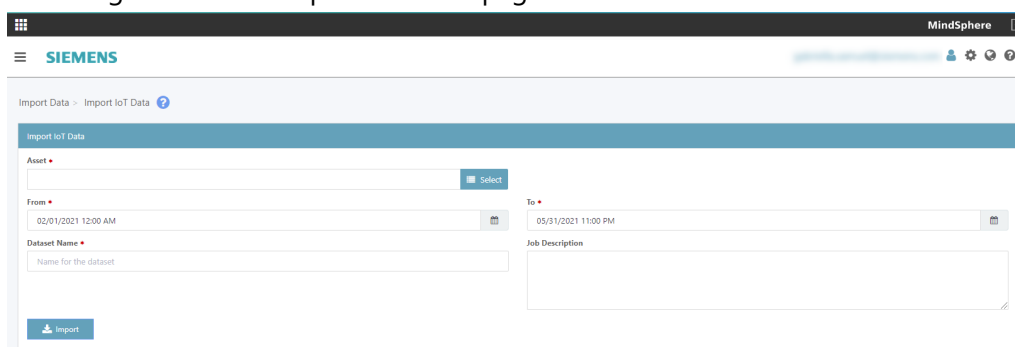
Contact your PrL administrator for more information about setting up access to your Amazon S3 bucket.

2.5 Connecting to IoT Data

This feature allows you to import your IoT data and then run predictive analysis on the data from your internet-connected devices. If you give a dataset a name that is already in use, your current import will overwrite the previously existing dataset.

Import IoT Data Page

This image shows the Import IoT Data page.



The screenshot shows the 'Import IoT Data' page within the Siemens MindSphere application. The page has a header with the Siemens logo and a user profile. The main content area is titled 'Import IoT Data' and contains several input fields: 'Asset' with a 'Select' button, 'From' date (02/01/2021 12:00 AM), 'To' date (05/31/2021 11:00 PM), 'Dataset Name' (placeholder: Name for the dataset), and 'Job Description'. An 'Import' button is at the bottom left.

How to Import IoT Data

Both shared assets and root assets are supported; however, not all root assets can store data, therefore you need to check that the asset you select either contains data or has a child asset that can store data.

Follow these steps to import your IoT data:

1. Click Import IoT Data on the **Predictive Learning** menu. *The Import IoT Data page opens.*
2. Click Select in the **Asset Type** field and select an asset from the list. *The asset type you selected appears in the field and the Asset field appears.*
3. Select an asset from the **Asset** field.
4. Click Select in the **Property Set** field and select a property set from the list. *The property set you selected appears in the field.*
5. Select **From** and **To** dates. Minimum time frame is one hour. Maximum time frame is one year.
6. Enter a name (no forward or backward slashes allowed) for the data import job.
7. Enter an optional **Job Description**.
8. Click **Import**. *A message that warns about the risk of overwriting a dataset displays and asks if you want to proceed.*
9. Click **Yes**. *A success message displays and the import name displays on the Manage Data Import, Manage Datasets, and Workspace pages.*

2.6 Managing Data Imports

The Manage Data Imports page displays a list of all the import jobs you created on the Import IoT Data and Import Product Intelligence Data pages. It shows the name, description, data source, status, error message (if the import failed), date created, and date of last update for each import job. You can delete any import jobs you no longer want on this page.

Manage Data Imports Page

This image shows the Manage Data Imports page.

Actions	Name	Description	Data Source	Status	Error Message	Created	Updated
	PRL_IOTImport_for_s hare_1675587029040	IOT data Job description	IoT	Finished		February 5, 2023 8:51 AM	February 5, 2023 8:52 AM
	PRL_IOTImport_for_s hare_1675586777886	IOT data Job description	IoT	Finished		February 5, 2023 8:47 AM	February 5, 2023 8:48 AM
	Foundation_IOTImport_ rt_E2E167557552219 9	IOT data Job description	IoT	Finished		February 5, 2023 5:39 AM	February 5, 2023 5:40 AM
	PRL_IOTImport_for_s hare_1675513492161	IOT data Job description	IoT	Finished		February 4, 2023 12:25 PM	February 4, 2023 12:27 PM
	PRL_IOTImport_for_s hare_1675512822081	IOT data Job description	IoT	Error	ServiceException - Please select the correct date range to get IoT data. Data and asset should exists for that range.timeout	February 4, 2023 12:15 PM	February 4, 2023 12:15 PM
	PRL_IOTImport_for_s	IOT data Job description	IoT	Finished		February 3, 2023 4:33 PM	February 3, 2023 4:33 PM

How to Delete Data Imports

Follow these steps to delete any data import jobs you no longer want:

1. Click Manage **Imports** on the Predictive menu and select **Manage Data Imports**. *The Manage Data Imports page opens.*
2. Find the import job you want to delete in the list and click the delete icon. *A message appears asking you to confirm the action.*
3. Click **Yes** to proceed. *The import job is deleted and no longer appears in the list.*

2.7 Managing Files and Folders

The Manage Files page allows you to upload your data files from your local machine to an S3 bucket, and then download the files from the S3 bucket to your local machine. You can also view a list and perform various actions on those files and folders.

About Files

Predictive Learning supports:

- CSV and JSON files
- Up to 100 MB in size
- Filenames must be unique within the same folder

Actions you can take with files:

- Cut
- Download
- Rename
- Delete

Personal and Shared Locations

Two locations are available for file uploads and downloads:

- Personal: only you can view and access files in a personal location.
- Shared: anyone who has a standard user role in the current tenant can view, download, and delete files from shared locations.

About Folders

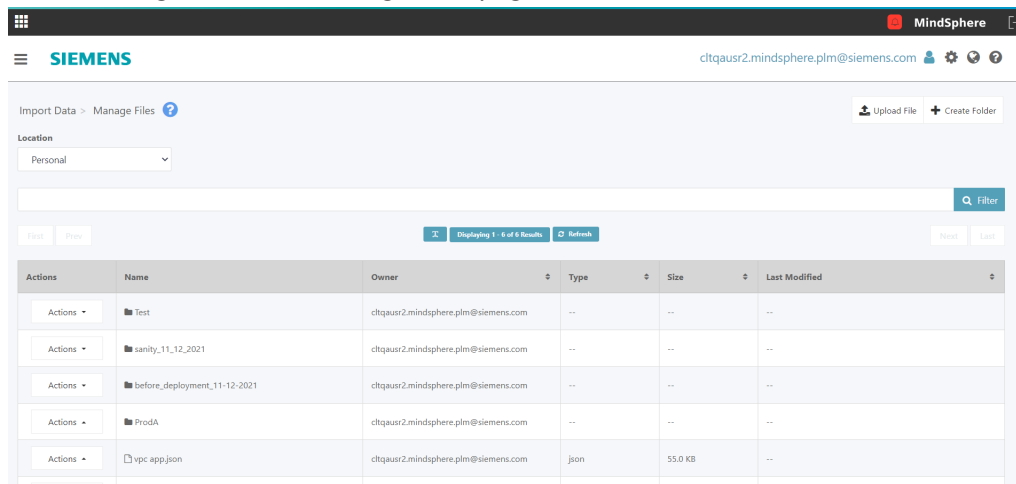
You can create multiple levels of folders on the Manage Files page and add files to each folder.

You can perform these actions on folders:

- Open
- Rename
- Delete

Manage Files Page Illustration

The following shows the Manage Files page.



How to Access Manage Files

Follow these steps to access the Manage Files page:

1. Log into Predictive Learning. The Main menu opens.
2. Click **Manage Files** on the **Import Data** tile. *The Manage Files page appears.*

How to Upload a File

Currently the uploading of CSV and JSON to Predictive Learning are supported.

Follow these steps to upload a file:

1. Navigate to the **Manage Files** page.
2. Select Personal or Shared from the **Location** field: Personal or Shared. Default is Personal. *The option you select displays in the field.*
3. Click the **Upload** button. *The Upload File pop-up window opens.*
4. Click the **Browse** button. *The Open pop-up window opens.*
5. Navigate to the file you want to upload, select it, and click **Open**. *The file name you select displays in the grid.*

6. Click **Upload**. *The file uploads to the specified destination and the progress bar indicates the percentage of completion. Once the upload is complete, a success message displays.*

How to Create a Folder

Follow these steps to create a folder:

1. *Navigate to the **Manage Files** page.*
2. Click **Create Folder**. *A pop-up window opens.*
3. Enter a name for the new folder.
4. Click **Submit**. *The folder is created and appears at the bottom of the grid.*

How to Search for a File

The search field allows you to look for a particular file already uploaded to Predictive Learning. Follow these steps to search for a file:

1. Navigate to the **Manage Files** page.
2. Enter a name or partial name in the **search** field.
3. Click the magnifying lens. *The system returns a list of matches from which you can select.*

How to Cut and Paste a File

Cut and paste works the same as any Windows application. It moves a file from one location to another, and deletes it from the original location.

Follow these steps to cut and paste a file:

1. Navigate to the **Manage Files** page.
2. Locate the file in the grid and click the **Actions** button.
3. Click **Cut**. *A Paste button appears next to Create Folder at the top of the page.*
4. Select the folder you want to move the file to in the grid and click the **Paste** button. *The file is moved to the new folder and deleted from the original location.*

How to Download a File

Follow these steps to download a file:

1. Access the **Manage Files** page. *The Manage Files page appears.*
2. Locate the file in the grid and click the **Actions** button.

3. Select **Download** from the **Actions** drop-down list. *The file is downloaded to your computer and the file name appears at the bottom of the browser window.*
4. Click the download and select an option. You can choose to open, always open, or show the file in a folder.
5. Save the file to a location of your choice.

How to Rename a File

Follow these steps to rename a file:

1. Access the **Manage Files** page. *The Manage Files page appears.*
2. Locate the file in the grid and click the **Actions** button. *A menu appears.*
3. Click **Rename**. *A pop-up window opens.*
4. **Provide the new name for the file** and click **Submit**. *The new file name displays in the grid.*

How to Delete a File

Follow these steps to delete a file:

1. Access the **Manage Files** page. *The Manage Files page appears.*
2. Locate the file in the grid and click the **Actions** button. *A menu appears.*
3. Click **Delete**. *A warning message appears.*
4. Click **OK** to delete the file. *The file is deleted and no longer appears in the grid.*

How to Delete a Folder

1. Navigate to the **Manage Files** page. *The Manage Files page appears.*
2. Locate the folder in the grid and click the **Actions** button. *A menu appears.*
3. Click **Delete**. *A warning message appears. If the folder contains files, the warning message indicates that.*
4. Click **OK** to delete the folder. *The folder is deleted and no longer appears in the table.*

2.8 Managing Datasets

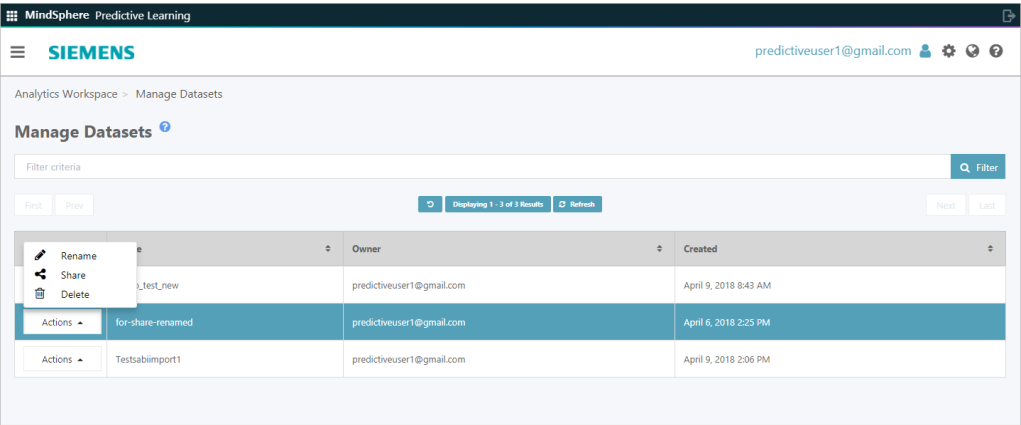
The Manage Datasets page displays the names of datasets that you own or that have been shared with you.

The Managing Datasets page displays the following for datasets:

- Date created
- Owner's name
- Actions: rename, share, or delete.

Manage Datasets Page

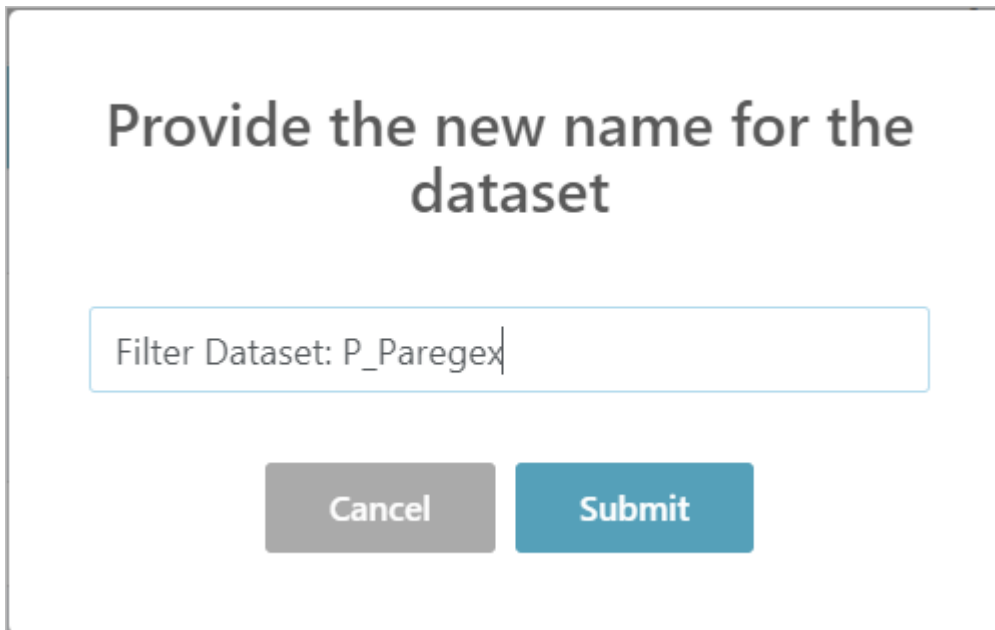
This shows the Manage Datasets page:



How to Rename a Dataset

Follow these steps to rename a dataset:

1. Select Manage Datasets from the **Predictive Learning** menu. *The Manage Datasets page opens.*
2. Select Rename from the **Actions** drop-down list in the row of the dataset you want to rename. *A dialog box displays the current dataset name.*



Provide the new name for the dataset

Filter Dataset: P_Paregex

Cancel Submit

3. Enter a name for the dataset and click **Submit**. *The dialog box closes and the new dataset name appears in the list.*

How to Share a Dataset

You can share a dataset that you own with any user in your tenant who has access to Predictive Learning. Once a dataset is shared, it appears in the shared-with user's Manage Datasets list. Look at the Owner column to determine whether a dataset was shared with you, or you are the owner. Only owners can share, rename, or delete a dataset.

Follow these steps to share a dataset:

1. Select Manage Datasets from the **Predictive Learning** menu. *The Manage Datasets page appears.*
2. Select Share from the **Actions** menu for the dataset you want to share. *The Share dataset pop-up window opens.*
3. Select the users that you want to share the dataset with from the **Contacts List**. *The selected users appear in the Shared With Contacts pane at the bottom.*
4. When you have made all of your selections, click the **Share** button. *A Success message displays and the dataset is shared with the users you specified.*

How to Delete a Dataset

Use caution about deleting datasets, because, if you delete a dataset that is referenced by a workspace configuration, you will lose access to the dataset and the workspace will not function properly.

Follow these steps to delete a dataset:

1. Click Manage Datasets on the **Predictive Learning** menu. *The Manage Datasets page opens.*
2. Select Delete from the **Actions** menu on the row that holds the dataset you want to delete. *A warning message asks if you are sure you want to delete the dataset.*
3. Click *Delete*. *The dataset is deleted and no longer appears in the list.*

2.9 Managing Spark Pipeline Models

The Manage Spark Pipeline Models page provides details on Spark models created in Predictive Learning (PrL). Models saved through the Notebook feature also appear on this page. This page will provide augmented functionality in the future, including the capability to deploy models to other locations.

Manage Spark Pipeline Models Illustration

This image illustrates the Manage Spark Pipeline Models page.

Name	Description	Type	Algorithm	Training Dataset	Accuracy	Created	Updated
Abalone Regression Model	Version 2018-01-10	Pipeline Model (Spark)	Gradient-boosted Tree (GBT) Regression	abalone.csv	90	April 10, 2018 3:00 PM	April 10, 2018 3:00 PM

How to Access Manage Spark Pipeline Models

You can view Spark pipeline models in PrL, however the ability to deploy Spark pipeline models has not yet been developed.

Follow these steps to view and manage Spark Pipeline Models:

1. Navigate to Application Launchpad and click the **Predictive Learning** icon. *The Predictive Learning landing page opens.*
2. Select Manage Spark Pipeline Models from the **Analytics Workspaces** menu. *The Manage Spark Pipeline Models page opens.*

2.10 Managing Analytics Workspaces

The Manage Analytics Workspaces page allows you to access and manipulate your datasets. On this page, you can create a new workspace where you can use out-of-the-box functions to view dataset characteristics, and perform transformations and analysis as needed.

Manage Analytics Workspaces Overview

You can perform multiple tasks on this page related to creating, processing, and analyzing PrL datasets. These tasks include:

[Accessing the Manage Analytics Workspaces Page](#)

[Starting and Stopping a Cluster](#)

[Specifying the S3 Bucket Location](#)

[Creating a New Workspace](#)

[Opening an Existing Workspace](#)

[Adding an Exploration Panel](#)

[View Distribution](#)

[Adding an Analysis Panel](#)

[Linear Regression Analysis](#)

[Adding a Transformation Panel](#)

[Selecting Columns to Carry Over](#)

[Transforming Data Using Joins](#)

[Filtering a Dataset](#)

[Filtering Workspace Data](#)

[Replacing Missing Values](#)

[Principal Component Analysis \(PCA\)](#)

[Normalizing Data With Standard Scaler](#)

[Running an Analysis](#)

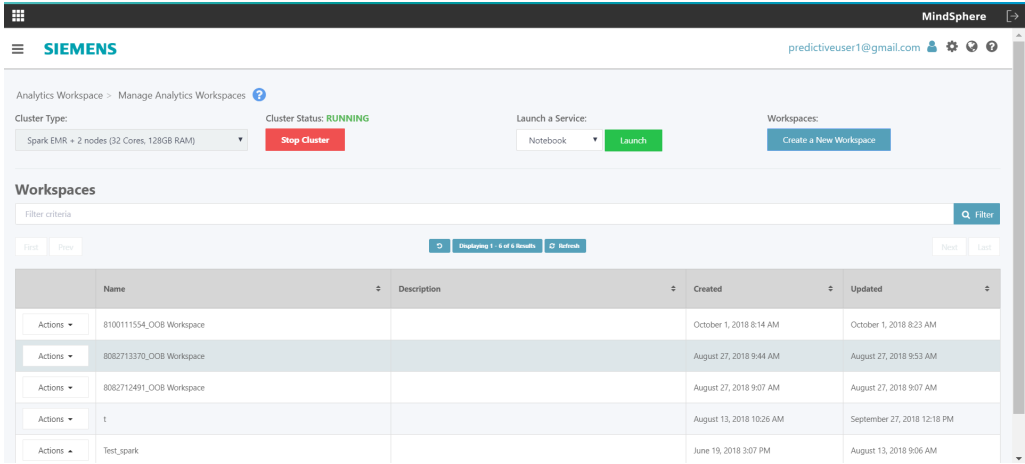
[Launching a Service](#)

[Making API Calls from Zeppelin Notebook](#)

2.11 Navigating to the Manage Analytics Workspaces Page

Manage Analytics Workspaces Page Illustration

This image shows the Manage Analytics Workspaces page.



How to Navigate to the Manage Analytics Workspaces Page

Follow these steps to navigate to the Manage Analytics Workspaces page where you can open an existing workspace or create a new one.

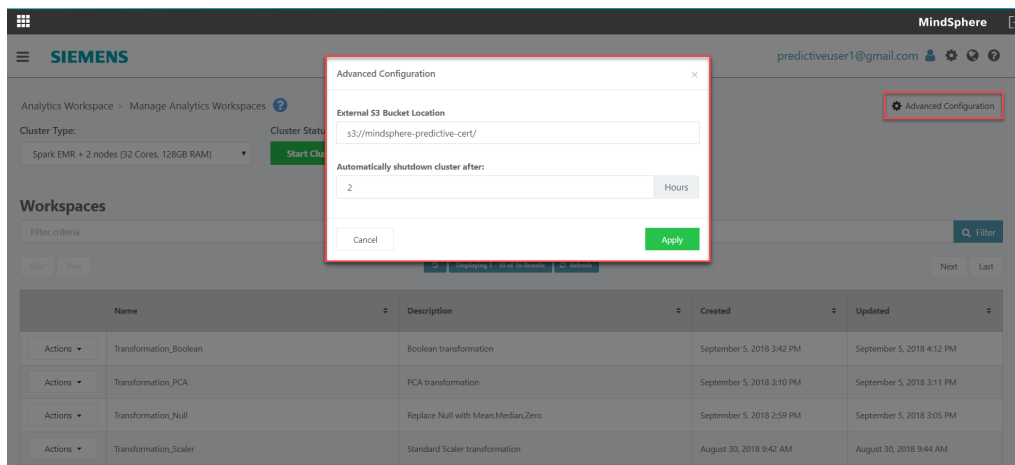
1. Click the **Predictive Learning** icon on the Launchpad. *The Predictive Learning landing page opens.*
2. Select the Manage Analytics Workspaces from the **Model Execution** menu. *The **Manage Analytics Workspaces** page opens.*
3. Select Open from the **Actions** drop-down menu in the row of the workspace you want to open. *The workspace opens.*

2.12 Using the Advanced Configuration Dialog

The Advanced Configuration dialog allows you to specify where your external data is stored in an S3 bucket, and the length of time before the running cluster automatically shuts down. The S3 bucket location you enter here is stored and used for future sessions. The auto shutdown time, however, is only saved for the current page. You will need to update it each time you access the Advanced Configuration dialog, unless you want to keep the two hour default time.

Advanced Configuration Dialog Box Illustration

This illustrates the Advanced Configuration dialog box.



How to Access the Advanced Configuration Dialog Box

Follow these steps to access the Advanced Configuration dialog box:

1. Access the Manage Analytics Workspaces page. *The Manage Analytics Workspaces page appears.*
2. Click the **Advanced Configuration** button. *The Advanced Configuration dialog box appears.*

For More Information

Go here to see more information on using the Advanced Configuration dialog box:

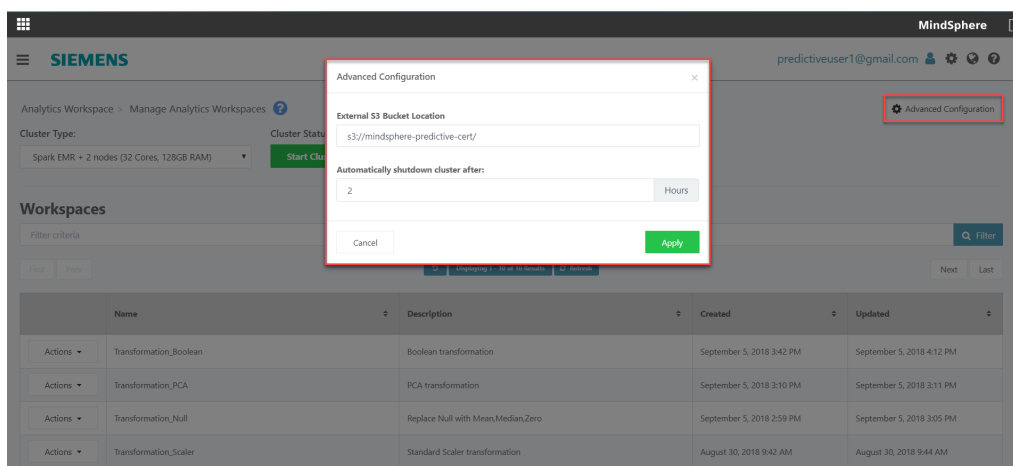
- [Specifying the S3 Bucket Location](#)
- [Defining Cluster Auto Shutdown Time](#)

2.13 Specifying the S3 Bucket Location

You can upload and store multiple file types to an S3 bucket location and then use the Notebook functionality to access and read those files.

Advanced Configuration Dialog Illustration

The following illustrates the Advanced Configuration Dialog.



How to Specify the S3 Bucket Location

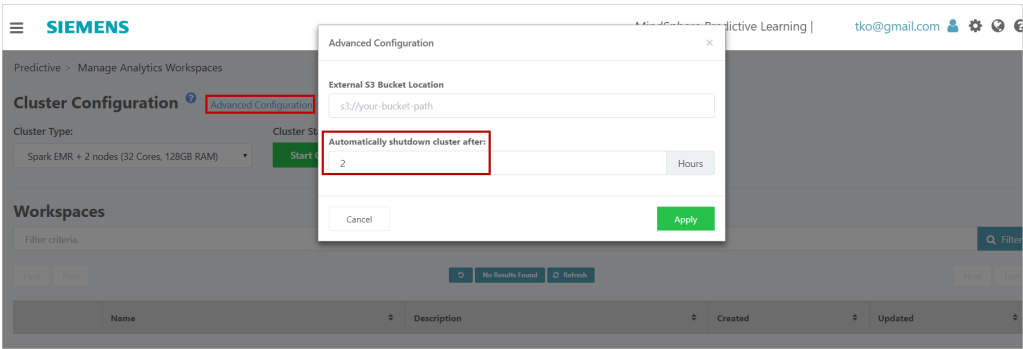
When you enter the path to the S3 bucket you want to use, your entry cannot be changed once the cluster is started.

Follow these steps to specify the S3 bucket location for use with the Notebook:

1. Access the Manage Analytics Workspaces page. *The Manage Analytics Workspaces page opens.*
2. Click the **Advanced Configuration** button. *The Advanced Configuration dialog box opens.*
3. Enter a valid path to your existing S3 bucket in the **External S3 Bucket Location** field and click **Apply**.
4. Click the **Start Cluster** button. *When the cluster is up and running, the button changes to Stop Cluster.*
5. Once the cluster is running, select a service from the **Launch a Service** drop-down list, and click the **Launch** button to open a Notebook. *The notebook opens.*
6. Follow the steps for how to access and read the S3 bucket contents, which is located in a section of the notebook.

2.14 Defining Cluster Auto Shutdown Time

You can define a time frame to automatically shut down the cluster to avoid unnecessary charges. You must do this before you start the cluster. The default is two hours. Changes to the shutdown time are only saved for the current page. They are not retained permanently. The Advanced Configuration dialog box is shown here.



How to Define Cluster Auto Shutdown Time

- Follow these steps to define a cluster auto shutdown time.
1. Select Manage Analytics Workspaces on the Predictive menu. *The Manage Analytics Workspaces page opens.*
 2. Click **Advanced Configuration** next to the Cluster Configuration heading. *The Advanced Configuration dialog box opens.*
 3. Enter a length of time to allow the cluster to run in the **Automatically shutdown cluster after Hours** field. Minimum run time is two hours, maximum is 24 hours. Default is two hours.
 4. Click **Apply**. Once you click Start Cluster, the cluster runs for the number of hours you enter, and automatically shuts down, unless you manually stop it before the time is up.

2.15 Starting a Cluster

Predictive Learning can involve a great deal of resource consumption. To maximize your usage and contain costs, starting the right cluster type is the most important step. Starting a cluster creates a new environment.

Select a cluster type that is large enough and has the processing power to run your dataset analysis without timing out. On the other hand, a 25-node cluster with 750 GB RAM is larger than necessary to process a moderately sized dataset.

It is also important to turn off pop-up blockers in your browser so all buttons function correctly. It can take several minutes for a newly started cluster to reach 'Running' status.

Cluster Types

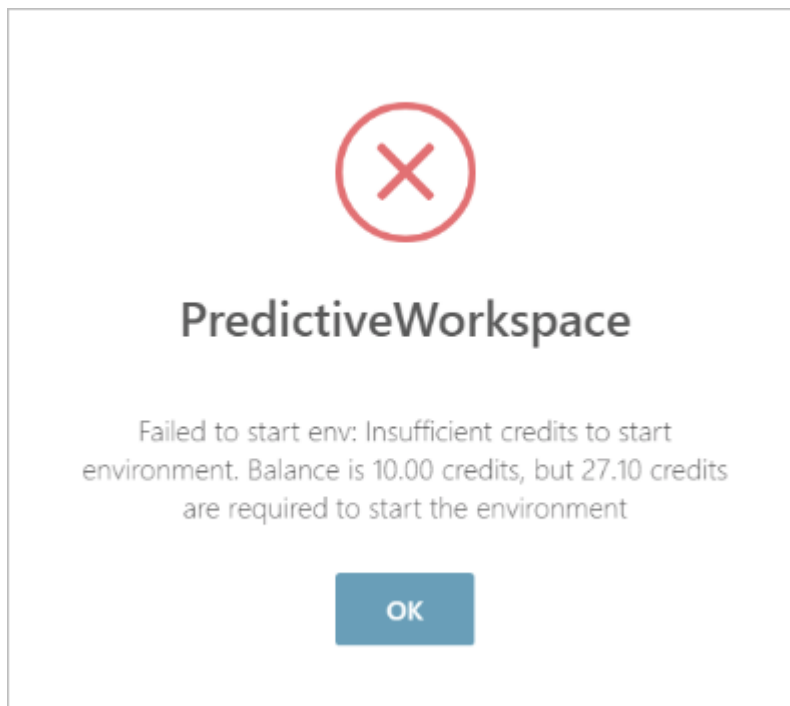
The information in this table describes the available cluster configuration types and their recommended uses.

Cluster type...	Node type...	Best suited for...
-----------------	--------------	--------------------

Cluster type...	Node type...	Best suited for...
Spark EMR + 5 nodes (80 Cores, 320GB RAM) Spark EMR + 10 nodes (160 Cores, 640GB RAM) Spark EMR + 25 nodes (400 Cores, 1600GB RAM))	General	General purpose tasks
Spark EMR + 2 nodes (32 Cores, 244GB RAM) Spark EMR + 5 nodes (80 Cores, 610GB RAM) Spark EMR + 10 nodes (160 Cores, 1220GB RAM) Spark EMR + 25 nodes (400 Cores, 3050GB RAM))	Memory optimized	Memory intensive tasks
Spark EMR + 2 node (12 Cores, 24 GB RAM) Spark EMR + 2 nodes (32 Cores, 60GB RAM) Spark EMR + 5 nodes (80 Cores, 150GB RAM) Spark EMR + 10 nodes (160 Cores, 300GB RAM) Spark EMR + 25 nodes (400 Cores, 750GB RAM)	Computer optimized	CPU intensive tasks
GPU EMR (1 GPU, 12 GPU memory, 4 Cores, 61GB RAM) GPU EMR (1 GPU, 16 GPU memory, 8 Cores, 61GB RAM) GPU EMR (4 GPU, 64 GPU memory, 32 Cores, 244GB RAM) GPU EMR (8 GPU, 128 GPU memory, 64 Cores, 488GB RAM)	GPU and AI optimized	Tasks that require GPU

If You Run Out of Credits

If you run out of credits while a cluster is running, the cluster will continue to run until it is shut down (manually or automatically). However, if you do not have enough credits for **at least one hour**, the system does not allow you to start a cluster, and displays the following message, which includes information on your remaining credits, and the number of credits the cluster configuration you want to run requires:



To purchase more credits, contact your system administrator.

How to Start a Cluster

Follow these steps to start a cluster:

1. Select Manage Analytics Workspaces from the Analytics Workspace menu. *The Manage Analytics Workspaces page opens.*
2. Select the cluster type most suited to the size of the dataset you are processing from the **Cluster Type** drop-down list.
3. Click **Start Cluster**. *The Cluster Status changes to "Starting". When the cluster is running, the status changes to "Running".*

Stopping Clusters

Once your analysis is complete, you can manually stop the cluster. If the cluster was configured for auto-shutdown when you started it, it will run through the shutdown time, then automatically terminate.

Clusters cost money as long as they are running, but even if you don't define an auto-shutdown time, the cluster will stop running after two hours by default. If a job is in progress when the cluster shuts down, manually or automatically, the job terminates at the time the cluster shuts down.

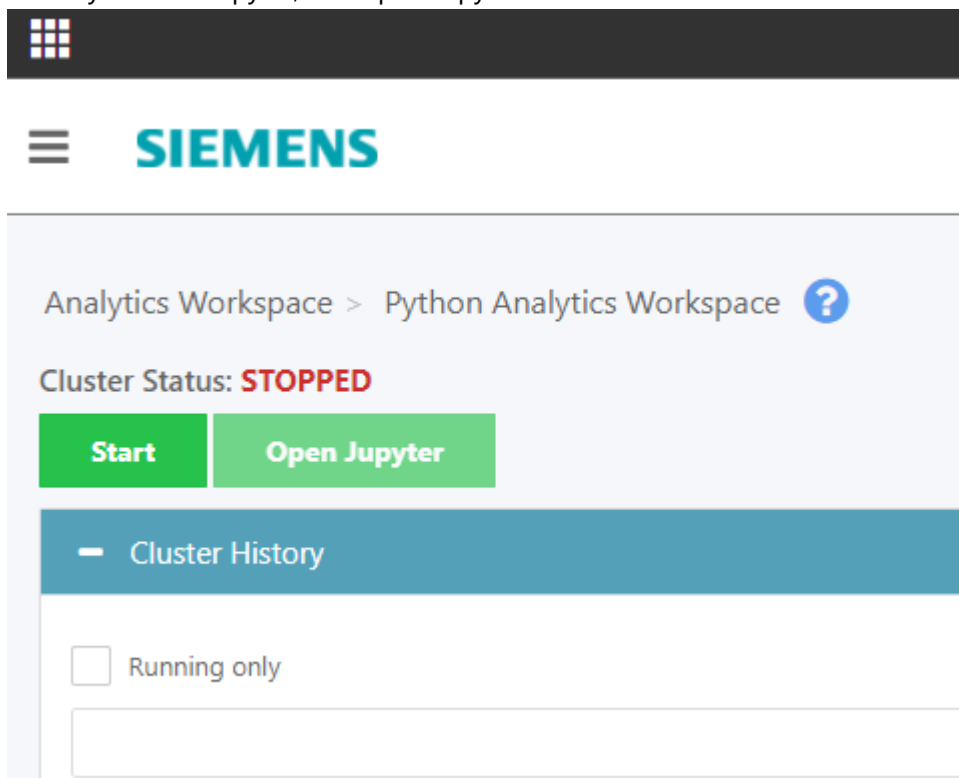
How to Stop a Cluster

Follow these steps to stop a cluster:

1. Select Manage Analytics Workspaces from the Analytics Workspace menu. *The Manage Analytics Workspaces page opens.*
2. Click the **Stop Cluster** button. *When the cluster stops, the cluster status message changes to "Stopped". This may take a few minutes.*

Predictive Learning Essentials

The Essentials package provides a single, non-clustered instance with Jupyter Notebook;. When you start Jupyter, the Open Jupyter button activates when a cluster is running.



2.16 Creating a Workspace

This section explains how to create a new workspace where you can explore and analyze your data, apply transformations, and produce datasets.

You can create multiple workspaces and run analyses simultaneously. Running clusters and analyses in PrL incur costs that accrue against the processing resources provisioned to your organization.

Actions Available on the Manage Analytics Workspaces Page

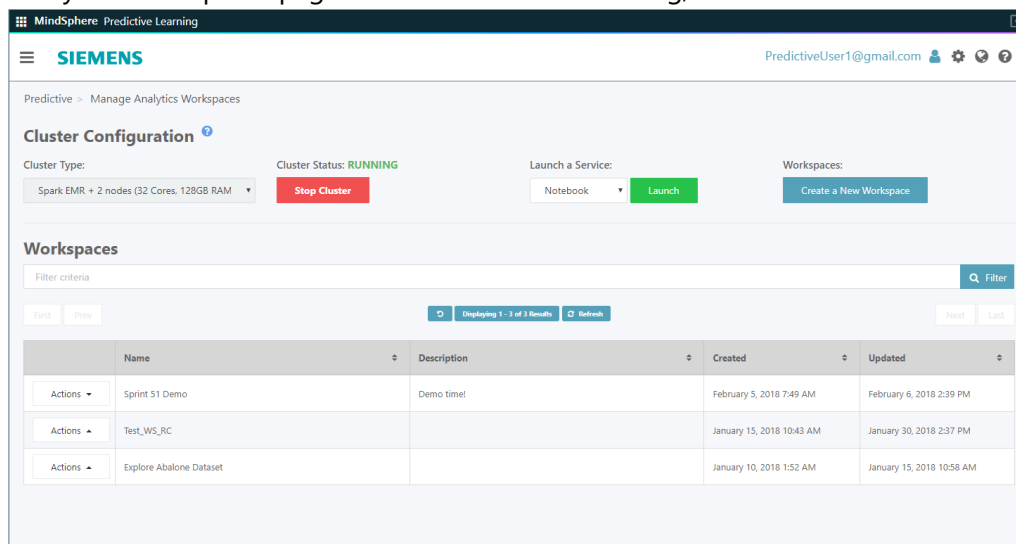
The Workspace page is where you analyze your dataset. On the Workspace page you can:

- Open an existing workspace in the current browser window, or in a new tab

- Save a workspace
- Add an exploration panel
- Add a transformation panel
- View a new dataset
- View a new schema
- View statistics about a dataset
- Save a dataset

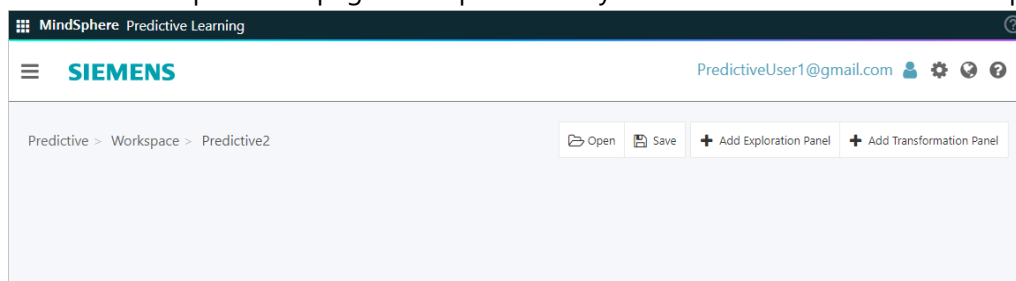
Manage Analytics Workspaces Page Illustration

This image shows the Launch a Service and Create a New Workspace buttons on the Manage Analytics Workspaces page. When no cluster is running, these two buttons are not active.



Blank Workspace Page Illustration

Here's an example of the page that opens when you click the Create a New Workspace button:



How to Create a New Workspace

Follow these steps to create a new workspace:

1. Click the Predictive Learning icon on the **Launchpad**. *The Predictive Learning landing page opens.*
2. Click the **Start Cluster** button. *When the cluster status changes to "Running", go to the next step.*
3. Click the **Create a New Workspace** button. *The Create a New Workspace dialog opens.*
4. Enter a name for the workspace in the **Name** field.
5. Enter an options description in the **Description** field.
6. Click **Create**. *The Workspace page opens and the new workspace name displays in the breadcrumb.*

2.17 Opening an Existing Workspace

You can open an existing workspace in the current browser window, or click Open in a New Tab to open the workspace in a new browser window. If you open the workspace in the current browser window where a workspace is already open, you will overwrite any work you had done on the first workspace, so use caution.

How to Open an Existing Workspace

Follow these steps to open an existing workspace:

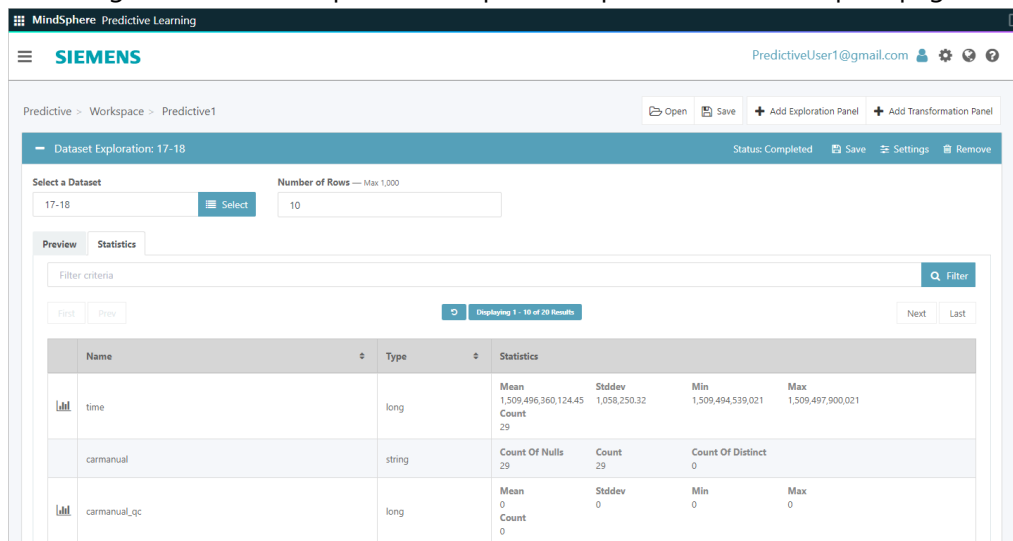
1. Navigate to the Manage Analytics Workspaces page and start a cluster. *The Manage Analytics Workspaces page opens.*
2. Select Open from the **Actions** drop-down list. *The Workspace opens.*
3. To open additional workspaces, click the **Open** button at the top of the page. *The Choose a Workspace to Open dialog box opens.*
4. Select the workspace you want to open. To keep the previously opened workspace open, click **Open in a New Tab**, otherwise, click **Open**. *The workspace you select opens.*

2.18 Adding an Exploration Panel

The exploration panel allows you to select a dataset to analyze. You can preview the data in the dataset, and view the schema. Once you know what your dataset contains, you can then define the transformations you want to apply to it. After performing your transformations, you can preview the new dataset, schema, and statistics, and save it as a new dataset.

Workspace Exploration Panel Illustration

This image shows an example of the exploration panel on the workspace page:



How to Add an Exploration Panel to the Workspace

Follow these steps to add an exploration panel to your workspace and begin your data exploration:

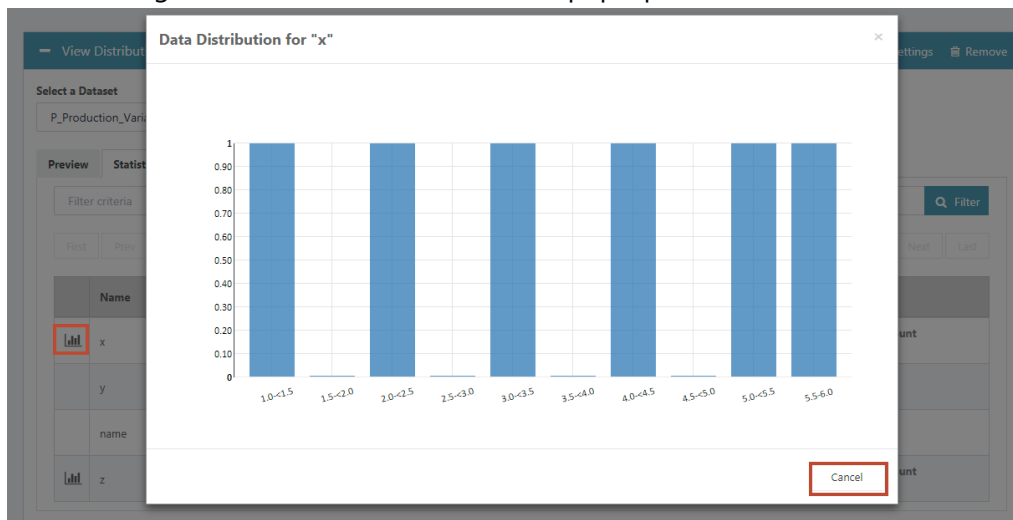
1. Click the **Add Exploration Panel** button on the Workspace page. *A Data Exploration panel opens.*
2. Click the **Select** button to choose a dataset. *A dialog box lists your saved datasets.*
3. Select a **dataset**. The first ten rows (default) of the dataset display in the workspace Preview tab.
4. To change the number of rows that display, enter a value in the **Number of Rows** field.
5. Click the **Schema** tab. *The schema defined for the dataset displays in the workspace.*
6. To remove the exploration panel, click **Remove** in the upper right corner. *The exploration panel closes.*
7. To add another exploration panel, click the **Add Exploration Panel** button. *Another exploration panel opens.*

2.19 Viewing Data Distributions

Viewing data distributions for numeric values can help you make decisions about performing transformations on the data, such as removing values or normalizing it.

View Data Distribution Illustration

The following illustrates the View Distribution pop-up window.



How to View a Data Distribution

Follow these steps to view the distribution of numeric data.

1. Select **Manage Analytics Workspaces** from the Analytics Workspace menu. *The Manage Analytics Workspaces page opens.*
2. Select a configuration from the **Cluster Type** drop-down list and click the **Start Cluster** button. *Wait for the cluster status to change to Running.*
3. Click the **Create a new Workspace** button. *The Workspace page opens.*
4. Click **Add Exploration Panel**. *A new Exploration panel opens.*
5. Select a dataset from the **Select a Dataset** drop-down list. *The Statistics tab displays information about the dataset.*
6. Click the **View Distribution** icon in the table row for the data distribution you want to view. *A pop-up window displays a histogram for that row.*
7. Click the **Preview** tab to preview the dataset.
8. Click **Cancel** to exit the window.

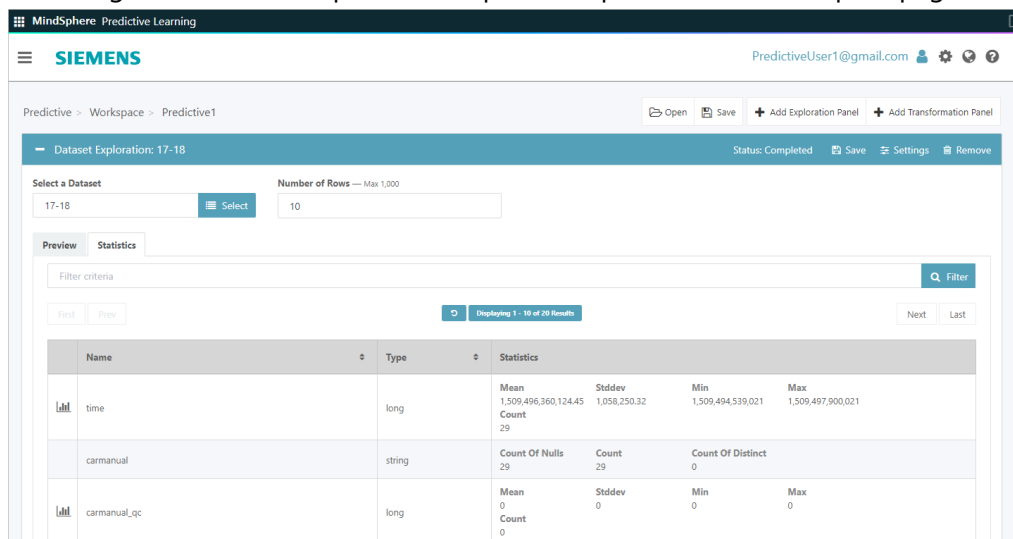
2.20 Adding an Exploration Panel

The exploration panel allows you to select a dataset to analyze. You can preview the data in the dataset, and view the schema. Once you know what your dataset contains, you can then define

the transformations you want to apply to it. After performing your transformations, you can preview the new dataset, schema, and statistics, and save it as a new dataset.

Workspace Exploration Panel Illustration

This image shows an example of the exploration panel on the workspace page:



How to Add an Exploration Panel to the Workspace

Follow these steps to add an exploration panel to your workspace and begin your data exploration:

1. Click the **Add Exploration Panel** button on the Workspace page. *A Data Exploration panel opens.*
2. Click the **Select** button to choose a dataset. *A dialog box lists your saved datasets.*
3. Select a **dataset**. The first ten rows (default) of the dataset display in the workspace Preview tab.
4. To change the number of rows that display, enter a value in the **Number of Rows** field.
5. Click the **Schema** tab. *The schema defined for the dataset displays in the workspace.*
6. To remove the exploration panel, click **Remove** in the upper right corner. *The exploration panel closes.*
7. To add another exploration panel, click the **Add Exploration Panel** button. *Another exploration panel opens.*

2.21 Linear Regression Analysis

Linear Regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables, usually denoted as X . In practice, linear regression can be used in two ways: prediction and feature engineering.

For prediction, linear regression can be used to fit a predictive model to an observed dataset of y and X values. When the model is developed, and a new value for X is given, the model can predict the value of y .

For feature engineering, when given a variable y , and a number of variables X_1, X_2, \dots, X_n (collectively denoted as X in matrix representation) that may be related to y are given, linear regression analysis can quantify the strength of the relationship between y and X , and help determine which X variable(s) may have no relationship with y , and identify subsets of X that may contain redundant information about y .

Linear regression models are often fitted using the least squares approach, but it can also be fitted using some other penalized version of the least squares loss functions as in Ridge Regression (L2) and LASSO (L1). In our system, it is controlled by setting the second regulation (alpha) parameter from 0 to 1.

Lambda and Alpha Values and Results

This table describes the type of penalized model that results, based on the values specified for the alpha and lambda options.

Lambda Value	Alpha Value	Result
0	Any value	No regularization. Alpha is ignored.
> 0	$= 0$	Ridge Regression
> 0	$= 1$	LASSO
> 0	$0 < \text{Alpha} < 1$	Elastic Net Penalty

Defining Input Parameters

This table describes the input parameters on the Fill in Parameters tab, which are required to run a linear regression on a dataset.

Field	Input Description
Prediction Column	Column in the dataset on which you want to make a prediction using the linear regression algorithm.
Input Columns	Columns in the dataset you want to use as supporting data points to calculate the prediction.

Field	Input Description
Split	The percentage of data used for training versus prediction. Minimum value is 0.1. An entry of 0.8 means 80% of the data will be used for training, and 20% for prediction.
Max Iteration	Linear Regression uses a Gradient Descent Iterative algorithm to find the optimum Solution space. This variable is the maximum limit on the number of iterations. If optimal parameters are found prior to reaching max Iterations, the algorithm stops iterating forward. Enter a number between 10 and 1000.
Seed	Unique random number. It is used to get the same results for multiple executions given the same seed number.
Alpha	Alpha is a parameter used for Regularization that linearly combines the L1 and L2 penalties of the lasso and ridge methods. For $\alpha = 0$, the penalty is an L2 penalty. For $\alpha = 1$, it is an L1 penalty. For α in $(0,1)$ the penalty is a combination of L1 and L2.
Lambda	Lambda is a parameter used for Regularization. It is used to avoid overfitting. The larger the lambda, the more the coefficients for regression are shrunk toward zero. When the value is $0 \leq \lambda < \infty$, regularization is disabled and ordinary linear models are fitted.
Metrics Tab	Displays the metric values for your analysis.
Training Metrics	The Metrics tab shows the values for the following Training metrics: <ul style="list-style-type: none"> - RootMeanSquareError - R-Square - MeanAbsoluteError - Co-efficients
Evaluation Metrics	The Metrics tab shows the values for the following Evaluation metrics: <ul style="list-style-type: none"> - RootMeanSquareError - R-Square - MeanAbsoluteError

Analysis Panel with Linear Regression Illustration

This image shows an example of the Fill in Parameters tab on the analysis panel, when using the Linear Regression algorithm.

The screenshot displays the 'Analysis: Linear Regression: LinearRegressionData' interface. At the top, the status is 'Completed' with buttons for 'Run', 'Settings', and 'Remove'. Below the dataset selection, the 'Fill in Parameters' tab is active. It contains the following fields:

- Prediction Column:** A dropdown menu showing 'y'.
- Input Columns:** A multi-select field with several columns selected.
- Split:** A numeric input field set to 0.7.
- Seed:** A numeric input field set to 10000.
- Alpha:** A numeric input field set to 0.
- Lambda:** A numeric input field set to 1000.
- Max Iterations:** A numeric input field set to 100.

Understanding Output Metrics

Once a job is complete, the Metrics tab displays the results of the analysis.

Linear Regression algorithm metrics include:

Training Metrics: The name of each feature column you selected as a data point and its

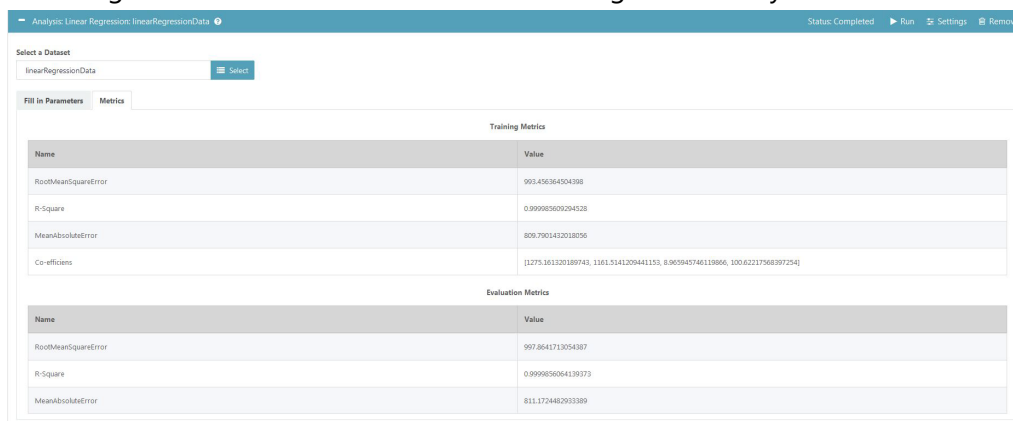
- numeric value.
 - RootMeanSquareError: Deviation of the residuals (prediction errors) used to verify experimental results.
 - R-Square: Statistical measure of how close the data are to the fitted regression line.
 - MeanAbsoluteError: A measure of the difference between two continuous variables.
 - Co-efficients: The constant represents the rate of change of one variable (y) as a function of changes in the other (x); it is the slope of the regression line.

Evaluation Metrics: The metrics used in the Linear Regression algorithm and their

- calculated values, which include:
 - R-Square: Statistical measure of how close the data are to the fitted regression line.
 - MeanAbsoluteError: A measure of the difference between two continuous variables.
 - RootMeanSquareError: Deviation of the residuals (prediction errors) used to verify experimental results.

Analysis Panel Metrics Tab Illustration

This image shows the Metrics tab from a linear regression analysis:



The screenshot shows a web interface for a linear regression analysis. At the top, there's a header bar with 'Analysis: Linear Regression: linearRegressionData' and a status 'Status: Completed'. Below this, there's a 'Select a Dataset' section with 'linearRegressionData' selected. The main area is divided into 'Fill in Parameters' and 'Metrics' tabs, with 'Metrics' being the active tab. It contains two tables: 'Training Metrics' and 'Evaluation Metrics'.

Training Metrics	
Name	Value
RootMeanSquareError	965.456364504398
R-Square	0.999985609204328
MeanAbsoluteError	809.7901432018856
Co-efficients	[1275.161320189743, 1361.5141209441153, 8.965945746119866, 100.62217568997254]

Evaluation Metrics	
Name	Value
RootMeanSquareError	997.8641713054387
R-Square	0.9999856064139373
MeanAbsoluteError	811.1724482933389

How to Run a Linear Regression Analysis

Follow these steps to run an analysis using linear regression:

1. Start a cluster.
2. Open or create a workspace.

3. Click the **Add Analysis Panel** button and select **Linear Regression** from the dialog box. *The Linear Regression analysis panel opens.*
4. Select the dataset you want to use from the **Select a Dataset** list.
5. Enter parameters in all (required) input fields on the **Fill in Parameters** tab. See the *Defining Your Input Parameters* section for details.
6. Click **Run**. *A Job Submitted success message displays, the panel updates as the job progresses, and the Metrics tab displays the results.*

2.22 Logistic Regression Analysis

Logistic Regression is a classification algorithm which is used to predict a variable that can take binary values, such as: 0 or 1. This algorithm is available out-of-the-box on the Add Analysis panel. It provides data scientists the ability to run the algorithm on their datasets without coding. Logistic Regression provides a less computationally expensive option for classification tasks, and is easier to interpret by non-data scientists.

Cases where Logistic Regression provides a good option include:

- Analyzing the sentiment of a given statement (Positive/Negative)
- Predicting who Bob is going to vote for (Democrat/Republican)
- Determining the probability that a student will enroll in a master's program based on his academic, extra-curricular activities, enrollment in online courses, projects, work experience etc., (Probability values of each classification instead of direct yes/no)
- Determining the probability of an employee staying in the current job (Loyal/Not Loyal)

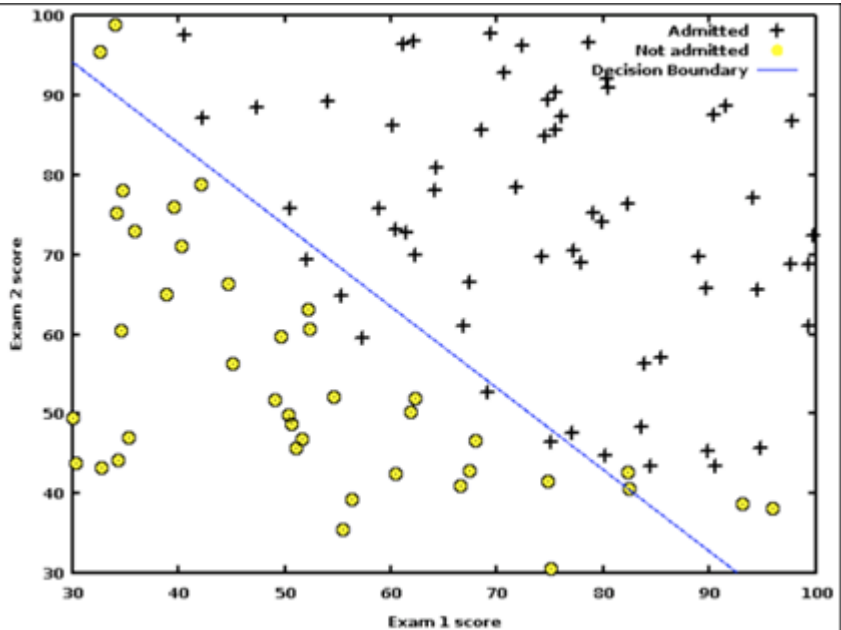
Logistic Regression Examples

An example where Logistic Regression might be useful is illustrated in the following table. In the data below, the target column is Income > \$100k. Based on the given details about an individual such as education level, type of industry, city of employment, and the job role, we can predict whether the person will earn more than \$100k annually or not.

Education	Industry	City	Role	Income > \$100K
Bachelors	Retail	Charlotte	Sales Executive	No
Bachelors	IT	New York	Software Engineer	No
Masters	IT	San Francisco	Data Scientist	Yes
Bachelors	Transportation	Durham	NA	No

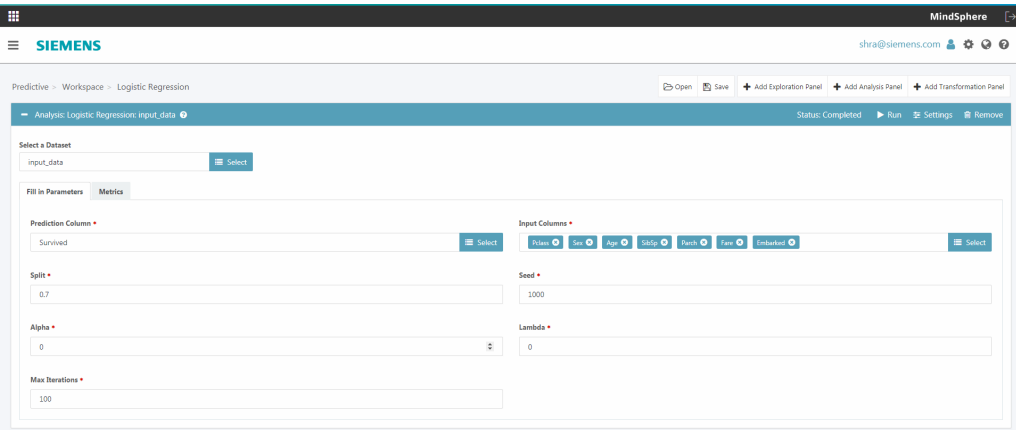
Education	Industry	City	Role	Income > \$100K
PhD	Education	Chicago	Professor	Yes
PhD	IT	New York	AI Research Scientist	Yes

A visual representation of the results of a Logistic Regression analysis is shown below. In this example, the data is almost linearly separable. While this analysis could be run using a non-linear classifier, such as Neural Network, the alternative is computationally very expensive.



Logistic Regression Analysis Panel Illustration

The following illustrates the Fill in Parameters tab on the Logistic Regression Analysis panel.



Definitions of Logistic Regression Input Parameters

The following table defines the input parameters for the Logistic Regression algorithm.

Parameter	Definition
-----------	------------

Parameter	Definition
Prediction Column	The dataset column that you want to use the logistic regression algorithm on to make a prediction.
Input Columns	Columns in the dataset that you want to use as supporting data points to calculate the prediction.
Split	The percentage of data used for training versus prediction. Minimum value is 0.1. An entry of 0.8 means 80% of the data will be used for training, and 20% for prediction.
Max Iteration	Logistic Regression uses a Gradient Descent Iterative algorithm to find the optimum Solution space. This variable is the maximum limit on the number of iterations. If optimal parameters are found prior to reaching max iterations, the algorithm stops iterating forward. Enter a number between 10 and 1000.
Seed	Unique random number. It is used to get the same results for multiple executions given the same seed number.
Alpha	Alpha is a parameter used for Regularization that linearly combines the L1 and L2 penalties of the Lasso and Ridge methods.
Lambda	Lambda is a parameter used for Regularization. It is used to avoid over fitting. The larger the lambda, the more the coefficients or regression are shrunk toward zero. When the value is 0, regularization is disabled and ordinary linear models are fitted.

Regularization Parameters

The following table describes the results of the various values for the Regularization parameters described above.

Lambda Value	Alpha Value	Results
0	Any value	No regularization - Alpha is ignored
> 0	= 0	Ridge Regression
> 0	= 1	LASSO
> 0	0 < Alpha < 1	Elastic Net Penalty

How to Access the Logistic Regression Panel

Follow these steps to access the Logistic Regression panel:

1. Navigate to the **Manage Analytics Workspaces** page. *The Workspaces window opens.*
2. Start a cluster, if one is not already running.

3. Open a Workspace or create a new one.
4. Click **Add Analysis Panel**. *The Select an Algorithm dialog box opens.*
5. Select **Logistic Regression** and click **OK**. *The Logistic Regression panel opens.*

How to Run a Logistic Regression Analysis

Follow these steps to run a Logistic Regression analysis:

1. On the Logistic Regression panel, select a dataset from the **Dataset** drop-down list.
2. Select the column to run the analysis on from the **Prediction Column** dialog.
3. Select the columns to include in the result set from the **Input Columns** dialog.
4. Enter a value between 0.1 and 1.0 in the **Split** field. This value determines what percentage of the data is used for training and what percentage is used in the actual analysis.
5. Enter a numeric value in the **Seed** field. This entry is used to identify the run so that the analysis can be repeated in the future.
6. Enter a numeric value in the **Alpha** field.
7. Enter a numeric value in the **Lambda** field.
8. Enter a number in the **Max Iterations** field to indicate the maximum number of times the analysis should run against this dataset.
9. Click the **Run** button at the top of the panel. *The status changes to "Running" and your analysis runs. Once the job has completed, you can view the results on the Metrics tab.*

Training Metrics Parameter Definitions

The following table defines the parameters used in the training metrics.

Parameter	Definition
AUC	Area under the ROC curve
ROC	The point in the ROC curve which gives the best combination of True Positive and False Positive rates
FMeasure	F1 score

Evaluation Metrics Parameter Definitions

The following table defines the parameters used in the evaluation metrics.

Parameter	Definition
Accuracy	Fraction indicating the correct predictions
True Positive Rate	Number of correct positive predictions
False Positive Rate	Number of incorrect positive predictions
True Negative Rate	Number of correct negative predictions
False Negative Rate	Number of incorrect negative predictions
FMeasure	F1 Score

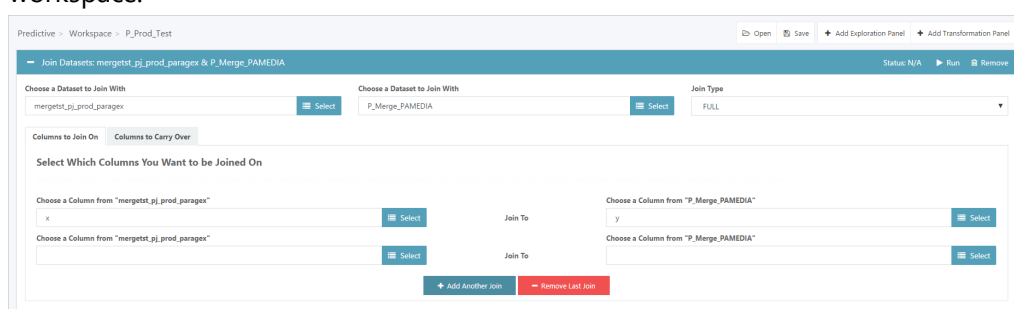
2.23 Adding a Transformation Panel

You can add multiple transformations to the workspace for the dataset. The available transformations are:

- Joins
- Filter
- Replace Missing Values
- Principal Component Analysis

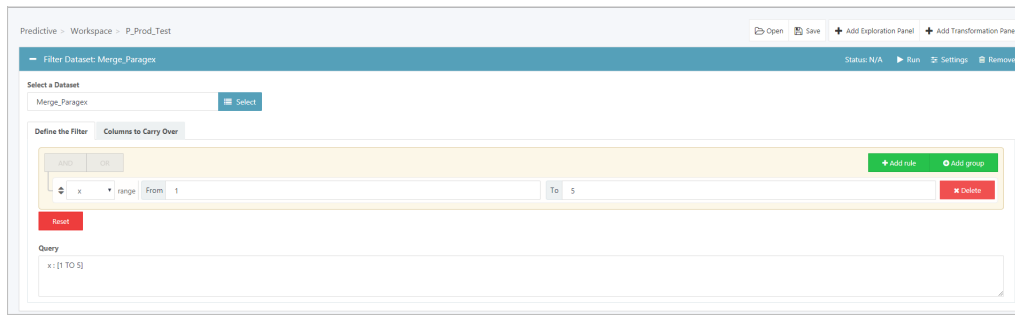
Join Transformation Panel Illustration

The following image shows a Join transformation panel, Columns to Join On tab, on the workspace.



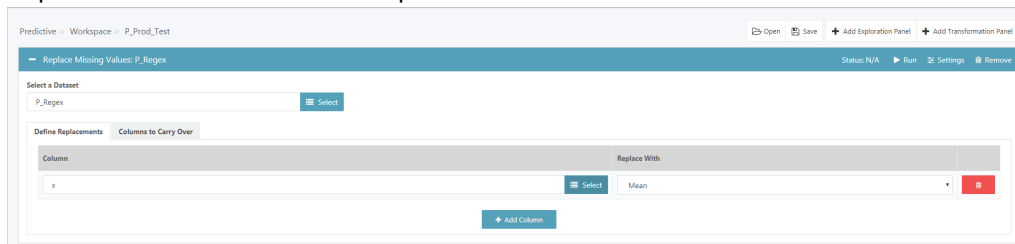
Filter Transformation Panel Illustration

The following image shows a Filter transformation panel, Define the Filter tab, on the workspace.



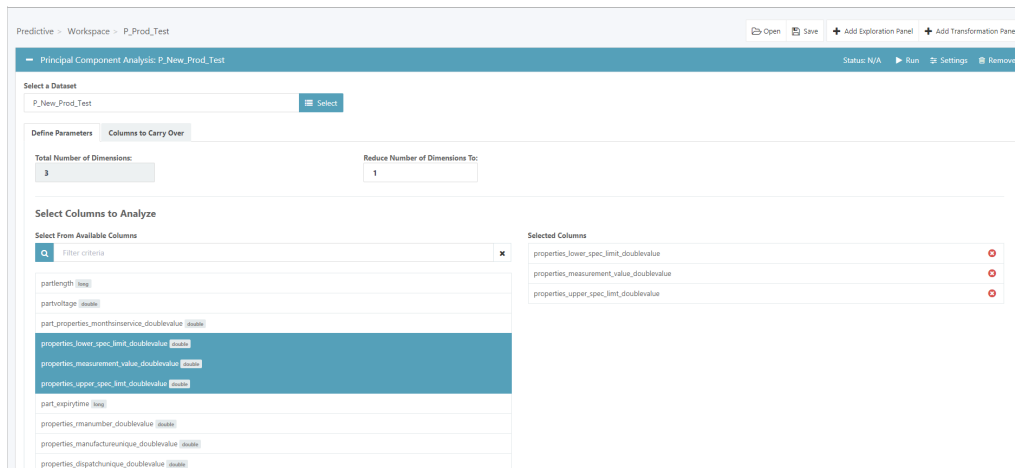
Replace Missing Values Transformation Panel Illustration

The following image shows the Replace Missing Values transformation panel, Define Replacements tab, on the workspace.



Principal Component Analysis Transformation Panel Illustration

The following image shows the Principal Component Analysis transformation panel, Define Parameters tab, on the workspace.



How to Add a Transformation Panel to the Workspace

Follow these steps to add a transformation panel to the workspace:

1. Click the **Add Transformation Panel** button at the top of the workspace page. A *Select Transformation dialog box* opens.
2. Select a transformation type from the list and click **Select**. A *transformation panel* opens.

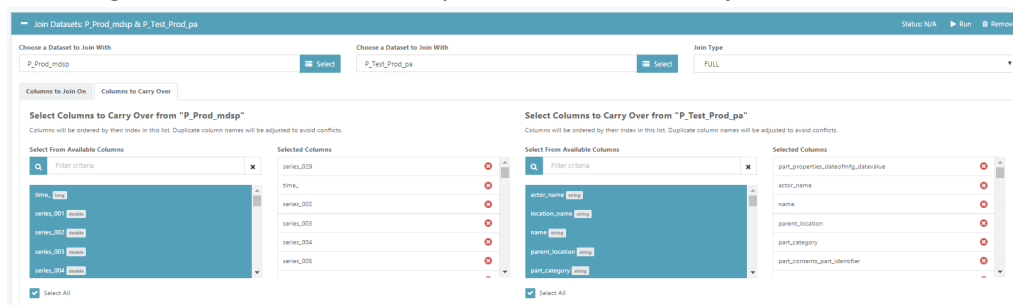
3. Click the **Remove** icon in the upper right corner of the panel to remove it. *The transformation panel closes and your selections no longer apply.*

2.24 Selecting Columns to Carry Over

Each transformation pane contains a Columns to Carry Over tab where you can specify the dataset columns to include in your transformation. This list determines the order in which the data is displayed in the preview. You can select all columns or just a few, and save your selections.

Columns to Carry Over Tab Illustration

This image shows a transformation panel, Columns to Carry Over tab, on the workspace.



How to Select Columns to Carry Over

Follow these steps to select columns to carry over for the transformation:

1. Click the **Columns to Carry Over** tab. *The Select Columns to Carry Over pane opens.*
2. Click a column to add it to the **Selected Columns** list. *The selected column appears in the list*
OR
3. Click **Select All** to choose all the columns to carry over. *All columns are added to the Selected Columns list.*
4. Click **Save** to save the column selections for the transformation.
5. If your cluster is running, click the **Run** button at the top of the Transformation panel to run the transformation. *The preview opens.*

2.25 Transforming Data Using Joins

When you join data together from different sources, you must specify which type of join to use. Predictive Learning uses four join types:

Inner Join: returns **all rows** in which the join condition is met, that is, at least one match occurs in **both** tables.

Left Join: returns **all rows** in which the join condition is met, that is, all rows from the **left** table, and row **matches** that occur from the **right** table.

Right Join: returns **all rows** in which the join condition is met, that is, all rows from the **right** table, and row matches that occur from the **left** table.

Full Outer Join: returns **all rows** in which the join condition is met, that is, when a match occurs in **one** of the tables.

Diagram of the four join types in Predictive Learning

Left joins and right joins are sometimes referred to as left full join and right full join, respectively. Here is a visual representation of the four join types:



How to Create a Join Transformation

Join transformations allow you to join two columns from two different datasets to form one new column. The two columns must be the same type, i.e., string-to-string or integer-to-integer. An even number of joins must be specified. A maximum of three joins per transformation panel are allowed.

Follow these steps to create a Join transformation:

1. On the Join Datasets panel, click **Select** and choose a dataset from the first field.
2. Click **Select** and choose a dataset from the second field.
3. Select Full, Inner, Left, or Right from the **Join Type** drop-down list. *The Columns to Join On and Columns to Carry Over tabs display.*
4. Click **Select** and choose the columns to join from each dataset from the ****Choose a Column from...**** fields. Both columns must be the same data type.
5. Click the **Add Another Join** button to add an additional join to the transformation. You can define a maximum of three joins per panel.
6. Click the **Columns to Carry Over** tab to specify the columns you want to include in the joined dataset.
7. If your dataset has a large number of columns, use the **Select All** option with caution since performance issues may arise. Select at least one column from each dataset, and click the **X** to delete a column.

8. Click the **Save** button at the top of the workspace page to save your transformations.
9. Click the **Run** button at the top of the Join Datasets panel to run the transformation. *The status message shows the current status of the job.*

2.26 Filtering a Dataset

Filtering the data in a Workspace gives you an opportunity to refine the dataset before saving it. See [Filtering Workspace Data](#) for information on using the query builder.

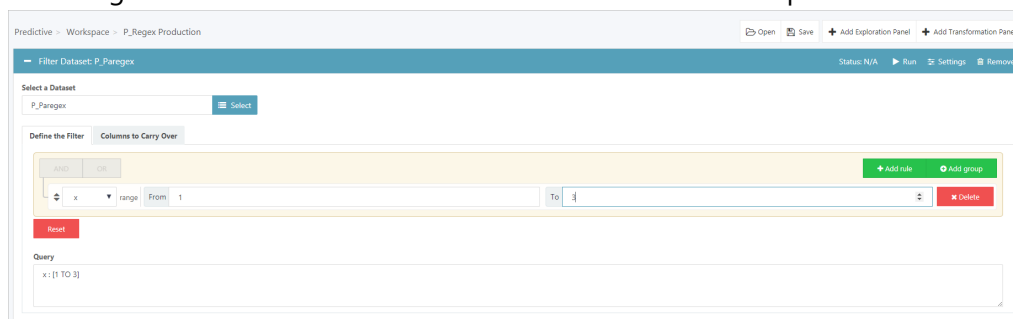
Filter Transformation Panel Illustration

This image shows the Filter Transformation panel when it first opens:



Define the Filter Tab Illustration

This image shows the Define the Filter tab on the Filter Dataset panel:



Columns to Carry Over Tab Illustration

This image shows the Columns to Carryover tab on the Filter Dataset panel:



How to Filter a Dataset

Follow these steps to define a workspace dataset filter:

1. Navigate to the Manage Workspace page and start a cluster. *When the cluster is up and running, the status changes to Running and the cluster button changes to Stop Cluster.*
2. Open an existing workspace or create a new one. *The workspace opens.*
3. Click the **+ Add Transformation** button. *The Select a Transformation dialog box opens.*
4. Select **Filter** and click the **Select** button. *A Filter Dataset panel opens.*
5. Click the **Select** button and select a dataset from the drop-down list. *The dataset loads in the panel.*
6. Use the query builder on the **Define the Filter** tab to select the parameters for the dataset filter. See the [Filtering Workspace Data](#) topic for information on using the interface.
7. Click the **Columns to Carryover** tab and select the columns to include in the filter. *The columns will be ordered by their index in the list. Duplicate names will be adjusted to avoid conflicts.*
8. Click the **Run** button at the top of the **Filter Dataset** panel. *A message appears that the transformation is running. When it has completed, the Run button changes to a Save as Dataset button.*
9. Click the **Save as Dataset** button to save the dataset with the filter. *The name defaults to the dataset name but can be changed.*

2.27 Filtering Workspace Data

The Filters panel allows you to create and use queries to filter the data before saving it to a dataset.

Filter Query Terminology

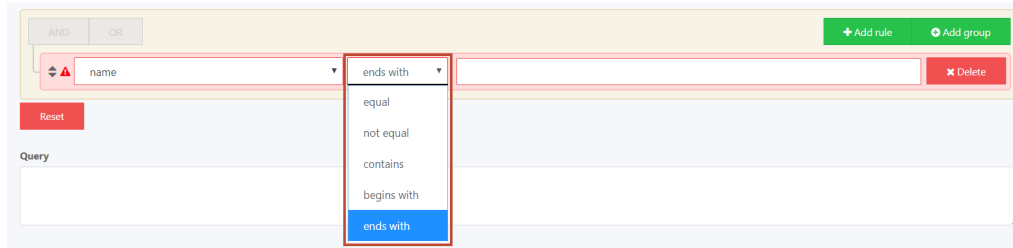
The Filters panel uses "rules" and "groups" to create queries and filter data. It may help to clarify certain terms used on the Filters panel.

- **Rules** are the individual expressions that are sent to the query engine.
- **Groups** organize these expressions so that the And/Or expressions can be applied appropriately. Groups allow you to set parentheses around the rules and expressions so they are executed in the correct logical order.

Additional Filter Operators

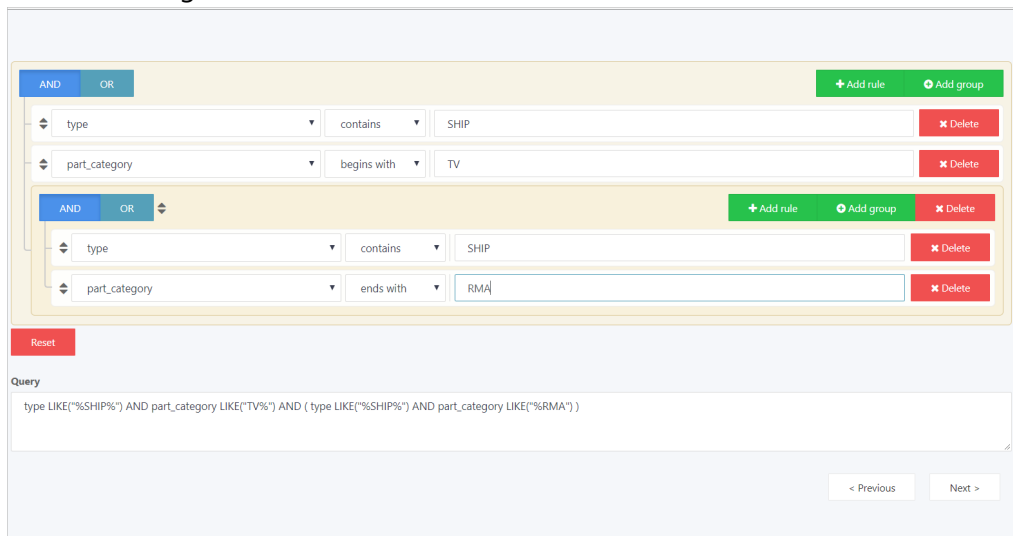
The query builder uses operators to filter records in various ways. Operators specify conditions in a statement and serve as conjunctions for multiple conditions in a statement. There are currently

five operators available in Predictive Learning as shown in the following illustration.



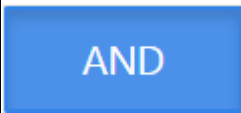

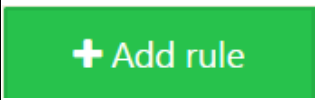
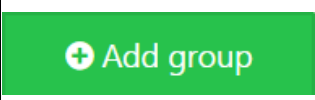
Filter Operators Example

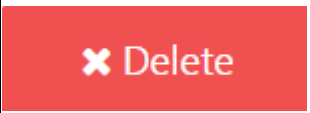
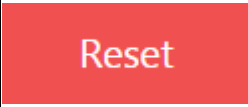




Operators can be used in conjunction with rules and groups to create complex queries as shown in the following illustration:



Filters Panel Buttons and Icons Descriptions

The following table describes the buttons and icons on the Filters panel.

Button/Icon...	Purpose...	Click this button to...
 	Operator selector	Set the operator for the rule or group. The operator must be consistent within a group.
	Add rule	Add a row and create a rule for the query.
	Add group	Add a row and create a group of nested items with a rule.

Button/Icon...	Purpose...	Click this button to...
	Delete row	Delete the row from the query.
	Reset Filters panel	Clear all selections and begin again.
	Cancel transaction	Cancel the transaction, discard all entries, and exit the Filters page.
	Go to previous page	Go back to the previous tab or page.
	Move row	Move the row up or down and rearrange the order.
	Open drop-down list	Open the drop-down list of options to select.

How to Filter Data

Data filtering can include multiple rules and groups. As you make selections in the query builder, the Query field responds to your input and, additionally, you can make changes in the Query field itself. If your syntax is correct, when you click out of the Query field, the query builder updates with the changes.

Follow these steps to filter your data before saving it to a dataset:

1. Click the **Filters** tab on any PrL page, or select **Next** of the Schema page, to navigate to the Filters page.
2. Select an **entry** from the first field's drop-down list.
3. Select an **operator** from the second field's drop-down list, and enter a value in the third field.
4. Select **And** or **Or** to specify the relationship between fields.
5. Click **Add Rule** to specify individual data elements to include.
6. Click **Add Group** to specify more than one data element and group them together.
7. Click **Next** or click the **Preview** tab to view the data with the filters you specified on the Filters page.

Filter Query Example

The following illustrates a sample query using the Filter query builder.

In this example, the top Group of rules are run. The result produced is then joined to the third rule by AND or OR.

Adds a group of rules to the query

Adds a rule to the query

AND OR

part_category equal Laptop

actor_name equal Collector SAT

A group containing two rules, "Rule 1" and "Rule 2".

AND OR

parent_location equal San Jose

"Rule 3"

Reset

Query

part_category = "Laptop" AND actor_name = "Collector SAT" AND (parent_location = "San Jose")

Your rule and filter selections are displayed in a Javascript-like syntax in the Query field.

< Previous Next >

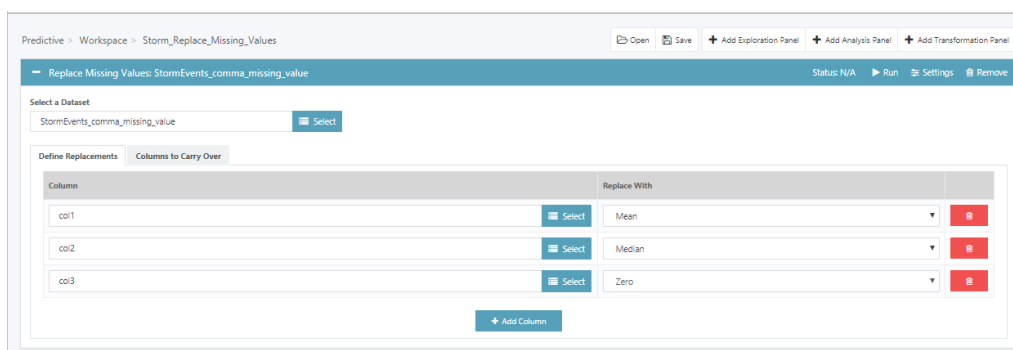
2.28 Replacing Missing Values

This transformation allows you to replace missing values in your dataset at the column level. Here are some important points about replacing missing values:

- You must specify a replacement type for each column in which you want to replace missing values.
- You can only define one replacement value for each column.
- This feature is not available for String and Boolean data types.
- The drop-down list contains these replacement options:
 - Mean (default)—replaces missing values with the **average** of all the values in the selected column; available for **double** data types only.
 - Median—replaces missing values with the **median** of all the values for the selected column; available for **double** data types only.
 - Zero—replaces missing values with **zero**; available for **all** numeric data types.
- After specifying the replacement values, select the columns to include in the dataset, preview the results, and register the dataset.

Replace Missing Values Transformation Panel

Here is an example of the Replace Missing Values transformation panel.



How to Replace Missing Values

Follow these steps to replace missing values in your dataset using the Replace Missing Values transformation:

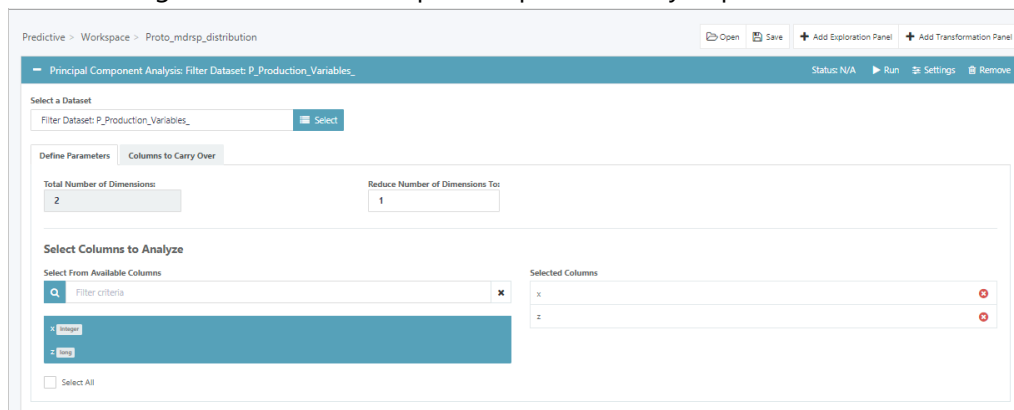
1. Navigate to the Manage Analytics Workspaces page and start a cluster. *The cluster starts and displays "Cluster Running".*
2. Click **Create a New Workspace** or open an existing workspace. *The Workspace opens on a new page.*
3. Click **Add Transformation Panel**. *The Select Transformation dialog box opens.*
4. Select **Replace Missing Values** and click **Select**. *A Replace Missing Values panel opens.*
5. Click **Select a Dataset** and select a dataset from the **Select a Dataset** pop-up window.
6. On the **Define Replacements** tab, click **Select** and choose a column.
7. Select an option (Mean, Median, Zero) for the **Replace With** field.
8. Click the **+Add Column** button, select a column and replacement value. Repeat as needed for other columns.

2.29 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical procedure. It uses an orthogonal transformation to convert a set of observations (of possibly correlated variables) into a set of values of linearly uncorrelated variables called principal components. It finds a rotation in which the first coordinate has the largest variance possible and each succeeding coordinate, in turn, has the largest variance possible. The columns of the rotation matrix are called principal components. PCA is widely used in dimensionality reduction. After finding the new coordinates, it is possible to choose fewer dimensions (where variation happens most) in the new coordinate system and project that data back to the reduced dimension coordinate system, while preserving most of the information about the data.

Principal Component Analysis Panel Illustration

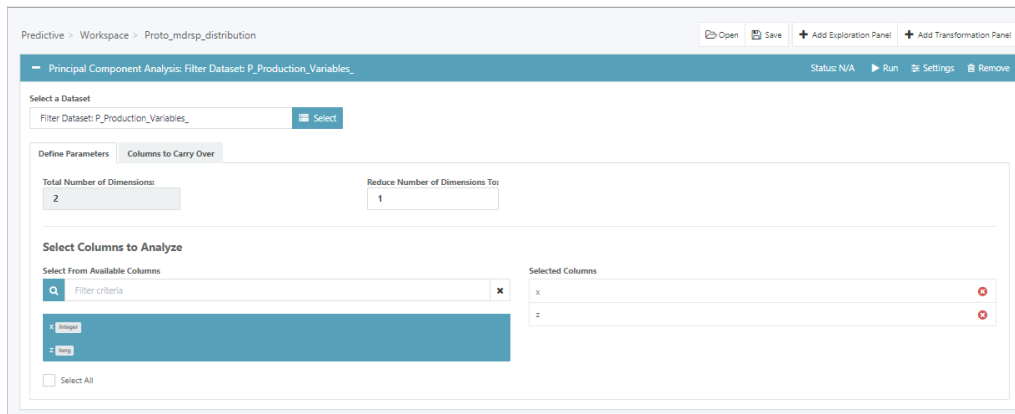
The following illustrates the Principal Component Analysis panel on the Predictive Workspace.



How to Use the Principal Component Analysis Transformation

Follow these steps to use the Principal Component Analysis transformation:

1. Access the Manage Analytics Workspaces page. *The Manage Analytics Workspaces page appears.*
2. Select a cluster configuration and start a cluster. *The cluster starts and the cluster status message changes to Running.*
3. Click **Create a New Workspace** or open an existing workspace. *The Workspace opens on a new page.*
4. Click **Add Transformation Panel**. *The Select Transformation dialog box opens.*
5. Select **Principal Component Analysis** from the list and click the **Select** button. *The Principal Component Analysis transformation panel displays.*
6. Click the **Select** button next to the **Select a Dataset** field and choose a dataset from the list.
7. Select the columns you want to include from the **Select Columns to Analyze** list, or click **Select All**. *The columns you select appear on the right, and the Total Number of Dimensions reflects your selections.*
8. Enter the number of columns you want to reduce the dataset to in the **Reduce Number of Dimensions** field.
9. Specify the columns you want to show on the **Columns to Carry Over** tab.
10. Click **Run** on the panel title bar. *The transformation runs and the new dataset displays.*



2.30 Normalizing Data With StandardScaler

StandardScaler is an out-of-the-box transformation tool for datasets. Scaling, also known as "normalization", helps improve the convergence rate during the optimization process, and also prevents features with very large variances from exerting excessive influence during model training.

It is highly recommended to transform null values before executing a StandardScaler Transformation because this tool ignores columns that contain null values, which can mean a fewer-than-expected number of rows in the resulting dataset.

StandardScaler Formula

The formula used by the Predictive Learning Scaler is:

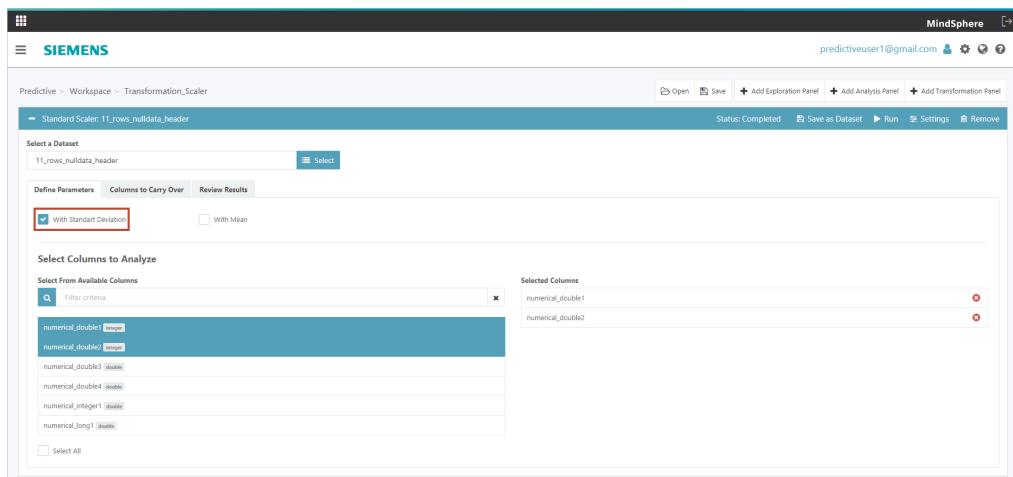
X: Input Column to transform

X_STS: Transformed Column

$$X_STS = X - \text{mean}(X) / \text{std_dev}(X)$$

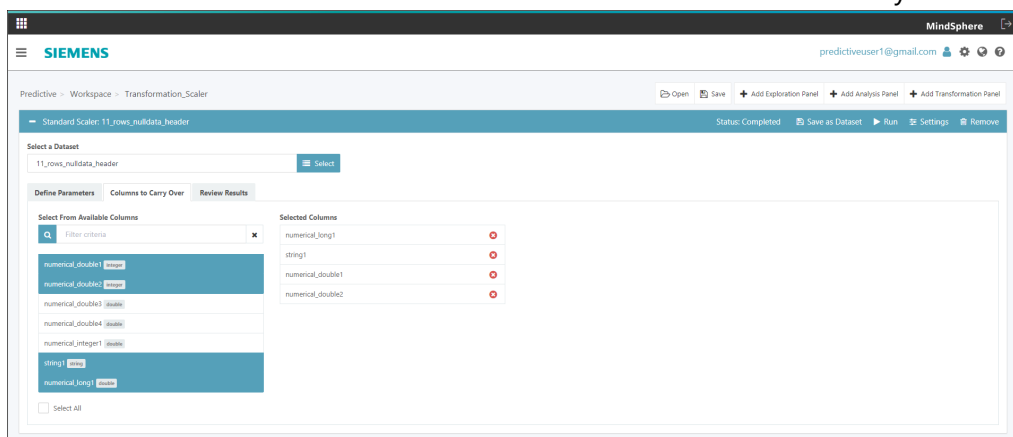
StandardScaler Transformation Panel Illustration

The following illustrates the Parameters tab in StandardScaler.



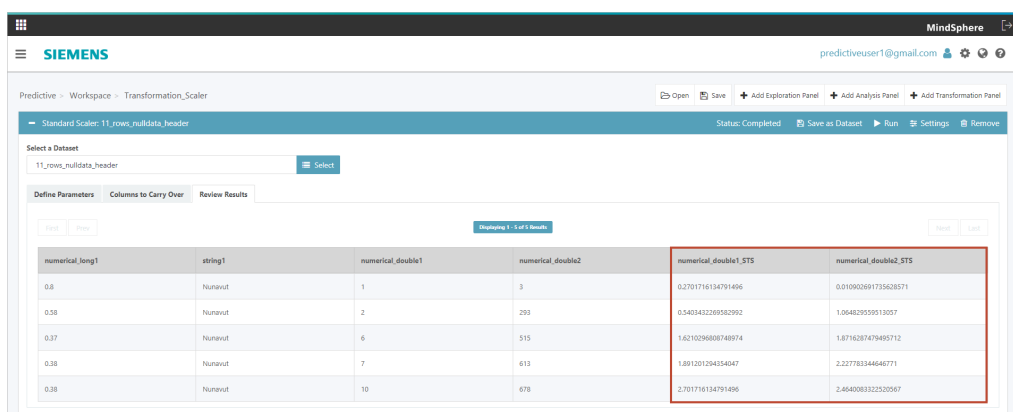
Columns to Carry Over Tab Illustration

This illustrates the StandardScaler Transformation Panel Columns to Carry Over tab.



Review Results Tab Illustration

This illustrates the Review Results tab.



How to Access the StandardScaler Transformation Panel

Follow these steps to access the StandardScaler Transformation panel:

1. Click on **Manage Analytics Workspaces** on the Predictive Learning menu. *The Cluster Configuration page opens.*
2. Start a cluster if one is not already running. *The cluster status changes to RUNNING once the cluster has started.*
3. Open an existing workspace or create a new one. *The Workspace opens.*
4. Click the **Add Transformation Panel** button at the top right. *The Select Transformation dialog box opens.*
5. Select **StandardScaler**. *A StandardScaler transformation panel opens.*

How to Use the StandardScaler to Transform Your Data

Follow these steps to normalize your data using the StandardScaler:

1. Click the **Select** button and choose a dataset from the **Select a Dataset** field. *A list of columns appears in the Columns to Analyze list.*
2. Select the **Define Parameters** tab and select **With Standard Deviation** (default), **With Mean**, or **both**. *Your selection determines the hyperparameter used in the StandardScaler algorithm.*
3. Select at least one column to perform the transformation on, or select the **Select All** check box for transforming all columns. *Your selections appear in the Selected Columns list.*
4. Select the **Columns to Carry Over** tab and select at least one column to include in the preview, or select the **Select All** check box to preview all columns.
5. Click **Run** to run the transformation now, or click **Save as Dataset**. *A save dialog opens.*
6. Accept the default dataset name or enter a new one in the **Dataset name** field.
7. Click the **Preview** tab. *The transformed columns are appended to the end of the results grid with the name <OriginalColumnName _STS>.*
8. Click **Save as Dataset**. *The transformed attributes are added to the original dataset and saved to the S3 location you specify.*

2.31 Running an Analysis

The last step in the Predictive Learning workflow is to run the analysis of the dataset using the transformations you have applied in your workspace.

How to Run an Analysis

Follow these steps to run an analysis in the Manage Analytics Workspace page:

1. Navigate to the **Manage Analytics Workspace** page.
2. Start a **cluster**.
3. Open a **data configuration**.
4. Add **exploration panels** and **transformations** to the data configuration.
5. Click **Save** to save the workspace.
6. Locate the analysis you want to run in the Join Transformation panel and click **Run**. *The status changes to "Running", the analysis runs, and the new dataset displays on the Preview tab of the Manage Analytics Workspace page.*

2.32 Launching a Service

Once you create a dataset and define transformations in the workspace, you can use the Predictive Learning notebook functionality and launch an external service. Introductory notebooks are provided with initial instructions to get you started. Refer to the documentation for the service for further instructions.

Predictive Learning supports the following services:

- Zeppelin Notebook
- Tensorboard
- Spark History
- Jupyter Notebook

How to Launch a Service

Follow these steps to launch a service:

1. Navigate to the **Manage Analytics Workspace** page. *The Cluster Configuration page appears.*
2. Select a **Cluster Type** and click **Start Cluster**. *When the cluster is running, the Cluster Status changes to RUNNING and the Launch a Service field and Launch button display.*
3. Open a **saved workspace**, with transformations and columns selected, and select a service from the **Services** drop-down list.
4. Click **Launch**. *An introductory page for the service you selected displays with instructions on getting started.*
5. Refer to the Help provided for the selected service for more information.

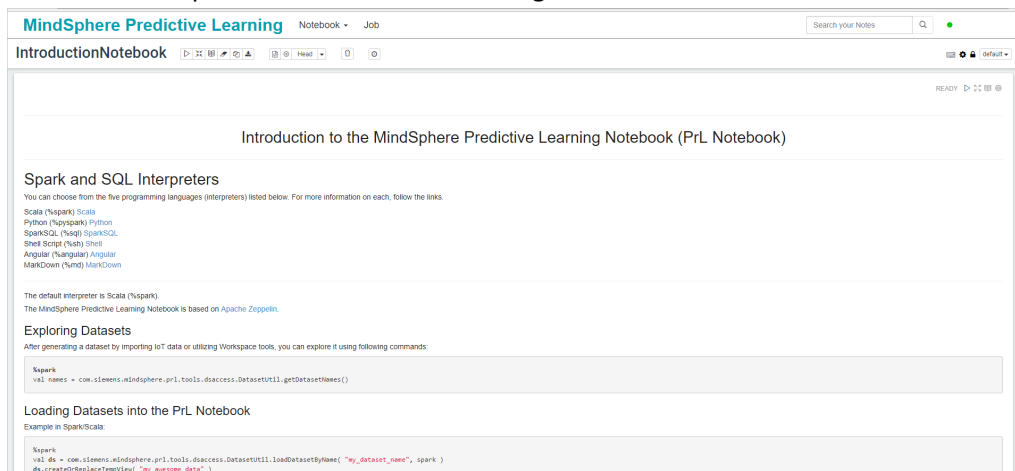
For the Apache Zeppelin Introduction Notebook

Once the Notebook has launched, click the **Introduction Notebook** for information on how to use and access the Notebook in the Predictive Learning environment.

Once launched, you can write your own script for performing an analysis. You can also create multiple notebooks, as needed.

PrL Introduction Notebook Illustration

This is an example of the Predictive Learning Introduction Notebook:



Jupyter Notebook

When selecting Jupyter Notebook in the Service drop down, a new browser tab will open, providing access to your previously saved notebooks.



2.33 Making API Calls from Zeppelin Notebook

The model development workspace in Predictive Learning is Zeppelin Notebook, which allows use of Public APIs, such as Data Exchange, Model Management, and IoT Time Series.

You can use APIs that your organization has access to in the Predictive Learning workspace.

Access Zeppelin Notebook from these pages:

- Manage Environments
- Manage Analytics Workspaces

Zeppelin Notebook includes both Python 2.7 and Python 3.6 versions.

Working with Zeppelin Notebook

This page covers the processes for using Zeppelin Notebook, including:

- Checking the Packages Installed in Zeppelin
- Best Practices
- Installing Your own Python Libraries
- Installing Your own R Libraries
- Updating Python Interpreters
- Changing the Python Interpreter version
- Calling Public APIs from Zeppelin Notebook
- Copying Data from Integrated Data Lake (IDL)

Go [here](#) for in-depth information on Public APIs.

Checking Your Configuration

When you start working with scripts, the environment requires a certain set of libraries. Some libraries required to run minimal services within the cluster come preinstalled and are available

[here](#)

Use these commands to examine installed packages from Zeppelin:

```
%sh
```

```
pip freeze --user
```

Best Practices for Zeppelin Notebook Performance

To optimize Zeppelin notebook performance and, because the custom actions you perform on the cluster are not stored between restarts, we recommend that you:

- Install the required packages as the **first** step in using the notebook.
- Execute the packages **each time** you start a cluster.

Installing Your own Python Libraries

Users are free to install any libraries are required using:

```
%python
```

```
import os
```

```
import requests
```

```

import json

dlpath = '/datalake/v3/generateAccessToken'

gw = os.environ['GATEWAY_ENDPOINT'] + '/gateway/'

# increment_value = 1

headers = {
    'Content-Type': 'application/json'
}
payload="{ \"subtenantId\": \"\" } "
dl_url = gw + dlpath
response = requests.post(dl_url, data=payload, headers=headers)
#print(response.status_code)
dl = json.loads(response.text)
os.environ["AWS_ACCESS_KEY_ID"] = dl['credentials']['accessKeyId']
os.environ["AWS_SECRET_ACCESS_KEY"] = dl['credentials']['secretAccessKey']
os.environ["AWS_SESSION_TOKEN"] = dl['credentials']['sessionToken']

```

If you require additional external sources for your project, please contact your organization's PrL administrator.

Once the instance is stopped, all libraries and modifications performed on the instance will be lost. You will need to run the installation paragraphs every time the note is imported into Zeppelin, to make sure everything is up to date on the machine when you start working. If the newly installed/updated packages are not in the list as they should, the Python interpreter should be restarted.

Installing R Libraries

Installing R packages is done in a %spark.r paragraph using R commands. The package will be installed with all the dependencies.

```

%spark.r
install.packages('ggvis', repos='https://ftp.fau.de/cran/')
libraries <- as.data.frame(installed.packages()[,c(1,3:4)])`
libraries

```

```

%spark.4
library(ggvis)
head(mtcars)

```

```

%spark.r
remove.packages('ggvis')

```

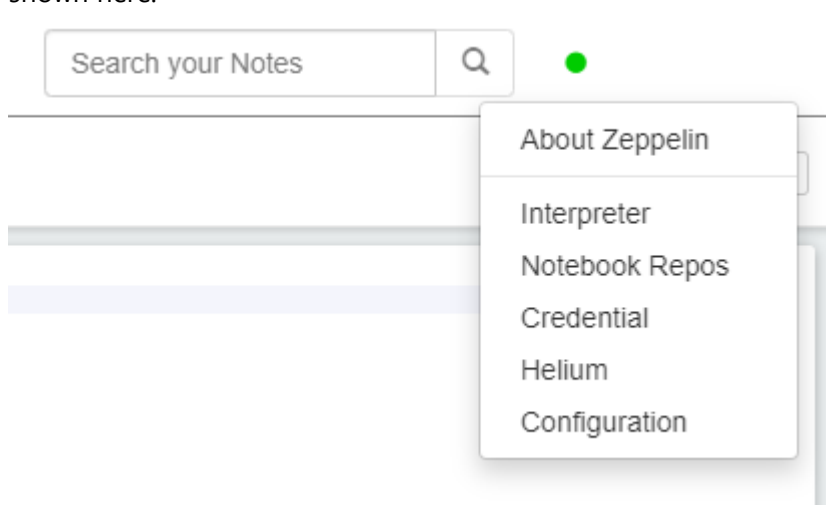
Make sure to select a proper mirror in order to minimize the time required to install the package. All the mirrors can be found [here](#).

Updating the Python Interpreters

A colored circle appears at the top of the Zeppelin Notebook screen that indicates the connectivity status to the server or interpreter. Green indicates connectivity. Red indicates the:

- Session is expired
- Cluster is stopped
- Interpreter is not responding

Click to the right of the circle to view a drop-down menu for customizing Zeppelin options, as shown here:



Changing the Interpreter to Python 3 does not affect the Python version used by the %pyspark Interpreter. The default python interpreter version used by %pyspark is Python 2 and, to change that setting, you must change the spark's zeppelin.pyspark.python setting from 'python' to 'python3'.

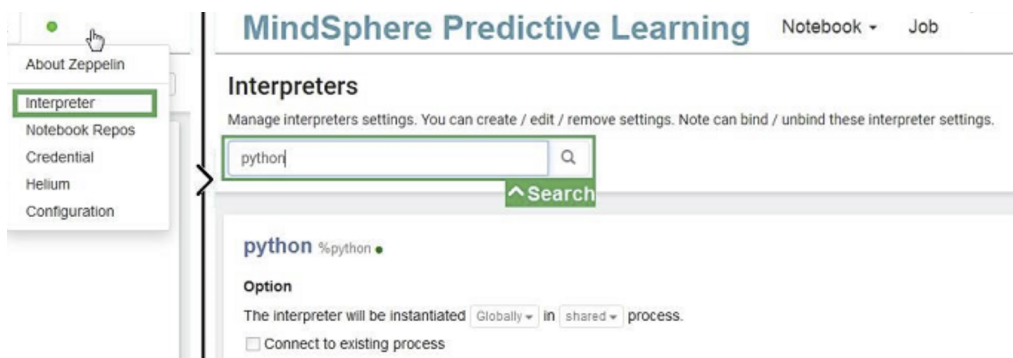
How to Change the Interpreter to Python 3

Predictive Learning provisions the cluster with both Python 2 and 3 Interpreter variants, however, it is required that you update the command line that executes the paragraph with the interpreter's settings. Python 2 is the default interpreter setting.

Follow these steps to change Python Interpreter to Python 3:

1. Navigate to Zeppelin Notebook.
2. Select Interpreter from the **About Zeppelin** drop-down list. *The Interpreters page opens.*
3. Enter "Python" in the **search** field and click **Edit**. The Edit section opens.
4. Enter Python3 in the **zeppelin.python** field and click **Save**. *A message asks you to confirm the change.*

5. Click **OK**. The *Python3* interpreter variant is set.



How to Call a Data Exchange API

This shows a sample API call from a Zeppelin Notebook using the Python interpreter and our internal gateway that handles the authentication procedures in a seamless manner:

```
%python
import os
import requests
import json
#get the proxy - we call it Gateway- URL
gw = os.environ['GATEWAY_ENDPOINT'] + '/gateway/'
#some paths to remember
DEpath = 'dataexchange/v3/'
dirs = 'directories/'
pub = '_PUBLIC_ROOT_ID'
response = requests.get(gw + DEpath + dirs + pub) #this will list the Public Directory from Data Exchange
#let's parse the response
allpub = json.loads(response.content)

#we only read the 'files'; this also contains the 'directories' child which is also iterable
for file in allpub['files']:
    print("File id: " + str(file['id']))
#uploading files work in a similar fashion; or working with other MDSP services** <br/>
```

This example calls the Predictive Learning Developer Data Exchange API and lists the contents of the public root folder.

How to load existing Datasets

When using Zeppelin notebooks, these variables can be read using the Zeppelin session instance variable named 'z', available under Spark and PySpark Interpreters:

```
//make sure you have the proper context set up at the beginning of your pa
ragraph, like %Spark
```

```
var inpf = z.get( inputFolder )
var outf = z.get( outputFolder )
var dsname = z.get( datasetName )
```

When you want to load the dataset you need in your Zeppelin notebook, you can use the built-in library functions, like this:

```
%spark
val names = com.siemens.mindsphere.prl.tools.dsaccess.DatasetUtil.getDatasetNames()
//make sure you also pass in the spark context
var ds = com.siemens.mindsphere.prl.tools.dsaccess.DatasetUtil.loadDatasetByName('my dataset', spark)
ds.createOrReplaceTempView( \"my_awesome_data\" )
```

Copying Data From IDL

Use the code below to obtain a temporary token (via the PrL gateway) and send a read request to the Integrated Data Lake (IDL) API . The IDL API permits only the **read** operation with a temporary access token. The path part `/data/ten=mytenant/` is fixed and cannot be changed.

Once the AWS temporary keys have been set up, you can use AWS CLI commands from a command line to perform read operations against the IDL bucket. The bucket name is indicated in Json format in the IDL response.

```
%%bash
content=$(curl --location --request POST $GATEWAY_ENDPOINT$'/datalake/v3/generateAccessToken' --header 'Content-Type: application/json' --data-raw '{ "subtenantId":"" } ')
#echo $content

`secret=$(jq -r '.credentials.secretAccessKey' <<< "${content}")`
`session=$(jq -r '.credentials.sessionToken' <<< "${content}")`
`accesskey=$(jq -r '.credentials.accessKeyId' <<< "${content}")`

export AWS_ACCESS_KEY_ID=$(echo "${accesskey}")
export AWS_SECRET_ACCESS_KEY=$(echo "${secret}")
export AWS_SESSION_TOKEN=$(echo "${session}")
aws s3 ls s3://datalake-integ-aaad/data/ten=tenantname/
```

2.34 Using Jupyter Notebook

You can use different Public APIs such as Data Exchange, Model Management, and IoT Time Series from Jupyter Notebook, the model development workspace for Predictive Learning. Your tenant must have valid access to these APIs in order to utilize them in the workspace environment. The Jupyter Notebook can be accessed from the:

- Manage Environments page
- Manage Analytics Workspaces page

The steps to accomplish this are outlined in the following procedures.

Go [here](#) for in-depth information on Public APIs.

Checking your Configuration

When working with scripts, your environment will require a certain set of libraries. Some libraries required to run the minimal services within the cluster come preinstalled and are available [here](#).

Run these commands to examine installed packages from Jupyter:

```
%pip freeze --user
```

We recommend you install the required packages at the beginning of the notebook and execute it each time the cluster has been started. Currently Predictive Learning does not store the custom actions you perform on the cluster between restarts.

Using Inputs from Job Executions

In the current implementation, all job executions require parameters; these parameters can be one of Data Exchange, IoT, Data Lake or Predictive Learning Storage (PrL Storage). For the first three, Job Manager will ensure copying the input into a temporary location that is available to your code. In Jupyter notebooks there are three variables available: *inputFolder*, *outputFolder* and *datasetName*. These can be read using the Jupyter magic command `%store` as in:

```
%store -r inputFolder #-r specifies a read
```

```
%store -r outputFolder
```

```
%store -r datasetName
```

The *datasetName* variable will only contain a value when you use the IoT input type. The *inputFolder* is prefilled by the job execution engine with a value pointing to the temporary location that holds the input files or data. That will be an S3 path on AWS or a blob storage on Azure. It does not contain the associated prefix like *s3://*. You can then use the *outputFolder* variable in a Jupyter notebook as in:

```
!aws s3 cp ./mylocalfile.txt s3://$outputFolder+ '/myfile.txt'
```

Always take into account that both *inputFolder* and *outputFolder* variables are remote storage paths, and not local folders, therefore most of the regular file functions will no work against it. However, the CLI and shell commands will use these as long as you correctly prefix them. For the Python Scala libraries that can work with remote storage services, Predictive Learning recommends checking the respective library's documentation; for example, The pandas Python library can save and read from AWS S3 storage.

Installing your own Python Libraries

Run these commands to install libraries:

```
#upgrade pip and install required libraries
%pip install --upgrade pip --user
%install requests --force-reinstall --upgrade --user
%pip install pandas --force-reinstall --upgrade --user
%install pyarrow --force-reinstall --upgrade --user
```

Not all external repositories are allowed. If you require additional external sources for your project, please contact your organization's PrL administrator.

Once the instance is stopped, all libraries and modifications performed on the instance will be lost. Due to that, you will need to run the installation paragraphs every time the note is imported into Jupyter, to make sure everything is up to date on the machine when you start working.

More About Jupyter Notebook

Jupyter is a powerful tool that allows multiple customizations and languages. These resources can help you explore further:

<https://jupyter.org/documentation>

<https://jupyter-notebook.readthedocs.io/en/stable/>

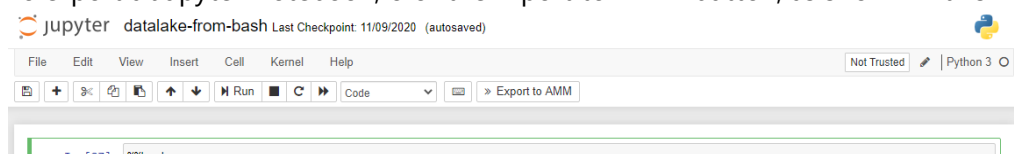
<https://ipython.readthedocs.io/en/stable/interactive/magics.html>

Exporting Notebooks to Analytical Models

When your model is ready to be deployed as a job you can easily move it into Model Management tool allowing it to be exposed to job execution, with no notebook export needed. Model versioning is not supported, which means that if you want to update a model that exists in Model Management, you must first export it from Jupyter Notebook.

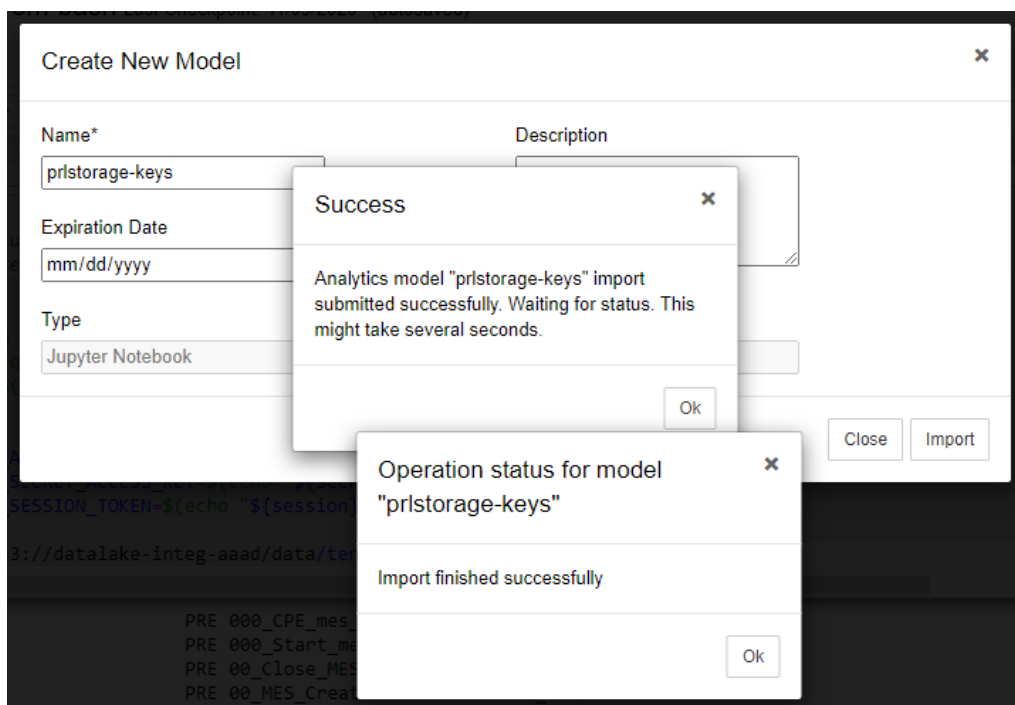
Jupyter Notebook Export Illustration

To export a Jupyter Notebook, click the Export to AMM button, as shown in this image:



Export Windows Example

Here is an example of the po-up windows that display for the export:



How to Export a Jupyter Notebook to an Analytical Model

Follow these steps to export a Jupyter Notebook:

1. Navigate to the Jupyter Notebook.
2. Click the **Export to AMM** button. *The Create New Model pop-up window displays.*
3. Enter the name of the new model in the **Name** field.
4. Click **Import**. *A Success message indicates the import has been submitted.*
5. Click **Ok**. *The Success message closes.*
6. Click **Ok** to close the 'Import finished successfully' message.

Calling a Data Exchange API

The code below shows an example of an API call from a Zeppelin Notebook using the Python interpreter and Predictive's internal gateway that handles the authentication procedures in a seamless manner:

```
import os
import requests
import json
#get the proxy - we call it Gateway- URL
gw = os.environ['GATEWAY_ENDPOINT'] + '/gateway/'
#some paths to remember
DEpath = 'dataexchange/v3/'
dirs = 'directories/'
```

```

pub = '_PUBLIC_ROOT_ID'
response = requests.get(gw + DEpath + dirs + pub) #this will list the Public Directory from Data Exchange
#let's parse the response
allpub = json.loads(response.content)
#we only read the 'files'; this also contains the 'directories' child which is also iterable
for file in allpub['files']:
    print("File id: " + str(file['id']))
#uploading file work in a similar fashion; or working with other MDSP services

```

This example shows the call to Predictive Learning Developer Data Exchange API and lists the contents of the public root folder.

Using exported IoT Datasets

We prepare the environments that you start with our `prlutils` library that handles reading of parquet dataset files into a Pandas Dataframe. Please use the bellow snippet to show the available datasets and then load a single dataset.

```

from prlutils import datasetutils
import boto3
import os
import json
import s3fs

du = datasetutils.DatasetUtils()
datasetnames = du.get_dataset_names()
print('Dataset names: ' + str(datasetnames))
#ds.shape

```

You will get a list of datasets like in:

```

['test_asset_2',
 'Last30DaysAsset2Filtered',
 'Last30DaysAsset2']

```

You can load the dataset you want with the:

```

ds = du.load_dataset_by_name(datasetnames[0])
ds

```

and check its data immediately:

	_time	time	flow	flow_qc	pressure	pressure_qc	temp	temp_qc	pk	rk	tll	_itime
0	1586060055730	NaN	105	NaN	35	NaN	52.0	NaN	None	NaN	NaN	1586166501255
1	1586060066750	NaN	105	NaN	36	NaN	66.0	NaN	None	NaN	NaN	1586166511032
2	1586060077330	NaN	116	NaN	36	NaN	51.0	NaN	None	NaN	NaN	1586166521100
3	1586060088248	NaN	117	NaN	37	NaN	64.0	NaN	None	NaN	NaN	1586166530858
4	1586060098822	NaN	109	NaN	36	NaN	57.0	NaN	None	NaN	NaN	1586166540960
...
974	1586304058495	NaN	169	NaN	0	NaN	0.0	NaN	None	NaN	NaN	1586262761348
975	1586304346027	NaN	149	NaN	0	NaN	0.0	NaN	None	NaN	NaN	1586264739121
976	1586304738584	NaN	149	NaN	0	NaN	0.0	NaN	None	NaN	NaN	1586264745255
977	1586304918584	NaN	149	NaN	0	NaN	0.0	NaN	None	NaN	NaN	1586264750490
978	1586305222970	NaN	149	NaN	0	NaN	0.0	NaN	None	NaN	NaN	1586264755822

979 rows × 12 columns

Then, you

can use your dataset just like with any other Pandas Dataframe:

```
filteredDataset = ds[ds['temp']>60]
print("Number of entries AFTER filtering: "+ str(filteredDataset.shape))
try:
    path = "s3://" + outputFolder
    filteredDataset.write.csv(path)
    filteredDataset.write.csv('s3://prl-storage-216273414971/prlteam/dat
a/')
except:
    print('Output folder is None.')
else:
    print('Filtered dataset written to outputFolder' + outputFolder)
```

If you encounter issues with using our library, make sure that the libraries used by our Datasets utility does not conflict with your previously installed Python libraries. Our utility library uses the following:

```
%pip install pyarrow fastparquet fss pec s3fs boto3 awscli
```

Copying Data From Integrated Data Lake (IDL)

Run the code below to obtain a temporary token (via the PrL gateway) and enable PrL to directly perform a read operation on the IDL API data bucket.

Once the AWS temporary keys have been set up, you can use AWS CLI commands to perform read operations against the Integrated Data Lake bucket.

The bucket name can be observed from the data lake's response, in a json format. The path part /data/ten=mytenant/ is fixed and cannot be changed.

Run the commands below to copy data from IDL:

Uploading Data to IDL

```

import os
import requests
import json
dlpath = '/datalake/v3/generateAccessToken'
gw = os.environ['GATEWAY_ENDPOINT'] + '/gateway/'
# increment_value = 1
headers = {
    'Content-Type': 'application/json'
}
payload="{ \"subtenantId\": \"\" } "
dl_url = gw + dlpath
response = requests.post(dl_url, data=payload, headers=headers)
#print(response.status_code)
dl = json.loads(response.text)
os.environ["AWS_ACCESS_KEY_ID"] = dl['credentials']['accessKeyId']
os.environ["AWS_SECRET_ACCESS_KEY"] = dl['credentials']['secretAccessKey']
os.environ["AWS_SESSION_TOKEN"] = dl['credentials']['sessionToken']

```

Run the following commands to upload data in Integrated Data Lake:

```

# upgrade pip and install required libraries
%pip install --upgrade pip --user
%pip install requests --force-reinstall --upgrade --user
%pip install pandas --force-reinstall --upgrade --user
%pip install pyarrow --force-reinstall --upgrade --user
import datetime
import requests
import os
import json
import re
HEADERS = {
    'Accept': '/*/*',
    'Accept-Encoding': 'gzip, deflate, br',
    'Connection': 'keep-alive',
    'Content-Type': 'application/json'
}
GATEWAY = os.environ['GATEWAY_ENDPOINT'] + '/gateway/'
OUTPUT_FOLDER = 'OUTPUT_FOLDER'
#Get a signed URL for down/upload of data. The function
#attempts for 5 times to obtain the URL and then raises an exception.
def getSignedURL(fileName, folder, attempt=0, upload=True):

```

```
if upload:
    IDLpath = 'datalake/v3/generateUploadObjectUrls'
else:
    IDLpath = 'datalake/v3/generateDownloadObjectUrls'
IDLFilePath = '/%s/%s' % (folder, fileName)
url = GATEWAY + IDLpath
body='{"paths": [{"path": "%s"}]}' % IDLFilePath
response = requests.post(url, headers=HEADERS, data=body)
try:
    return json.loads(response.text)['objectUrls'][0]['signedUrl']
except KeyError:
    if attempt < 5:
        attempt += 1
        return getSignedURL(fileName, attempt, upload)
    else:
        raise Exception('Failed to get a signed URL')
!echo "This is a test!" >> test.txt
fileName='test.txt'
signedURL = getSignedURL(fileName, OUTPUT_FOLDER)
requests.put(signedURL, headers=HEADERS, data=fileName)
```

Reading data from IoT

Run the following commands to read data from IoT sources:

```
%pip install --upgrade pip --user
%pip install requests --force-reinstall --upgrade --user
%pip install awscli --force-reinstall --upgrade --user
%pip install pandas sklearn seaborn matplotlib joblib --user
import json
import io
import os
import datetime
import time
from dateutil import parser
import random
from threading import Thread
import requests
import pandas as pd
import tempfile
def read_iot(entity_id = "<<iot_entity_id_GUID>>",
             aspect_name = "<<aspect_name>>",
             tenant = "tenantname",
             max_results = 2000, #max is 2000
```

```

        from_dt = "2020-06-01T13:09:37.029Z",
        to_dt = "2020-07-01T08:02:27.962Z",
        variable = "pressure",
        sort = "asc"):
    if variable is not None:
        url = "?from=" + from_dt + "&to=" + to_dt + "&sort=" + sort + "&limit=" + str(max_results) + "&select=" + variable
    else:
        url = "?from=" + from_dt + "&to=" + to_dt + "&sort=" + sort + "&limit=" + str(max_results)
    #this is the IoT Timeseries API base URL
    TSpath = 'iottimeseries/v3/timeseries'
    #this is the Predictive Gateway URL that handles authentication for you
r API calls
    gw = os.environ['GATEWAY_ENDPOINT'] + '/gateway/'
    headers = {
        'Content-Type': 'application/json'
    }
    iot_url = gw + TSpath + "/" + entity_id + "/" + aspect_name + url
    response = requests.get(iot_url, headers=headers)
    return response
import pandas as pd
import tempfile
start = datetime.datetime.utcnow() - datetime.timedelta(days=70)
end = start + datetime.timedelta(days=30)
response = read_iot(entity_id = "<<iot_entity_id_GUID>>",
                    aspect_name = "<<aspect_name>>",
                    tenant = "tenantname",
                    max_results = 2000, #max is 2000
                    from_dt = start.strftime('%Y-%m-%dT%H:%M:%S.%f')[:-3] + 'Z',
                    to_dt = end.strftime('%Y-%m-%dT%H:%M:%S.%f')[:-3] + 'Z',
                    sort = "asc",
                    variable = None)
if response.status_code == 200:
    f = tempfile.TemporaryFile()
    f.write(response.content)
    f.seek(0)
    #we read the IoT data into a Pandas DataFrame
    data = pd.read_json(f.read())
    f = tempfile.TemporaryFile()
    f.write(response.content)
    f.seek(0)
    print(data.shape)

```

```
else:
    print(response.status_code)
    print(response.content)
```

2.35 Using GPU

GPU instance support is provided for GPU-capable code and frameworks. It can significantly boost the performance of many machine learning training or inference algorithms. GPU instance support is available for PrL workspaces and environments used by jobs.

The provided GPU instances are non-clustered and provide up to 8 GPU computation capabilities per instance.

Run the commands below to view the driver and CUDA toolkit details:

```
nvcc --version
```

```
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2018 NVIDIA Corporation
Built on Wed_Apr_11_23:16:29_CDT_2018
```

Cuda compilation tools, release 9.2, V9.2.88

```
nvidia-smi
```

+-----+ Driver Version: 396.26 +-----+																		
NVIDIA-SMI 396.26																		
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+																		
GPU Name		Persistence-M		Bus-Id		Disp.A		Volatile Uncorr. ECC										
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage		GPU-Util		Compute M.										
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+																		
0	Tesla K80		Off	00000000:00:1E.0 Off				0										
N/A	37C	P0	65W / 149W	0MiB / 11441MiB		99%		Default										
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+																		
+-----+ GPU Memory +-----+																		
Processes:																		
GPU	PID	Type	Process name				Usage											
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+																		
No running processes found																		
+-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+ +-----+																		

Once you know what your system's defaults are for GPU, install any GPU-enabled framework you require, and try out your GPU code:

```
import torch
cuda0 = torch.device('cuda:0')
torch.ones([2, 4], dtype=torch.float64, device=cuda0)
```

2.36 Viewing Usage Metrics

The Usage Metrics page lets you see how many Predictive Learning compute hours remain for your organization, and lists your individual transactions and the number of compute hours you have used.

Usage Metrics Illustration

The following image shows the Usage Metrics page.

Analytics Workspace > Usage Metrics

Compute Hours Remaining: 29,955.1 —Total for Organization

From: Sunday, July 1, 2018 12:00 AM To: Wednesday, October 31, 2018 11:59 PM

First Prev [Displaying 1 - 15 of 20 Results] Refresh Next Last

Instance Used	Compute Hours Used	Transaction Time
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	7.5	Tuesday, October 9, 2018 3:27 PM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	5	Friday, October 5, 2018 12:36 PM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	2.5	Friday, September 7, 2018 4:16 PM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	12.5	Friday, September 7, 2018 3:49 PM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	2.5	Friday, September 7, 2018 11:38 AM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	22.5	Thursday, September 6, 2018 4:29 PM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	7.5	Wednesday, September 5, 2018 4:35 PM
Spark EMR + 2 nodes (32 Cores, 60GB RAM)	7.5	Friday, August 31, 2018 4:06 PM
Spark EMR + 2 nodes (32 Cores, 128GB RAM)	7.8	Thursday, August 30, 2018 10:54 AM

How to View Your Usage Metrics

Follow these steps to view your usage metrics and remaining compute hours:

1. Log into Predictive Learning and select **Usage Metrics** on the Predictive Learning menu. *The Usage Metrics page appears.*
2. Specify a date range in the **From** and **To** fields. Default is the last 90 days.
3. *Your individual instances and compute hours used are displayed in the grid along with the date and time the transactions were initiated.*
4. *The total number of compute hours remaining for your organization appears at the top of the page.*

Need Assistance?

Your compute hours are updated based on the package your company purchases. If your compute hours do not appear on the Usage Metrics page, or are not what you expected, contact Support (GTAC) for assistance:

https://www.plm.automation.siemens.com/en_us/support/gtac/gtac-hours.shtml.

To purchase more hours, contact Predictive Learning Sales:

<https://www.plm.automation.siemens.com/store/en-us/mindsphere/contactus.html>.

2.37 Manage Environment Configurations

This functionality allows an organization's administrators to create environment configurations based on a configuration template. The environment configurations allow an organization's users

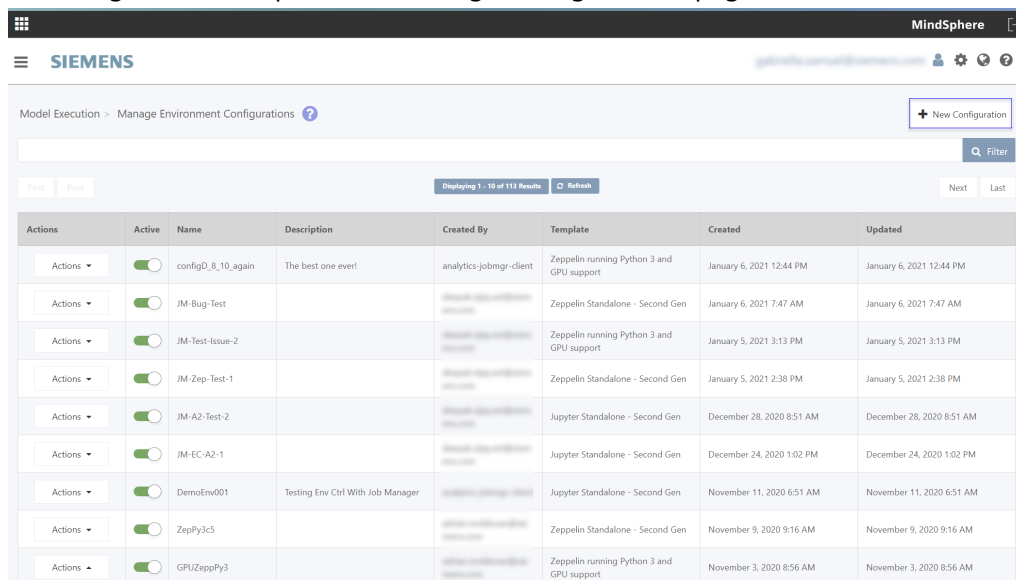
to start and stop environments for planned jobs, and administrators to view, update, and delete configurations.

Currently, the Model Execution template provides these configurations:

- c5.xlarge
- m4.4xlarge
- c4.4xlarge
- r4.4xlarge
- p2.xlarge (for GPU templates)
- p3.2xlarge (for GPU templates)

Manage Configurations Page Illustration

This image is an example of the Manage Configurations page:



Actions	Active	Name	Description	Created By	Template	Created	Updated
Actions ▾	<input checked="" type="checkbox"/>	config0_8_10_again	The best one ever!	analytics-jobmgr-client	Zeppelin running Python 3 and GPU support	January 6, 2021 12:44 PM	January 6, 2021 12:44 PM
Actions ▾	<input checked="" type="checkbox"/>	JM-Bug-Test			Zeppelin Standalone - Second Gen	January 6, 2021 7:47 AM	January 6, 2021 7:47 AM
Actions ▾	<input checked="" type="checkbox"/>	JM-Test-Issue-2			Zeppelin running Python 3 and GPU support	January 5, 2021 3:13 PM	January 5, 2021 3:13 PM
Actions ▾	<input checked="" type="checkbox"/>	JM-Zep-Test-1			Zeppelin Standalone - Second Gen	January 5, 2021 2:38 PM	January 5, 2021 2:38 PM
Actions ▾	<input checked="" type="checkbox"/>	JM-A2-Test-2			Jupyter Standalone - Second Gen	December 28, 2020 8:51 AM	December 28, 2020 8:51 AM
Actions ▾	<input checked="" type="checkbox"/>	JM-EC-A2-1			Jupyter Standalone - Second Gen	December 24, 2020 1:02 PM	December 24, 2020 1:02 PM
Actions ▾	<input checked="" type="checkbox"/>	DemoEnv001	Testing Env Crl With Job Manager		Jupyter Standalone - Second Gen	November 11, 2020 6:51 AM	November 11, 2020 6:51 AM
Actions ▾	<input checked="" type="checkbox"/>	ZepPy3c5			Zeppelin Standalone - Second Gen	November 9, 2020 9:16 AM	November 9, 2020 9:16 AM
Actions ▾	<input checked="" type="checkbox"/>	GPUZeppPy3			Zeppelin running Python 3 and GPU support	November 3, 2020 8:56 AM	November 3, 2020 8:56 AM

How to Access the Manage Environment Configurations Page

Follow these steps to access the Manage Configurations page:

1. Navigate to the **Predictive Learning** page.
2. Select Manage Environment Configurations from the **Model Execution** menu. *The Manage environment Configurations page opens.*

New Configuration Dialog Illustration

This illustrates the dialog for creating a new environment configuration:

New Configuration ×

Template *

Select

Name *

Description

Cancel Save

Select a Template Window Illustration

This illustrates the dialog for selecting a template for an environment configuration:

Select a Template ×

Q

Filter criteria

×

Prev

Displaying 1 - 9 of 9 Results

Next

Docker GPU Standalone
Docker Standalone
Essentials Jupyter - Second Gen
Jupyter Standalone - Second Gen
Jupyter running Python 3 and GPU support
Zeppelin Standalone - Second Gen
Zeppelin running Python 3 and GPU support
[DEPRECATION WARNING] Essentials Jupyter
[DEPRECATION WARNING] Zeppelin Standalone

Clear

Close

How to Create a New Environment Configuration

Follow these steps to create a new environment configuration:

1. On the Manage Configurations page, click the **+ New Configuration** button. A *New Configuration dialog box opens*.
2. Click Select in the **Template** field. *The Select a Template pop-up window opens*.
3. Click the **template** you want to use for the new configuration. *The pop-up window closes and the template displays in the New Configuration pop-up window*.
4. Enter a unique name for the configuration in the **Name** field.
5. Enter a description in the **Description** field (optional).
6. Select an entry from the **Instance Type** drop-down list.
7. Click **Save**. *The environment configuration is saved and its name displays in the Environment Configurations list on the Manage Configurations page*.

How to Open an Environment Configuration

After navigating to the Manage Environment Configurations page, here are some simple ways to open an existing environment configuration:

- Enter the first few characters of an environment configuration name in the **Filter** field. The table refreshes with names that match your entry. Locate the one you want and click the **Action** button, and select **Open**.
- Scroll through the list of environment configuration names and when you locate the one you want, click the **Action** button and select **Open**.

How to Delete an Environment Configuration

After navigating to the Manage Environment Configurations page, here are some simple ways to delete an existing environment configuration:

- Enter the first few characters of an environment configuration name in the **Filter** field. The table refreshes with names that match your entry. Locate the one you want and click the **Action** button, and select **Delete**.
- Scroll through the list of environment configuration names and when you locate the one you want, click the **Action** button and select **Delete**. Click **OK** in response to the 'Are you sure?' warning message.

2.38 Manage Environments

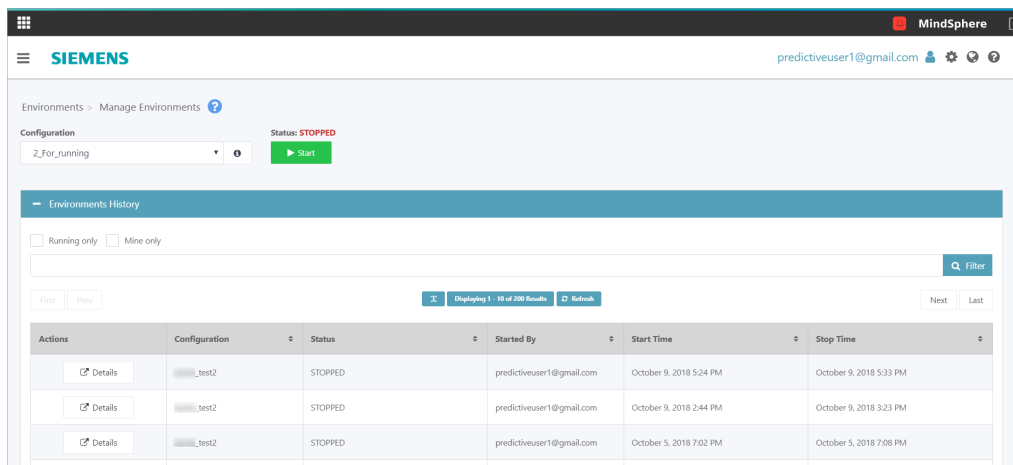
The Manage Environments page allows you to select and start or stop an environment to run your model on. The drop-down list contains all of the environment configurations created and saved by your administrator on the Manage Configurations page. You can start one environment per configuration for each user.

The "Running Only" and "Mine Only" check boxes allow you to filter the list of available configurations to your specifications. The Environments History grid shows the Configuration name, Status, Started By, Start Time, and Stop Time for each configuration displayed.

Currently, the only action available on the grid is "Details" which opens the environment configuration dialog box. You can also open this dialog box by clicking the "i" icon next to the Start/Stop button. The Start/Stop button starts and stops the environment configuration and the current status is displayed above the button.

Manage Environments Illustration

The following illustrates the Manage Environments page.



How to Access the Manage Environments Page

Follow these steps to access the Manage Environments page:

1. Log in to Predictive Learning. *The Home page appears.*
2. Click **Manage Environments** on the **Environments** tile on the Predictive Learning menu. *The Manage Environments page appears.*

How to Start an Environment

Follow these steps to start an environment for your model:

1. Access the Manage Environments page. *The Manage Environments page appears.*
2. Select a configuration from the drop-down list in the **Configuration** field. Click the "i" next to the drop-down arrow to view the details of the configuration.
3. Click the **Start** button. *Once the environment is up and running, the status message changes to "running".*

How to Stop an Environment

Follow these steps to stop an environment that is already running:

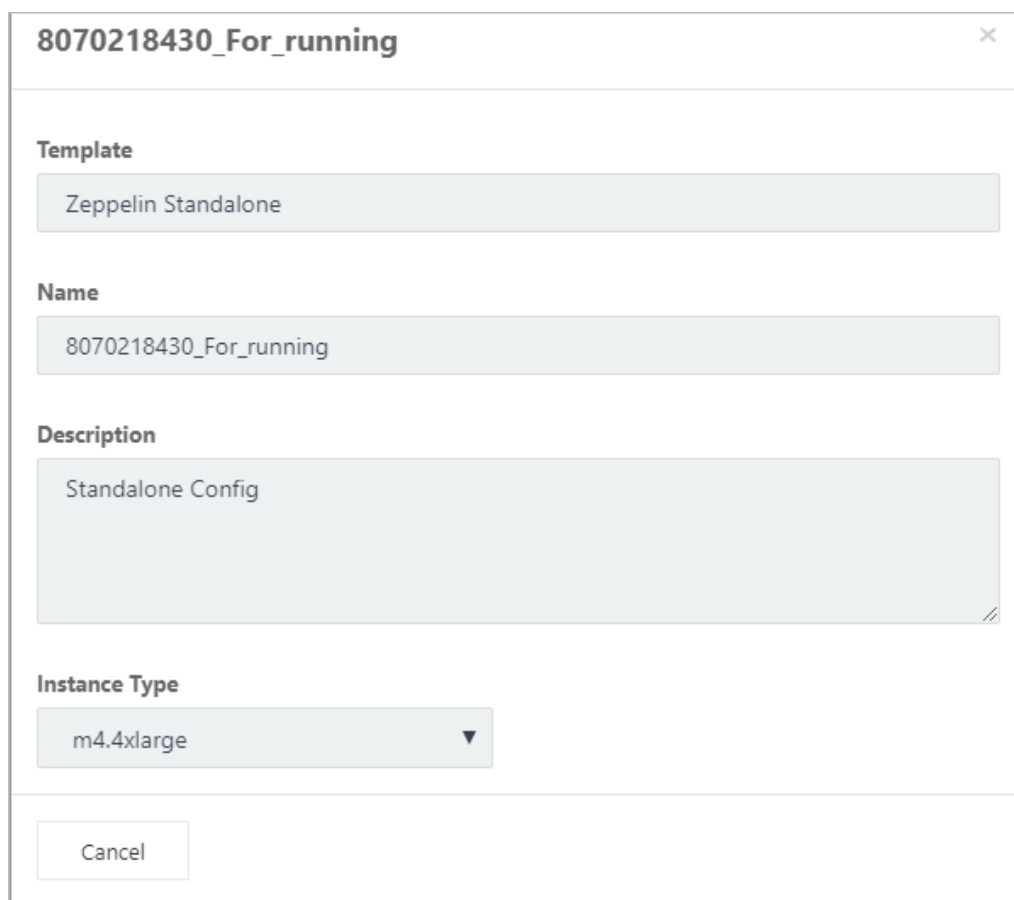
1. Access the Manage Environments page. *The Manage Environments page appears.*
2. Select a configuration from the drop-down list in the **Configuration** field. Click the "i" next to the drop-down arrow to view the details of the configuration.
3. Click the **Stop** button. *Once the environment has stopped, the status message changes to "stopped".*

How to Open an Environment Configuration

You can access your saved environment configurations by clicking the Actions button on the Manage Environments page.

Follow these steps to open an environment configuration:

1. Access the Manage Environments page. *The Manage Environments page appears.*
2. If you have a long list of saved environment configurations, you can search for the configuration using the **Filter** field.
3. Find the configuration you want to open in the grid and click the **Actions** button in that row. *An action menu opens.*
4. Click **Open Environment** to open the started environment (Jupyter or Zeppelin notebook). This action is not available for Docker environments.
5. Click **Details** in the **Actions** menu if you want to review the details of your environment.



The screenshot shows a modal window titled "8070218430_For_running" with a close button (X) in the top right corner. The modal contains the following fields:

- Template:** A text field containing "Zeppelin Standalone".
- Name:** A text field containing "8070218430_For_running".
- Description:** A text area containing "Standalone Config".
- Instance Type:** A dropdown menu showing "m4.4xlarge" with a downward arrow.
- Buttons:** A "Cancel" button is located at the bottom left of the modal.

2.39 Managing Analytical Models (External)

Manage Analytical Models allows you to upload models developed outside of Predictive Learning and store them in an external S3 bucket. A model can contain multiple versions, which may include different file types and belong to different authors. Use this function to create and manage your external models and versions.

Actions available on this page include:

- Create New Model
- View List of Models (with last version information)
- Open Model Details
- Edit Model Details
- Create New Version
- Download Model
- Delete Model

Please note the following limitations of Manage Analytical Models:

- Model file size is limited to at least 10 bytes but no more than 100 MB
- Model file type is not validated
- Any model uploaded can be accessed by any user in your tenant

Manage Analytical Models (External) Illustration

The following illustrates the Manage Analytical Models page:

Actions	Name	Last Version	Type	Author	Description	Creation Date	Expiration Date
Actions -	Test Zepplin for	1.0	Zepplin Notebook	@siemens.com	Test Zepplin for Chuck, simple print	October 9, 2018 5:35 AM	October 3, 2019 8:34 AM
Open Details Edit Details	Boost Model	1.0	Zepplin Notebook	@siemens.com	Using xgboost for classification	September 28, 2018 2:22 PM	December 31, 2019 10:20 AM
New Version Download Model	nieTest4	2.0	Text file	@siemens.com	come back again	September 25, 2018 5:11 PM	December 12, 2018 1:09 PM
Delete Model	nieTest13333	1.0	JSON file	@siemens.com	test2333	September 21, 2018 2:38 PM	December 28, 2018 10:35 AM
Actions -	blabla	1.0	Text file	@siemens.com	blabla	September 25, 2018 2:18 PM	September 20, 2018 10:14 AM
Actions -	super predictions model	2.0	JSON File	predictiveuser1@gmail.com		September 21, 2018 3:32 PM	May 29, 2018 11:26 AM
Actions -	testnine	1.0	JSON File	@siemens.com		October 9, 2018 12:05 PM	March 9, 2019 8:04 AM
Actions -	MDS-AMM	3.0	JSON File	@siemens.com	Test	September 25, 2018 6:54 PM	

How to Access Manage Analytical Models (External)

Follow these steps to access Manage Analytical Models (External):

1. Log on to **Predictive Learning**. *The main page appears.*
2. On the **Analytical Model Management** tab, click **Manage Analytical Models (External)**. *The Manage Analytical Models (External) page appears displaying a list of models uploaded to Predictive Learning.*

How to View a List of Analytical Models

The Manage Analytical Models (External) page lists all models uploaded to Predictive Learning for your tenant, and provides the following information about each:

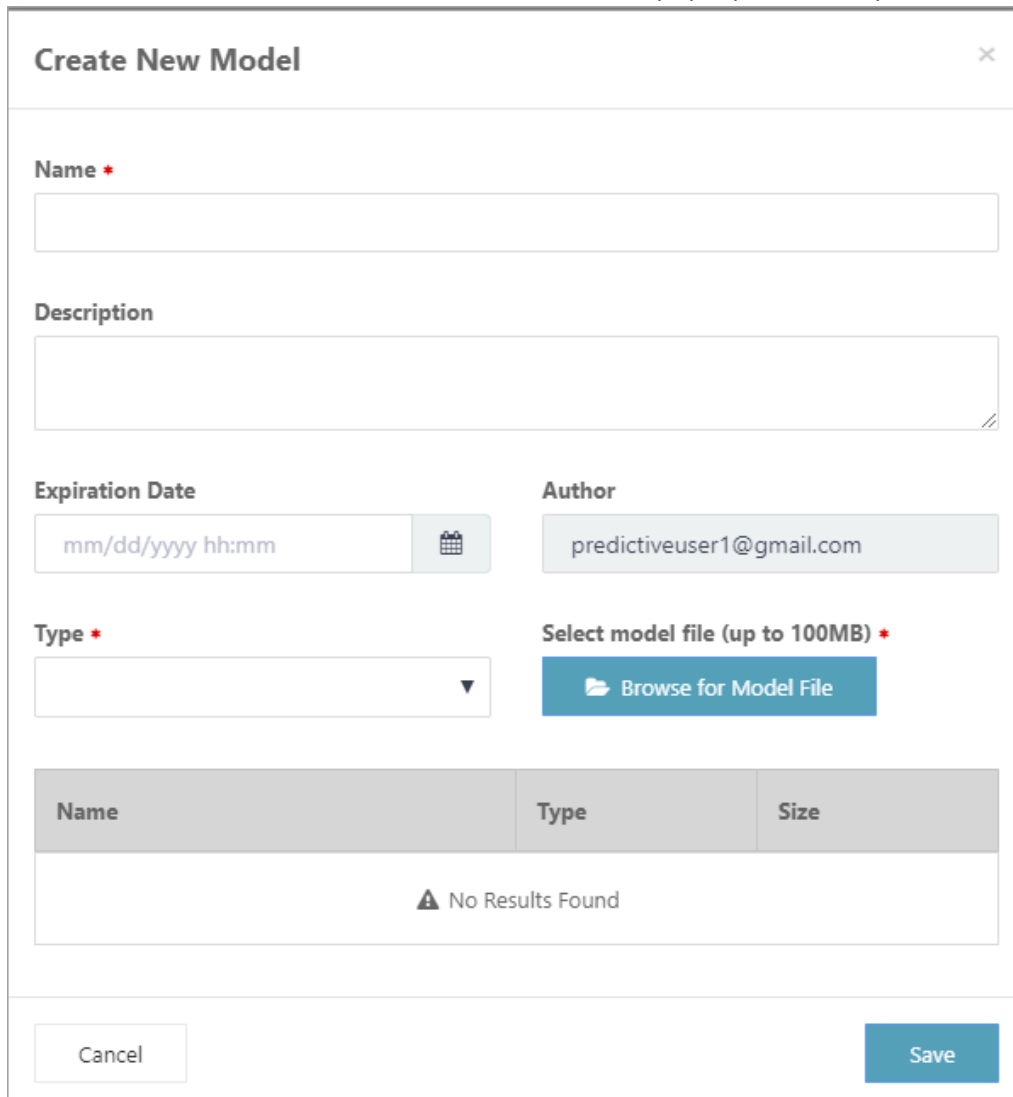
- Name
- Last Version
- Type
- Author
- Description
- Creation Date
- Expiration Date

How to Create a New Model

Follow these steps to create and upload a new model to Predictive Learning:

1. Access the **Manage Analytical Models (External)** page.

2. Click the **New Model** button. A *Create New Model pop-up window opens.*



The 'Create New Model' pop-up window contains the following fields and controls:

- Name ***: A text input field.
- Description**: A text area for a brief description.
- Expiration Date**: A date and time picker showing 'mm/dd/yyyy hh:mm'.
- Author**: A read-only field displaying 'predictiveuser1@gmail.com'.
- Type ***: A dropdown menu.
- Select model file (up to 100MB) ***: A button labeled 'Browse for Model File'.
- Table**: A table with columns 'Name', 'Type', and 'Size'. It currently displays 'No Results Found'.
- Buttons**: 'Cancel' and 'Save' buttons at the bottom.

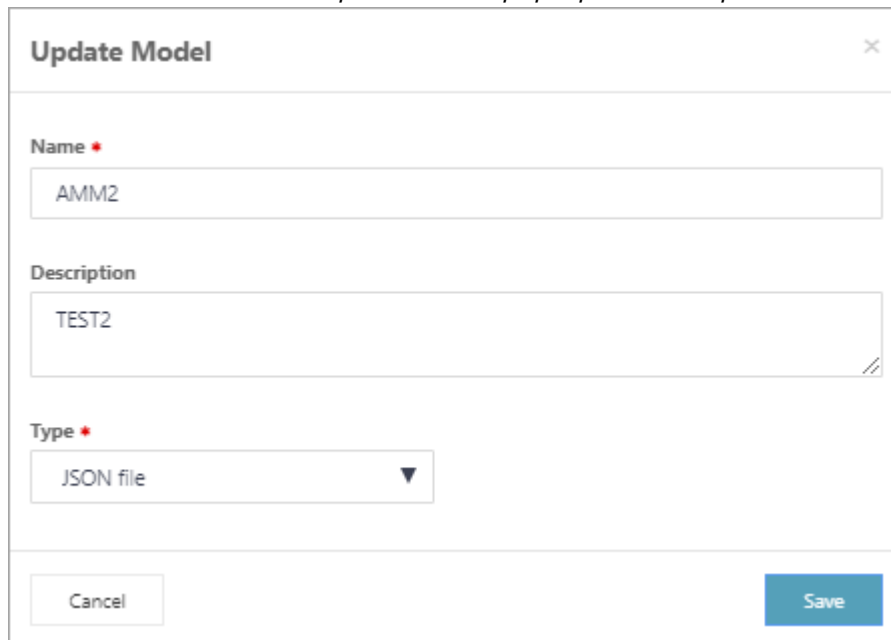
1. Enter a **Name** for the new model.
2. Enter a brief **Description** of the model.
3. Select an **Expiration Date**, if desired.
4. The **Author** field displays your user name and cannot be changed.
5. Select a **Type** from the drop-down list. Current options include: Zeppelin Notebook JSON file XML file Text file
6. Click the **Browse for Model File** button and select a file to upload from your local drive. File size cannot exceed 100 MB. *The file name you select appears in the grid.*
7. Click **Save**. *The file uploads to your external S3 bucket and appears in the list on the Manage Analytical Models (External) page.*

Models can also be automatically created from Jupyter Notebooks using the Export to AMM button located on Jupyter's user interface top menu.

How to Edit Model Details

Follow these steps to edit the model details:

1. Access the **Manage Analytical Models (External)** page.
2. Click the **Actions** button. *A menu opens.*
3. Select **Edit Details**. *The Update Model pop-up window opens.*



The 'Update Model' pop-up window contains the following fields:



- Name ***: Text input field containing 'AMM2'.
- Description**: Text area containing 'TEST2'.
- Type ***: Dropdown menu showing 'JSON file'.
- Buttons**: 'Cancel' and 'Save' buttons at the bottom.

4. You can change the model **Name**, **Description**, or **Type**.
5. Click **Save** when finished. *The model is saved with your changes.*

How to Download a Model

Follow these steps to download a model:

1. Access the **Manage Analytical Models (External)** page.
2. Click the **Actions** button. *A menu opens.*
3. Select **Download Model**. *A download message appears at the bottom of the window and the file downloads to your local drive.*

Actions ▴	API DEV 3	1.0	XML 3
Actions ▴	11111111111111	1.0	JSON file
<div> <div>  super predictionsjson ▴ </div> <div>  AMM2.json ▴ </div> </div>			

4. Click the download arrow for options to access the file.

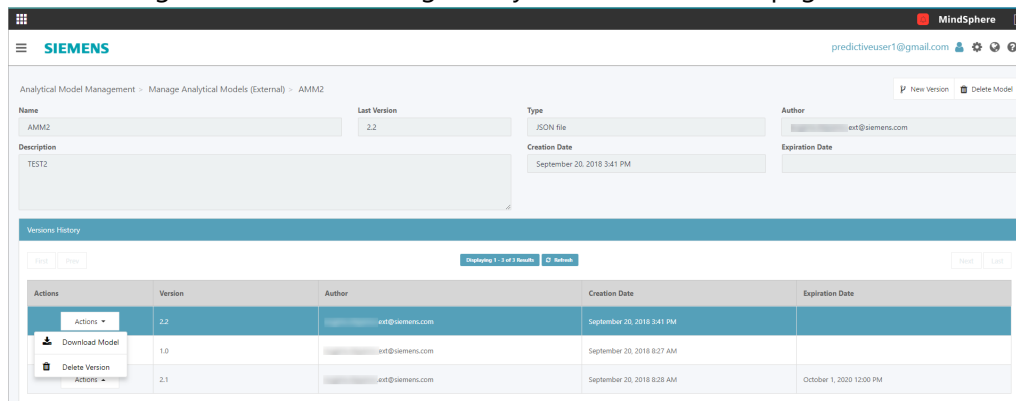
2.40 Managing Analytical Models Details

The Manage Analytical Models Details page opens in a new tab when accessed from the Manage Analytical Models (External) page. On this page, you can:

- Create a New Version
- Delete a Model
- Download a Model Version
- Delete a Version

Manage Analytical Models Details Page Illustration

The following illustrates the Manage Analytical Models Details page:



How to Open Model Details

Follow these steps to open and view model details:

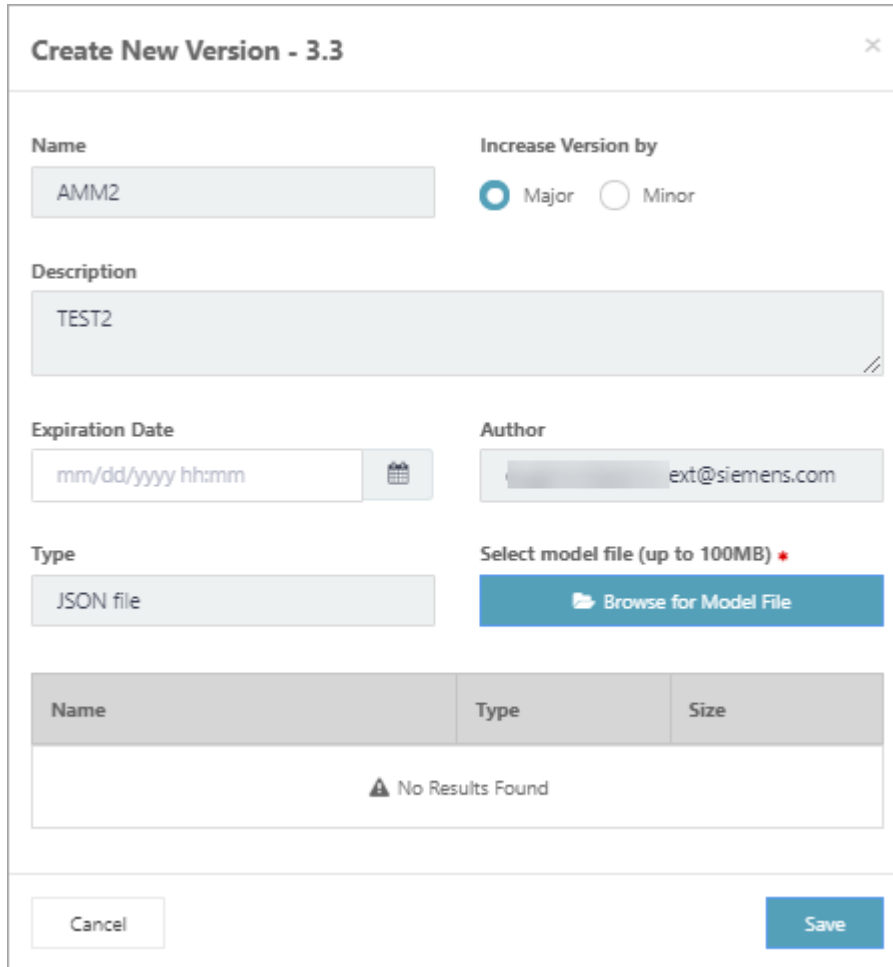
1. Access the **Manage Analytical Models (External)** page.
2. Select a model in the grid and click the **Actions** button. *A menu opens.*
3. Select **Open Model Details**. *The Details page opens in a new tab for the selected model.*

How to Create a New Version

Follow these steps to create a new model version:

1. Access the **Manage Analytical Models Details** page. *The page opens in a new tab.*

2. Click the **New Version** button. *The Create New Version pop-up window opens.*



The 'Create New Version - 3.3' pop-up window contains the following fields and controls:

- Name:** A text field containing 'AMM2'.
- Increase Version by:** Radio buttons for 'Major' (selected) and 'Minor'.
- Description:** A text area containing 'TEST2'.
- Expiration Date:** A date-time picker showing 'mm/dd/yyyy hh:mm'.
- Author:** A text field containing 'ext@siemens.com'.
- Type:** A dropdown menu showing 'JSON file'.
- Select model file (up to 100MB):** A blue button labeled 'Browse for Model File'.
- Table:** A table with columns 'Name', 'Type', and 'Size'. It currently displays 'No Results Found'.
- Buttons:** 'Cancel' and 'Save' buttons at the bottom.

3. The **Name**, **Description**, **Author** and **Type** fields are populated and cannot be changed.

4. Select **Major** or **Minor** for the version increment. Default is Major.

5. Select an **Expiration Date** if desired.

6. Click the **Browse for Model File** button and select a file to upload. Files cannot exceed 100MB. *The file name you select appears in the grid.*

7. Click **Save**. *The new version is saved.*

How to Download a Version

Follow these steps to download a model version:

1. Access the **Manage Analytical Models Details** page for a model. The page displays.

2. Click the **Actions** button. *A menu opens.*

3. Select **Download a Version**. *A download message appears at the bottom of the window and the file downloads to your local drive.*

4. Click on the download arrow for options to access the file.

How to Delete a Model

Follow these steps to delete a model:

1. Access the **Manage Analytical Models Details** page for a model. *The page displays.*
2. Click the **Delete Model** button at the top of the page. *A warning message appears.*
3. Click **OK** to proceed. *The model and all related versions are deleted.*

How to Delete a Version

Follow these steps to delete a model version:

1. Access the **Manage Analytical Models Details** page.
2. Click the **Actions** button in the row for the version you want to delete. *A menu opens*
3. Select **Delete Version**. *A warning message appears.*
4. Click **OK** to proceed. *The model version is deleted.*



You cannot delete a version if only one exists for the model.

See [Manage Analytical Models \(External\)](#) for more information.

2.41 Managing Docker Models

Docker Image Overview and Constraints

In addition to supporting Python (2, 3) and R models developed in Jupyter or Notebook, Docker is among the types of models PrL supports. Docker models have the advantage of being able to run any custom code, in any program language, and also Linux distribution preferred by users. The default operating system for all other types of models is the AWS AMI Linux distribution. There are few constraints related to the data ingestion and persistence functions in the docker image setup and, specifically the Docker image persisted in Model Management has these constraints:

- Data will be consumed from the `/data/input` folder.
- Data that is to be persisted, will be written in the `/data/output` folder.

These folders will be correctly setup for automated execution by the Job Manager service, which will retrieve the data for the job and persist it in `/data/input`, as well as data written in the

/data/input folder, and will place it into the designated persistence service, such as Data Exchange, Predictive Learning Storage, or Integrated Data Lake (IDL).

About Creating a Docker Image to Use in Predictive Learning

If you want to create your own Docker image to hold your code or model, you will require at a very minimum a Dockerfile. Usually, you 'inherit' one of the public images that provides minimal support for your code or model. Here's a short example:

```
ARG BASE_CONTAINER=python:3.9-slim-bullseye
FROM $BASE_CONTAINER
```

```
USER root
```

```
RUN ["mkdir", "/tmp/input"]
RUN ["mkdir", "/tmp/output"]
RUN chmod 777 -R /tmp
RUN ["mkdir", "/data"]
RUN ["mkdir", "/data/input"]
RUN ["mkdir", "/data/output"]
RUN chmod 777 -R /data
RUN ["mkdir", "/iot_data"]
RUN ["mkdir", "/iot_data/input"]
RUN ["mkdir", "/iot_data/output"]
RUN ["mkdir", "/iot_data/datasets"]
RUN chmod 777 -R /iot_data
RUN ["mkdir", "/prl_storage_data"]
RUN chmod 777 -R /prl_storage_data
```

```
RUN pip install awscli
RUN apt-get update
RUN apt-get install wget -y
RUN apt-get install curl -y
RUN apt-get install jq -y
```

```
COPY . .
```

```
ENTRYPOINT ["python3", "./my_python_script.py"]
```

The lines that create folders 'RUN ["mkdir", ...]' will create the proper folders for Job Manager to copy in input files, or to copy from results. If you do not pass in any inputs or outputs to your container, then, these are not needed. In addition, if you want your Docker image to contain additional libraries, you can install these here using 'RUN apt-get install ...'. These commands depend on your operating system, and they should be adapted to each. For detailed instructions on how to design your Dockerfile please check [Dockerfile reference](#).

Persisting a Docker Image in Model Management

Follow these steps to create a new Docker model:

1. Access the Manage Analytical Models Details page. The page opens in a new tab.
2. Click the New Version button. The Create New Version pop-up window opens.
3. From the Type drop-down list, select Docker Image.

This updates the dialog window, and displays these Docker-relevant controls:

- A Generate Token button
- A text field in which users must provide a complete Docker image repository and tag version.

Importing a Model

When importing an existing model, the process begins with the "Import a Model/Develop a New Model" pop-up Window:

Develop a New Model

Import a Model

Name *

Select a Name →

Description

Expiration Date *

dd/mm/yyyy hh:mm

Author

.. @siemens.com

Type *

Select model file(up to 50MB) *

↑

 Browse for Model File

Select a Model Type →

Select a File →

Name	Type	Size
⚠ No Results Found		

Cancel

Save

Follow these steps to import a model:

1. Click "Add/Develop Model" on the Landing or Models list page. The Import/Develop a Model pop-up window displays.
2. Make sure you are on the "Import a Model" tab.
3. Enter a name and description (optional).
4. Select an expiration date from the Calendar pop-up window.
5. Select a model type from the Type drop-down list, or select "Browse" to locate and select a model file.
6. Click "Save". Your imported model displays in the Models table.

Importing Docker Images

If you select the model Type to "Docker Image" the "Browse for Model File" button will be replaced with the "Generate Token" button. This is required due to way Docker images can be imported in the application. In general, Docker images are developed locally or, it can be imported from an external source.

The screenshot shows the 'Import/Validate Model' dialog box with the 'Import Model' tab selected. The dialog has a title bar with a close button (X). Below the title bar are two tabs: 'Validate Model' and 'Import Model'. The 'Import Model' tab contains the following fields and controls:

- Name ***: A text input field.
- Select a Name →**: A red text prompt below the Name field.
- Description**: A text input field.
- Expiration Date ***: A date and time picker showing 'dd/mm/yyyy hh:mm'.
- Author**: A text input field containing '@siemens.com'.
- Type ***: A dropdown menu showing 'Docker Image'.
- Generate Token for Docker**: A blue button labeled 'Generate Token'.
- Image repository URI (with tag) ***: A text input field.
- Buttons**: 'Cancel' and 'Save' buttons at the bottom right.

Clicking the "Generate Token" will provide you with a temporary session credentials that will allow you to upload the Docker image to our Docker registry. We require this in order to allow secure and high-performance on any usages of your Docker image. After you upload your

Steps for uploading a Docker Image in Model Management:

- Please save this URI, it will be required to associate the image with an analytical model:

7e324168-432f-4526-ae89-3d5e995bf397

Copy

- On your local environment tag your model using:


```
docker tag <local_image_ID> 7e324168-432f-4526-ae89-3d5e995bf397.dkr.ecr.eu-central-1.amazonaws.com/integ/modelmanagement/r...repository/7e324168-432f-4526-ae89-3d5e995bf397
```

Copy

- On your local environment use the following command to log in your docker to the image repository (password expires at 2023-07-19T15:25:04.000):


```
docker login -u AWS -p eyJyYXlsb2FkljoitXINTKUwQnhLR25TbkdwazdTMOF3NWxKNTJNVWlxMXV1TzUQUit1bUhYDZdzWDBCN3ZxekpFUGZodksvbHl0ZW8wL1F0cWkzTS9yZEt2amRsMetFL2FsdJlkYjVTMxRjdjZxcnZqWkJldjZGWI1Mk9mRzdmcUNNThaaZCV3BWBdVQM1VgZ09FMjU2dDZTRlI3ZO5PWEFlUG9VOHJEEdEtEQOURYVIU1NjlldmhSaVVvOXc3anlnSDI3SGRmMFErZlNAbnRuJWwwRT293ZHVQOQRU21aSJfGZUp6bjITOWWlWXR2THRWrmN1R1Z0cjhySXJNVHgZOHdlD0lwZjhOaW8vdUxpbeYrRGdyUKJnQWdZYkZDOTA3L3da5DcxL3ZXVmdFMm5kYSs3b3hh50ptVlFaMnh6b244TWlncmxOUUYzMjYwZS9RR2hiUjBuN3VWQmNOelQ2dzhNUlVaZytlaTkUktMbXRdV1BZYlBsVvhQbdDsbskVOZXFbzjOXFGM3RyRE96zbNbWZVjPYXd0eIJcy82Q1NKdFI1WHI2R3IFRS5SVdrOXNxUC9LU8U8rUktLaFc4M2ZaV2FFZ1d6b01uSkVvL2tlcnFQL29mZHFWbDQvemZ3b0F4WHI3bVR3cXJreWVCY1Isk1MxbWF4RXJ5RjFyb2s3SjZVJUbuZk2RWIEdjYOSXR4THEOWWcwQjJGaWlZMjNOSW11Wm13deJPUOFDbWwMEFS5EsaZ0AAAE1VLdVyaNEkhaWhtL2l2b2E7LkdGMjIDOG95dGF5bGVhZC9kZ0F5Z0FEMjUuaEF0QWw1L2Rlc3RlFmRmRkE0Q3E0
```

Copy

- On your local environment use the next command to upload your image:


```
docker push 7e324168-432f-4526-ae89-3d5e995bf397.dkr.ecr.eu-central-1.amazonaws.com/integ/modelmanagement/r...repository/7e324168-432f-4526-ae89-3d5e995bf397
```

Copy

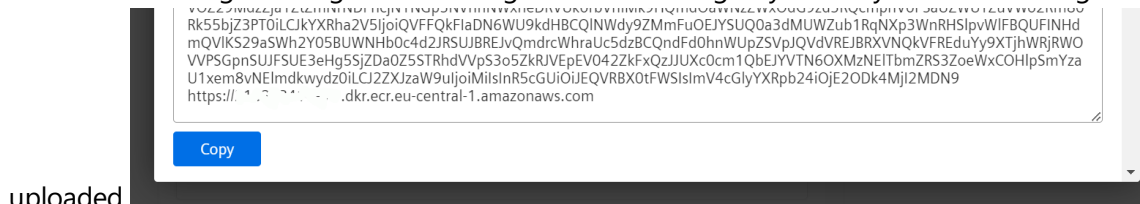
After credentials were generated, you have 24 hours to create a Docker model in Model Management (through the "New Model" dialog), where you have the use the correct repository and tag URI.

1. This is meant for reference only, our system designates an URI that will tell you where your Docker image will be uploaded. This is immutable and attempting to change it, will make our

system unaware of where you have uploaded your Docker image

2. Tag your local image with the instructions from this step. You need to replace with your local IMAGE_ID, that you find by using "docker images" command; the can be found under the "IMAGE_ID" column

3. login to our Docker registry using the command provided at this step. You can expand the textbox containing the long session string to reveal the registry where your Docker image will be



uploaded

4. after you get a successful login at the above step, you can start "pushing" (uploading) your local Docker image to our registry using the command from this step.

Now you can close the pop-up.

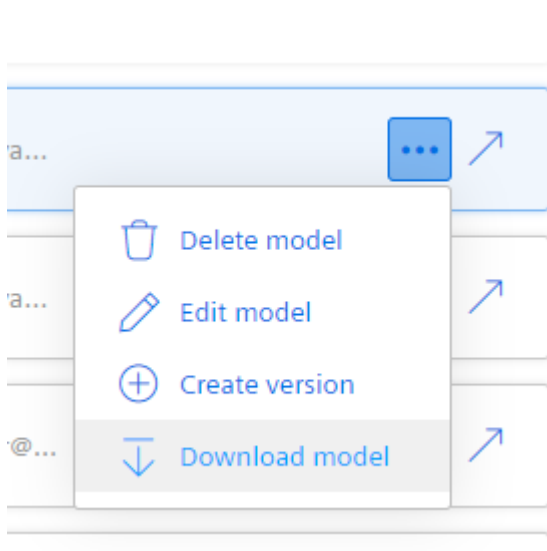
Please note that your local image might have a "tag" that is usually the string that follows after the URL and is separated by a colon, like in "URL:tag". The tag is helpful to denote versions for example, like "v1.0.1" or "final-v1.0". If your tagged image at step 2 above includes this tag, then, after closing the pop-up, you need to paste the URL stated at step 1 above, in the pop-up, in the "Image Repository URI (with tag)" field, including the tag, as in the picture below.

Make sure that you click "Save" only after pushing your Docker image has been finished.

Clicking "Save" will instruct the system to verify the Docker's image existence in our registry and its validity.

Downloading a Docker Image

You can download a previously uploaded Docker image by using similar steps as the ones above. Instead of pushing you will be able to download (pull) a Docker image once you have a valid temporary session with our Docker registry. From the Models list, click the "..." button and use the "Download model" action menu. This will not download the actual image but the access session in the form of a JSON file. From the JSON file you can depict the keys needed to login to our registry.



Using Docker CLI, you can proceed using a similar "docker login -u AWS -p " where and are provided in the downloaded file. Once you logged in, you can use "docker pull " where is also provided in the downloaded JSON file.

Provided JSON file contains two types of authentication:

1. first part for Docker compliant CLIs under the "credentials" key, that contains "user", "password"; these can be used with Docker CLI to connect to our registry
2. second part, "providerCredentials" containing "accessKey", "secret" and "sessionToken" for AWS CLI For the second option you can use the AWS CLI tools to interact with your image. It provides additional -but limited to AWS ECR- functionality than Docker CLI (e.g. docker image scanning); the list of capabilities can be explored directly from the AWS CLI once you logged in the registry.

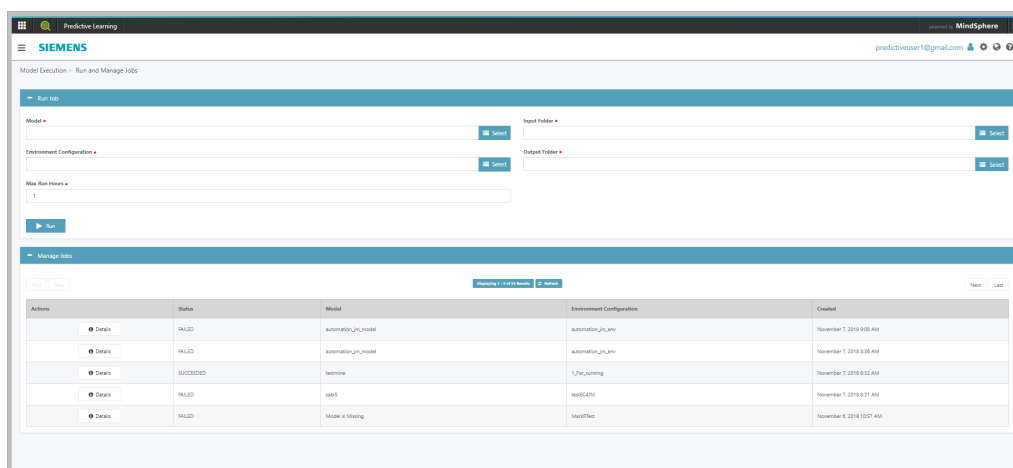
2.42 Running and Managing Jobs

Use the Run and Manage Jobs page to select a model and execute the job. On the top half of the page, you can select a model, environment configuration, and input and output folders. You can also specify the amount of time you want the job to run before terminating it to save unnecessary processing costs.

The bottom half of the page shows the jobs you have run with their status, model name, environment configuration, and date and time created. You can also view the details by clicking the Details button.

Run and Manage Jobs Page Illustration

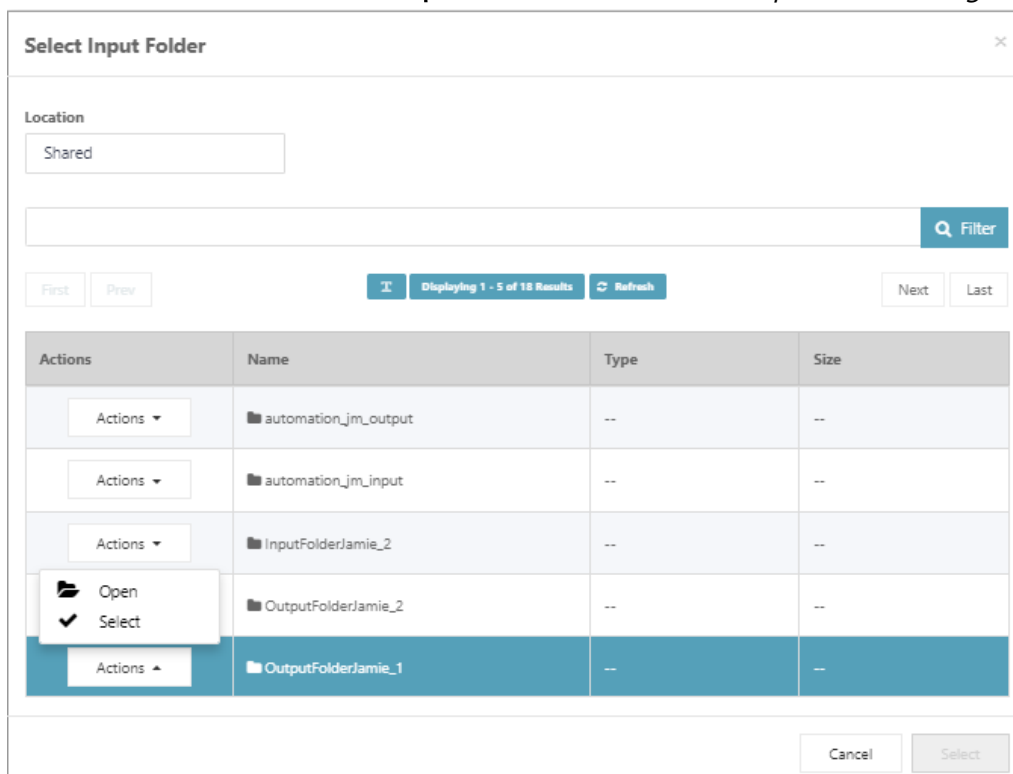
The following image shows the Run and Manage Jobs page.



How to Run a Job

Follow these steps to run a model job:

1. Click Run and Manage Jobs on the **Model Execution** menu. *The Run and Manage Jobs page opens.*
2. Select an entry from the **Model** field. Required.
3. Select an entry from the **Environment Configuration** field. Required.
4. Enter the number of hours you want the job to run before terminating in the **Max Run Hours** field. This prevents jobs that never completed successfully from using unnecessary processing time.
5. Click the **Select** button in the **Input Folder** field. *The Select Input Folder dialog box opens.*



6. Choose an entry for the **Input Folder** field from the list, click the **Actions** button for that row, and click **Select**.
7. You can also open a folder to view its contents by clicking on the **Actions** button and selecting **Open**.
8. Click the **Select** button in the **Output Folder** field. *The Select Output Folder dialog box opens.*
9. Choose an entry for the **Output Folder** field from the list, or open a folder to view its contents by clicking on the **Actions** button and selecting **Open**.
10. Click **Run**. *If the job executes successfully, a Job Created message displays.*

How Job Manager Uses Your Input Parameters

In Predictive Learning, all jobs require one of these parameters:

- Data Exchange
- Internet of Things (IoT)
- Integrated Data Lake (IDL)
- Predictive Learning Storage (PrL Storage).

For the first three parameters above, Job Manager stores the copied input in a temporary location and ensures the location is available to your code, when the code executes.

Job Manager will copy the entire folders (recursively) and data files from Data Exchange, IoT and Data Lake. For the IoT exports the actual content will be a parquet file, which consists of a subset of multiple files. When you write scripts that you plan to be executed in a job and you require the input parameters, always use the variables provided by the job execution contexts:

inputFolder, outputFolder and datasetName. You can find more details about how to use these parameters in the *Using Inputs from Job Execution* section.

Job Details Illustration

The bottom half of the Run and Manage Jobs page lists all the jobs you have run, and allows you to open and view job details. The table lists the status of each job, the model name, environment configuration used, and the date and time the job was created.

You can view more information on the Job Details page, pictured here.

Job Details [X]

Status FAILED	Max Run Hours 1
Model automation_jm_model	Environment Configuration automation_jm_env
Created November 7, 2018 9:08 AM	Created By priteam
Input Folder automation_jm_input	Output Folder automation_jm_output

Message

```
Failed to run paragraphs. [Traceback (most recent call last):
  File "/tmp/zeppelin_pyspark-3914218504365322953.py", line 367, in <module>
    raise Exception(traceback.format_exc())
Exception: Traceback (most recent call last):
  File "/tmp/zeppelin_pyspark-3914218504365322953.py", line 355, in <module>
    exec(code, _zcUserQueryNameSpace)
  File "<stdin>", line 15, in <module>]
```

Cancel

How to View Jobs and Details on the Manage Jobs Page

Follow these steps to view the details of a job on the Run and Manage page.

1. Navigate to the **Run and Manage Jobs** page. *The page opens.*
2. Click the **Details** button in the **Actions** column for the job you want to view. *The Job Details dialog box opens.*
3. View the information and click **Cancel** to close.

2.43 Running Docker Containers as Jobs

You can execute Docker images in Predictive Learning (PrL), just as you can any model stored in PrL.

About Docker Images

Please read through the following important points about pushing Docker images to Predictive Learning or when executing them:

- Do not store access data such as usernames, tokens, and secrets within Docker images, even though the storage location itself is secure.
- The Docker containers you create have limited external connectivity; for example, they can only download Python or R libraries; therefore, you should include in the Docker image all the extra data and libraries they require.

- When using Docker containers remember the EC gateway(Proxy to invoke Public API's) can only be accessed from within Predictive Learning, and not directly from your local environment.
- Exposing ports from your Docker container has no effect because they run in isolation and other components cannot access them.

Additional Code Required in Dockerfiles

Depending on the type of input your container requires, the Dockerfile used in building the container requires a few extra lines of code to allow the environment to prepare input data, and to extract outputs.

For Data Exchange or Data Lake Input

The Dockerfile for Data Exchange or Data Lake input requires the following:

```
RUN ["mkdir", "/data"]  
RUN ["mkdir", "/data/input"]  
RUN ["mkdir", "/data/output"]  
RUN chmod 777 -R /data
```

For IoT Input

The Dockerfile for IoT input requires the following:

```
RUN ["mkdir", "/iot_data"]  
RUN ["mkdir", "/iot_data/input"]  
RUN ["mkdir", "/iot_data/output"]  
RUN ["mkdir", "/iot_data/datasets"]  
RUN chmod 777 -R /iot_data
```

For PrL Storage Inputs or Outputs

The Dockerfile for PrL storage inputs and outputs requires the following:

```
RUN ["mkdir", "/prl_storage_data"]  
RUN chmod 777 -R /prl_storage_data
```

2.44 Running Scheduled Jobs

Navigate to PrL scheduling functionality by selecting Manage Jobs with Schedule from the Model Execution menu on the Predictive Learning (PrL) home page. The top half of the page is where

you create a schedule for a job and specify where PrL reads the data from and where PrL is to write the data and other details of the job. In the lower half of the page, a table displays the scheduled jobs along with details about each. The Actions drop-down list for each job allows you to start and/or stop a job, and a details link that you can click to see all of the details for a job.

About Input Types

PrL supports input from Data Exchange, IoT, and Data Lake. Here are important details for each of the input sources:

- **Data Exchange:** select an input and output folder; input folder data is read from the local path of the instance; Job Manager writes the output to the Data Exchange folder you specify.
- **IoT Data:** select an IoT data export previously defined on the Manage Sources page.
- **Data Lake:** specify an IDL path defined in the Manage Sources page.
- **Predictive Learning Storage:** specify a remote storage path remote storage path that Job Manager can use for logs, errors, and outputs; these paths can be defined in the 'Managing_Job_Sources.htm'(file title is 'Managing Sources').

About Output Types

PrL supports output to Data Lake and Predictive Learning Storage. Here are important details for each of the output destinations:

- **Data Lake:** select a path previously defined on the Manage Sources page; job manager utilizes that path as a log and stores standard console/error data there.
- **Predictive Learning Storage:** define a remote storage path that Job Manager can use to logs, errors, and temporary outputs.

The screenshot shows the 'Create Schedule' form in the Siemens MindSphere interface. The breadcrumb trail is 'Model Execution > Manage Jobs with schedule'. The form is titled 'Create Schedule' and contains the following fields and controls:

- Name ***: A text input field with the placeholder 'Schedule name'.
- Input Type ***: A dropdown menu currently showing 'IoT Data'.
- Model ***: A text input field with a 'Select' button to its right.
- Input IoT Data ***: A text input field with a 'Select' button to its right.
- Environment Configuration ***: A text input field with a 'Select' button to its right.
- Output Type ***: A dropdown menu currently showing 'Predictive Learning Storage'.
- Max Run Hours ***: A text input field with the value '1'.
- Days to run ***: A text input field with the value '7'.
- Schedule ***: A text input field with a 'Change' button to its right.
- Create**: A blue button with a play icon at the bottom left.

How to Schedule a Job Part 1

All fields in the Create Schedule feature are required. Follow these steps to schedule a job:

1. Select **Manage Jobs with Schedule** from **Model Execution** menu on the PrL home page. *The Create and Manage Schedules page displays.*
2. Enter a name for the job you are scheduling in the **Name** field.
3. Select an input type from the **Input Type** drop-down list.
4. Click the **Select** button in the **Model** field. *The Select Model pop-up window opens.*
5. Select a **model** from the list and click **Select**. *The Select Model pop-up window closes.*
6. Click the **Select** button in the **Input Location** field. *The input location pop-up window opens.*
7. Select an input location from the list and click **Select**. *The input location pop-up window closes.*

How to Schedule a Job Part 2

1. Click the **Select** button in the **Environment Configuration** field. *The Select Environment Configuration pop-up window opens.*
2. Select an output type from the **Output Type** drop-down list.
3. Enter numbers in the **Max Run Hours** and **Days to run** fields.
4. Click the **Select** button in the **Output Location** field. *The Output Location pop-up window opens.*

5. Select a location and click **Select**. *The Output Location pop-up window closes.*
6. Click the **Change** button in the **Schedule** field. *The Configure Schedule Rule pop-up window opens.*
7. Select a frequency from the **Schedule Rule** drop-down list and click **Apply**. *The Configure Schedule Rule pop-up window closes.*
8. Click **Create**. *A success message displays and the job is displayed at the top of the **Manage Schedules** table.*

2.45 Managing Sources

Predictive Learning ensures a certain level of genericity is maintained in model execution. Only sources with previously defined input and output parameters can be added, and the following source types are supported:

- IoT
- Predictive Learning Storage
- Data Lake

Adding a New Source

To add a new data source, click the **+ Add New** button as shown in the following example:

Model Execution > Manage Sources

Type

- IoT
- Predictive Learning Storage
- Data Lake

Filter

Displaying 1 - 10 of 87 Results Refresh

Actions	Source	Window (h)	Name	Created By	Created
Open	IOT	1	-iot-ds	@siemens.com	May 25, 2021 7:49 AM
Open	IOT	5	turbine	@siemens.com	May 25, 2021 7:37 AM
Open	IOT	5	iot_123	@siemens.com	May 25, 2021 7:27 AM
Open	IOT	5	pri-ui-app1	@siemens.com	May 25, 2021 6:27 AM
Open	IOT	5	jobmanagersource	@siemens.com	May 25, 2021 3:50 AM
Open	IOT	5	pri-ui	@siemens.com	May 19, 2021 7:19 AM
Open	IOT	1	test12345s	@siemens.com	March 26, 2021 8:03 AM
Open	IOT	1	iotsource-test	@siemens.com	March 26, 2021 7:52 AM

About Scheduling an IoT Data Export

If your model execution requires periodic exports and access to IoT data, the New Source window provides the fields for defining the export. The data export executes at the time you specify and you can also define the properties, and the period of time to go back in history, starting from the export execution time. When the schedule is run, the system writes the exported data to the instance that carries the job execution.

New Source ×

Name

Asset * Select **Property Set *** Select

Last
 ▼

Name *

Description

About Integrated Data Lake (IDL) Source Paths

For exports that require direct access to a specific path in the data lake, the New Source window provides the fields for defining that path, but the path must already exist in the data lake. If the path doesn't already exist, you must log in to the Integrated Data Lake and create it; there's no way to create the path from outside IDL. When the IDL path is used as an output parameter in scheduling a job, the system uploads the resulting data files to the path you set in the New Source dialog box.

New Source ×

Name

Storage Path

Name *

Description

About Predictive Learning Storage Sources

Predictive Learning storage (PrL storage) functions similarly to IDL, except that, with PrL storage, you can define the storage path from within the New Source pop-up window, and the system copies the output data to the path you define.

PrL storage is available for all jobs, so there is no need to define it specifically as an input source, as it is always active. Users are free to copy data from PrL storage as needed.

New Source

Name

PrL Storage Source

Storage Account

prl-storage-

Storage Path

/

/data/

sensor1/input

Name *

Schedule Job PrL Storage Source

Description

Cancel

Save

2.46 Predictive Learning (PrL) API

This storage API is provided for users who require the ability to upload and download large files and have the files available to work with Predictive Learning. The PrL API is a simple remote storage repository that allows all operations using a temporary token and session.

Access to the API

Access to the storage path and API is provided for PrL clusters, environments, and jobs. AWS and Azure CLI can interact with the PrL API using these commands:

Access

```
aws s3 ls s3://prl-storage-<digits>/<tenant>/data/mypath/
```

Uploading files

```
aws s3 cp /tmp/myfile.txt s3://prl-storage-<digits>/data/<tenant>/data/mypath/myfile.txt
```

Downloading files

```
aws s3 cp s3://prl-storage-<digits>/<tenant>/data/mypath/myfile.txt /tmp/myfile.txt
```

External Access to the API

To access the storage API from outside Predictive Learning, you must obtain a temporary token from the PrL API and use the provided credentials, through either a credentials file store, or through environment variables.

Use a POST request to obtain a token from the API:

```
curl --location --request POST $GATEWAY_ENDPOINT$/prlstorage/v3/generateAccessToken' --header 'Content-Type: application/json'
```

Which results in:

```
{
  "credentials": {
    "accessKeyId": "<access_key>",
    "secretAccessKey": "<secret>",
    "sessionToken": "<session>",
    "expirationTime": "2019-11-19T14:33:21.000"
  },
  "storageAccount": "prl-storage-<account>",
  "storagePath": "/<mytenant>/data/"
}
```

Alternatively, you can extract the needed keys if you store the request result in a bash variable that can be parsed with the following commands:

```
%%bash
content=$(curl --location --request POST $GATEWAY_ENDPOINT$/prlstorage/v3/generateAccessToken' --header 'Content-Type: application/json')
secret=$(jq -r '.credentials.secretAccessKey' <<< "${content}")
session=$(jq -r '.credentials.sessionToken' <<< "${content}")
accesskey=$(jq -r '.credentials.accessKeyId' <<< "${content}")
export AWS_ACCESS_KEY_ID=$(echo "${accesskey}")
export AWS_SECRET_ACCESS_KEY=$(echo "${secret}")
export AWS_SESSION_TOKEN=$(echo "${session})")
```

When PrL storage is used as an input or output source for a job, the root tenant path is available to the Docker container being executed. Both input and output is mounted in the container's

/data/input and **/data/output** paths.

Open Source Software

3

3.1 Open Source Software

Features of Predictive Learning In accordance with to the [GNU Lesser General Public License](#), we are making the source code of the following open source resources available:

Library Name	Version with source code
awscli	1.18.14
boto	2.49.0
docutils	0.15.2
scikit-image	0.13.1
scikit-learn	0.19.1
python36-sagemaker-pyspark	1.2.1
PyYAML	5.3.1
requests	2.24.0