BI Office

R & Data Mining Guide
Version 6.5

A. Overview 3
B. Installing R 3
C. Enabling R 3
i. Enable R Environment 3
ii. Local/Remote Host 3
iii. Timeout 3
iv. Use shared folder 3
v. Libraries Management 3
D. General Usage of R capabilities 3
E. Forecasting 4
i. Overview 4
ii. Basic Usage 4
iii. Wizard 4
iv. Manage Forecasting 4
F. Clustering 4
i. Overview 4
ii. Wizard 5
iii. Manage Clustering 5
G. Prediction 5
i. Overview 5
ii. Wizard 5
iii. Manage Predictions 6
iv. Manage Predictive Models 6

A. Overview

The R engine in BI Office is designed to allow end users to perform data mining operations from inside BI Office against their data models. In the current version, this includes running forecasts on a time series, clustering (grouping) of items and predictions of any selected attribute (measure or hierarchy).

The R packages and configuration are accessible from the administrative console.

The following guide outlines any necessary details for the R data mining engine.

B. Installing R

- The BI Office installation is installing a vanilla version of R Environment on the application server that can be used as a local host environment.
- You can choose to install Revolution R Open or Microsoft R Open on the application server and use them instead. The R version to be used is set in the registry, by the R Environment being installed (HKLM\SOFTWARE\R-core).
- You can install any R Environment on a remote Windows or Linux server, and set BI Office to connect to it, as explained in the "Enabled R" section.
    - You need to install and activate **RServe** package to connect to a remote server.
    - If using a remote R server, be sure to check access between the R server and the application server.

C. Enabling R

In the administrative console, under settings, there is an R Environment configuration.

i. Enable R Environment

- This must be turned on in order to allow access to data mining tools for the end users.

ii. Local/Remote Host

- Determines where the R Engine is installed. By default, an R engine is installed on the application server (local host) and can be used after the setup.
- A remote host could be used, by setting its IP and PORT numbers, and running the **RServe** package on the remote machine.

iii. Timeout

- The amount of seconds for a mining operation to timeout.

iv. Use shared folder

- This enables the mining operation to use the "File Upload" folder to transport data throughout the process.
- If a remote host with a Linux OS is being used, this should be unchecked, due to Windows to Linux compatibility issues.

v. Libraries Management

- Default Download URL: the default repository that the libraries (packages) should be installed from.
- The list of libraries
    - Library name
    - Download URL: an alternative repository to download from
    - Default URL: should use the default or the specific repository
    - Type: on what operation should the library be loaded ("general" is loaded for all operations)

D. General Usage of R capabilities

The deferent R capabilities are used to study and produce conclusions on the existing data models, and add new data to these models.

The R Engine uses R scripts to do so, and these scripts will be auto-generated for the end users, then allowing them to change the scripts as the wish. The scripts are built through a wizard, in which the end users can choose what they want to do, while the script is changed accordingly.

The user's wizard-design and script can be saved for reuse and sharing, or just run once without saving them, but still creating the result output.

E.Forecasting

i. Overview

The forecasting is accessible through the Analytics menu in the Data Discovery client tool. This tool will forecast future values of a time series, such as Sales over the next 6 months, according to the current displayed data. The forecast looks at the row data as a time line and forecasts values for the columns.

- The output is new calculated values at the end of the rows of the data. The output is not saved, but calculated and dynamically each query run.
- The forecasting is attached to the report's query until the end user removes it.

ii. Basic Usage

There are 6 forecasting algorithms built in to the system, which can be accessed quickly through the forecast split button

- The current attached forecast can be removed using "Remove Forecast" under the forecasting split button.
- Note that Standard and Seasonal forecasts cannot be run on time periods of years (only quarters, months and days).

iii. Wizard

The forecast engine can be configured for more sophisticated options using Advanced Forecasting, under the forecasting split button.

Algorithm

The algorithm can be set to any of the built in options, or any user saved ones.

- The type of periods of data of the time series can be set (Years, Months, etc.)
- The number of future periods of data can be set: how many values into the future will be forecasted.

R Script

The script is built dynamically, to use the user's algorithm configurations. The user can also choose to customize the script as he wishes.

- The script will be running as written, after an 'input' vector will be initialized with the existing time series values.
- After the script is run, a vector 'mean' will be assumed to contain the forecasted values.

Final

A name (caption) and description can be set for the customized forecast design, and it can be saved for reuse.

- The saved forecast will be accessible in other reports of the same data source.

iv. Manage Forecasting

The saved forecasting designs can be rerun, edited and deleted through this option under the forecasting split button.

F.Clustering

i. Overview

The clustering is accessible through the Analytics menu in the Data Discovery client.

This tool is used to divide items into clusters (groups) according to their attributes. For example, dividing products into 3 groups according to color and sub-category.

- The outputs are:
    - Custom sets, each contains the items of the group calculated by the R Engine.
    - Custom members, each aggregated from the items for a matching custom set.

ii. Wizard

The clustering can be configured using Advanced Clustering, under the clustering split button.

Algorithm

The algorithm can be set to any of the built in options, or any user saved ones.

- Note that Kmeans algorithm can only process numbers, so no categories to group by can be selected while using it.
- The number of clusters (sets) can be set be the user.

Items to Group

The hierarchy the will be divided into sets. The values of this hierarchy will be contained in the result sets.

Categories/Numbers to Group by

Hierarchies and measures associated with the "Items to Group", which will be used as attributes by the algorithm to determine which item belongs in which set.

- There must be at least 1 attribute to group by. Mclust and Kmodes algorithms need at least 2.

R Script

The R Script that will be run can be edited by the user, to allow any clustering algorithm to be used.

- The script will be running as written, after an 'input' data-frame (table of values) will be initialized with the selected items and attributes.
- After the script is ran, a vector 'output' will be assumed to contain the numbers of clusters for each item. The first value is the number of the first item and so on.

Final

A name (caption) and description can be set for the clustering design, and it can be saved for reuse.

- The name will be set to be the name of the result custom sets, followed by the number of the cluster (My Cluster #1, My Cluster #2).
- The saved clustering design will be accessible in other reports of the same data source.

iii. Manage Clustering

The saved clustering designs can be rerun, edited and deleted through this option under the clustering split button.

G. Prediction

i. Overview

The prediction is accessible through the Analytics menu in the Data Discovery.

NOTE: The prediction can only be used on Tabular data models created BI Office, on which the end user has permissions to process them.

This tool will create a new hierarchy or measure, with values calculated by the R Engine.

- At first a predictive model is created by studying existing data. The predictive model is a black box calculator:
  - It studies an attribute X of the input items, and learns how to calculate it according to attributes A, B, C of the items. For example, learning Sales of transactions, according to price and discount.
- Then the predictive model is used to calculate the X attribute for the existing data it was created from, allowing to compare the result prediction with the original values.
- If the predictive model is saved, it can be used to predict values of type X on items with attributes A, B, C from other data models, as described below.

ii. Wizard

The prediction can be configured using Advanced Prediction, under the prediction split button.

Algorithm

The algorithm can be set to any of the built in options, or any user saved designs or predictive models.

- If a "Categorical Items" output is selected, hierarchy values will be processed and a new hierarchy will be created with the result values. The new hierarchy will be associated with the dimension of the original hierarchy.

- If a “Numeric Values” output is selected, measure values will be processed and a new measure will be created with the result values.
- If a saved design is selected, the full process will be running: creating the predictive model and then determining the predicted values.
- If a saved predictive model is selected, the first step will be skipped, using the saved model to predict the new values.
- The first part, creating the predictive model, can be set to use only a sub-set of the data. After, all the items will be used to predict each of their values.

What to Predict

The hierarchy or measure that the prediction is set on. If the prediction should calculate sales, then sales measure should be selected.

Items to Predict for

- The key (unique value identifier) hierarchy of the dimension that the predicted values are associated with. If the prediction should calculate sales, then transactions’ row-id should be selected.
- This key will be used to attach the results back to the dimension, as a new hierarchy or measure.

Categories/Numbers to Predict by

Hierarchies and measures associated with the “Items to predict for”, which will be used as attributes used by the algorithm to learn how the predicted values should be calculated.

- There must be at least 1 attribute to predict by.

R Script

The R Script that will be run can be edited by the user, to allow any prediction algorithm to be used. The script will be running as written, after an ‘input’ data-frame (table of values) will be initialized with the selected items and attributes.

- In case an existing predictive model is selected to run the prediction, then the ‘fit’ object is loaded with it before the script starts.
- After the script is ran, a vector ‘output’ will be assumed to contain the predicted value for each item. The first value is of the first item and so on.
- If the prediction output type is Categorical, then the results should be set from the deferent possible values (‘Yes’, ‘No’ if the prediction is whether a customer will porches).
- If the prediction output type is Numeric, then the results should be set to the relevant numbers (sales in dollars).
- If the script is calculation the predictive model, and not using an existing one, then the ‘fit’ object should hold it in the end of the run, so it could be saved for later use.

Final

A name (caption) and description can be set for the prediction design and predictive model, and it can be saved for reuse.

- The name will be set to be the name of the result hierarchy or measure.
- The saved prediction design will be accessible in other reports of the same data source.
- The saved predictive model will be accessible in other reports of any data source that has the needed hierarchies and measures:
    - Items to Predict for
    - Categories/Numbers to Predict by

Note that the “What to Predict” is not needed to use the predictive model. A model used to calculate ‘sales’ of ‘transactions’, using ‘price’ and ‘discount’ could be used by a data source that has ‘transactions’, ‘price’ and ‘discount’, but no ‘sales’, and it will create a new measure of ‘predicted sales’.

iii. Manage Predictions

The saved prediction designs can be rerun, edited and deleted through this option under the predictions split button.

iv. Manage Predictive Models

The saved predictive models can be rerun and deleted through this option under the predictions split button.