Project Columns

Updated: August 1, 2015

Selects columns to include or exclude from a dataset in an operation

Category: Data Transformation / Manipulation (https://msdn.microsoft.com/en-us/library/azure/dn905863.aspx)

Module Overview

You can use the **Project Columns** module to choose a subset of columns to use in downstream operations. This can be useful if you want to use only a few columns for a specific operation or if you want to generally reduce the size of the dataset.

For example, **Project Columns** is the tool to use in these scenarios:

- You need to "block off" text columns in the dataset so you can apply a math operation to all numeric columns.
- When applying feature selection, you want to pass through only a subset of the categorical feature columns.
- You have many numerical columns to normalize and want to apply a different normalization operation to different columns, so you create a data projection for each group of columns.

Note

This module does not reorder output columns based on the order of the column list value selected in the input.

For example, if you specify columns by using an order such as (4,3,2,1), the projected columns will appear in the output according to the original dataset order (1,2,3,4).

Examples

For examples of how to use **Project Columns**, see these sample experiments in the Model Gallery (http://gallery.azureml.net/):

- The Breast cancer detection (http://go.microsoft.com/fwlink/?LinkId=525726) sample uses **Project Columns** to remove a trailing empty column, remove a column with duplicate data, and to project training and test sets.

- In the Flight delay prediction (http://go.microsoft.com/fwlink/?LinkId=525725) sample, **Project Columns** is used to exclude all string columns and to exclude columns by name.

- In the Prediction of student performance (http://go.microsoft.com/fwlink/?LinkId=525727) sample, **Project Columns** is used to get all temporal features, and to exclude multiple columns.

- In the Compare Regressors (http://go.microsoft.com/fwlink/?LinkId=525731) sample, **Project Columns** is used to exclude the column, **num-of-doors**, because it is the wrong data type for the math operation that follows.

Technical Notes

In relational algebra, a *projection* is a unary operation, which is written as a set of attribute names. The result of a projection is the set of those attributes, with other attributes discarded.

In database terms, a *projection* is a function (such as a Transact-SQL or LINQ statement) that takes a table as input and produces a related output. Technically speaking a view is not always a projection, because a view that is a projection can use data from only one table, and selection conditions (WHERE statements) cannot be applied.

Expected Input

Name	Type	Description
Dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Input dataset

Module Parameter

Name	Range	Type	Default	Description
Select columns	any	ColumnSelection		Select columns to keep in the projected dataset.

Output

Name	Type	Description
Results dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Output dataset

Exceptions

For a list of all exceptions, see Machine Learning Module Error Codes (https://msdn.microsoft.com/en-us/library/azure/dn905910.aspx).

Exception	Description
Error 0001 (https://msdn.microsoft.com/en-us/library/azure/dn905993.aspx)	An exception occurs if one or more specified columns of the dataset couldn't be found.
Error 0003 (https://msdn.microsoft.com/en-us/library/azure/dn906003.aspx)	An exception occurs if one or more input datasets are null or empty.

See Also

Data Transformation / Manipulation (https://msdn.microsoft.com/en-us/library/azure/dn905863.aspx)
A-Z List of Machine Learning Studio Modules (https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx)