Reference Guide

alteryx

Alteryx and Microsoft R Integration

v 1.4, February 2017

Overview

This document describes the integration of Microsoft R ScaleR technology and proprietary functions for scaling predictive analytics with Alteryx.

The integration is implemented through the use of an XDF (.xdf) file, which allows a number of Alteryx predictive modeling tools (Boosted Model, Count Regression, Decision Tree, Forest Model, Gamma Regression, Lift Chart, Linear Regression, Logistic Regression, Score, and Stepwise) to make use of Microsoft R ScaleR functions.

ScaleR functions are available in both Microsoft R Client, a free version of Microsoft R, and Microsoft R Server, an enterprise class server version of Microsoft R. With Microsoft R Client, the data to be processed must fit in local memory, and processing is limited up to two threads for ScaleR functions. Microsoft R Server adds support for parallel and chunked data processing and the data does not have to fit in local memory with the ScaleR functions. See https://msdn.microsoft.com/en-us/microsoft-r/index for more information.

Integrating Microsoft R with Alteryx Designer

Install Microsoft R before installing the Alteryx Designer Predictive tools.

1. Install the current version of Microsoft R.
2. Go to http://downloads.alteryx.com/predictive.html and install **Alteryx Predictive Tools for Microsoft R** for the version of Microsoft R that has been installed on the machine.
3. Verify that Microsoft R was installed correctly by verifying the **XDF Input** tool and the **XDF Output** tool are available in the In/Out tool category in Alteryx Designer.

Using the XDF tools in Alteryx Designer

Alteryx uses the XDF Input tool and the XDF Output tool to read and write .xdf files in Alteryx.

The XDF Output tool takes an Alteryx data stream and writes it to an .xdf file either in Alteryx's temporary directory or to a user specified permanent location on disk. In addition to writing the .xdf file, an XDF metadata stream is also produced. The metadata stream provides downstream predictive tools with information about the underlying metadata describing the data, along with information that enables a predictive tool to determine the location of the relevant .xdf file.

When the input into a predictive modeling tool is the metadata stream from an XDF Output tool, the predictive modeling tool identifies the input as being an .xdf file and uses the appropriate ScaleR modeling function. If the input into an Alteryx predictive modeling tool is a standard Alteryx data stream, as opposed to an XDF metadata stream, then the appropriate open source R function is used. The configuration of the predictive modeling tool is the same in both cases, and it is the type of input to the tool that determines the use of the ScaleR or open source R modeling function.

An XDF metadata stream consists of the number of fields the user has selected to be included in the .xdf file in the XDF Output tool and two data records. This small amount of data is enough to properly populate the user interface of downstream predictive modeling tools. Additional metadata is conveyed via a JSON string that contains information about the compute context (with a keyword of “Context” and a value of “XDF” to work with a standalone .xdf file) and the path to the .xdf file to be used in the analysis (with a keyword of “File.Loc” and a string of the full file path to the .xdf file as the value) which is contained in the “Source” field of the metadata. The JSON string is repeated for each field to ensure that it is available to a downstream modeling tool regardless of what fields are included in the model. The metadata is read by a modeling tool, allowing for model estimation.

Saving the .xdf file to disk is useful if the user plans on creating several Alteryx workflows using the same .xdf file. In subsequent workflows that use the .xdf file, the XDF Input tool is used to read needed information into Alteryx. If the XDF data will only be used with predictive modeling tools (so no further data cleansing or blending will be done by the user), then only the XDF metadata stream for the .xdf file is read into Alteryx, saving considerable time. Otherwise, the full .xdf file is read into Alteryx. Whether only the metadata stream or the full .xdf file is read into Alteryx is specified by the user via an option in the XDF Input tool. The tool makes use of Alteryx’s capability of reading data into R in chunks, and the user can specify the number of rows to read in each chunk. The tool defaults to 256,000 records, which was selected to make use of the number of rows in an Alteryx block (64,000 rows) and the optimal number of rows to use with ScaleR functions (between 200,000 and 300,000).

Scoring with XDF data uses the same mechanism as scoring in the traditional open source R Alteryx environment, with the exception that if the data to be scored is in an .xdf file, the scored values will be appended to this file. Otherwise, the ScaleR models will append predicted values to the Alteryx data stream used for scoring. What enables scoring to scale is Alteryx’s ability to read and write data between R and Alteryx one chunk at a time, regardless of whether or not the model used in scoring was created using ScaleR or open source R tools. Alteryx has successfully scored over 160MM records using only open source R tools.