

www.0xdata.com

H2O on Excel

May 26, 2014

# H2O on Excel

# Introduction

H2O is a powerful open source math platform developed by 0xdata. The goal of the machine learning engine is to bring distributed and parallelized algorithms into the workflow of engineers and data scientists without the growing pains. H2O is offered as an easy to integrate modular component to the infrastructure already in place. It is accessible by widely used languages like R and JSON via REST API and can ingest data from HDFS, S3, SQL, NoSQL, or from a local disk. When working with excel the HTTP call to H2O will elicit a response in XML format that Microsoft excel can parse in both 32-bit and 64-bit version (does not have a JSON reader).

This whitepaper is made available for excel users that is already familiar with excel features and functions. A cross between the H2O web interface and R’s manipulative API, excel will have point and click features as well as R like functions available for calculation in the cells. Excel is the frontend client with functions written in VBA so a macro enabled Microsoft Excel is necessary to run the worksheet.

# Perquisites

* 64-bit Java 1.6 +

# Overview

* Get and install R
* Download [H2O](http://0xdata.com/downloadtable/)
* Have Microsoft Excel

# Launch H2O and/or Initialize Connection

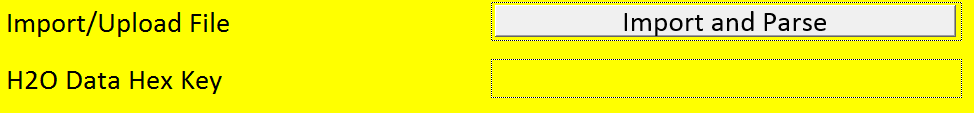
* If no instance of H2O is launched yet:
  1. Input IP address and port you wish to launch at.
  2. Choose the number of instances and the size of each instance.
  3. Choose the version of h2o you’ll like to use, by default excel chooses master build in your GIT target repository.
  4. Hit “Launch Instance.” A command prompt will load for each instance you indicate you want, wait for message “Cloud of size <instance number> formed” before closing pop up window “Launching H2O – Wait for all nodes to launch before closing.”
* If H2O has already been launched, either previously via command prompt, on R, or on a Hadoop cluster:
  1. Input IP address and port you wish to connect to, instance number and heap size can be bypassed.
  2. Hit “Launch Instance,” to connect.

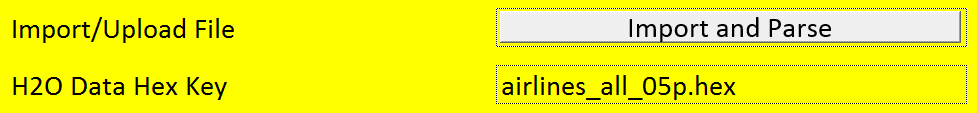
# Import Data and Load Summary

1. Either find your file using “Data File” prompt or input the path to the data file in entry “File Path.”
2. Hit “Import and Parse,” and you can check the status of the parse down by the lower left hand corner:

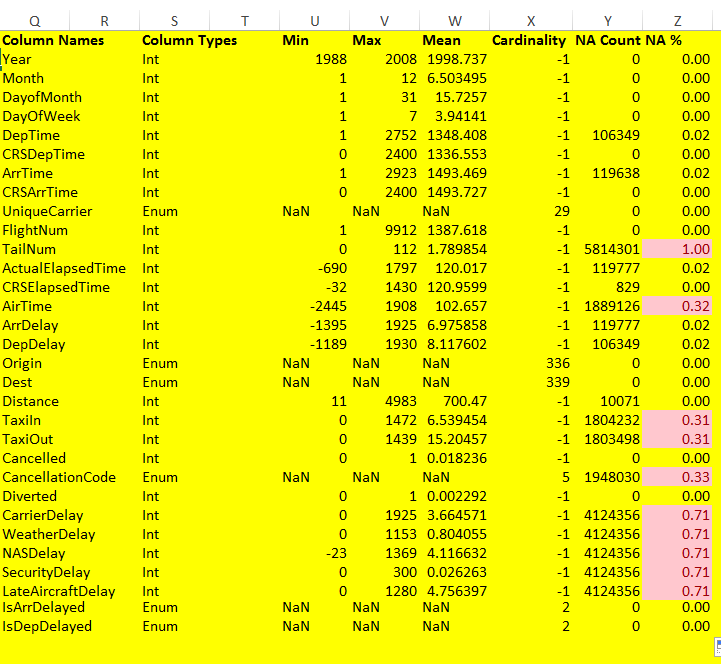


1. The entry “H2O Data Hex Key” should automatically fill in with the destination key of the hex file now sitting in H2O.



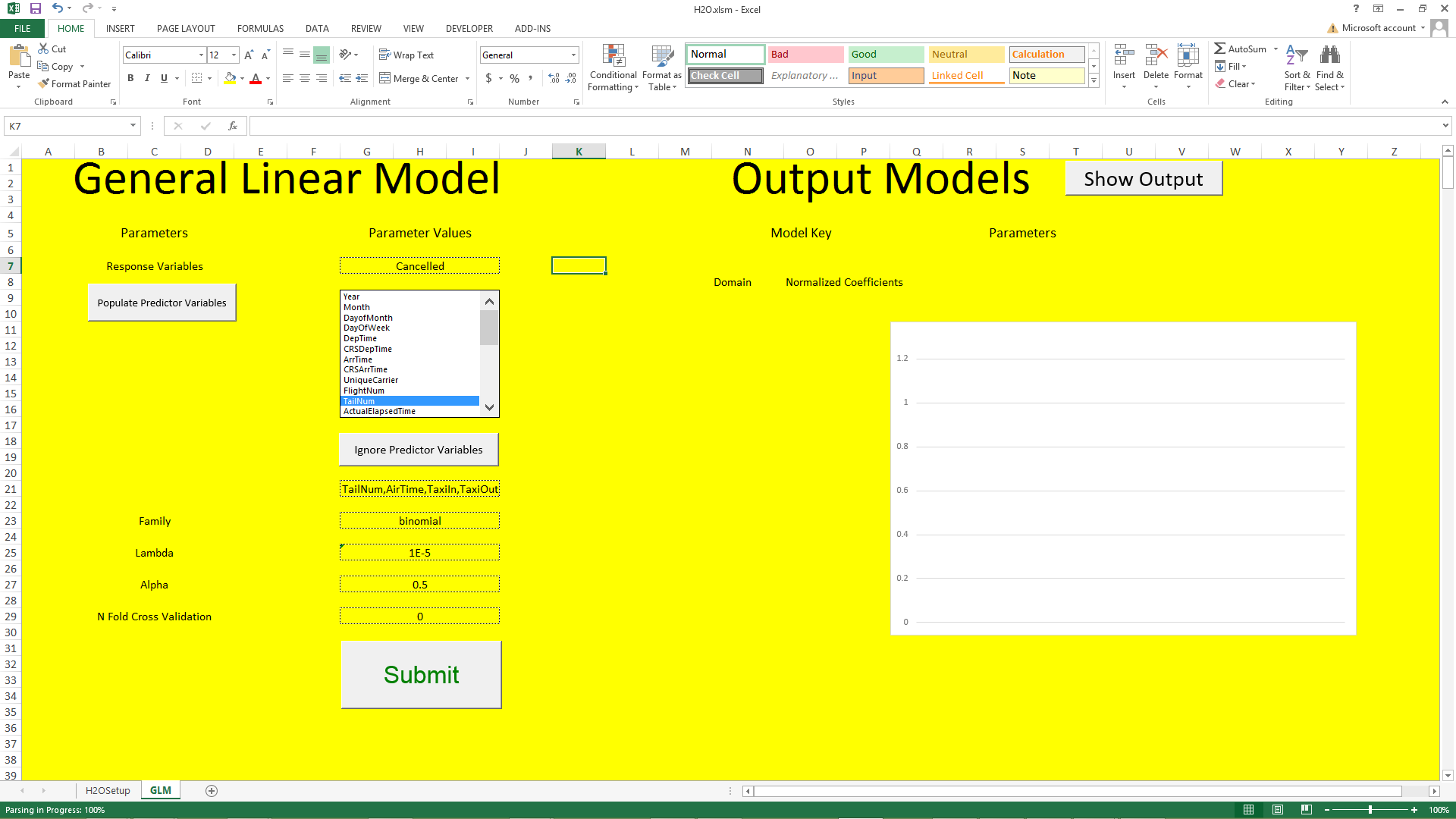


1. After which “Generate Summary” will list out all the columns in the dataset and some general statistics. All columns except “NA%” are taken directly from H2O’s inspect page, only “NA%” is calculated using the “NA Count” column and dividing it by the number of rows in the data set which can be calculated using function =/h2onumrow(h2oHexKey) :



Note: The red highlighted values indicate large number of NA values in the column and suggest ignoring such columns.

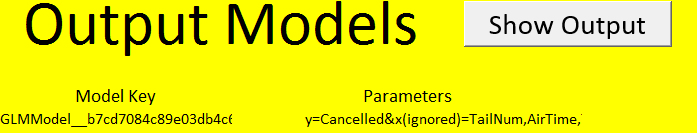
# Build GLM Model



1. Response Variable should auto-populate, choose the response from the drop down. Then hit “Populate Predictor Variables,” which will eliminate the response variable from the Predictor Variables as well as highlight the variables with high NA counts to ignore.
2. Select any other variables you’ll like to ignore before hitting “Ignore Predictor Variables.”
3. Fill in other information such as the family of distribution, lambda, alpha, and the number of cross validation you want to run. Once satisfied submit your parameters and wait for the GLM to finish running:



1. The Model Key and Parameters should auto populate with the destination key of the model and the parameters that went into building the model respectively.



1. Choose to “Show Output” to show the coefficients from a GLM model, from here the values are all numeric and strings that user can manipulate to create different visuals, the default is a simple bar graph:

