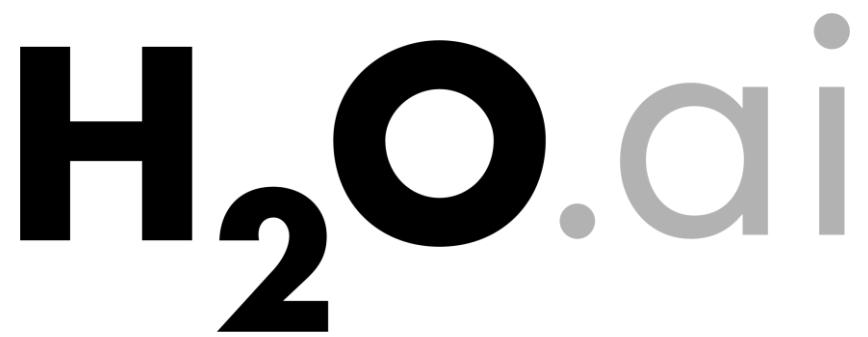


Machine Learning with H₂O



Jo-fai (Joe) Chow

Data Scientist

joe@h2o.ai

@matlabulous

R Addicts Paris Meetup
1st December, 2016

Agenda

- Introduction to H₂O
 - About H₂O.ai
- Our Open Source Products
 - Overview
 - H₂O Platform
 - Steam
 - Live Demo
 - H2O + R + Web + Steam



H₂O.ai

About Me

- Civil (Water) Engineer
 - 2010 – 2015
 - Consultant (UK)
 - Utilities
 - Asset Management
 - Constrained Optimization
 - Industrial PhD (UK)
 - Infrastructure Design Optimization
 - Machine Learning + Water Engineering
 - Discovered H2O in 2014
- Data Scientist
 - From 2015
 - Virgin Media (UK)
 - Domino Data Lab (Silicon Valley)
 - H₂O.ai (Silicon Valley)

I ❤️ R



Jo-fai Chow

woobe

Civil Engineer turned Data Scientist

 H2O.ai

 United Kingdom

 jofai.chow@gmail.com

 <http://www.jofaichow.co.uk/>

 Joined on Aug 10, 2012

Organizations



Overview

Repositories 41

Stars 372

Followers 119

Following 30

Popular repositories

blenditbayes

Code used in my blog "Blend it like a Bayesian!"

 R  73  81

deapr

An R package to streamline the training, fine-tuning and predicting processes for deep learning based on 'darch' and 'deepnet'.

 R  40  14

rPlotter

Wrapper functions that make plotting in R a lot easier for beginners.

 R  29  4

rCrimemap

This is the next generation of CrimeMap!

 R  22  9

rugsmaps

This app is my submission to the visualization contest held by Revolution Analytics.

 R  19  18

rApps

Repository for my R (Shiny) web applications.

 R  16  36



Crime Data Visualisation

INTRODUCTION
This ShinyApp allows you to download and visualise crime data in England, Wales & Northern Islands from data.police.uk. The data is made available under the Open Government License. For more information, see my original blog post.

USAGE
Simply enter a location of your choice (e.g. Oxford), choose the first month for data collection (e.g. Jan 2012), decide how many months of data you need and then click "update". There are some more settings available for you to customise the plots. Scroll down and try them out!

READY?
Continue to scroll down and modify the settings. Come back and click this when you are ready to render new plots.
[Update Graphics and Tables](#)

BASIC SETTINGS
Enter a location of interest:

Examples: London, Wembley Stadium, M16 GRA etc.
First Month of Data Collection:

Length of Analysis (Months):

Note: Data is available from Dec 2010 to Sep 2013. There is inconsistency in 2010-2011 records so I have omitted them for now. It takes longer to render the plots when you increase this number.

MAP SETTINGS
Choose Facet Type:
 none
 choropleth
Choose Google Map Type:
 roadmap
 satellite
 High Resolution?
 Black & White?
Zoom Level (Recommended - 14):

DENSITY PLOT SETTINGS
Alpha Range:



My First Data Viz & Shiny App Experience
[CrimeMap \(2013\)](#)

Revolutions
Daily news about using open source R for big data analysis, predictive modeling, data science, and visualization since 2008

[« How to integrate R with your calendar](#) | [Main](#) | [Entering the field as a data scientist with certification »](#)

August 21, 2014

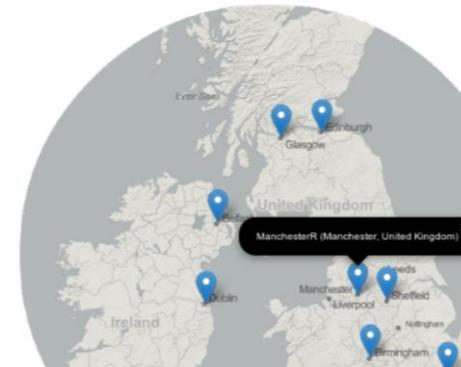
Revolution Analytics' User Group Map Contest has a Winner

by Joseph Rickert

We are pleased to announce that [Jo-fai Chow](#) is the winner of the Revolution Analytics contest. Jo-fai's entry, which was implemented as a [Shiny project](#), may be viewed by clicking on the figure below.

R User Groups Around the World

[About](#) [Maps](#) [Data](#) [More](#) ▾



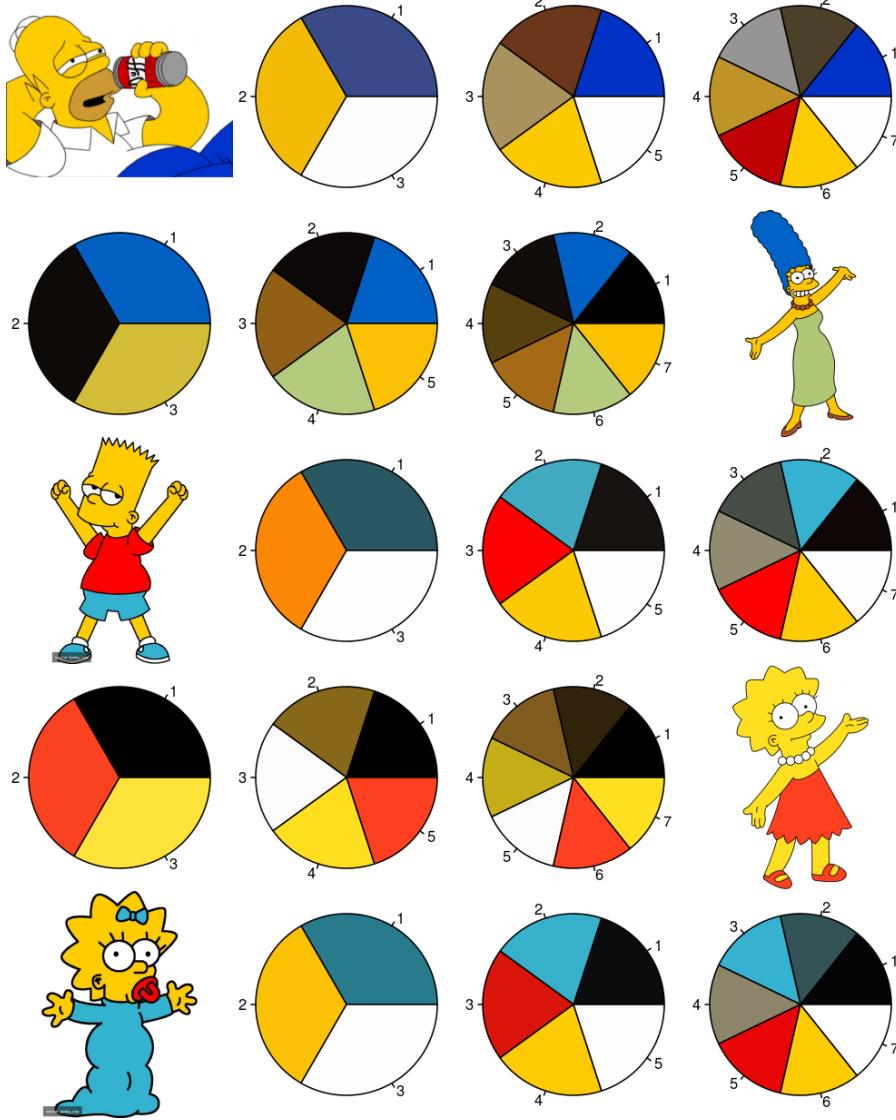
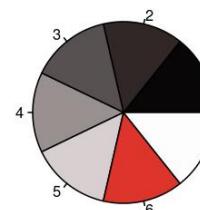
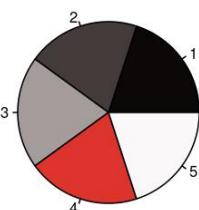
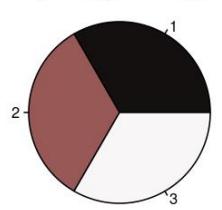
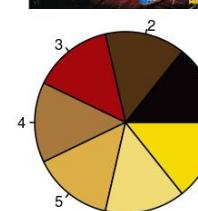
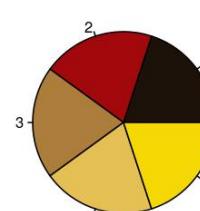
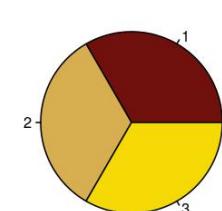
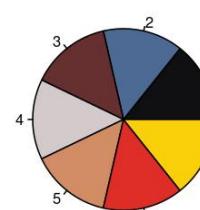
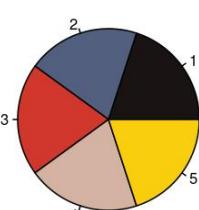
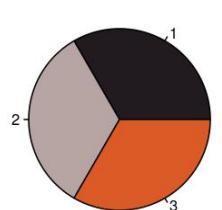
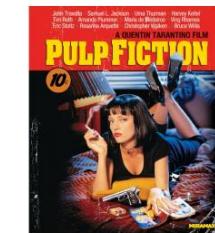
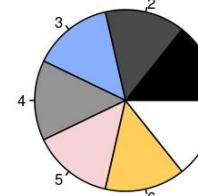
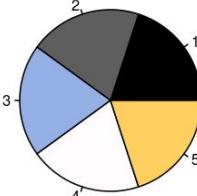
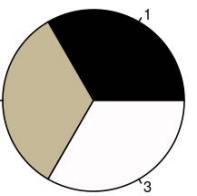
 **Jo-fai (Joe) Chow**
@matlabulous

Thank you very much @RevolutionR
@revodavid @RevoJoe #iloveR
bit.ly/rugsmaps #Shiny #rMaps



Revolution Analytics' Data Viz Contest
[RUGSMAPS \(2014\)](#)

I ❤ R



Developing R Packages for Fun
[rPlotter](#) (2014)



The screenshot shows a blog post on the Domino Data Lab website. The left sidebar features the Domino Data Lab logo with the tagline "At the intersection of data science and engineering." It includes links to the "Domino App Site" and social media icons for Twitter and Email. The main content area has a dark header with the date "19 Sep 2014". Below the date are social sharing buttons for Facebook (Like 0), Twitter (Tweet 21), and Google+ (g+1 4). The title of the post is "How to use R, H2O, and Domino for a Kaggle competition". A note below the title states, "Guest post by Jo-Fai Chow". Another note says, "The sample project (code and data) described below is available on Domino." If you're in a hurry, there are three tutorial links: "Tutorial 1: Using Domino", "Tutorial 2: Using H2O to Predict Soil Properties", and "Tutorial 3: Scaling up your analysis". The introduction section explains that this post is a sequel to TTTAR1 and describes it as a machine learning case study based on a Kaggle competition.

19 Sep 2014

Like 0 Tweet 21 g+1 4

How to use R, H₂O, and Domino for a Kaggle competition

Guest post by [Jo-Fai Chow](#)

The sample project (code and data) described below is [available on Domino](#).

If you're in a hurry, feel free to skip to:

- Tutorial 1: [Using Domino](#)
- Tutorial 2: [Using H₂O to Predict Soil Properties](#)
- Tutorial 3: [Scaling up your analysis](#)

Introduction

This blog post is the sequel to [TTTAR1](#) a.k.a. [An Introduction to H₂O Deep Learning](#). If the previous blog post was a brief intro, this post is a proper machine learning case study based on a recent [Kaggle competition](#): I am leveraging [R](#), [H₂O](#) and [Domino](#) to compete (and do pretty well) in a real-world data mining contest.

R + H₂O + Domino for Kaggle
[Guest Blog Post for Domino & H₂O \(2014\)](#)

About H₂O.ai

What exactly is H₂O?

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	<ul style="list-style-type: none">• H2O Open Source In-Memory AI Prediction Engine• Sparkling Water• Steam
Mission	Operationalize Data Science, and provide a platform for users to build beautiful data products
Team	70 employees <ul style="list-style-type: none">• Distributed Systems Engineers doing Machine Learning• World-class visualization designers
Headquarters	Mountain View, CA



H₂O.ai

Scientific Advisory Council



Dr. Trevor Hastie

- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*



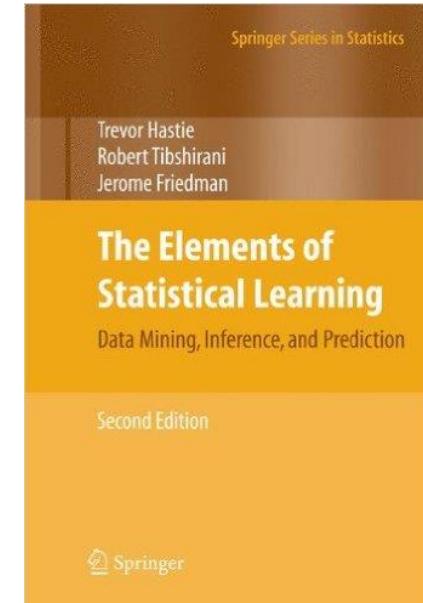
Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*



Dr. Steven Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*



A large, semi-transparent image of an underwater scene with bright yellow sunlight rays filtering down through dark blue water.

H₂O is an open source platform
empowering business transformation

Bring AI To Business Empower Transformation

Financial Services, Insurance and Healthcare as Our Vertical Focus



Community as Our Foundation

Users In Various Verticals Adore H₂O



Hospital Corporation of America™



H₂O.ai

H2O In Action

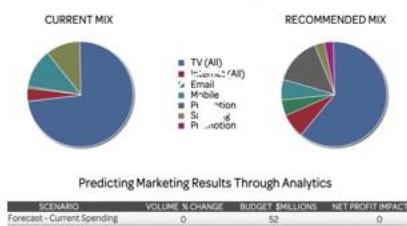
www.h2o.ai/customers

Capital One



Capital One uses H2O open source machine learning for various use cases.

MarketShare



H2O predictive analytics helps boost the impact and results of digital marketing.

Kaiser



Kaiser uses H2O machine learning to save lives.

Zurich Insurance



Zurich turned to H2O as a strategic differentiator for commercial insurance.

Progressive



Progressive uses H2O predictive analytics for user-based insurance.

Comcast



Comcast uses H2O to improve customer experience.

Hospital Corporation of America



HCA uses H2O to predict patient outcomes in real-time.

McKesson



McKesson discusses the adoption of artificial intelligence in healthcare.

Macy's



Macy's uses H2O for personalized site recommendations.

Transamerica



Transamerica turns to H2O to develop a product recommendation platform for insurance.

Paypal



Paypal turned to H2O Deep Learning for fraud detection and customer churn.

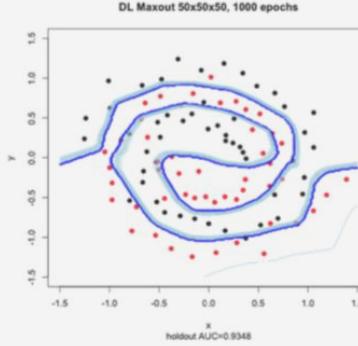
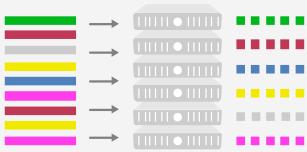
eBay



eBay chose H2O for open source machine learning.

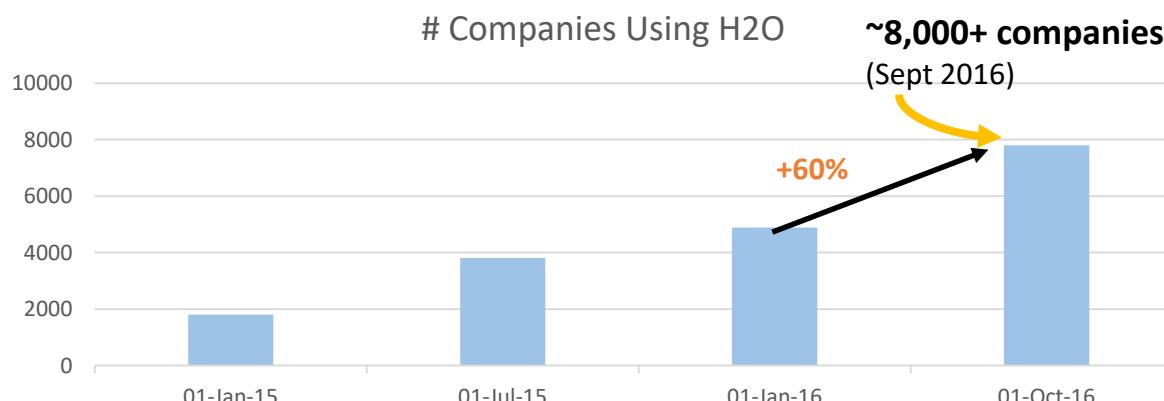
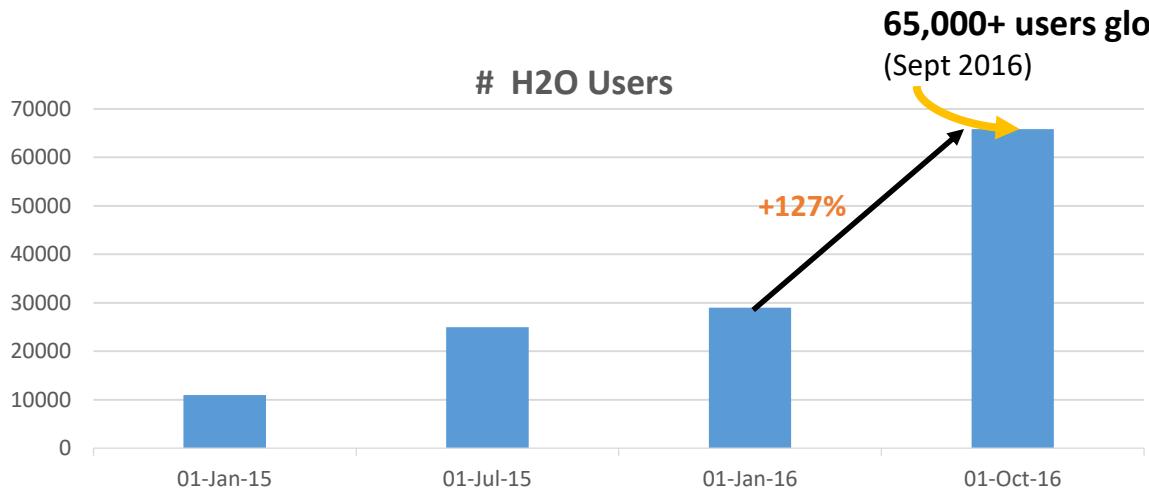
H₂O.ai

H₂O.ai Makes A Difference as an AI Platform

Open Source	Big Data Ecosystem	Flexible Interface	Smart and Fast Algorithms
 <ul style="list-style-type: none">• 100% open source	 	    H₂O Flow	
Scalability and Performance	Rapid Model Deployment	GPU Enablement	Cloud Integration
 <ul style="list-style-type: none">• Distributed In-Memory Computing Platform• Distributed Algorithms• Fine-Grain MapReduce	<ul style="list-style-type: none">• Highly portable models deployed in Java (POJO) and Model Object Optimized (MOJO)• Automated and streamlined scoring service deployment with Rest API 		  

H₂O Community Growth

Tremendous Momentum Globally



* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

Large User Circle

- 65,000+ users from ~8,000 companies in 140 countries. Top 5 from:

1. United States
2. India
3. Japan
4. Germany
5. United Kingdom

H₂O Community Support

Google forum – h2osteam

The screenshot shows the Google forum interface for the group "h2osteam". The sidebar on the left includes sections for Groups, My groups, Home, Starred, Favourites, Recently viewed, Recent searches, and Recently posted to. A yellow callout box highlights the "Favourites" section with the text "Click on a group's star icon to add it to your favourites". The main content area displays a list of topics under the heading "H2O Open Source Scalable Machine Learning - h2osteam Shared publicly". Topics include "When is Steam going to be released?", "H2O Python Modules", "H2O Installation", "PySparkling launch problem with Python 2.6 or older", "Predicted Values", and "Combining holdout predictions, while keep_cross_validation_predictions parameter is active in Python". A note at the bottom encourages users to shift their energy toward building the new community website.

community.h2o.ai

Please try

The screenshot shows the H2O community website at https://community.h2o.ai/index.html. The sidebar on the right lists categories such as Algorithms, Announcements, Artificial Intelligence, Deep Water, Demos, H2O, Java, Machine Learning, Python, R, Source Code, Sparkling Water, Steam, Tools, and Troubleshooting. A yellow callout box highlights the "Sparkling Water" section with the text "Release 08/30" and "We are happy to announce that Sparkling Water 2.0 release is almost here. On September 1, 2016 we will release Sparkling Water 2.0. Download info is coming soon." The main content area shows a feed of posts under "All Posts", including topics like "When is Steam going to be released?", "H2O Python Modules", "H2O Installation", "PySparkling launch problem with Python 2.6 or older", "Predicted Values", and "Combining holdout predictions, while keep_cross_validation_predictions parameter is active in Python".

H₂O for Kaggle Competitions

CIFAR-10 Competition
Winners: Interviews with Dr.
Ben Graham, Phil Culliton, &
Zygmunt Zajac

Triskelion | 01.02.2015

[READ MORE](#)

“I did really like H2O’s deep learning implementation in R, though - the interface was great, the back end extremely easy to understand, and it was scalable and flexible. Definitely a tool I’ll be going back to.”

Kaggle challenge
2nd place winner
Colin Priest

for creating this corpus. ,
do not contain Spanish sent-
is a widespread major langu-
reason was to create a corp-
tasks. These tasks are com-

Completed • Knowledge • 161 teams

Denoising Dirty Documents

Mon 1 Jun 2015 – Mon 5 Oct 2015 (3 months ago)

[READ MORE](#)

“For my final competition submission I used an ensemble of models, including 3 deep learning models built with R and h2o.”

H₂O.ai

H₂O for Academic Research

European Journal of Operational Research

Available online 22 October 2016

In Press, Accepted Manuscript — Note to users



Innovative Applications of O.R.

Deep neural networks, gradient-boosted trees, random forests:
Statistical arbitrage on the S&P 500

Christopher Krauss^{1,a}, Xuan Anh Do^{1,a}, Nicolas Huck^{1,b}.

Received 15 April 2016, Revised 22 August 2016, Accepted 18 October 2016, Available online 22 October 2016

Highlights

- Latest machine learning techniques are deployed in a statistical arbitrage context.
- Deep neural networks, gradient-boosted trees, and random forests are considered.
- An equal-weighted ensemble of these techniques produces the best performance.
- Daily returns are substantial though declining over time.
- The system is especially effective at times of financial turmoil.

<http://www.sciencedirect.com/science/article/pii/S0377221716308657>

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > physics > arXiv:1509.01199

Search or Article-id (Help | Advanced search) All papers ▾ Go!

Physics > Physics and Society

Inferring Passenger Type from Commuter Eigentravel Matrices

Erika Fille Legara, Christopher Monterola

(Submitted on 25 Aug 2015)

A sufficient knowledge of the demographics of a commuting public is essential in formulating and implementing more targeted transportation policies, as commuters exhibit different ways of traveling. With the advent of the Automated Fare Collection system (AFC), probing the travel patterns of commuters has become less invasive and more accessible. Consequently, numerous transport studies related to human mobility have shown that these observed patterns allow one to pair individuals with locations and/or activities at certain times of the day. However, classifying commuters using their travel signatures is yet to be thoroughly examined. Here, we contribute to the literature by demonstrating a procedure to characterize passenger types (Adult, Child/Student, and Senior Citizen) based on their three-month travel patterns taken from a smart fare card system. We first establish a method to construct distinct commuter matrices, which we refer to as eigentravel matrices, that capture the characteristic travel routines of individuals. From the eigentravel matrices, we build classification models that predict the type of passengers traveling. Among the models explored, the gradient boosting method (GBM) gives the best prediction accuracy at 76%, which is 84% better than the minimum model accuracy (41%) required vis-à-vis the proportional

Download:

- PDF
- Other formats (license)

Current browse context: physics.soc-ph
< prev | next >
new | recent | 1509

Change to browse by: cs cs.CY physics physics.data-an stat stat.AP stat.ML

References & Citations

- INSPIRE HEP (refers to | cited by)
- NASA ADS

Bookmark (what is this?)



<https://arxiv.org/abs/1509.01199>

H_2O
democratizes
artificial intelligence & big data science

Our Open Source Products

100% Open Source. Big Data Science for Everyone!

H₂O.ai Offers AI Open Source Platform Product Suite to Operationalize Data Science with Visual Intelligence



Visual Intelligence and UX Framework For Data Interpretation and Story Telling on top of Beautiful Data Products

100% Open Source



In-Memory, Distributed
Machine Learning
Algorithms with Speed and
Accuracy

Deep Water

State-of-the-art
Deep Learning on GPUs with
TensorFlow, MXNet or Caffe
with the ease of use of H2O

Spark + H₂O
SPARKLING
WATER

H2O Integration with Spark.
Best Machine Learning on
Spark.

Steam

Operationalize and
Streamline Model Building,
Training and Deployment
Automatically and Elastically

H₂O.ai Offers AI Open Source Platform Product Suite to Operationalize Data Science with Visual Intelligence

This Talk + Live Demos

100% Open Source



Deep Water

In-Memory, Distributed
Machine Learning
Algorithms with Speed and
Accuracy

State-of-the-art
Deep Learning on GPUs with
TensorFlow, MXNet or Caffe
with the ease of use of H2O

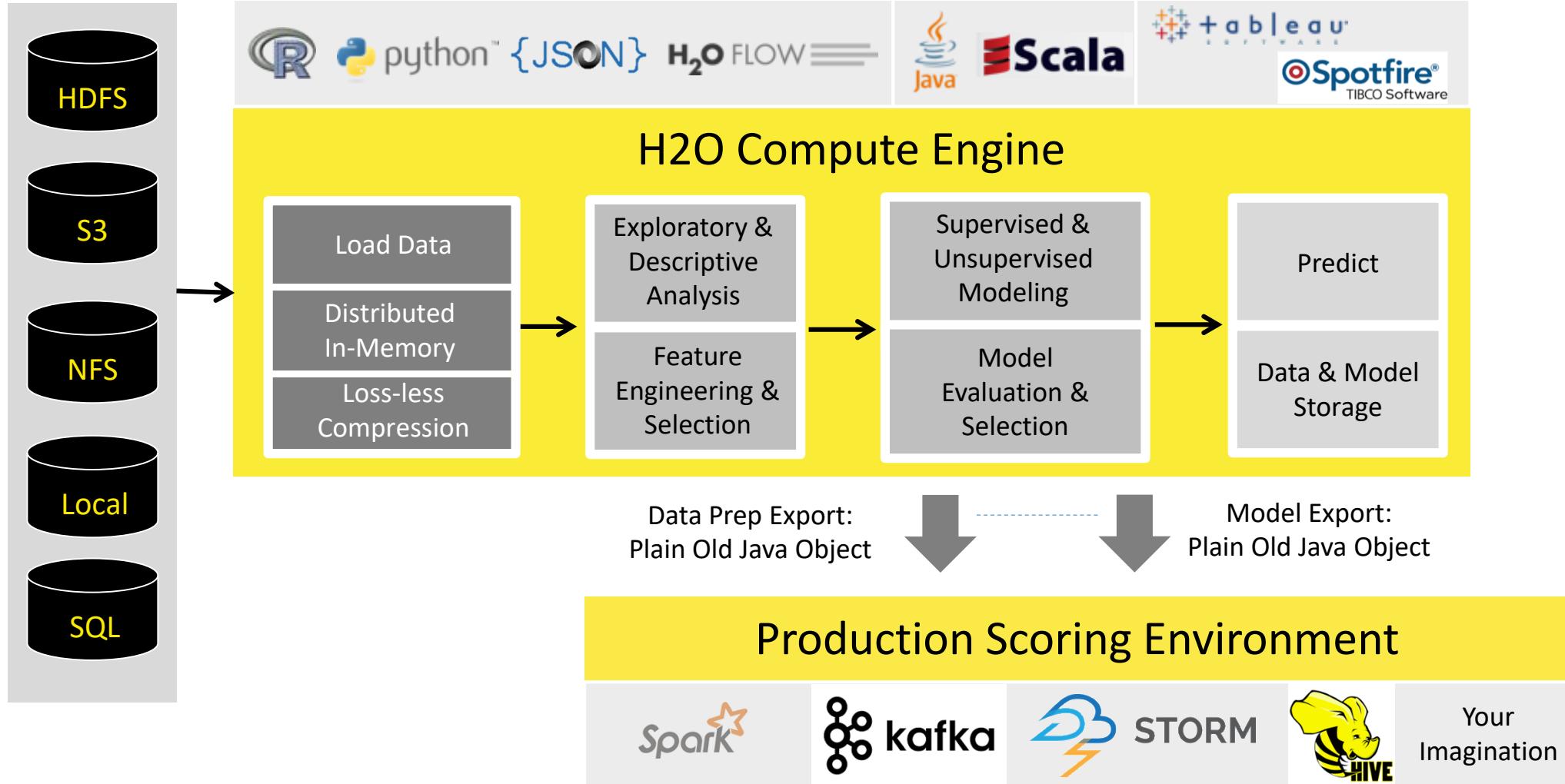
Spark + H₂O
SPARKLING
WATER

H2O Integration with Spark.
Best Machine Learning on
Spark.

Steam

Operationalize and
Streamline Model Building,
Training and Deployment
Automatically and Elastically

High Level Architecture



Algorithms Overview

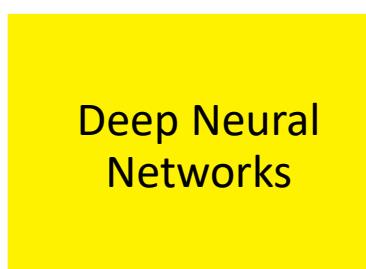
Supervised Learning



- **Generalized Linear Models:** Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Naïve Bayes**



- **Distributed Random Forest:** Classification or regression models
- **Gradient Boosting Machine:** Produces an ensemble of decision trees with increasing refined approximations

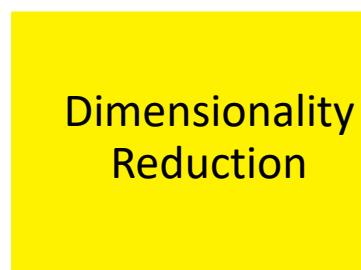


- **Deep learning:** Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

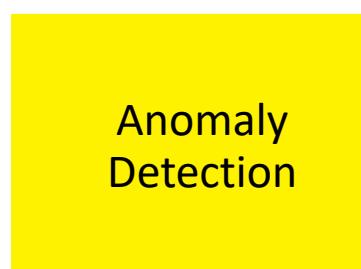
Unsupervised Learning



- **K-means:** Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k



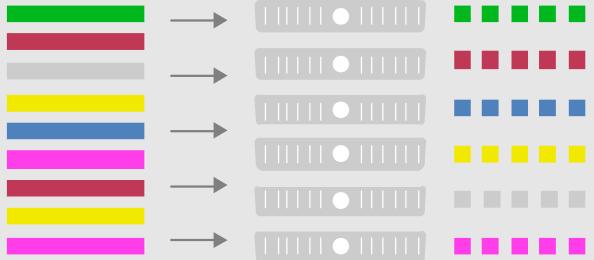
- **Principal Component Analysis:** Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data



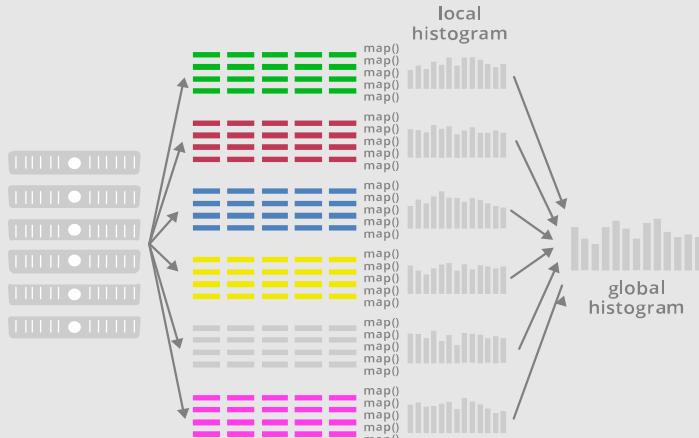
- **Autoencoders:** Find outliers using a nonlinear dimensionality reduction using deep learning

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable
Distributed Histogram Calculation for GBM

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H₂O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- “Out-of-box” functionalities for all algorithms (**NO MORE SCRIPTING**) and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

H₂O Deep Learning in Action

116M rows, 6GB CSV file
800+ predictors (numeric + categorical)

airlines_all_selected_cols.hex

Actions: View Data, Split..., Build Model..., Predict, Download, Export

Rows	Columns	Compressed Size
116695259	12	2GB



Job

Run Time 00:00:36.712

Remaining Time 00:00:17.188

Type Model

Key Q deeplearning-dd2f42f7-81f7-42e8-9d98-e34437309828

Description DeepLearning

Status RUNNING

Progress 69%

Iterations: 12. Epochs: 0.628821. Speed: 2,243,735 samples/sec. Estimated time left: 21.849 sec

Actions View, Cancel Job

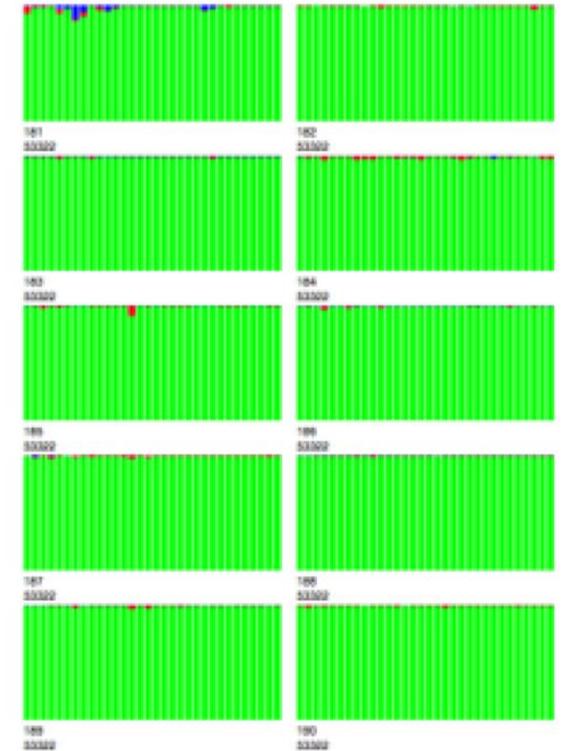
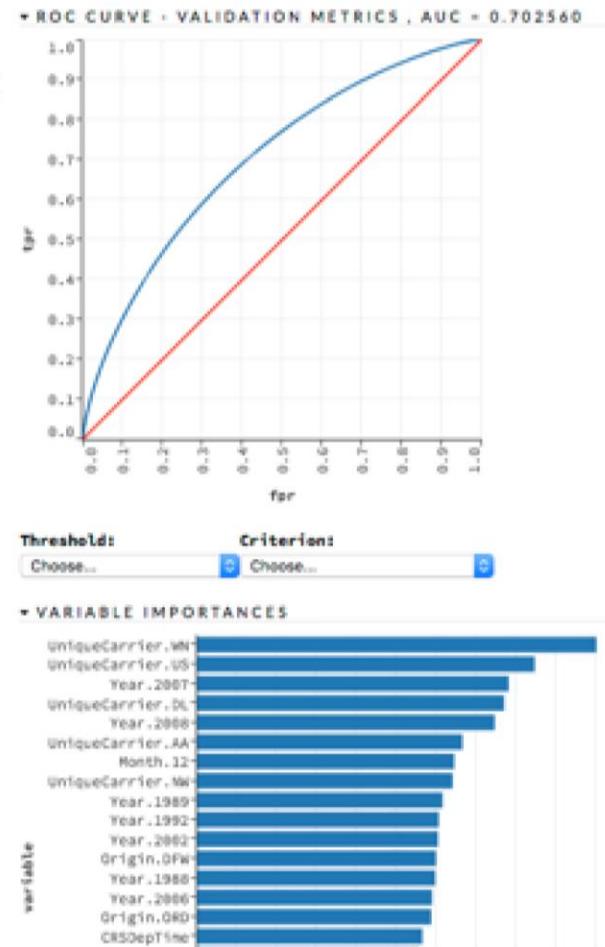
* OUTPUT - STATUS OF NEURON LAYERS (PREDICTING ISDELAYED, 2-CLASS CLASSIFICATION, BERNoulli DISTRIBUTION, CROSSENTROPY LOSS, 17,462 WEIGHTS/BIASES, 221.3 KB, 106,585,385 TRAINING SAMPLES, MINI-BATCH SIZE 1)

layer	units	type	dropout	l1	l2	mean_rate	rate_RMS	momentum	weight_RMS	mean_weight	weight_RMS	mean_bias	bias_RMS
1	887	Input	0										
2	20	Rectifier	0	0	0	0.0493	0.2020	0	-0.0021	0.2111	-0.9139	1.0036	
3	20	Rectifier	0	0	0	0.0157	0.0227	0	-0.1833	0.5362	-1.3988	1.5259	
4	20	Rectifier	0	0	0	0.0517	0.0446	0	-0.1575	0.3068	-0.8846	0.6046	
5	20	Rectifier	0	0	0	0.0761	0.0844	0	-0.0374	0.2275	-0.2647	0.2481	
6	2	Softmax	0	0	0	0.0161	0.0083	0	0.0741	0.7268	0.4269	0.2056	

H₂O.ai

Deep Learning Model

real-time, interactive
model inspection in Flow



10 nodes: all
320 cores busy

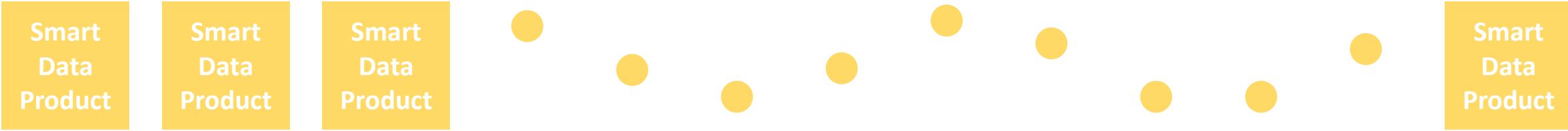


Steam

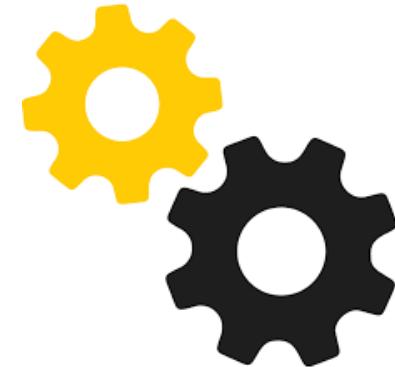


Steam

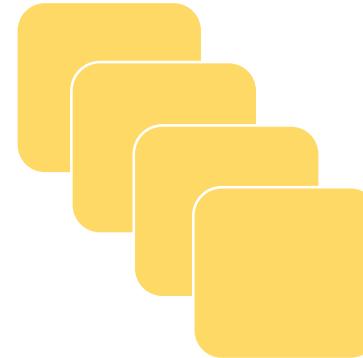
Automated Platform to Build and Scale Smart Data Products



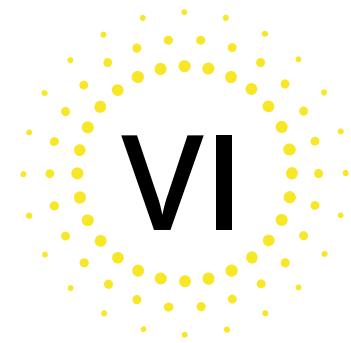
AI – Machine Learning



Automation



Scalability

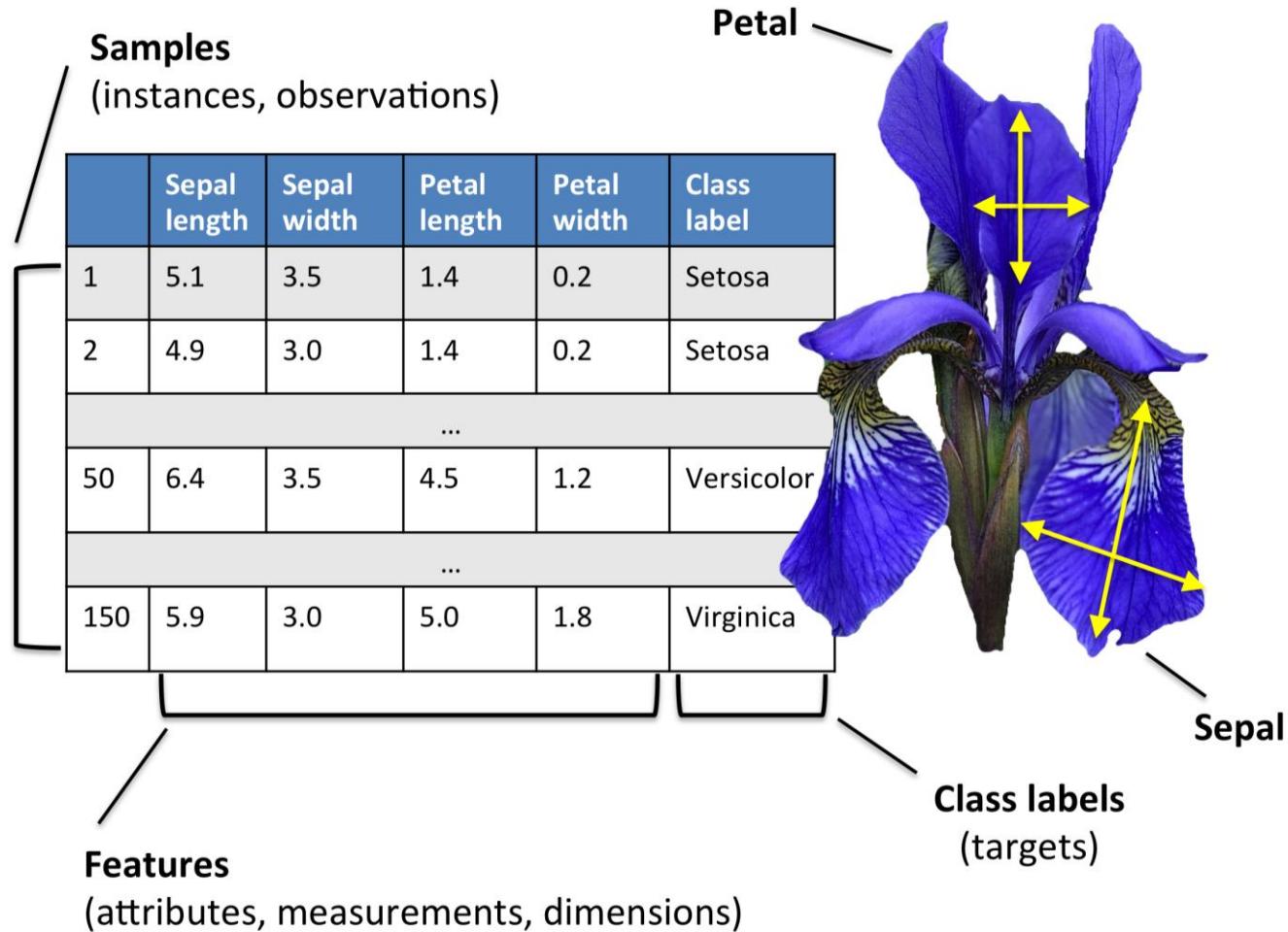


Visual Intelligence

H2O in Action: H₂O + R + Web + Steam

Quick Demo (5 mins)

Simple Demo – Iris



H₂O + R

The screenshot shows an RStudio interface with a file named "h2o_iris_demo.R". The code in the script is as follows:

```
1 # -----
2 # Build a simple classification model using iris dataset
3 # -----
4
5 # Start and connect to a local H2O cluster
6 library(h2o)
7 h2o.init(nthreads = -1)
8
9 # Import data from a R data frame
10 data(iris)
11 d_iris <- as.h2o(iris)
12
13 # Define Targets and Features
14 target <- "Species"
15 features <- setdiff(colnames(d_iris), c("Species"))
16
17 # -----
18 # Train a H2O Model
19 # -----
20
21 # Train three basic H2O models
22 model_drf <- h2o.randomForest(x = features,
23                                y = target,
24                                model_id = "iris_random_forest",
25                                training_frame = d_iris)
26
27 model_gbm <- h2o.gbm(x = features,
28                        y = target,
29                        model_id = "iris_gbm",
30                        training_frame = d_iris)
31
32 model_dnn <- h2o.deeplearning(x = features,
33                                 y = target,
34                                 model_id = "iris_deep_learning",
35                                 training_frame = d_iris)
36
```

Please try

Demo Time

R, Flow and then Steam

Key Learning Resources

- Help Documentations
 - docs.h2o.ai
- Meetups
 - bit.ly/h2o_meetup
- YouTube Channel
 - bit.ly/h2o_youtube



H2O, Sparkling Water, and Steam Documentation

[Getting Started](#) [Data Science Algorithms](#) [Languages](#) [Tutorials, Examples, & Presentations](#) [For Developers](#) [For the Enterprise](#)

Getting Started

H2O

[What is H2O?](#)
[H2O User Guide](#)
[Recent Changes](#)
[Open Source License \(Apache V2\)](#)

[Quick Start Video - Flow Web UI](#)
[Quick Start Video - R](#)
[Quick Start Video - Python](#)

[Download H2O](#)

Sparkling Water

[What is Sparkling Water?](#)
[Sparkling Water Booklet](#)
[PySparkling Readme](#)
[RSparkling Readme](#)
[Open Source License \(Apache V2\)](#)

[Quick Start Video - Scala](#)
[Quick Start Video - Python](#)

[Download Sparkling Water](#)

Steam

[What is Steam?](#)
[Steam User Guide](#)
[Recent Changes](#)
[Open Source License \(AGPL\)](#)

[Download Steam](#)

Questions and Answers

[FAQ](#)
[Community Forum](#)
[h2ostream Google Group](#)
[Issue Tracking \(JIRA\)](#)
[Gitter](#)
[Stack Overflow](#)
[Cross Validated](#)

[For Supported Enterprise Customers](#)
[Enterprise Support via Web | Email](#)

Data Science Algorithms

Supervised Learning

Generalized Linear Modeling (GLM)	Tutorial	Booklet	Reference	Tuning
Gradient Boosting Machine (GBM)	Tutorial	Booklet	Reference	Tuning
Deep Learning	Tutorial	Booklet	Reference	Tuning
Distributed Random Forest	Tutorial	Booklet	Reference	Tuning
Naive Bayes	Tutorial	Booklet	Reference	Tuning
Ensembles (Stacking)	Tutorial	Booklet	Reference	Tuning

Unsupervised Learning

Generalized Low Rank Models (GLRM)	Tutorial	Reference
K-Means Clustering	Tutorial	Reference
Principal Components Analysis (PCA)	Tutorial	Reference

AI Open Source Platform

Operationalize Data Science with Visual Intelligence

Meetup Talk
Yesterday

bit.ly/h2o_paris_2_slides

Visual Intelligence and UI
Story Telling on top of Big Data

100% Open Source

3rd talk this evening
by Jakub (Kuba)



In-Memory, Distributed
Machine Learning
Algorithms with Speed and
Accuracy

Deep Water

State-of-the-art
Deep Learning on GPUs with
TensorFlow, MXNet or Caffe
with the ease of use of H2O

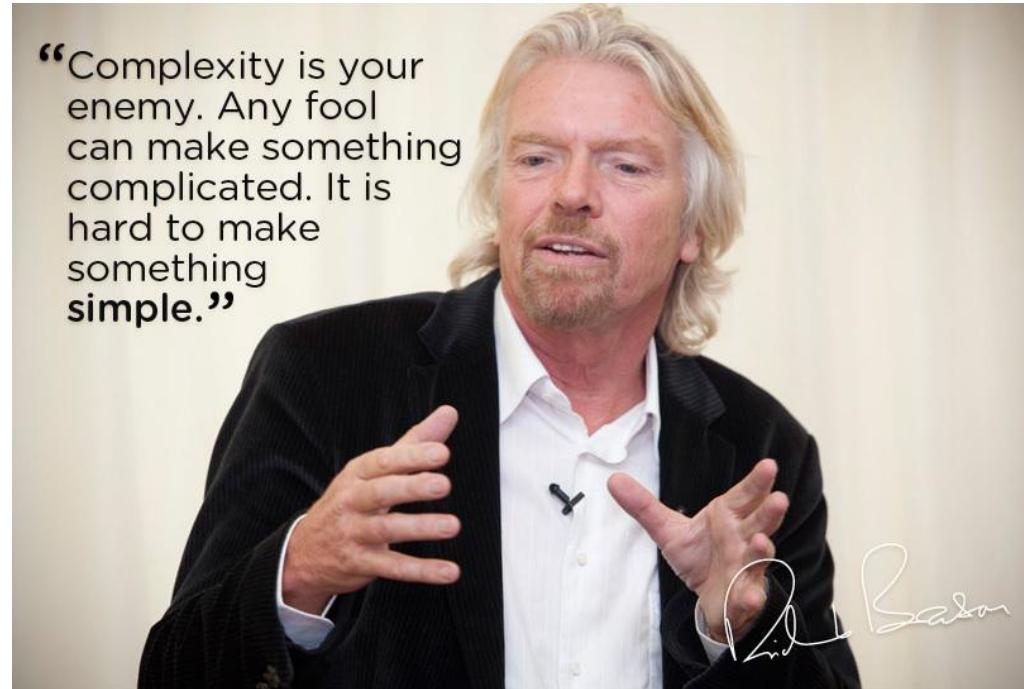

SPARKLING
WATER

H2O Integration with Spark.
Best Machine Learning on
Spark.

Steam

Operationalize and
Streamline Model Building,
Training and Deployment
Automatically and Elastically

H₂O's Mission



“Complexity is your enemy. Any fool can make something complicated. It is hard to make something simple.”

Making Machine Learning Accessible to Everyone

Photo credit: Virgin Media

H₂O.ai

Merci beaucoup!

- Organizers & Sponsors
 - Diane, Vincent, Barthelemy, François, Julie and Timeri
 - NUMA
- Code, Slides & Documents
 - bit.ly/h2o_meetups
 - docs.h2o.ai
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe



H₂O.ai

Extra Slides (Iris Demo)

H2O Iris Demo x jo-fai

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Iris Demo

Import Files...
Upload File...
Split Frame...
Merge Frames...
List All Frames
Impute...

CS Expression...

Ready localhost:54321/flow/index.html#

Connections: 0 H2O

iris.csv Show all X

?

H2O Iris Demo x jo-fai

localhost:54321/flow/index.html

H2O FLOW Flow Cell Date Model Score Admin Help

Iris Demo

CS Expression...

Upload Dataset...

Choose file iris.csv

Cancel Upload

?

Ready

Connections: 0 H2O

iris.csv

Show all

A B P C D E F G H I J K L M N O P Q R S T U V W X Y Z

Iris Demo



Setup Parse

PARSE CONFIGURATION

Sources

ID

Parser

Separator

Column Headers Auto

First row contains column names

First row contains data

Options Enable single quotes as a field quotation character

Delete on done

EDIT COLUMN NAMES AND TYPES

Search by column name...

1	Sepal.Length	<input type="button" value="Numeric ▾"/>	5.1	4.9	4.7	4.6	5	5.4	4.6	5	4.4
2	Sepal.Width	<input type="button" value="Numeric ▾"/>	3.5	3	3.2	3.1	3.6	3.9	3.4	3.4	2.9
3	Petal.Length	<input type="button" value="Numeric ▾"/>	1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4
4	Petal.Width	<input type="button" value="Numeric ▾"/>	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.2	0.2
5	Species	<input type="button" value="Enum ▾"/>	setosa								

[◀ Previous page](#) [▶ Next page](#)



Iris Demo



47ms

iris_from_csv

Actions: [View Data](#) [Split...](#) [Build Model...](#) [Predict](#) [Download](#) [Export](#) [Delete](#)

Rows

150

Columns

5

Compressed Size

1KB

COLUMN SUMMARIES

label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
Sepal.Length	real	0	0	0	0	4.3000	7.9000	5.8433	0.8281	3	
Sepal.Width	real	0	0	0	0	2.0	4.4000	3.0573	0.4359	3	
Petal.Length	real	0	0	0	0	1.0	6.9000	3.7580	1.7653	3	
Petal.Width	real	0	0	0	0	0.1000	2.5000	1.1993	0.7622	3	
Species	enum	0	50	0	0	0	2.0	.	.	3	Convert to numeric

Previous 20 Columns

Next 20 Columns

CHUNK COMPRESSION SUMMARY

FRAME DISTRIBUTION SUMMARY



localhost:54321/flow/index.html#

H2O FLOW Flow Cell Data Model Score Admin Help

Iris Demo ✓

Actions Impute Inspect

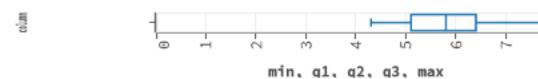
Summary: Sepal.Length

Actions Impute Inspect

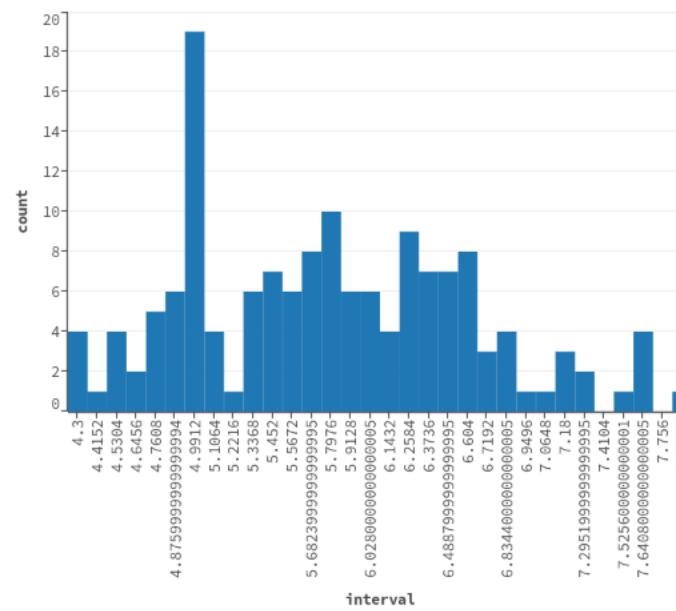
CHARACTERISTICS



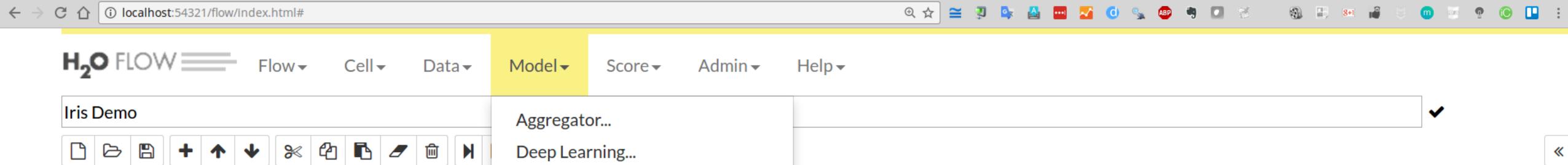
SUMMARY



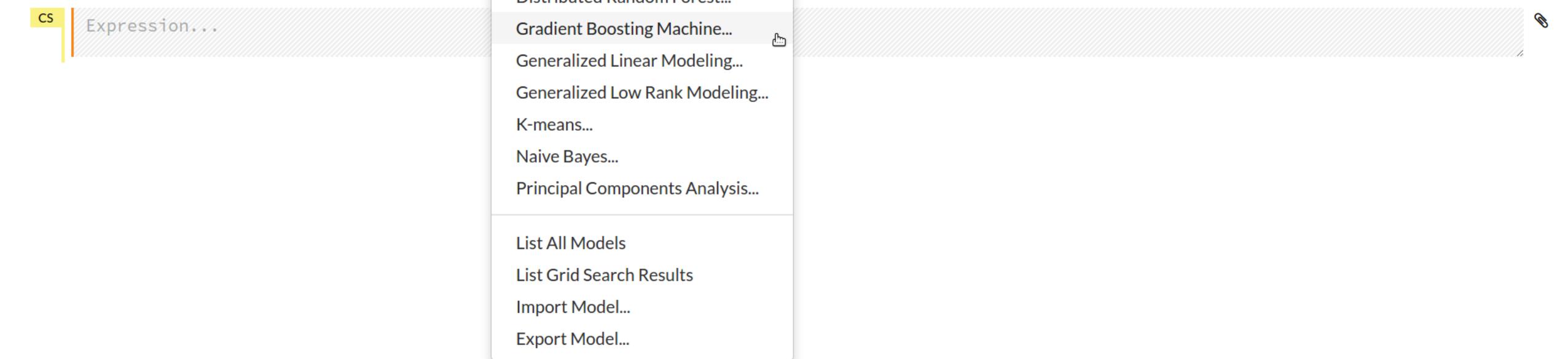
DISTRIBUTION



localhost:54321/flow/index.html#



The screenshot shows the H2O Flow interface. At the top, there's a navigation bar with tabs for 'Flow', 'Cell', 'Data', 'Model' (which is currently selected and highlighted in yellow), 'Score', 'Admin', and 'Help'. Below the navigation bar is a toolbar with various icons for file operations like opening, saving, and deleting. A sidebar on the left is titled 'Iris Demo' and contains a section labeled 'Expression...' with a 'CS' indicator. The main workspace is currently empty. A context menu is open over the workspace, listing several modeling options: Aggregator..., Deep Learning..., Distributed Random Forest..., Gradient Boosting Machine..., Generalized Linear Modeling..., Generalized Low Rank Modeling..., K-means..., Naive Bayes..., Principal Components Analysis..., List All Models, List Grid Search Results, Import Model..., and Export Model... .



The screenshot shows the H2O Flow interface. At the top, there's a navigation bar with tabs for 'Flow', 'Cell', 'Data', 'Model' (which is currently selected and highlighted in yellow), 'Score', 'Admin', and 'Help'. Below the navigation bar is a toolbar with various icons for file operations like opening, saving, and deleting. A sidebar on the left is titled 'Iris Demo' and contains a section labeled 'Expression...' with a 'CS' indicator. The main workspace is currently empty. A context menu is open over the workspace, listing several modeling options: Aggregator..., Deep Learning..., Distributed Random Forest..., Gradient Boosting Machine..., Generalized Linear Modeling..., Generalized Low Rank Modeling..., K-means..., Naive Bayes..., Principal Components Analysis..., List All Models, List Grid Search Results, Import Model..., and Export Model... .

Iris Demo



Expression...

CS buildModel "drf"

192ms

Build a Model

Select an algorithm: **Distributed Random Forest** ▾

PARAMETERS

GRID?

<i>model_id</i>	DRF-Iris-Demo	Destination id for this model; auto-generated if not specified.
<i>training_frame</i>	iris_from_csv ▾	Id of the training data frame (Not required, to allow initial validation of model parameters).
<i>validation_frame</i>	(Choose...)	Id of the validation data frame.
<i>nfolds</i>	0	Number of folds for N-fold cross-validation (0 to disable or >= 2).
<i>response_column</i>	Species	Response variable column.
<i>ignored_columns</i>	Search...	

Showing page 1 of 1.

<input type="checkbox"/> Sepal.Length	REAL
<input type="checkbox"/> Sepal.Width	REAL
<input type="checkbox"/> Petal.Length	REAL
<input type="checkbox"/> Petal.Width	REAL
<input type="checkbox"/> Species	ENUM(3)

H2O FLOW

Flow ▾ Cell ▾ Data ▾ Model ▾ Score ▾ Admin ▾ Help ▾

Iris Demo



nbins_top_level	1024	For numerical columns (real/int), build a histogram of (at most) this many bins at the root level, then decrease by factor of two per level
nbins_cats	1024	For categorical columns (factors), build a histogram of this many bins, then split at the best point. Higher values can lead to more overfitting.
r2_stopping	1.7976931348623157e+	r2_stopping is no longer supported and will be ignored if set - please use stopping_rounds, stopping_metric and stopping_tolerance instead. Previous version of H2O would stop making trees when the R^2 metric equals or exceeds this
stopping_rounds	0	Early stopping based on convergence of stopping_metric. Stop if simple moving average of length k of the stopping_metric does not improve for k:=stopping_rounds scoring events (0 to disable)
stopping_metric	AUTO	Metric to use for early stopping (AUTO: logloss for classification, deviance for regression)
stopping_tolerance	0.001	Relative tolerance for metric-based stopping criterion (stop if relative improvement is not at least this much)
max_runtime_secs	0	Maximum allowed runtime in seconds for model training. Use 0 to disable.
checkpoint		Model checkpoint to resume training with.
col_sample_rate_per_tree	1	Column sample rate per tree (from 0.0 to 1.0)
min_split_improvement	0.00001	Minimum relative improvement in squared error reduction for a split to happen
histogram_type	AUTO	What type of histogram to use for finding optimal split points
categorical_encoding	AUTO	Encoding scheme for categorical features

EXPERT

build_tree_one_node	<input checked="" type="checkbox"/>	Run on one node only; no network overhead but fewer cpus used. Suitable for small datasets.
sample_rate_per_class		Row sample rate per tree per class (from 0.0 to 1.0)
binomial_double_trees	<input checked="" type="checkbox"/>	For binary classification: Build 2x as many trees (one per class) - can lead to higher accuracy.
col_sample_rate_change_per_level	1	Relative change of the column sampling rate for every level (from 0.0 to 2.0)

GRID?

Build Model



Ready

Connections: 0

H2O



Flow ▾ Cell ▾ Data ▾ Model ▾ Score ▾ Admin ▾ Help ▾

Iris Demo



col_sample_rate_change_per_level 1

Relative change of the column sampling rate for every level (from 0.0 to 2.0)

Build Model

CS buildModel 'drf', {"model_id": "DRF-Iris-Demo", "training_frame": "iris_from_csv", "nfolds": 0, "response_column": "Species", "ignored_columns": [], "ignore_const_cols": true, "ntrees": 50, "max_depth": 20, "min_rows": 1, "nbins": 20, "seed": -1, "mtries": -1, "sample_rate": 0.6320000290870667, "score_each_iteration": false, "score_tree_interval": 0, "balance_classes": false, "max_confusion_matrix_size": 20, "max_hit_ratio_k": 0, "nbins_top_level": 1024, "nbins_cats": 1024, "r2_stopping": 1.7976931348623157e+308, "stopping_rounds": 0, "stopping_metric": "AUTO", "stopping_tolerance": 0.001, "max_runtime_secs": 0, "checkpoint": "", "col_sample_rate_per_tree": 1, "min_split_improvement": 0.00001, "histogram_type": "AUTO", "categorical_encoding": "AUTO", "build_tree_one_node": false, "sample_rate_per_class": [], "binomial_double_trees": false, "col_sample_rate_change_per_level": 1}

1.1s

Job

Run Time 00:00:00.183

Remaining Time 00:00:00.0

Type Model

Key [DRF-Iris-Demo](#)

Description DRF

Status DONE

Progress 100%

Done.

Actions [View](#)



Ready

Connections: 0

H2O



Model

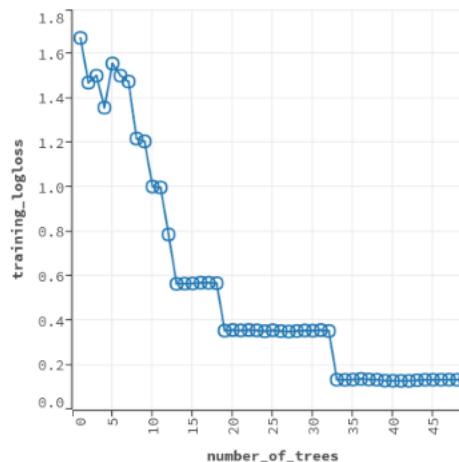
Model ID: DRF-Iris-Demo

Algorithm: Distributed Random Forest

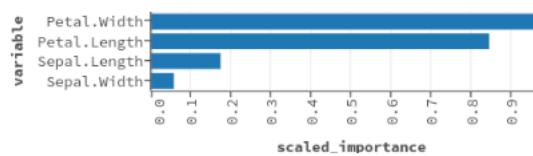
Actions: Refresh Predict... Download POJO Download Model Deployment Package Export Inspect Delete

MODEL PARAMETERS

SCORING HISTORY - LOGLOSS



VARIABLE IMPORTANCES



TRAINING METRICS - CONFUSION MATRIX VERTICAL: ACTUAL; ACROSS: PREDICTED

	setosa	versicolor	virginica	Error	Rate
setosa	50	0	0	0	0 / 50
versicolor	0	46	4	0.0800	4 / 50
virginica	0	4	46	0.0800	4 / 50
Total	50	50	50	0.0533	8 / 150

Extra Slides (Steam Demo)

-  Projects
-  Services
-  Clusters
-  Users

-  Support
-  Logout

WELCOME TO H₂O STEAM

Fast, Distributed Data Science For Teams

[Start A New Project](#)

STEAM

-  Projects
-  Services
-  Clusters
-  Users

-  Support
-  Logout

1. Select H2O Cluster

Select an H2O cluster to import models and datasets from.

CLUSTER	DATASETS	MODELS	Connect
joe	N/A	N/A	<button>Connect</button>

... Or Connect To A New H2O Cluster

Connect to a H2O cluster where your existing models and data sets are located.

localhost 54321

STEAM

- Projects
- Services
- Clusters
- Users

- Support
- Logout

1. Select H2O Cluster



H2O_started_from_R_joe_eon283
localhost:54321
[use a different cluster](#)

2. Select Dataframe

3. Select Model Category

4. Pick Models To Import

Models in a project must share the same feature set and response column to enable comparison. By default, Steam picks the most optimized model format for you to import. Advanced users can choose your own model type [here](#).

MODEL	RESPONSE COLUMN	CATEGORICAL	
iris_deep_learning	Species	Multinomial	<input checked="" type="checkbox"/> Select for Import
iris_random_forest	Species	Multinomial	<input checked="" type="checkbox"/> Select for Import
iris_gbm	Species	Multinomial	<input checked="" type="checkbox"/> Select for Import

5. Name Project

STEAM

- Projects
- Services
- Clusters
- Users
- Support
- Logout

PROJECTS

[CREATE NEW PROJECT](#)

All Projects

Steam Iris Demo

Multinomial

2016-11-23 23:04

STEAM < Projects

Steam ...

Models

Deployment

Configurations

Collaborators

?

→

Home > Projects > 5 > Models

MODELS

IMPORT MODELS

filter models

F	MODEL	MSE	LOGLOSS	R ²	ACTIONS
	iris_random_forest Created at: 2016-11-23 11:04:38 Num of Observations: 150 Cluster: H2O_started_from_R_joe_eon283	0.034834	0.122869	0.947749	view model details (mouse over) label as ▼ deploy model delete model
	iris_gbm Created at: 2016-11-23 11:04:40 Num of Observations: 150 Cluster: H2O_started_from_R_joe_eon283	0.002838	0.018819	0.995744	view model details label as ▼ deploy model delete model
	iris_deep_learning Created at: 2016-11-23 11:04:37 Num of Observations: 150 Cluster: H2O_started_from_R_joe_eon283	0.127301	0.577885	0.809048	view model details label as ▼ deploy model delete model

1 - 3 of 3 models

localhost:9000/#/projec x

localhost:9000/#/projects/5?_k=oqlpn

STEAM <Projects

Steam ...

Models Deployment Configurations Collaborators

filter models

MODEL

iris_random_forest
Created at: 2016-11-23
Num of Observations: 1
Cluster: H2O_started_fro

iris_gbm
Created at: 2016-11-23
Num of Observations: 1
Cluster: H2O_started_fro

iris_deep_learning
Created at: 2016-11-23
Num of Observations: 1
Cluster: H2O_started_fro

1 - 3 of 3 models

IMPORT MODELS

DEPLOY IRIS_GBM

CONFIGURE SERVICE Steam automatically selects a port that's not in use based on the port range set by your admin.

Service name steam_iris

Preprocessing Script None (Default)

Deploy Cancel

localhost:9000/#/projects/... Wed, Nov 23, 2016 11:10 pm

STEAM < Projects

Steam ...

Models

Deployment

Configurations

Collaborators

?

→

192.168.1.80:41788

localhost:9000/#/projects/5/deployment?_k=0nrxw3

Home > Projects > 5 > Deployment

DEPLOYMENT

UPLOAD NEW PACKAGE

DEPLOYED SERVICES

PACKAGING

steam_iris @ 192.168.1.80:41788 started	
Model 7	
Status OK	

localhost:9000/#/project Steam :: Prediction Service Joe-fai

Prediction Service Steam

Select input parameters, OR enter your own custom query string to predict

MODEL INPUT PARAMETERS

Parameters

1. Sepal.Length	1.2
2. Sepal.Width	0.6
3. Petal.Length	0.8
4. Petal.Width	1.1

Query String

The parameters above gets automatically built into a REST API query string. You can also input your own string if that's easier for you.

[http://192.168.1.80:41788/predict? Sepal.Length=1.2&Sepal.Width=0.6&Petal.Length=0.8&Petal.Width=1.1](http://192.168.1.80:41788/predict?Sepal.Length=1.2&Sepal.Width=0.6&Petal.Length=0.8&Petal.Width=1.1)

PREDICT **CLEAR**

BATCH PREDICTION *OPTIONAL

Select a Batch JSON file

PREDICTION RESULTS

Model Predictions

Predicting setosa

Index	Labels	Probability
0	setosa	0.7998
1	versicolor	0.1513
2	virginica	0.0489

Model Runtime Stats

Service started	2016-11-23 23:10:20 UTC
Uptime	28 s

MORE STATS

Page Footer: /media/SUPPORT/Repo/H... joe@asus-zbp /media/SUP... Steam :: Prediction Service ...

Page Footer: Wed, Nov 23, 2016 11:11 pm

A screenshot of a web browser window. The address bar shows the URL `192.168.1.80:41788/predict?Sepal.Length=1.2&Sepal.Width=0.6&Petal.Length=0.8&Petal.Width=1.1`. The page content displays a JSON response:

```
{"labelIndex":0,"label":"setosa","classProbabilities":[0.7998234476072545,0.15127335891610785,0.04890319347663747]}
```

. The browser has three tabs open: "localhost:9000/#/projec", "Steam :: Prediction Serv", and "192.168.1.80:41788/pre". A cursor icon is visible on the right side of the JSON output.

The Classic REST API Service