

Deep Dive into H2O (Amsterdam)

Part I - Driverless AI Hands-on Training



Jo-fai (Joe) Chow

Data Science Evangelist /
Community Manager

joe@h2o.ai

@matlabulous

Download → [https://bit.ly/
H2OAMS20181204](https://bit.ly/H2OAMS20181204)

Please Read:

- You do not need to download anything.
- We have a cloud lab environment for you.
- All you need is your web browser
(Google Chrome recommended).
- This slide deck is available online
<https://bit.ly/H2OAMS20181204>

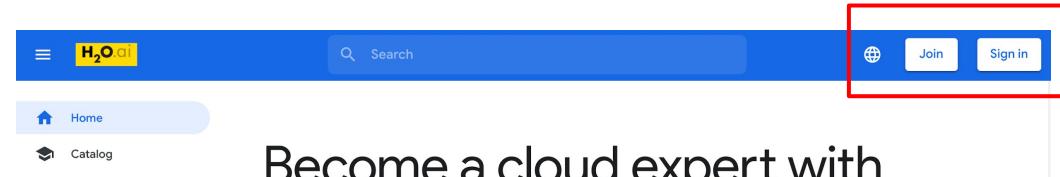
Driverless AI Training (9am - 12pm) Agenda

- **Introduction**
 - Setting Up Qwiklabs
 - About H2O.ai
 - Driverless AI Overview
- **Hands-on Experiment**
 - Credit Card Dataset
 - Automatic Visualisation
 - Machine Learning Workflow
 - Model Interpretability
- **Goals**
 - Import data from file system
 - Generate smart visualization
 - Build complex model ensembles
 - Explain model outputs
- **Bonus**
 - Try your own data
 - Work with H2O team (so we can better understand your business problems)

Setting Up Qwiklabs

Setting Up Qwiklabs

1. Go to **h2oai.qwiklabs.com**
2. **Join** (create an account with your own email) or **Sign in**
3. Search for “**Driverless**”
4. Select “**Introduction to Driverless AI (1 GPU)**”



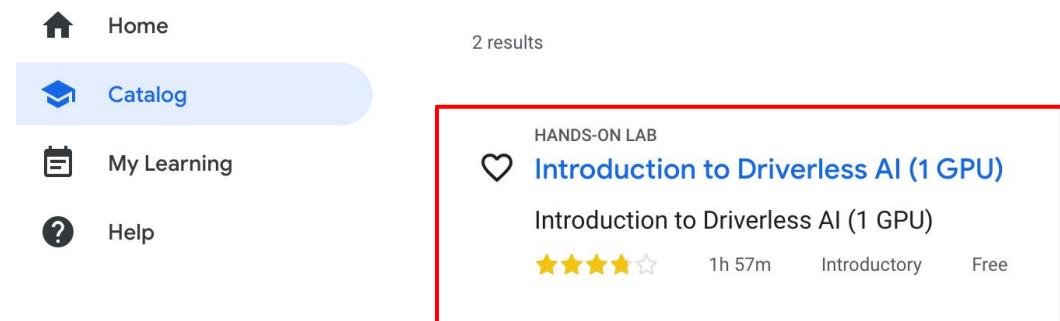
Become a cloud expert with hands-on training.

We give you temporary credentials to Google Cloud Platform and Amazon Web Services, so you can learn the cloud using the real thing – no simulations. From 30-minute individual labs to multi-day courses, from introductory level to expert, instructor-led or self-paced, with topics like machine learning, security, infrastructure, app dev, and more, we've got you covered.



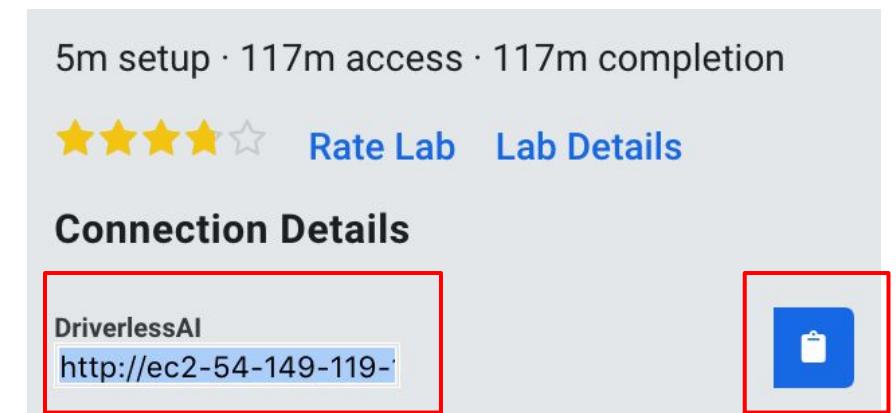
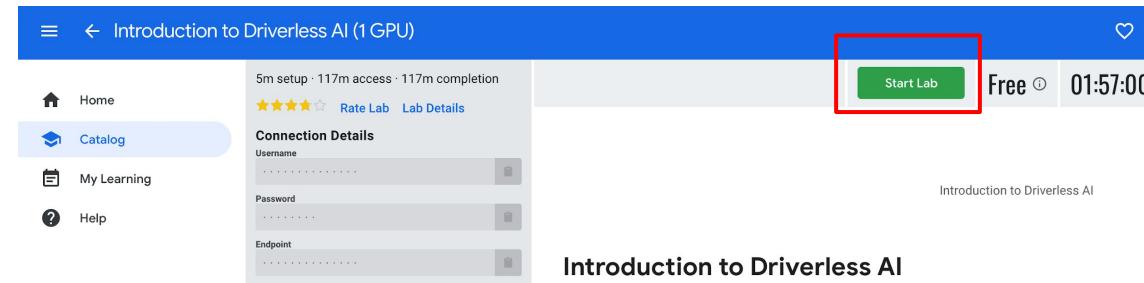
Welcome, Joe!

We give you temporary credentials to Google Cloud Platform and Amazon Web Services, so you can learn the cloud using the real thing – no simulations. From 30-minute individual labs to multi-day courses, from introductory level to expert, instructor-led or self-paced, with topics like machine learning, security, infrastructure, app dev, and more, we've got you covered.



Setting Up Qwiklabs

5. Click on **Start Lab** and wait (it may take several minutes)
6. When the lab is ready, you will see a link to the Driverless AI instance. Click on the **clipboard icon** to copy the URL. Go to this URL in a **new tab**.
7. You may see a “502 Bad Gateway” error message. It is harmless. Wait for another minute and try again.



Setting Up Qwiklabs

8. Scroll to down and click “I agree to these terms”.

9. Enter Username: h2oai

10. Enter Password: h2oai

PLEASE READ THIS H2O.AI DRIVERLESS AI EVALUATION AGREEMENT ("AGREEMENT") CAREFULLY BEFORE USING THE EVALUATION SOFTWARE OFFERED BY H2O.AI, INC. ("H2O.AI"), BY CLICKING THE "AGREE" (OR SIMILAR) BUTTON ON AN ONLINE ORDER FORM, BY USING THE EVALUATION SOFTWARE IN ANY MANNER, OR BY SIGNING AN ORDER FORM WHICH REFERENCES THESE EVALUATION TERMS AND CONDITIONS (AS APPLICABLE) YOU OR THE ENTITY YOU REPRESENT ("LICENSEE") AGREE THAT YOU HAVE READ AND AGREE TO BE BOUND BY AND A PARTY TO THE TERMS AND CONDITIONS OF THIS AGREEMENT TO THE EXCLUSION OF ALL OTHER TERMS. YOU REPRESENT AND WARRANT THAT YOU ARE AUTHORIZED TO BIND LICENSEE. IF THE TERMS OF THIS AGREEMENT ARE CONSIDERED AN OFFER, ACCEPTANCE IS EXPRESSLY LIMITED TO SUCH TERMS.

H2O.ai DRIVERLESS AI Evaluation Agreement

WHEREAS, H2O.ai is willing to supply, within the protection of a confidential relationship, the software, services and related materials provided in connection with this Agreement (collectively, the "Evaluation Software") to Licensee solely for internal evaluation purposes and not for any production use ("Evaluation");

NOW, THEREFORE, In consideration of the foregoing and the mutual covenants hereinafter set forth, the parties hereby agree as follows:

1. Use of Evaluation Software. Subject to the terms of this Agreement, H2O.ai hereby grants to Licensee a personal, nontransferable, nonsublicensable, nonexclusive, revocable license to access and use the Evaluation Software only in accordance with all documentation supplied by H2O.ai solely for Licensee's internal Evaluation purposes during the term of this Agreement. H2O.ai shall at all times retain all title to and ownership of the Evaluation Software and all intellectual property rights relating thereto. Licensee agrees to use the Evaluation Software only in the ordinary course of its Evaluation. Licensee shall not (and shall not allow any third party to): (a) decompile, disassemble, or otherwise reverse engineer any portion of the Evaluation Software; (b) remove, alter or obscure any product identification, copyright or other notices contained on or in the Evaluation Software; (c) disclose, provide, distribute, resell, lease, lend or allow access to the Evaluation Software to any third party; (d) use the Evaluation Software for timesharing or service bureau purposes, or otherwise for the benefit of any third party; (e) copy, modify, adapt or create a derivative work of any part of the Evaluation Software; (f) use the Evaluation Software in excess of any limitations provided by H2O.ai; (g) use the Evaluation Software to help develop any competitive product or service; (h) remove or export the Evaluation Software or any direct product thereof from the United States. H2O.ai

9. Term; Termination. This Agreement shall become effective upon Licensee's first access to or use of the Evaluation Software ("Start Date"). This Agreement may be terminated by either party for any reason or no reason upon written notice to the other party, or immediately upon notice of any breach by Licensee of the provisions of this Agreement, and in any case will terminate twenty-one (21) days from the Start Date, unless extended by H2O.ai in writing. Upon termination, the license granted hereunder will terminate and Licensee shall promptly cease accessing the Evaluation Software, and shall return any and all documents, notes and other materials regarding the Evaluation Software to H2O.ai, including, without limitation, all copies and extracts of the foregoing, but the terms of this Agreement will otherwise remain in effect.

I AGREE TO THESE TERMS

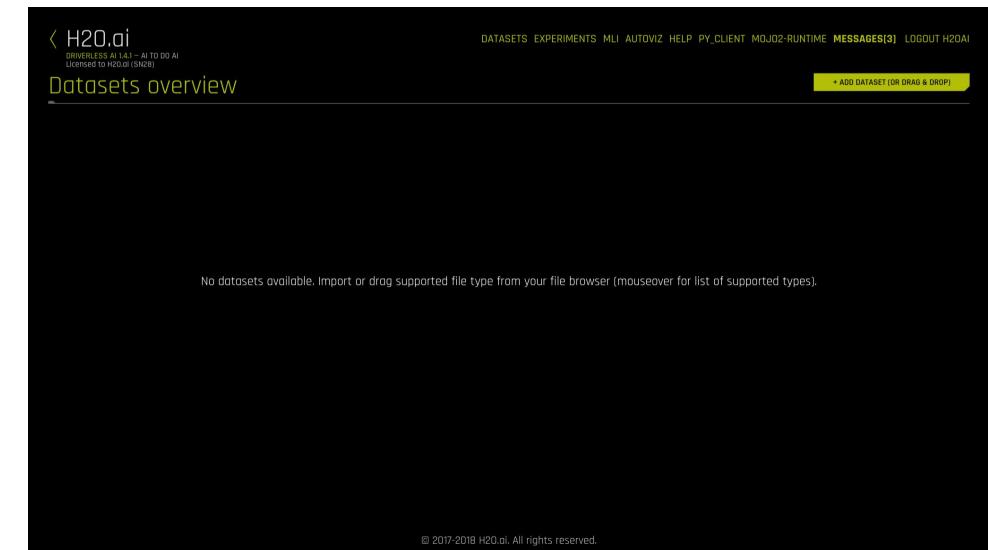
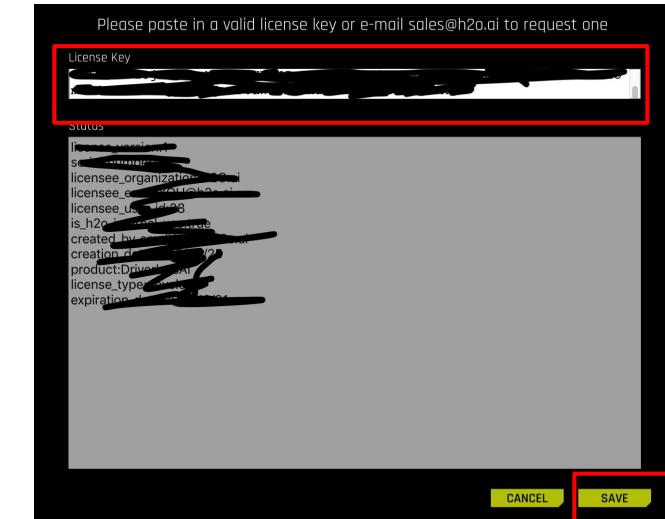


Setting Up Qwiklabs

11. It will show a warning message about the license.
12. **Copy and paste the following licence key. Click Save.**

```
K8O11BsW2uJwDlvZxScRC95k_2rsI2jyZ0z_5XaCsqDtbcdITGRphzGYa9Z2bRwen1jaPIYY8vOVdm3_4zkMqNZARi9VeBvgiuldKTeVnAFD2bxYPPX5rACwRPX_QFsgl7GJ-B6KYOALUph3wUKslmtvwlrjQM_aAz09JWKi-UbaUhE2XI67rw03oYu284uTAS3tgTjvLJGrLxn1aNWxpJrYoEo2wy4aHMNoBT53q7uRkgW6J5qkdppZnhEKjvxLPUWr4mFefY3RPEHU6 _QL31rcMGWkjMh2iKAZo3kbpSs7J0odKnQqPTy7HtAL1Qc-Nc3xyNMTYOh5wmf0W8GxpY2Vuc2VfdmVyc2lvbjoxCnNlcmlhbF9udW1iZXi6Mjg4ODYKbGljZW5zZWVfb3JnYW5pemF0aW9uOkgyTyBRd2IrbGFicmxpY2Vuc2VIX2VtYWlsOnRvbWsrbbWVldHVwcXdpa2xhYKBoMm8uYWkKbGljZW5zZWVfdXNlc9pZDoyODg4Ngppc19oMm9faW50ZXJuYWxfdXNI0mZhHNICmNyZWFOZWRfYnlfZW1haWw6cGF0cmlijay5tb3jhbkBoMm8uYWkKY3JIYXRpb25fZGF0ZToyMD E4LzExLzMwCnByb2R1Y3Q6RHJpdmVybGVzc0FJcmxpY2Vuc2VfdHlwZTpwaWxvdApleHBpcmF0aW9uX2RhdGU6M jAxOC8xMi8wNgo=
```

13. If you see this screen, congrats, you are now ready to run Driverless AI experiments :)



About H2O.ai

H2O.ai Overview

Company	Founded in Silicon Valley in 2012 Funded: \$75M Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
Products	<ul style="list-style-type: none">• H2O Open Source Machine Learning (14,000 organizations)• H2O Driverless AI – Automatic Machine Learning
Leadership	Leader in Gartner MQ Machine Learning and Data Science Platform
Team	120 AI experts (Kaggle Grandmasters, Distributed Computing, Visualization)
Global	Mountain View, London, Prague, India



H2O.ai HQ Mountain View



We're hiring data scientists. Ask Petra for more details.



H2O.ai
Prague Office



H2O.ai Product Suite



In-Memory, Distributed
Machine Learning Algorithms
with H2O Flow GUI



H2O AI Open Source Engine
Integration with Spark



Lightning Fast machine learning
on GPUs

DRIVERLESSAI

Automatic feature engineering,
machine learning and
interpretability

Steam

Secure multi-tenant H2O clusters

Worldwide Recognition in the H2O.ai Community

Open source
community

222 OF FORTUNE
THE 500
 H₂O

8 OF TOP 10
BANKS

7 OF TOP 10
INSURANCE COMPANIES

4 OF TOP 10
HEALTHCARE COMPANIES

Paying Customers



"H2O.ai's reference customers gave it the highest overall score for sales relationship and overall service and support" - Gartner MQ 2018

H2O.ai is a **Leader** in the 2018 Gartner Data Science and Machine Learning Platforms Magic Quadrant

- Technology leader with most completeness of vision
- Recognized for the mindshare, partner network and status as a **quasi-industry standard** for machine learning and AI
- H2O.ai customers gave the highest overall score among all the vendors for sales relationship and account management, customer support (onboarding, troubleshooting, etc.) and overall service and support

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms



Get the
Gartner
Magic
Quadrant
[here](#)

H₂O.ai

Partner Ecosystem



TRACE3



World Wide Technology, Inc.



HW Vendors

Cloud Providers

Strategic
Partners

Value Added
Resellers

System
Integrators

Data Stores



cloudera





H2O AI World London October 2018

Jo-fai (Joe) Chow @matlabulous · Oct 29

It is amazing to see 400 people trying out @h2oai #DriverlessAI at the same time!!!
#H2OAIWorld the #momentum is real – at London Hilton On Park Lane



H2O.ai Meetup Groups

Members

102,646

Groups

42

Countries

20



London Artificial Intelligence & Deep Learning

London, United Kingdom

Public group

7,410 members



Organized by
Ian Gomez and 2 others

Part of H2O Artificial Intelligence and Machine Learning – 42 groups



Share:



Contact Joe Chow
joe@h2o.ai

If you want to ...

- Give a talk about AI / machine learning use case (it is a great opportunity to promote your work)
- Host a joint meetup with H₂O.ai

A collage of various 360-degree selfie photos from around the world, including landmarks like the Louvre and the Eiffel Tower, and people in different settings like a subway station and a beach.

Meetup Tomorrow (Dec 5th)

Hilton Amsterdam Centraal Station (6-9pm)

Wednesday, December 5, 2018

5 DEC

Driverless AI Day Celebration: New Features in Driverless AI and H2O-3

Hosted by Jo-fai Chow
From Amsterdam Artificial Intelligence & Deep Learning
Public group ?

You're going 26 people going

✓ X

Share: [f](#) [t](#) [in](#) [d](#)

Organizer tools ▾

Happy #DriverlessAI Day!



Wednesday, December 5, 2018
6:00 PM to 9:00 PM
[Add to calendar](#)

DOUBLETREE BY HILTON AMSTERDAM CENTRAAL STATION
Oosterdokstraat 4 · Amsterdam
How to find us
Look for rooms Manchester 1 + 2 & Birmingham 1 + 2

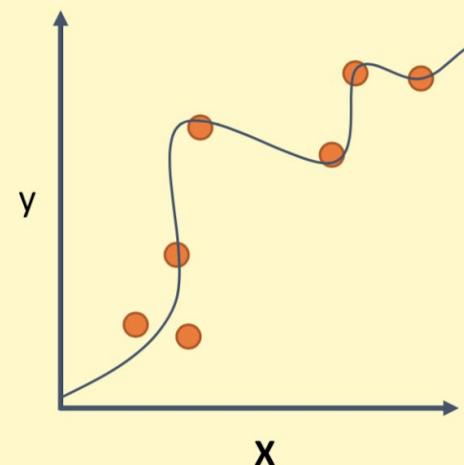


Driverless AI Overview

Supervised Learning

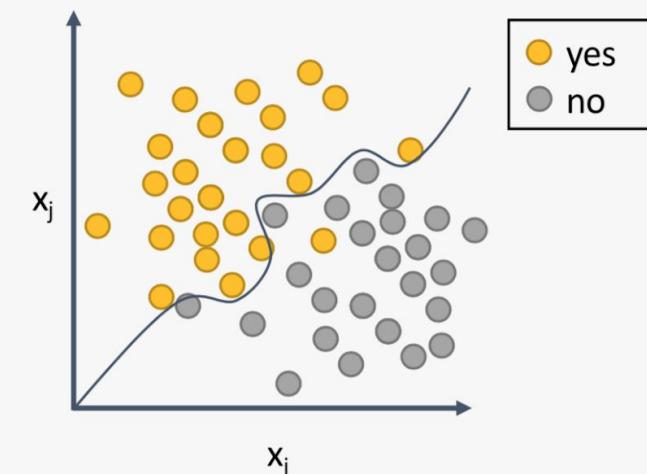
Regression:

How much will a customer spend?



Classification:

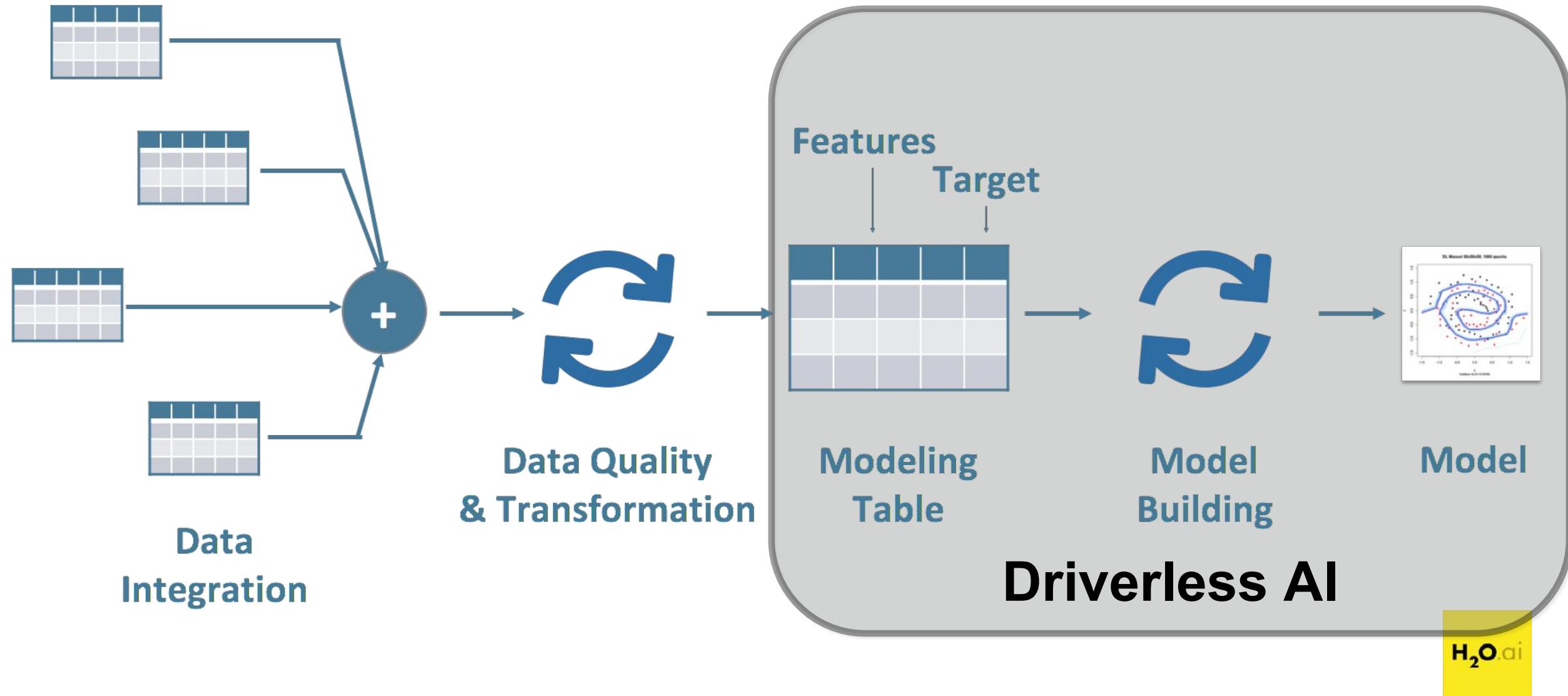
Will a customer churn?



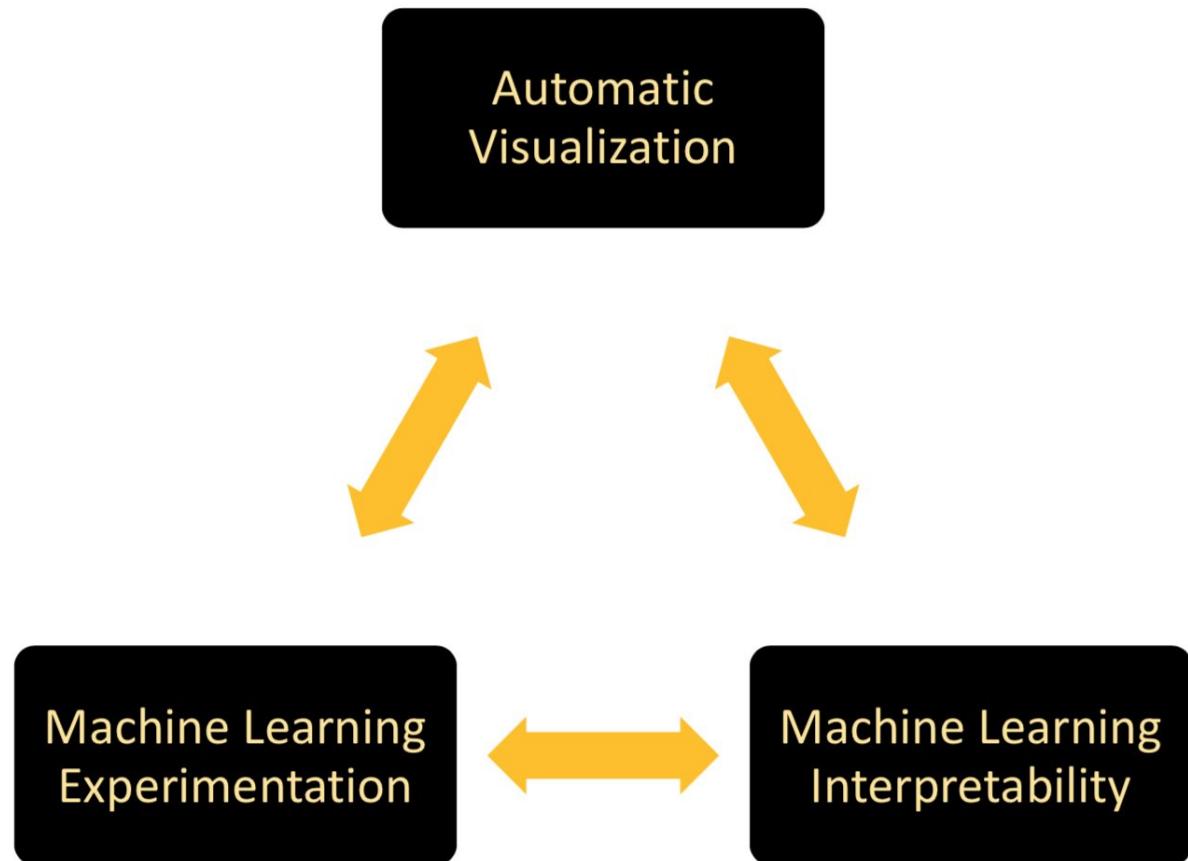
Problems Addressed by Driverless AI

- Supervised Learning
 - Regression
 - Classification
- Tabular Structured Data
 - Numeric
 - Categorical
 - Time / Date
 - Text
 - Missing Values
- Identically and Independently Distributed (iid) rows
- Time-series
 - Single time-series
 - Grouped time-series
 - e.g. Store - Department - Item
 - Time-series with gaps between training and test set to account for time to deploy

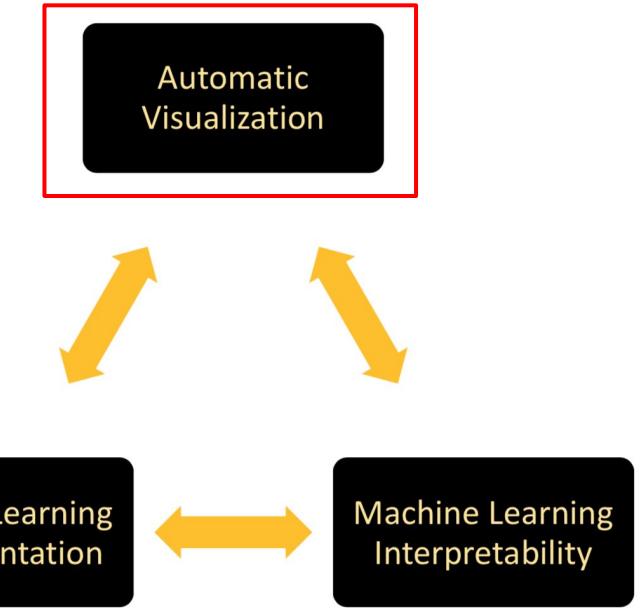
Driverless AI: Automates Data Science and ML Workflows

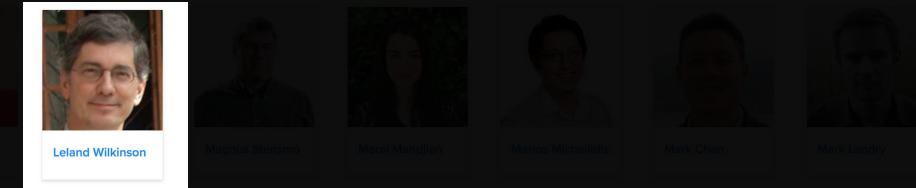
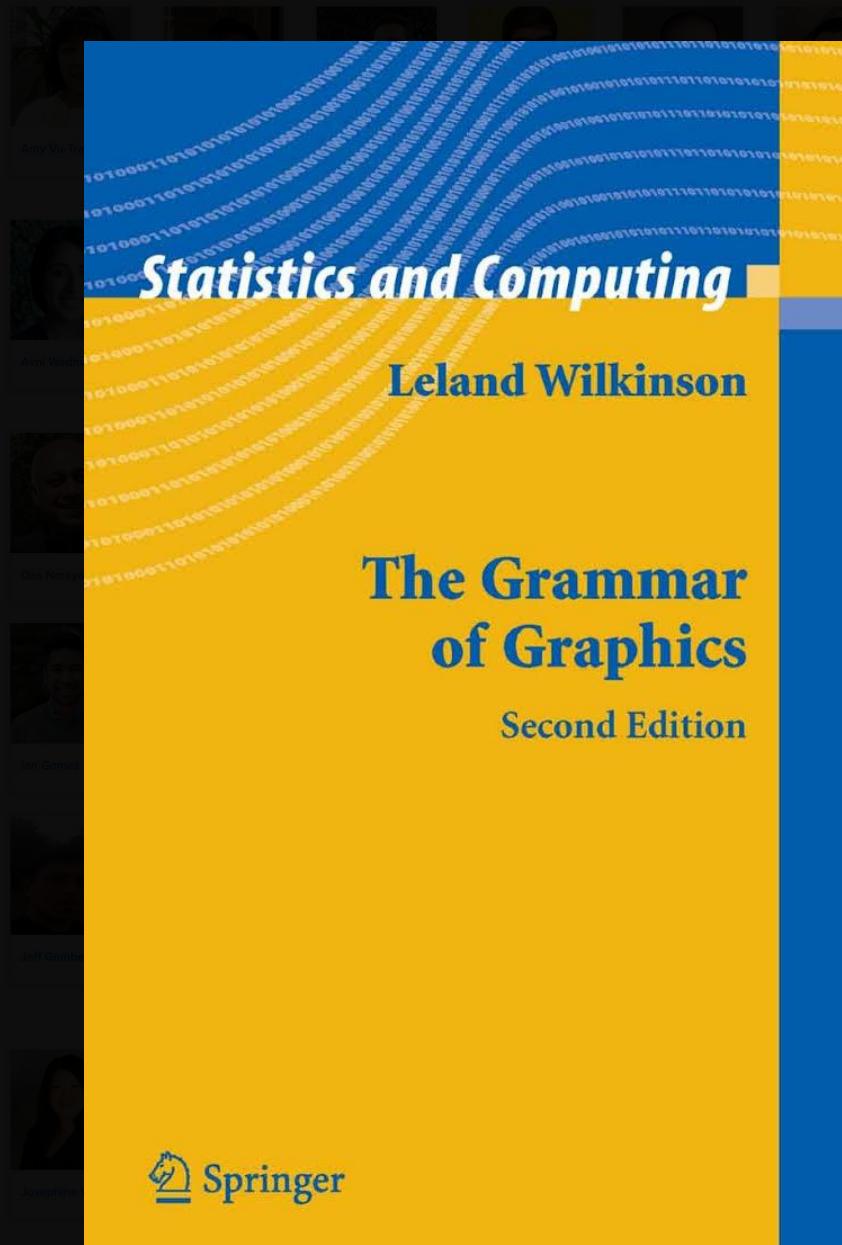


Driverless AI Components

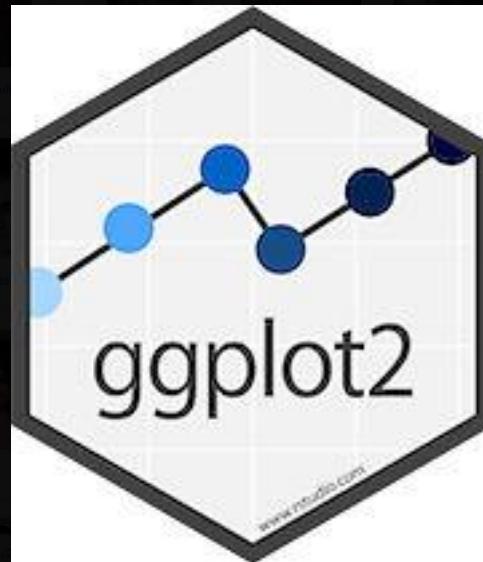


Driverless AI: Automatic Visualization





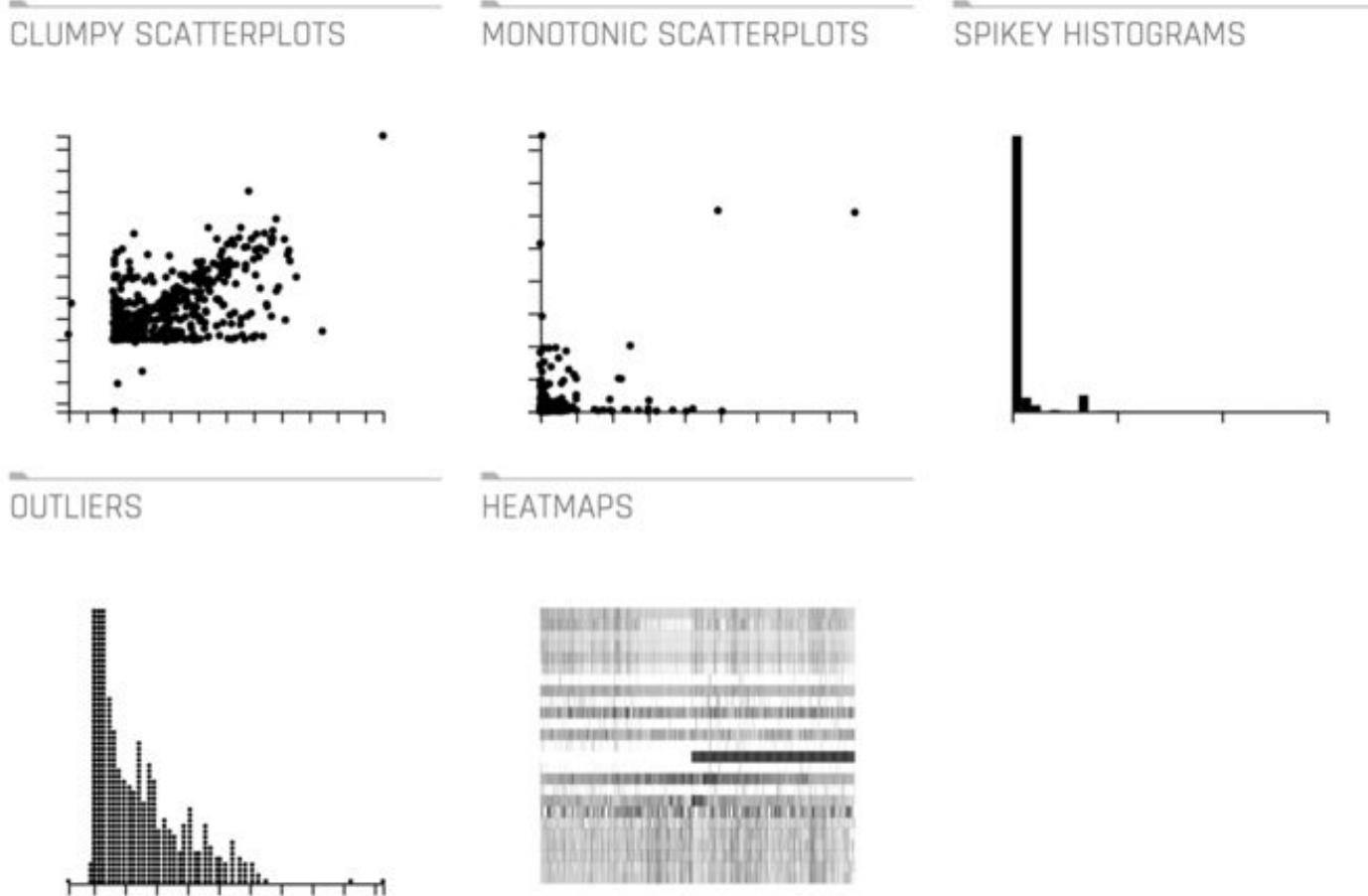
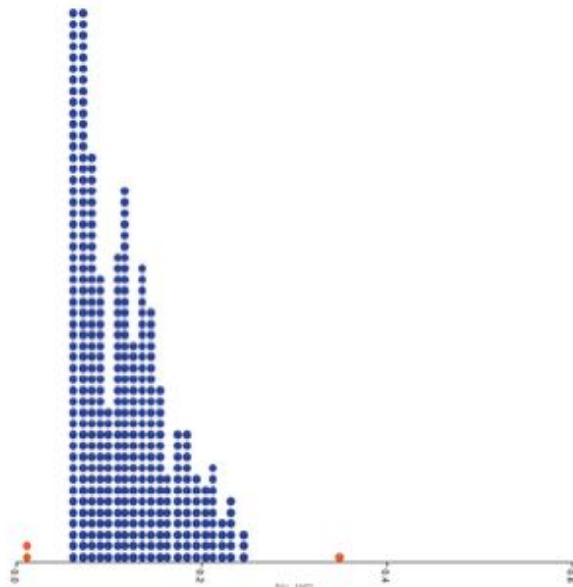
Origin of R Package `ggplot2`



Automatic Visualization

H2O.ai

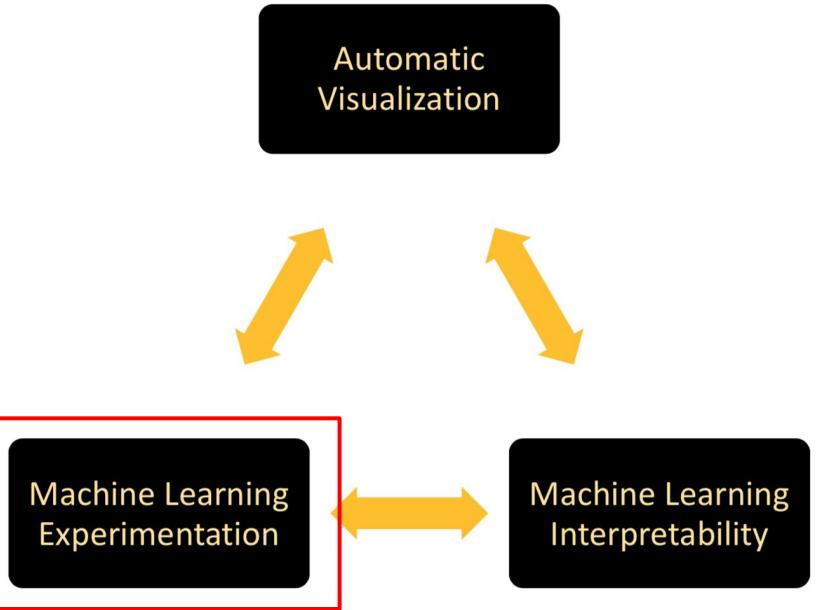
Automatic Scagnostics and other visualizations to generate the most relevant visualizations for each dataset



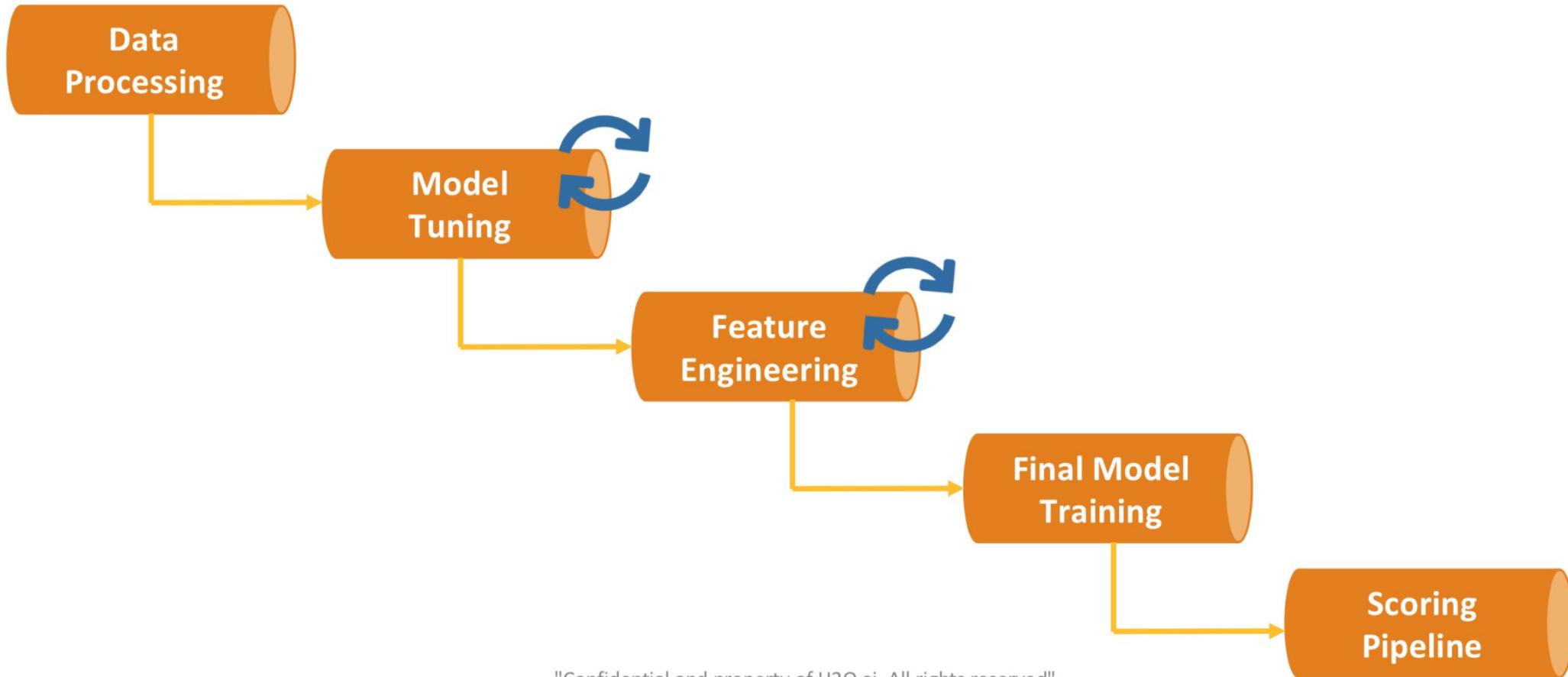
"Confidential and property of H2O.ai. All rights reserved"

H₂O.ai

Driverless AI: Machine Learning Experimentation

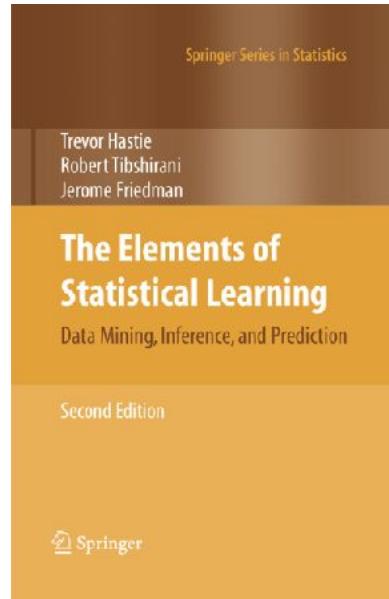
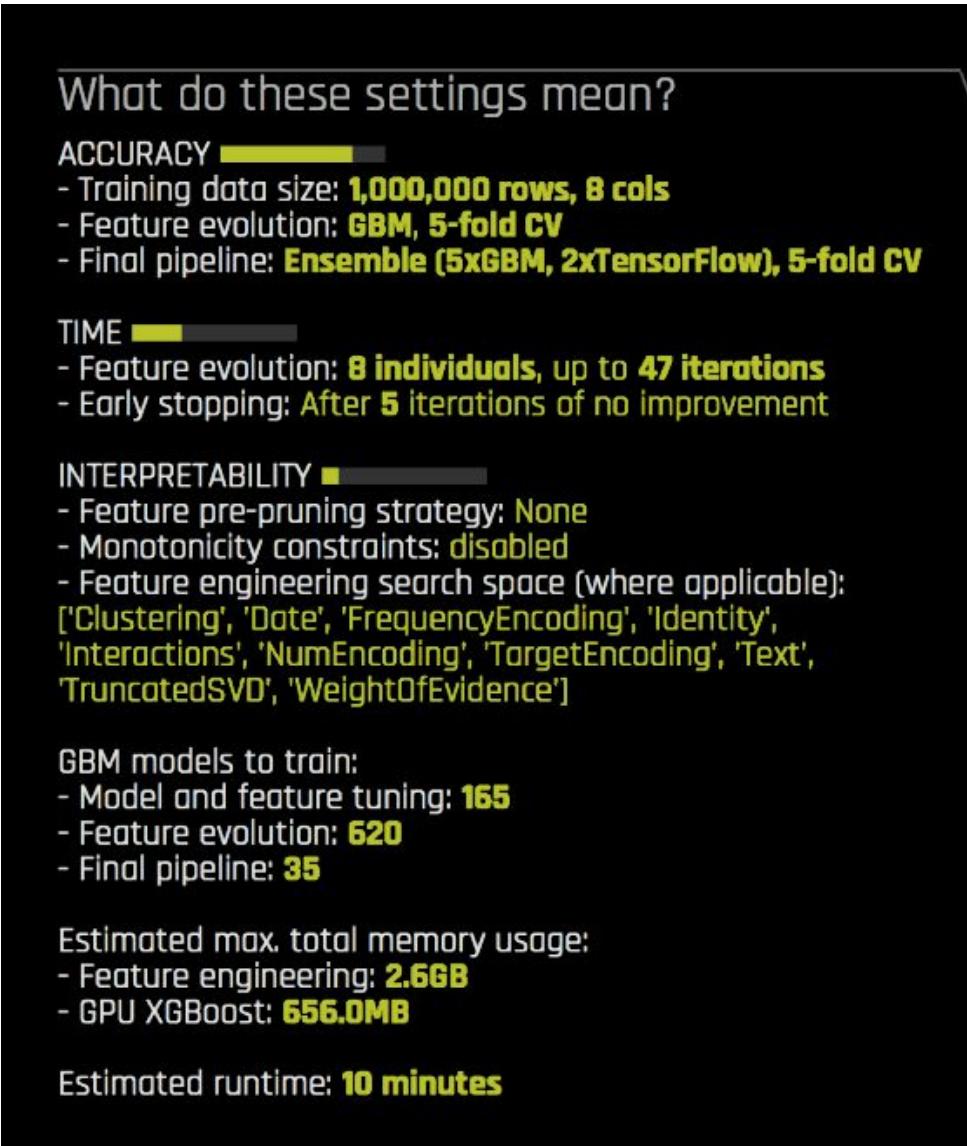


Driverless AI: Machine Learning Workflow



"Confidential and property of H2O.ai. All rights reserved"

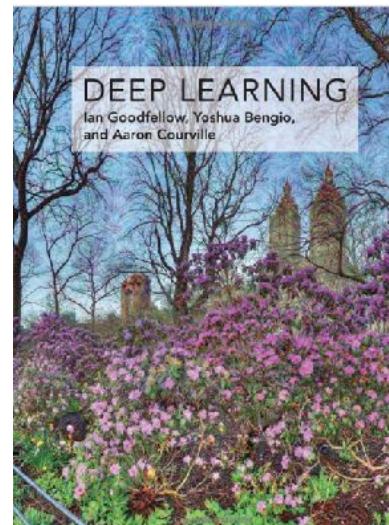
Statistical Learning vs Deep Learning - We Do Both!



GLM/CART/RF/GBM/XGBoost
K-Means/PCA/SVD

Typically better for structured data
(CSV, SQL, Transactional)

<https://web.stanford.edu/~hastie/Papers/ESLII.pdf>



TensorFlow Deep Learning

Typically better for unstructured data
(Images, Video, Audio, Text)

<http://www.deeplearningbook.org>



Kaggle Grandmasters (and their Highest Rank)

 **113**
Grandmasters

 **980**
Masters

 **3,339**
Experts

 **46,135**
Contributors

 **33,242**
Novices

About 80,000 Kagglers

H₂O Team

H₂O.ai

48th

1st

33rd

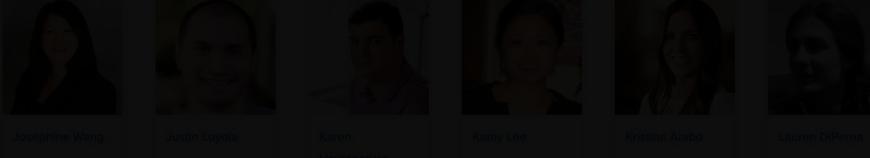
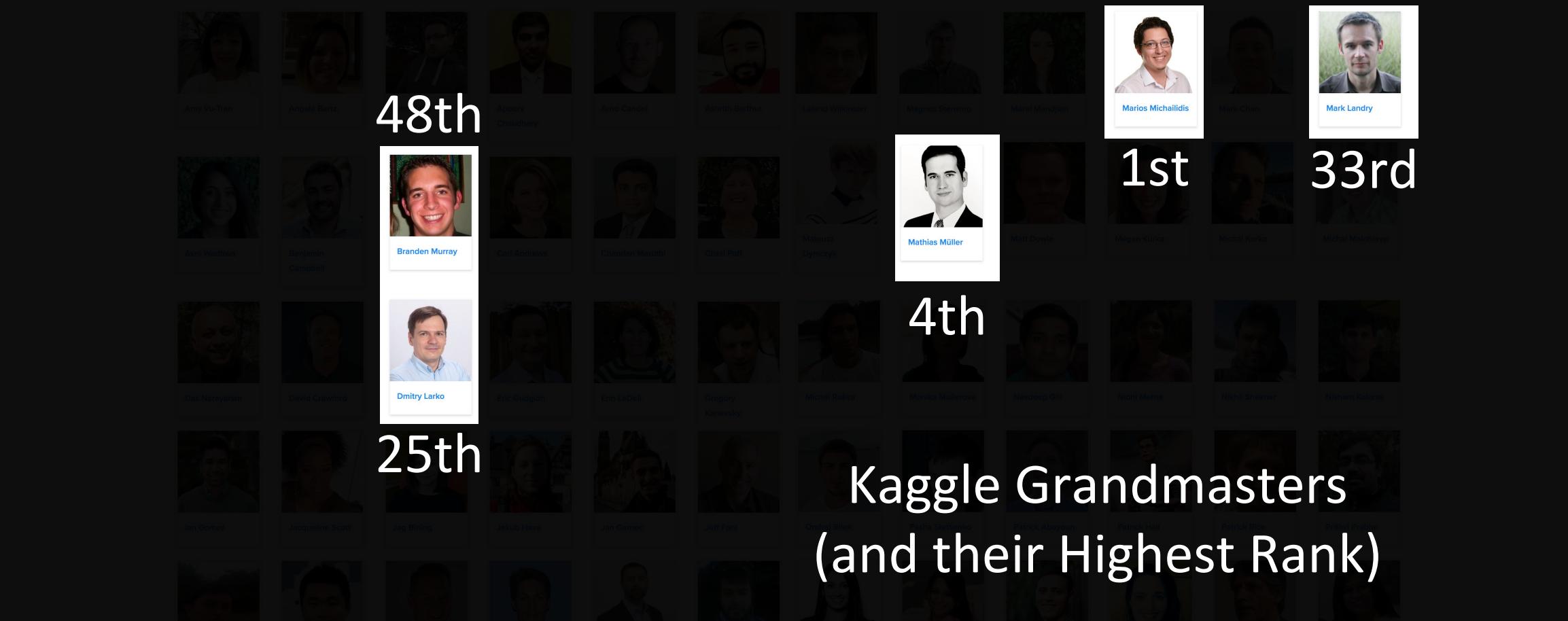
Branden Murray

Mathias Müller

Dmitry Larko

25th

4th



13th



Amy Vu-Tran

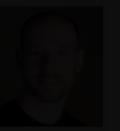


Angela Barz

48th



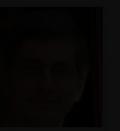
Apoorv Chaudhary



Arno Candel



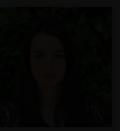
Ashwin Banbur



Leland Wilkinson



Magnus Stensmo



Maral Mandjidian



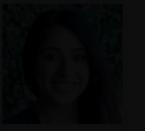
Marios Michailidis



Mark Chan



Mark Landry



Avni Wadhwa



Benjamin Campbell



Branden Murray



Carl Andrews



Chandan Manocha



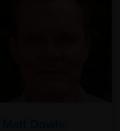
Chen Poff



Mateusz Dymczyk



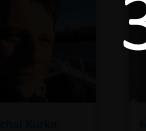
Mathias Müller



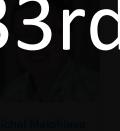
Matt Dowle



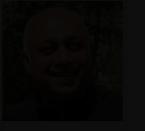
Megan Kurka



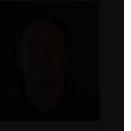
Michael Kurka



Michal Malohlava



Das Narayanan



David Crawford



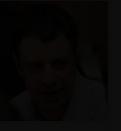
Dmitry Larko



Eric Quiggin



Erin LeDell



Gregory Kanhevsky



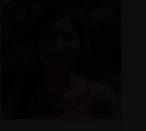
Michal Raksa



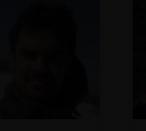
Monika Müllerova



Navdeep Gill



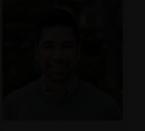
Nishi Mehta



Nisha Shukhar



Nishant Kalonia



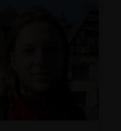
Ian Gomez



Jacqueline Scott



Jia Bining



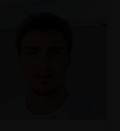
Jinhui Han



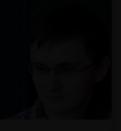
Jan Gamec



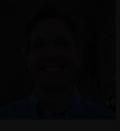
Jeff Feltl



Ondrej Blazek



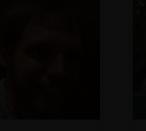
Pasha Slatoustov



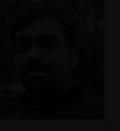
Patrick Abeyoum



Patrick Hall



Patrick Rice



Prithvi Prabhu



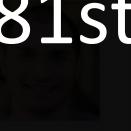
Jo-Fai Chow

Hoping to get closer to them at some point ...

181st



Josephine Wang



Justin Loyola



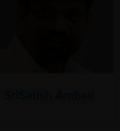
Karen Heyerophyton



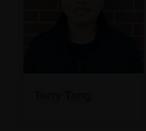
Kathy Lee



Kristina Arabo



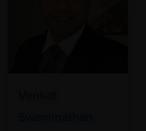
Lauren DiPerna



Srinath Ambati



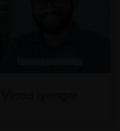
Terry Tang



Tom Kraljevic



Venkat Swaminathan



Venkatesh Yeduv



13th



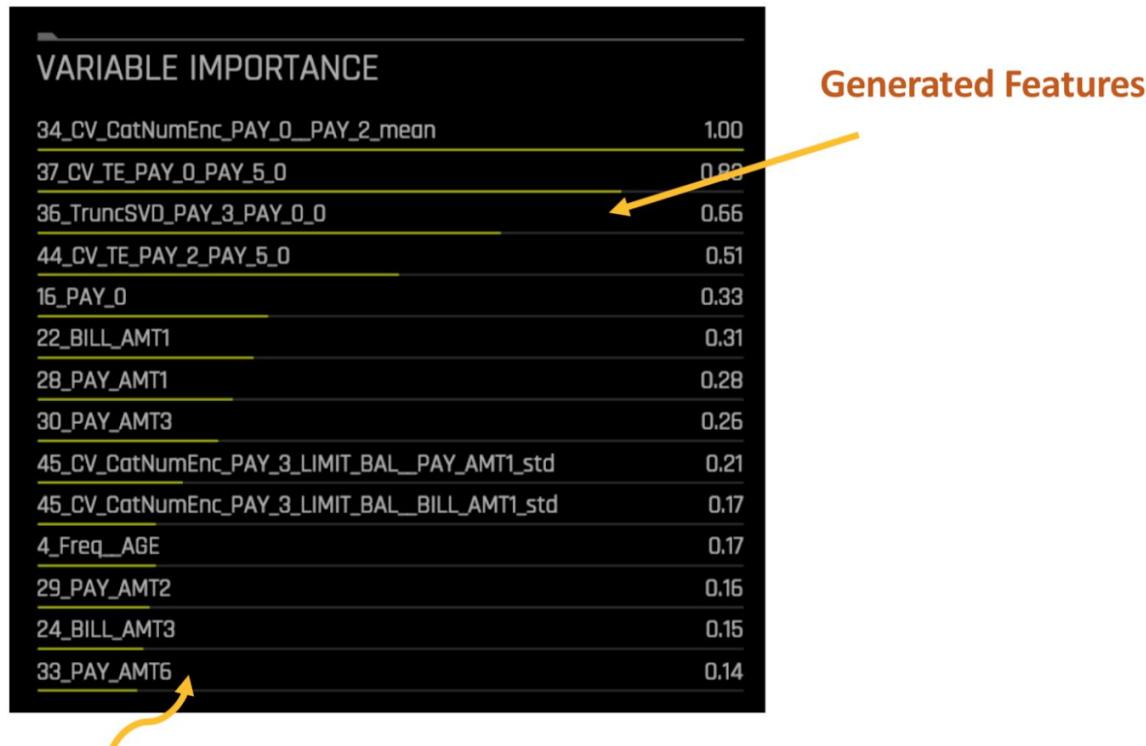
Wien Pham



Wendy Wong

H₂O TeamH₂O.ai

Automatic Feature Engineering: Kaggle Grandmasters' Recipes



Feature Transformations

- Automatic Text Handling
- Frequency Encoding
- Cross Validation Target Encoding
- Truncated SVD
- Clustering and more

Original Features

Copyright 2018 H2O.ai Inc. All rights reserved.

Accuracy

- Automatic feature engineering to increase accuracy - AlphaGo for AI
- Automatic Kaggle Grandmaster recipes in a box for solving wide variety of use-cases
- Automatic machine learning to find and tune the right ensemble of models

Driverless AI: top 5% in Amazon Kaggle competition

Driverless AI produces feature engineering pipeline (“more columns”) for downstream use

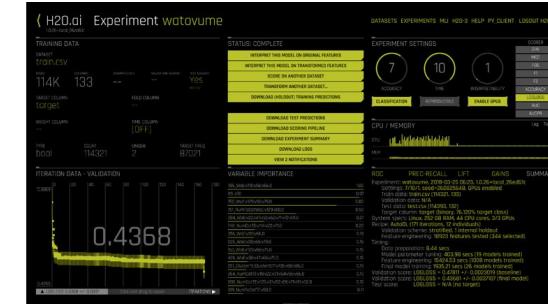


Amazon.com - Employee Access

Predict an employee's access needs
\$5,000 · 1,687 teams · 4 years ago

Driverless AI: 80th place (out of 1687 - top 5%)

Driverless AI: Top-10 in BNP Paribas Kaggle competition



single run, fully automated: 2h on DGX Station! 6h on PC

BNP Paribas Cardif Claims Management

Can you accelerate BNP Paribas Cardif's claims management process?

\$30,000 · 2,926 teams · 2 years ago

Submission and Description	Private Score	Public Score
sub.csv 2 minutes ago by Arno Candel 940b9f 7/10/1 cv 0.4354 finished after 172 iters	0.42945	0.43156

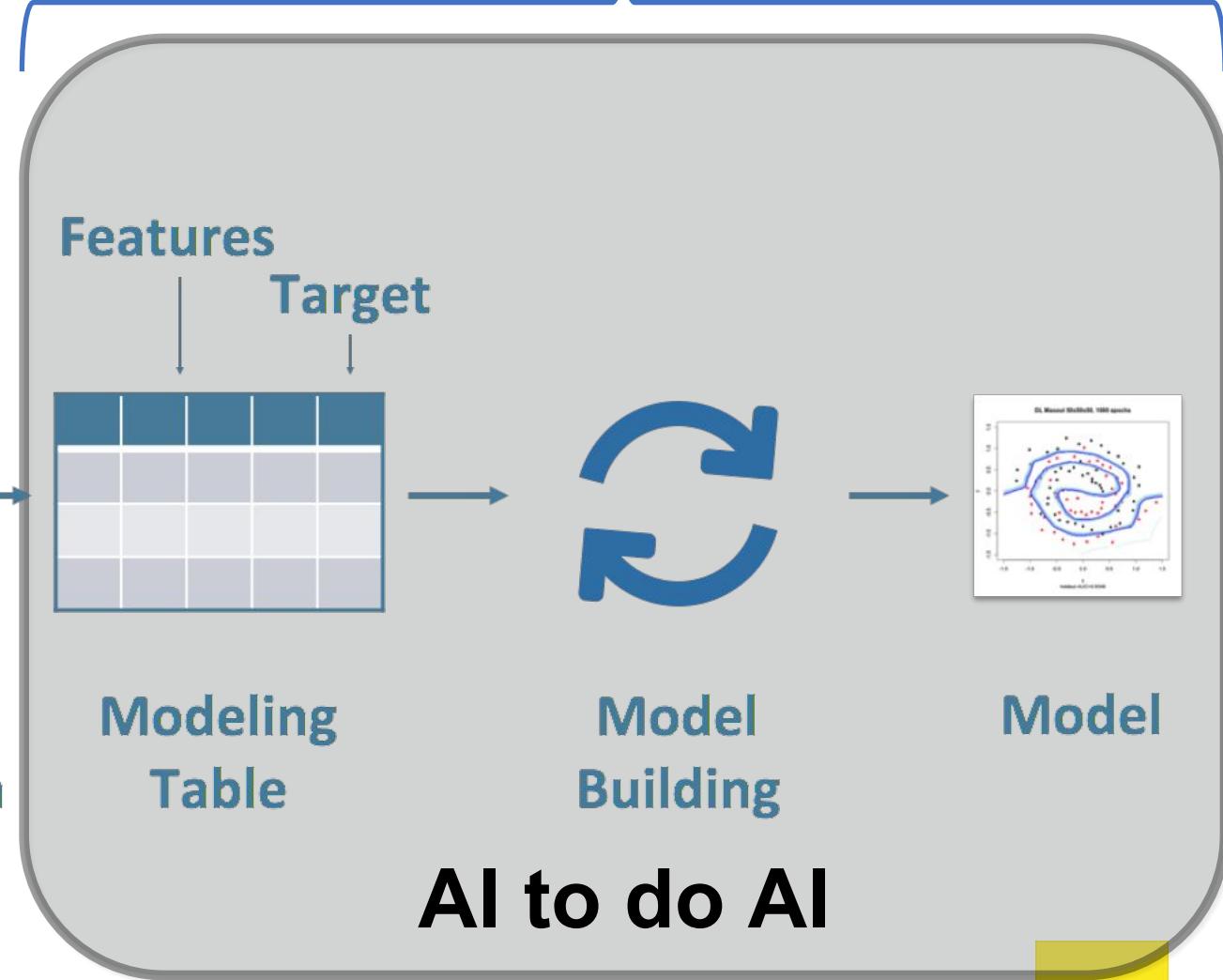
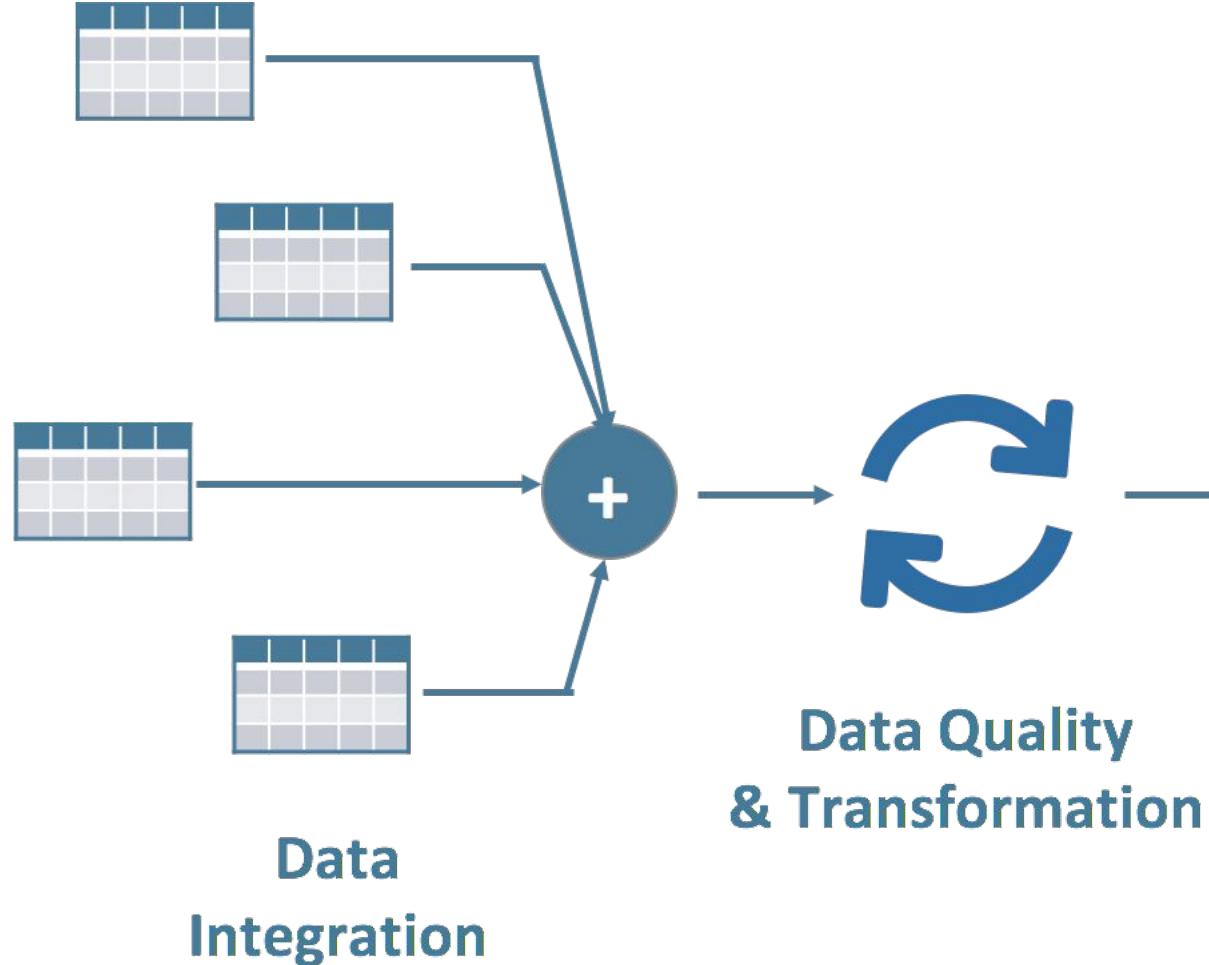
Driverless AI: 10th place in private LB at Kaggle (out of 2926)

2 months for Grandmasters — 2 hours for Driverless AI

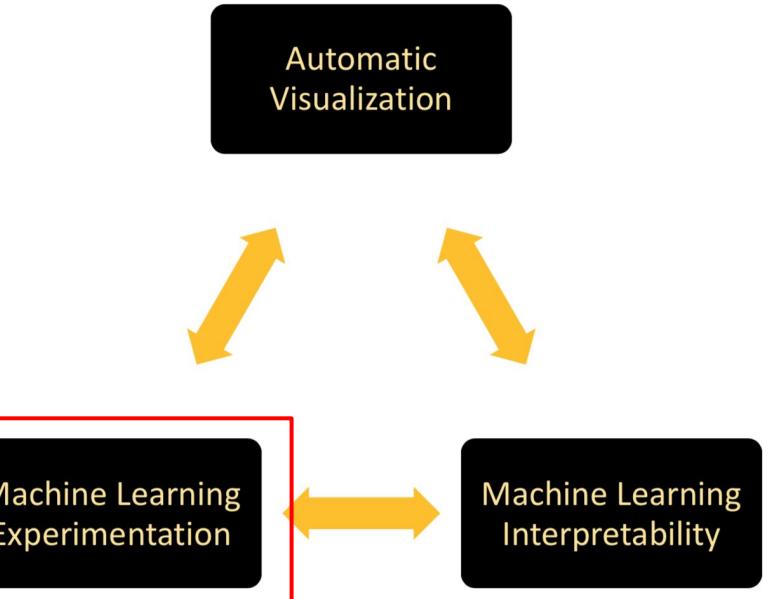
#	△pub	Team Name	Kernel	Team Members	Score ⚡	Entries	Last
1	—	Dexter's Lab			0.42037	198	2y
2	—	escalated chi			0.42079	162	2y
3	—	Exploding Kittens			0.42182	124	2y
4	—	Branden Nickel utility			0.42259	251	2y
5	—	the flying burrito brothers			0.42450	264	2y
6	—	n_m			0.42535	4	2y
7	—	PAFY			0.42557	310	2y
8	—	KAME			0.42688	121	2y
9	—	Jack (Japan)			0.42744	22	2y
10	▲ 1	Dmitry & Bohdan			0.43000	192	2y
11	▲ 1	Li-Der			0.43096	56	2y
12	▲ 2	BK3M2PRS			0.43089	338	2y
13	—	x2x4x8			0.43107	55	2y
14	—	Frenchies			0.43146	134	2y
15	▲ 1	Ains			0.43168	55	2y
16	▼ 1	maze runners			0.43262	164	2y
17	—	BLR-2			0.43313	129	2y
18	▲ 3	no one			0.43317	88	2y

H₂O.ai

Deployment: Auto Generated Pipelines



Driverless AI: Machine Learning Interpretability



Interpretability

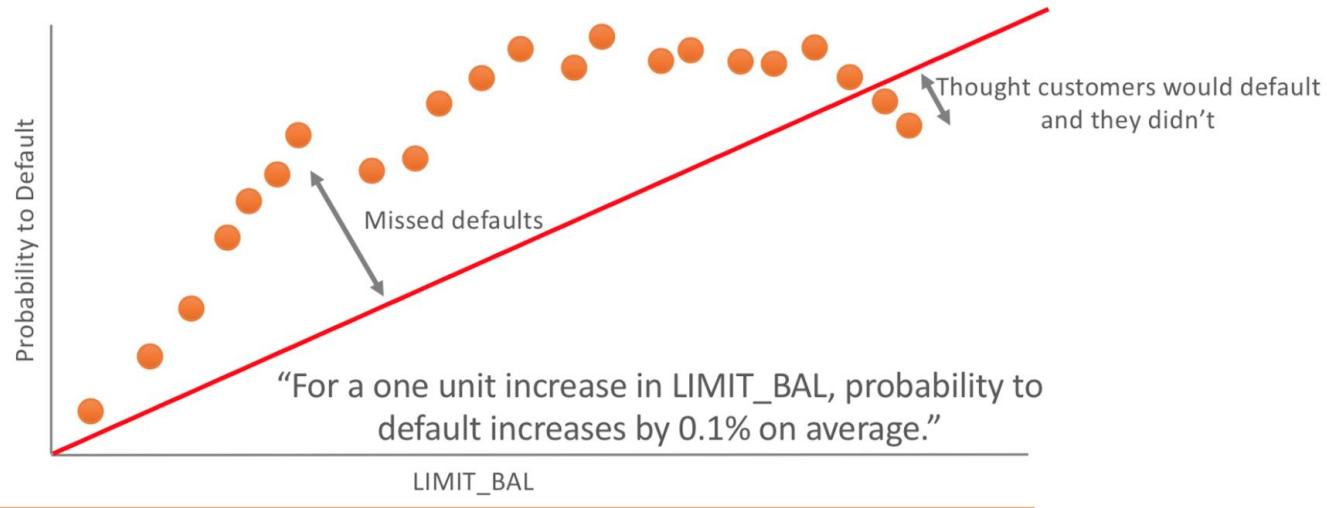
- Interpretability for debugging, not just for regulators
- Get reason codes and model interpretability in plain english
- K-Lime, LOCO, partial dependence and more



Linear Models vs. Machine Learning

Linear Models
Exact explanations for approximate models.

Machine Learning
Approximate explanations for exact models.



Approximate Reasoning

Local Reason Codes

k-LIME Local Attributions	Variable	with value	is associated with	DEFAULT_PAYMENT_NEXT_MONTH
Top Positive Local Attributions				
	PAY_1	2	increase of	0.34
	PAY_5	2	increase of	0.06
	PAY_3	2	increase of	0.06
Skipped 8 additional attributions, click to view all ...				
Top Negative Local Attributions				
	PAY_AMT3	3000	decrease of	0.01
	BILL_AMT5	24930	decrease of	0.01
	BILL_AMT1	21024	decrease of	0.01

Why will someone Default?

The fact that they haven't paid in 2 months increases their likelihood by 34%

Their Pay Amount is \$3,000 **decreases their likelihood by 1%**

Driverless AI: Real-World Use Cases

H2O Driverless AI Delivers **Value** in Every Industry



Financial Services

+6%
Accuracy

Matched 10 years of Machine
Learning Expertise



ArmadaHealth®

Healthcare

Near
perfect
scores

Increased customer
satisfaction



Marketing

2.5X
Performance

Outperforms alternative
digital marketing



Manufacturing

1 month
savings

Accurately predicting
supply chain

www.h2o.ai/customer-stories/



Predicting dosing levels for new medications with Driverless AI



Improving manufacturing process with Driverless AI



Driving Marketing Performance with Machine Learning

Hands-on Experiment: Credit Card Example

Credit Card Example

- **Dataset:**
 - information on default payments, demographic factors, credit data, history of payment, etc.
 - Source: www.kaggle.com/uciml/default-of-credit-card-clients-dataset
- **File System (Cloud Instance on Qwiklabs):**
 - /data/Kaggle/CreditCard/CreditCard-train.csv (for training models)
 - /data/Kaggle/CreditCard/CreditCard-test.csv (for making new predictions)
- **Our Goal:**
 - Predict whether someone will default on their credit card payment.

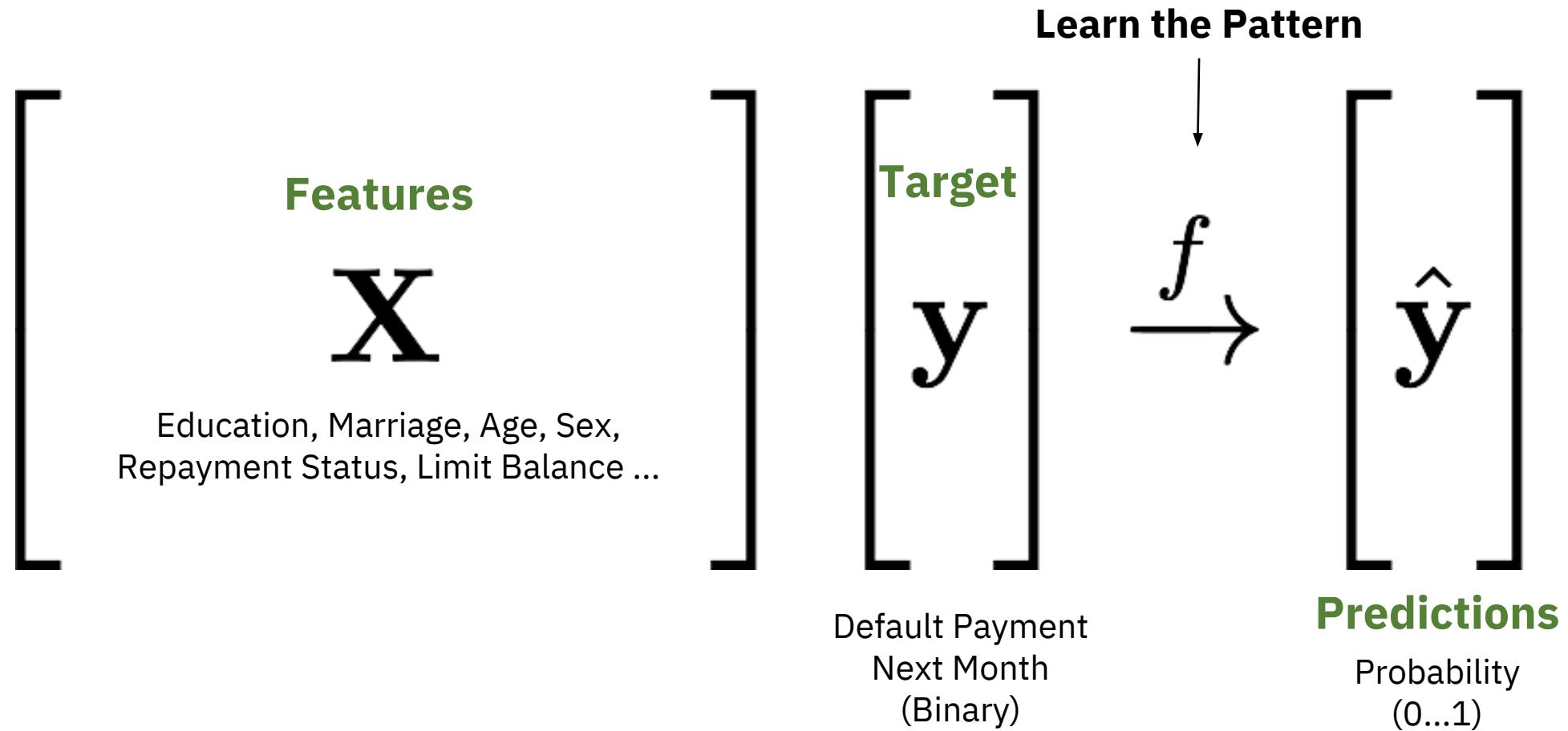
Credit Card Example

Column Name	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_x {0,2,3,4,5,6}	Repayment status in August, 2005 – April, 2005
BILL_AMTx {1, ..., 6}	Amount of bill statement in September, 2005 – April, 2005 (NT dollar)
PAY_AMTx {1, ..., 6}	Amount of previous payment in September, 2005 – April, 2005 (NT dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

Credit Card Example

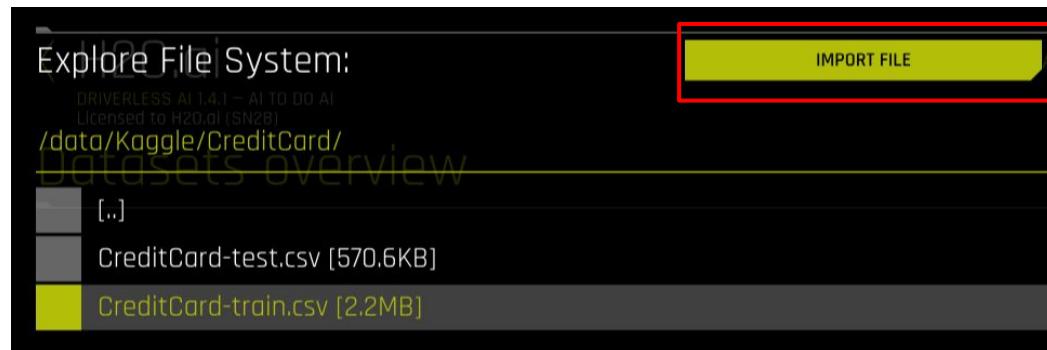
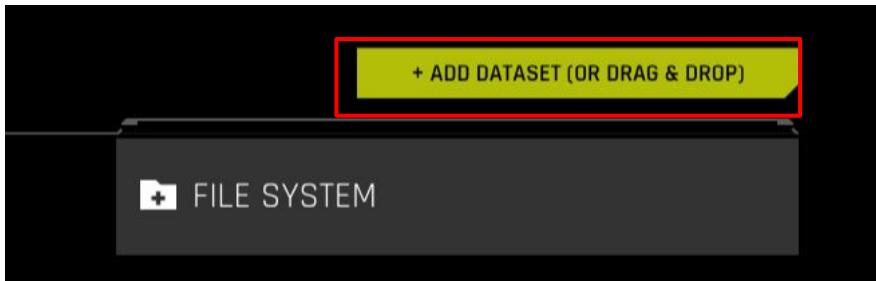
LIMIT_BAL	EDUCATION	AGE	PAY_1	PAY_2	BILL_AMT1	PAY_AMT1	DEFAULT_PAYMENT NEXT MONTH
120,000	university	26	-1	2	2,682	0	1
90,000	university	34	0	0	29,239	1,418	0
50,000	university	37	1	0	46,990	2,000	0
50,000	university	37	2	0	8,617	2,000	0
50,000	graduate	57	3	0	64,400	2,500	0

Learning from Credit Card Data



**Hands-on Experiment:
It is time to get our hands dirty!**

Credit Card Data



Datasets overview					
	SIZE	ROWS	COLUMNS	STATUS	
CreditCard-test.csv /data/Kaggle/CreditCard/	574KB	6000	25	[Click for Actions]	
CreditCard-train.csv /data/Kaggle/CreditCard/	2MB	23999	25	[Click for Actions]	

Credit Card Data

2. Click on **CreditCard-train.csv**.

You will see four available options:

- a. Details
- b. Visualize
- c. Predict
- d. Delete



3. Let's start with **Details**

Dataset Details: CreditCard-train.csv

DRIVERLESS AI 1.4.1 - AI TO DO AI
Licensed to H2O.ai (SN2B)

DATASETS EXPERIMENTS ML AUTOVIZ HELP PY_CLIENT MOJO2-RUNTIME N

DATASET ROWS

LOGOUT H2O X

Start typing to filter out columns

+ ADD DATASET (OR DRAG & DROP)



Dataset Details: CreditCard-train.csv

/data/Kaggle/CreditCard/CreditCard-train.csv

DATASETS EXPERIMENTS ML AUTOVIZ HELP PY_CLIENT MOJO2-RUNTIME **DATASET OVERVIEW** LOGOUT H2O X

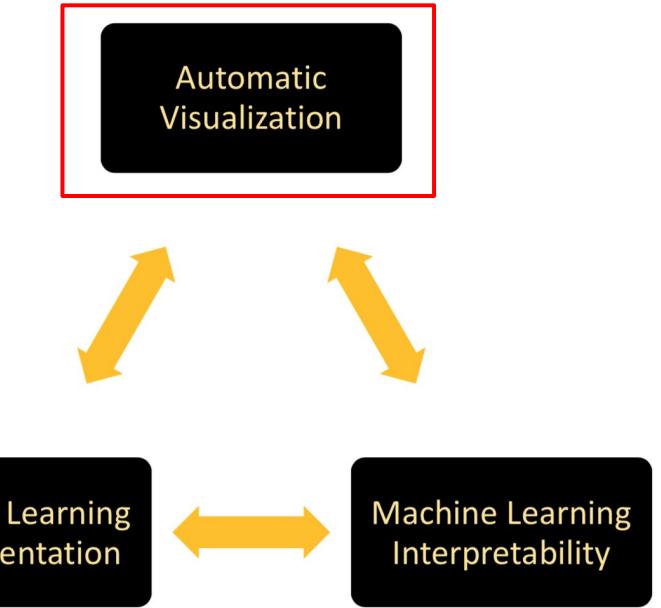
DRIVERLESS AI (AI) - AI TO DO AI
Licensed to H2O.ai (SN2B)

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	689	0	0	0
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100
5	50000	1	2	1	57	-1	0	-1	0	0	0	8517	5670	35935	20940	19146	19131	2000	36681	10000	9000
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007
11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	2870	1000	6500	6500	6500	6500
14	70000	1	2	2	30	1	2	2	0	0	2	65802	57369	55701	66782	36137	36894	3200	0	3000	3000
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	0	632	
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44005	46905	46012	2007	3582	0	3601

1 2 3 4 5 6 7 8 9 10 ... 1044 Next Page

Rows 1-23 of 23999 total

Hands-on Experiment: Visualize



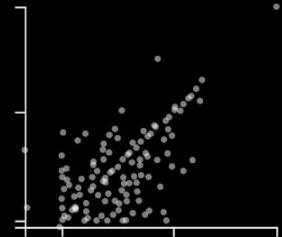
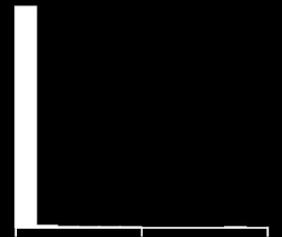
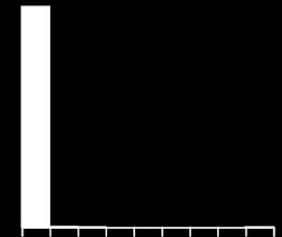
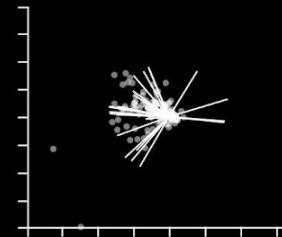
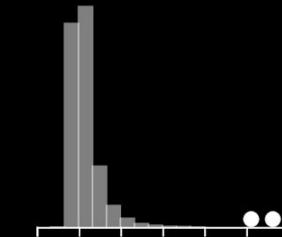
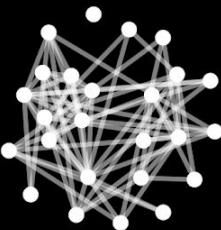
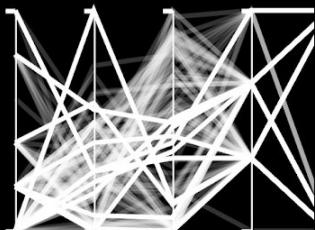
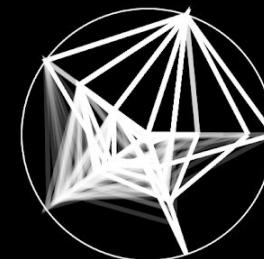
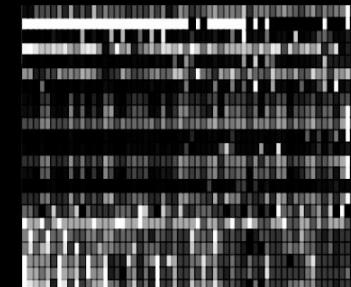
Datasets → CreditCard-train.csv → Visualize

The screenshot shows the H2O.ai interface. At the top, there is a navigation bar with links: DATASETS, EXPERIMENTS, MLI, AUTOVIZ, HELP, PY_CLIENT, MOJO2-RUNTIME, **MESSAGES[3]**, LOGOUT, and H2OAI. Below the navigation bar, there is a search bar with the placeholder "Search datasets, experiments, messages". On the left, there is a sidebar with a back arrow, the H2O.ai logo, and the text "DRIVERLESS AI 1.4.1 – AI TO DO AI Licensed to H2O.ai (SN28)". The main area is titled "Visualizations". A table lists a single dataset entry:

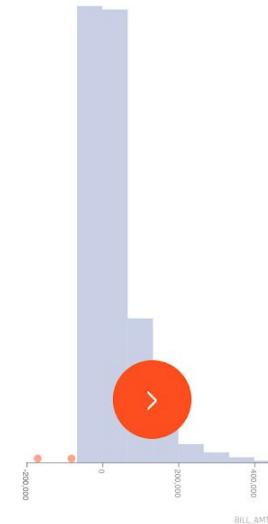
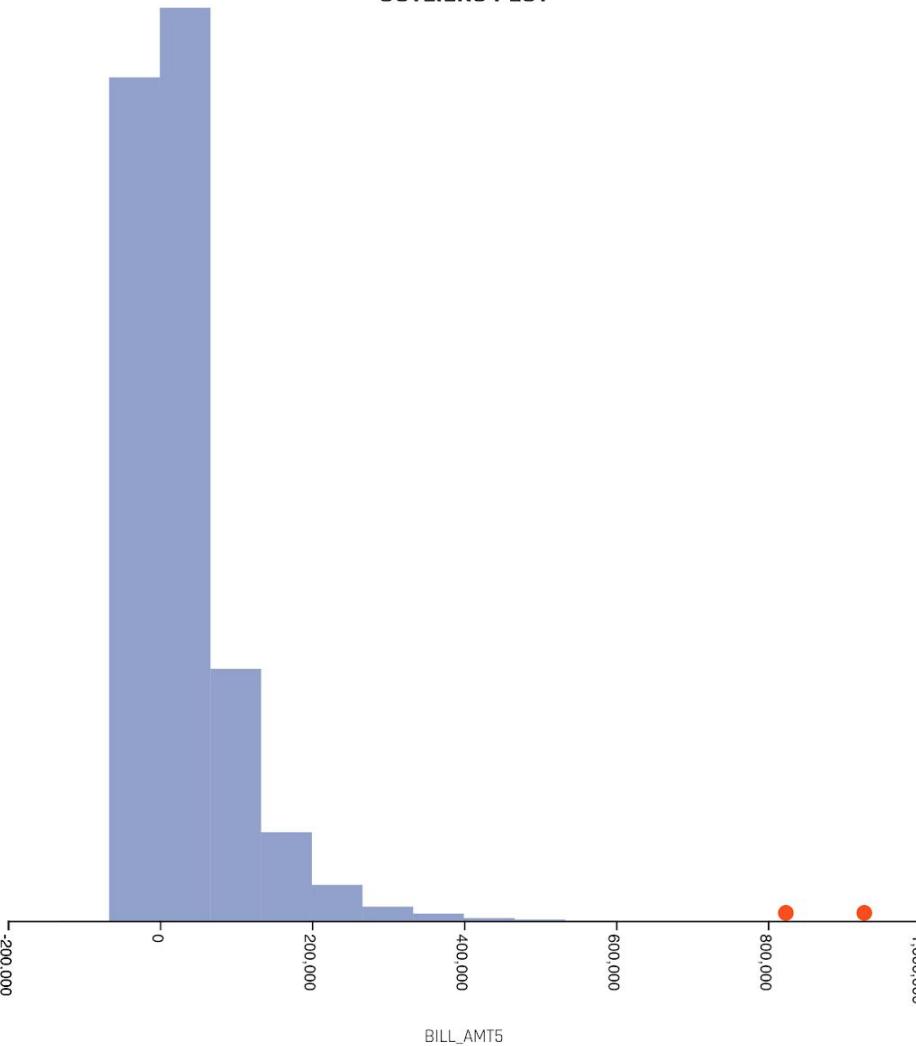
Dataset	Message	Status	Time
CreditCard-train.csv	Visualization ready	Done	00:00:03

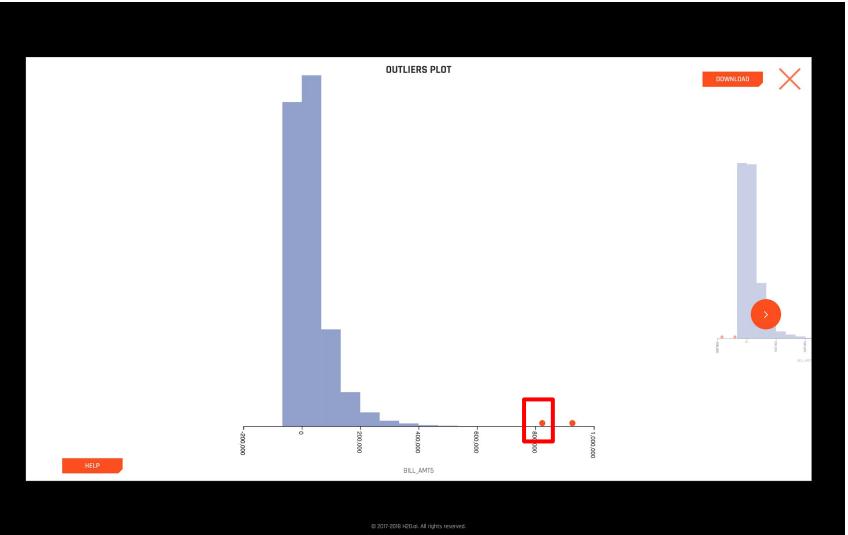
Below the table, there are buttons for "+ NEW VISUALIZATION", "SELECT", and "SORT BY ▾".

Visualizations for: CreditCard-train.csv

[CORRELATED SCATTERPLOTS](#)[SKEWED HISTOGRAMS](#)[GAPS HISTOGRAMS](#)[BIPLOTS](#)[OUTLIERS](#)[CORRELATION GRAPH](#)[PARALLEL COORDINATES PLOT](#)[RADAR PLOT](#)[DATA HEATMAP](#)

OUTLIERS PLOT

[DOWNLOAD](#)[HELP](#)

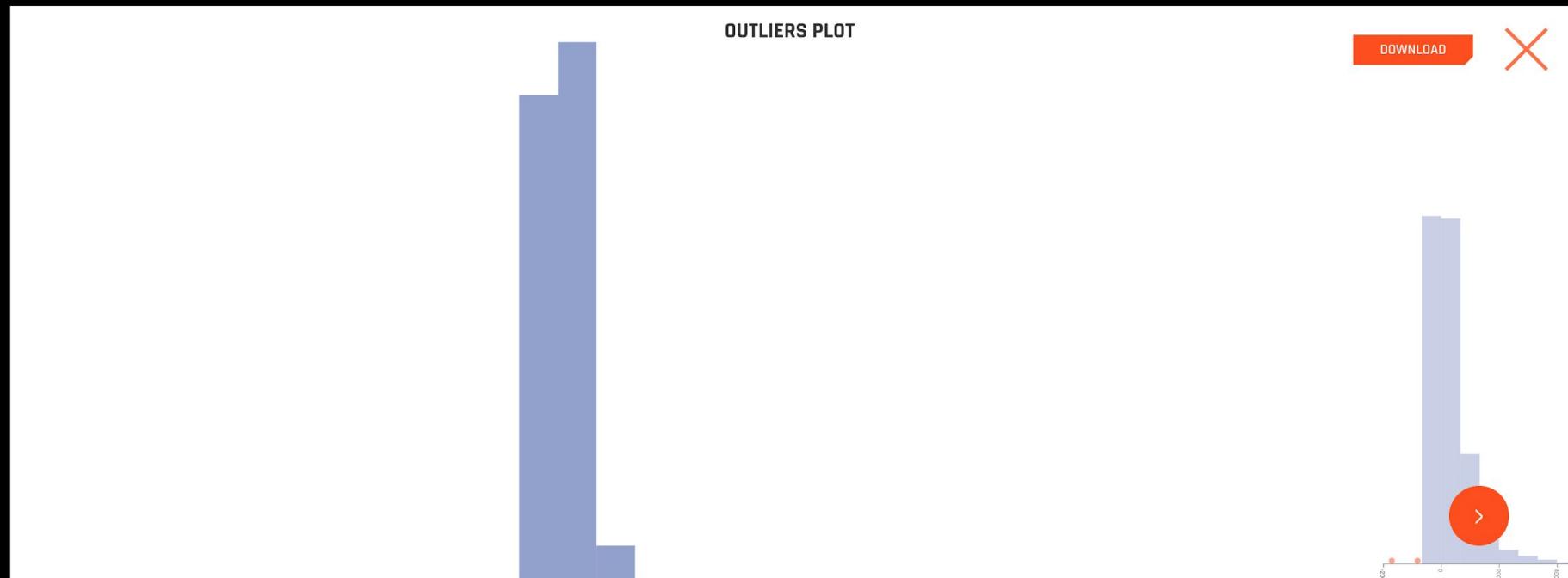
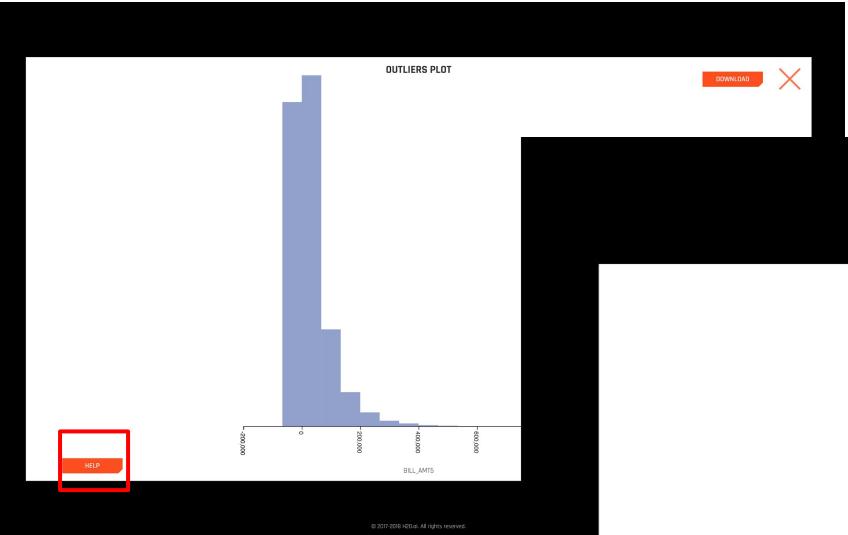


X

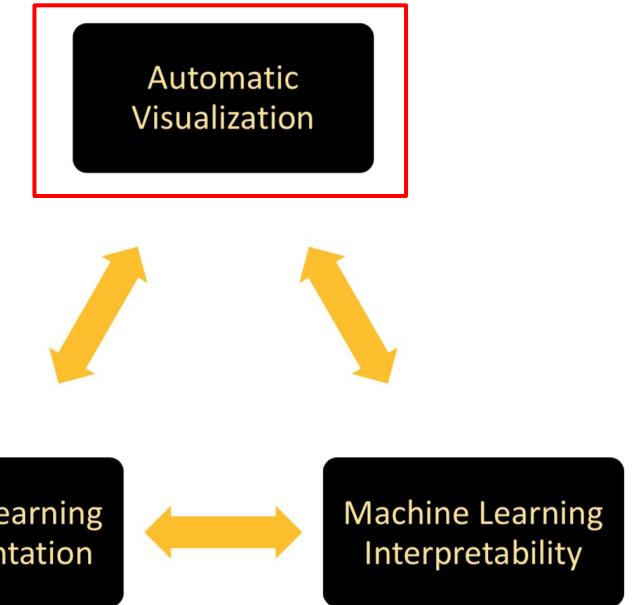
ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	P/
20893	550000	1	1	1	2	35	2	2	2	2	2	0	539092	552234	565550	572805	823540	501370

Rows 1-1 of 1 total

© 2017-2018 H2O.ai. All rights reserved.

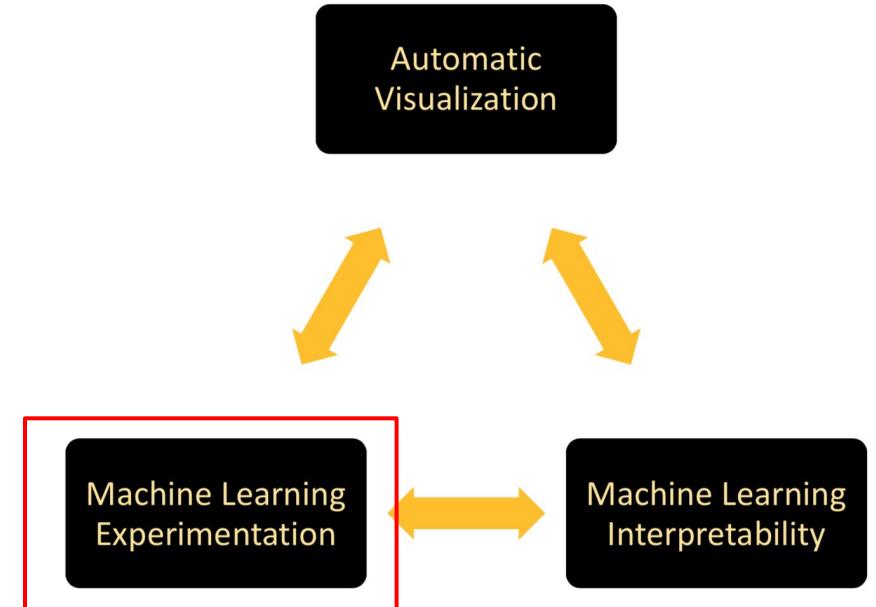


Variables with anomalous or outlying values are displayed as red points in a dot plot. Dot plots are constructed using an algorithm in Wilkinson, L. (1999). "Dot plots." *The American Statistician*, 53, 276–281. Not all anomalous points are outliers. Sometimes the algorithm will flag points that lie in an empty region (i.e., they are not near any other points). You should inspect outliers to see if they are miscodings or due to some other mistake. Outliers should ordinarily be eliminated from models only when there is a reasonable explanation for their occurrence.

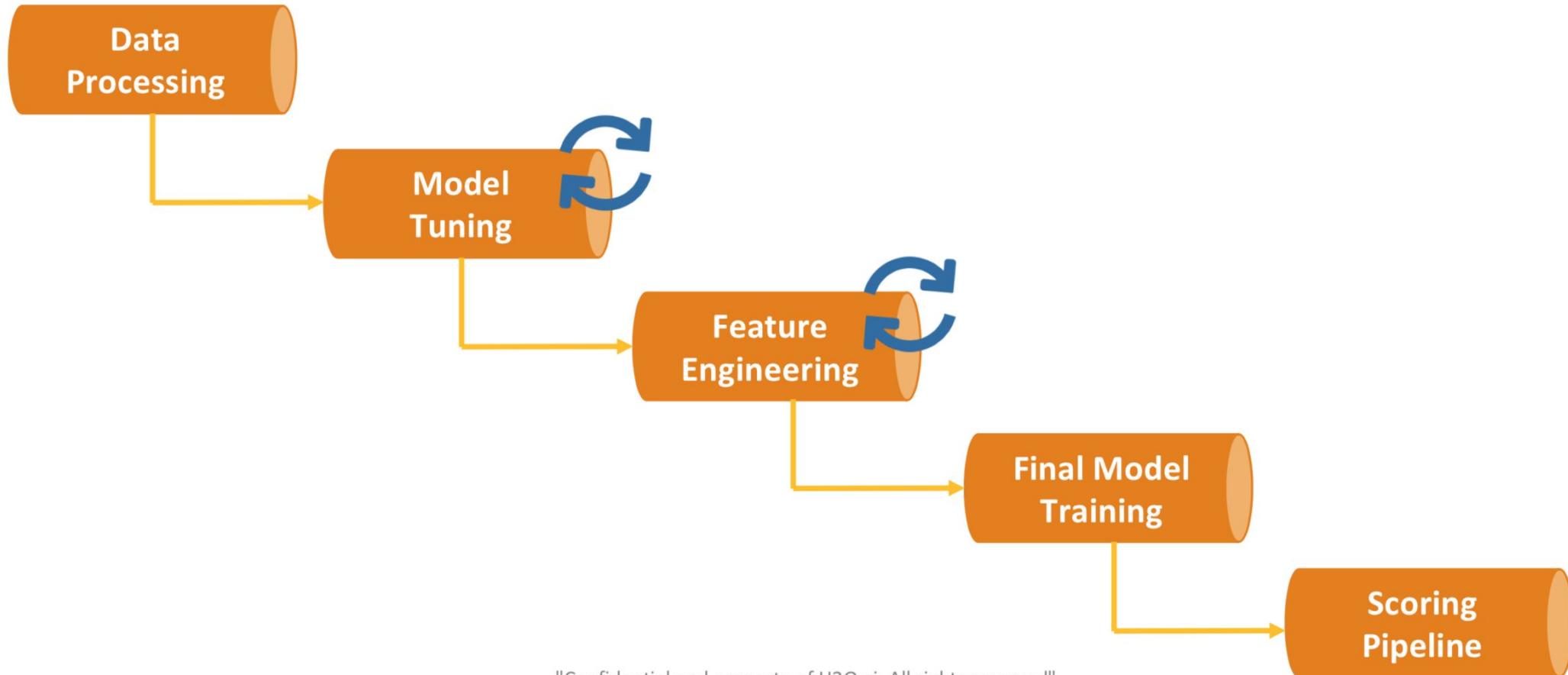


Hands-on Experiment: Visualize - Explore and Play with the Graphs

Hands-on Experiment: Predict



Driverless AI: Machine Learning Workflow



"Confidential and property of H2O.ai. All rights reserved"

Datasets → CreditCard-train.csv → Predict

The screenshot shows the H2O.ai Experiment interface. At the top left is the logo 'H2O.ai Experiment'. To its right are navigation links: DATASETS, EXPERIMENTS, MLI, AUTOVIZ, HELP, PY_CLIENT, MOJO2-RUNTIME, MESSAGES[3], LOGOUT, and H2OAI. Below these are tabs for TRAINING DATA and ASSISTANT. Under TRAINING DATA, it says 'DATASET' and shows 'CreditCard-train.csv'. Below this are sub-tabs: ROWS, COLUMNS, DROPPED COLS, VALIDATION DATASET, and TEST DATASET. A large central modal window has a dark background. It contains the text 'First time using Driverless AI? Click Yes to get a tour!' followed by a descriptive paragraph: 'Our interactive tour will help you become an expert grand master data scientist in 5 minutes!'. At the bottom of the modal are three buttons: 'YES' (highlighted in yellow), 'NOT NOW' (highlighted in white), and 'NO'.

© 2017-2018 H2O.ai. All rights reserved.

TRAINING DATA

DATASET

CreditCard-train.csv

ROWS

24K

COLUMNS

25

DROPPED COLS

--

VALIDATION DATASET

--

TEST DATASET

--

ASSISTANT

TARGET COLUMN

Select target column

FOLD COLUMN

--

WEIGHT COLUMN

--

TIME COLUMN

[OFF]

Select target column

DRIVERLESS AI 1.4.1 - AI TO DO AI
Licensed to h2o.ai (SN28)
Start typing to filter out items

- AGE
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_5
- PAY_6
- BILL_AMT1
- BILL_AMT2
- BILL_AMT3
- BILL_AMT4
- BILL_AMT5
- BILL_AMT6
- PAY_AMT1
- PAY_AMT2
- PAY_AMT3
- PAY_AMT4
- PAY_AMT5
- PAY_AMT6
- default payment next month

DATASETS EXPERIMENTS MLI AUTOVIZ HELP PY_CLIENT MOJO2-RUNTIME MESSAGES(3) LOGOUT H2O X

TRAINING DATA

DATASET

CreditCard-train.csv

ROWS

24K

COLUMNS

25

DROPPED COLS

--

VALIDATION DATASET

--

TEST DATASET

--

TARGET COLUMN

Select target column

FOLD COLUMN

--

WEIGHT COLUMN

--

TIME COLUMN

[OFF]

TRAINING DATA

ASSISTANT

DATASET
CreditCard-train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
24K	25	--	--	--

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE
bool

COUNT
23999

UNIQUE
2

TARGET FREQ
5369

EXPERIMENT SETTINGS

ACCURACY: 6

TIME: 4

INTERPRETABILITY: 6

CLASSIFICATION **REPRODUCIBLE** **ENABLE GPUS**

EXPERT SETTINGS

SCORER

- GINI
- MCC
- F05
- F1
- F2
- ACCURACY
- LOGLOSS
- AUC**
- AUCPR

LAUNCH EXPERIMENT

Machine Learning Experimentation



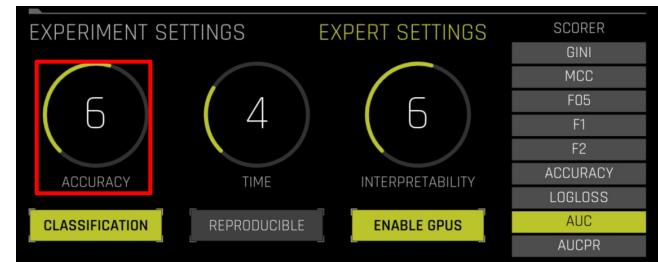
Accuracy



Time



Interpretability



Accuracy

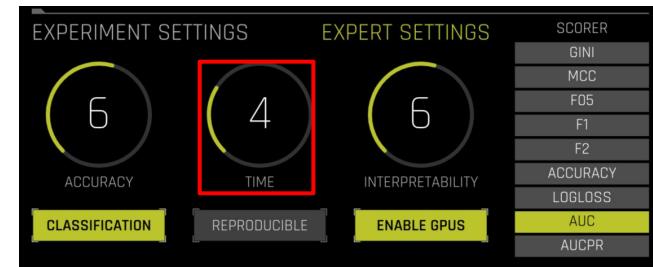
Accuracy	Max Rows x Cols	Ensemble Level	Target Transformation	Parameter Tuning Level	Num Folds	Only First Fold Model	Distribution Check
1	100K	0	False	0	3	True	No
2	1M	0	False	0	3	True	No
3	50M	0	True	1	3	True	No
4	100M	0	True	1	3-4	True	No
5	200M	1	True	1	3-4	True	Yes
6	500M	2	True	1	3-5	True	Yes
7	750M	<=3	True	2	3-10	Auto	Yes
8	1B	<=3	True	2	4-10	Auto	Yes
9	2B	<=3	True	3	4-10	Auto	Yes
10	10B	<=4	True	3	4-10	Auto	Yes

<http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/running-experiment.html#accuracy>

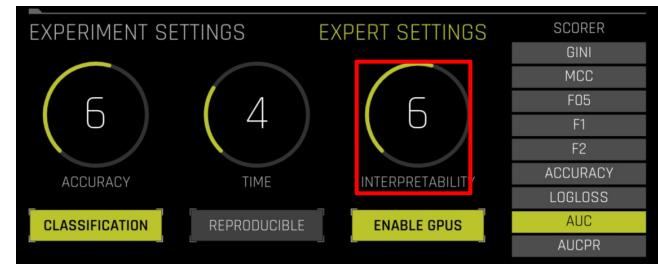
Time

This specifies the relative time for completing the experiment (i.e., higher settings take longer). Early stopping will take place if the experiment doesn't improve the score for the specified amount of iterations.

Time	Iterations	Early Stopping Rounds
1	1-5	None
2	10	5
3	30	5
4	40	5
5	50	10
6	100	10
7	150	15
8	200	20
9	300	30
10	500	50



<http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/running-experiment.html#time>

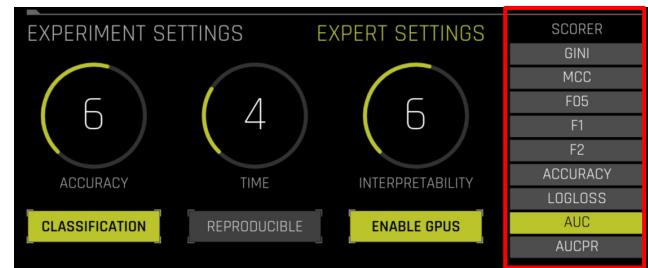


Interpretability

Interpretability	Ensemble Level	Target Transformation	Feature Engineering	Feature Pre-Pruning	Monotonicity Constraints
1 - 3	<= 3			None	Disabled
4	<= 3	Inverse		None	Disabled
5	<= 3	Anscombe	Clustering (ID, distance) Truncated SVD	None	Disabled
6	<= 2	Logit Sigmoid		Feature selection	Disabled
7	<= 2		Frequency Encoding	Feature selection	Enabled
8	<= 1	4 th Root		Feature selection	Enabled
9	<= 1	Square Square Root	Bulk Interactions (add, subtract, multiply, divide) Weight of Evidence	Feature selection	Enabled
10	0	Identity Unit Box Log	Date Decompositions Number Encoding Target Encoding Text (TF-IDF, Frequency)	Feature selection	Enabled

<http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/running-experiment.html#interpretability>

Scorer (Metrics)



Classification

SCORER
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR

Precision

Recall

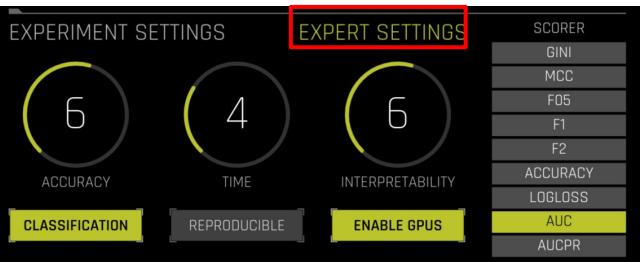
Best For
Imbalanced
Data

Regression

SCORER
GINI
R2
MSE
RMSE
RMSLE
RMSPE
MAE
MAPE
SMAPE

<http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/scorers.html>

Expert Settings



Expert Experiment Settings

XGBoost GBM models	<input type="button" value="AUTO"/>	<input type="button" value="ON"/>	<input type="button" value="OFF"/>
TensorFlow models (alpha)	<input type="button" value="AUTO"/>	<input type="button" value="ON"/>	<input type="button" value="OFF"/>
Max allowed feature shift (AUC) before dropping feature	0.6		
Time-series lags override, e.g. [7, 14, 21]			
Make MOJO scoring pipeline	<input type="button" value="DISABLED"/>		
Generate Holiday Features	<input type="button" value="ENABLED"/>		
LightGBM support	<input type="button" value="AUTO"/>	<input type="button" value="ON"/>	<input type="button" value="OFF"/>
RuleFit support (alpha)	<input type="button" value="AUTO"/>	<input type="button" value="ON"/>	<input type="button" value="OFF"/>
Enable Target Encoding	<input type="button" value="ENABLED"/>		
Probability to create non-target lag features	0.1		
Feature Brain Level (0..10)	2		
Random seed	1234		
XGBoost GLM models	<input type="button" value="AUTO"/>	<input type="button" value="ON"/>	<input type="button" value="OFF"/>
Data distribution shift detection	<input type="button" value="ENABLED"/>		
Time-series lag-based recipe	<input type="button" value="ENABLED"/>		
Make Python scoring pipeline	<input type="button" value="ENABLED"/>		
Smart imbalanced sampling (binary)	<input type="button" value="ENABLED"/>		
Max. original features selected	10000		

SAVE

CANCEL

Expert Experiment Settings

Max. pipeline features (-1 = auto)	-1
Max. feature interaction depth	8
Use Tensorflow in tuning/evolution	<input type="button" value="DISABLED"/>
Max. TensorFlow epochs	100
Min. DAI iterations	0
Max. learning rate for tree models	0.05
#GPUs/Experiment (-1 = all)	-1
Number of cores to use (0 = all)	0
GPU starting ID (0..visible #GPUs - 1)	0
Max. number of rows for feature evolution data splits (not for final pipeline)	1000000
Max. allowed fraction of uniques for integer and categorical cols	0.95
Threshold for string columns to be treated as text (0.0 - text, 1.0 - string)	0.3
Enable TensorFlow NLP	<input type="button" value="ENABLED"/>
Max. number of trees	3000
#GPUs/Model (-1 = all)	1
Feature engineering effort (0..10)	5

CANCEL

Let's Try 6-2-6 with AUC (with MOJO Enabled in Expert Settings)

The screenshot shows the H2O AI Experiment Assistant interface. On the left, the 'TRAINING DATA' section displays the dataset 'CreditCard-train.csv' with 24K rows and 25 columns. The target column is 'default payment next'. In the 'EXPERIMENT SETTINGS' section, three circular progress bars are shown: Accuracy (6), Time (2), and Interpretability (6). Below these are buttons for 'CLASSIFICATION', 'REPRODUCIBLE', and 'ENABLE GPUs'. At the bottom is a large green 'LAUNCH EXPERIMENT' button. On the right, the 'ASSISTANT' section shows a 'TEST DATASET' field set to 'Yes' (highlighted with a red box) and a 'SCORER' dropdown menu where 'AUC' is selected (also highlighted with a red box).

You can also add the Test dataset.
It will be used for making predictions.

This is a detailed view of the 'Expert Experiment Settings' dialog. It contains various configuration options for different machine learning models and pipelines. Key settings include:

- XGBoost GBM models: AUTO (ON), LightGBM support: AUTO (ON), XGBoost GLM models: AUTO (ON).
- TensorFlow models (alpha): OFF.
- RuleFit support (alpha): OFF.
- Data distribution shift detection: ENABLED.
- Max allowed feature shift (AUC) before dropping feature: 0.6.
- Enable Target Encoding: ENABLED.
- Time-series log-based recipe: ENABLED.
- Probability to create non-target log features: 0.1.
- Make Python scoring pipeline: ENABLED (highlighted with a red box).
- Generate Holiday Features: DISABLED.
- Feature Brain Level (0..10): 2.
- Random seed: 1234.
- Smart imbalanced sampling (binary): ENABLED.
- Max. original features selected: 10000.

At the bottom are 'SAVE' and 'CANCEL' buttons.

< H2O.ai Experiment CreditCard

DRIVERLESS AI 1.4.1 – AI TO DO AI
Licensed to H2O.ai (SN28)

TRAINING DATA

DATASET
CreditCard-train.csv

ROWS
24K

COLUMNS
25

DROPPED COLS
--

VALIDATION DATASET
--

ASSISTANT

TEST DATASET
Yes

CreditCard-test.csv

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE
bool

COUNT
23999

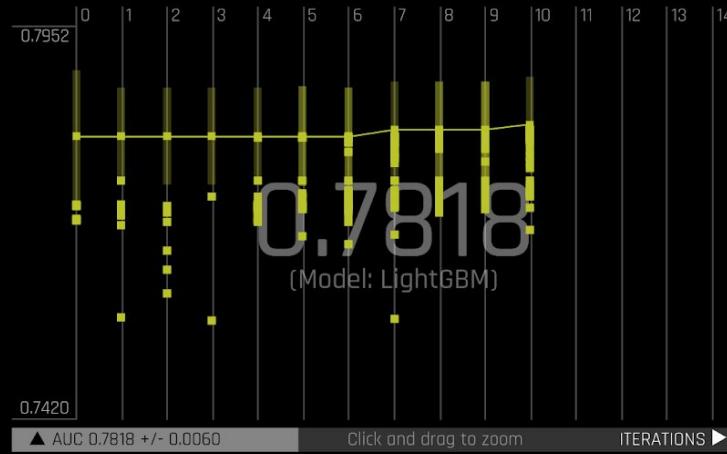
UNIQUE
2

TARGET FREQ
5369

SCORED 112/125 MODELS ON 2587 FEATURES.
LAST SCORED [LIGHTGBM, XGBOOST]



ITERATION DATA - VALIDATION



VARIABLE IMPORTANCE

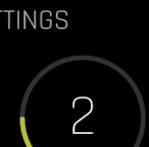
26_InteractionAdd:PAY_0:PAY_2	1.00
30_NumToCatTE:PAY_0:PAY_2.0	0.85
33_ClusterTE:ClusterID81:AGE:BILL_AMT1:EDUCATION:PAY_...	0.36
35_NumToCatWoE:BILL_AMT1:PAY_0.0	0.29
10_PAY_0	0.22
14_PAY_5	0.12
37_NumToCatTE:LIMIT_BAL:PAY_5:PAY_AMT5.0	0.11
24_InteractionSub:PAY_0:PAY_AMT3	0.11
24_InteractionSub:BILL_AMT1:LIMIT_BAL	0.10
32_ClusterTE:ClusterID49:BILL_AMT3:PAY_4:PAY_AMT4.0	0.08
24_InteractionSub:PAY_2:PAY_AMT3	0.08
25_InteractionDiv:BILL_AMT1:BILL_AMT2	0.08
25_InteractionDiv:AGE:PAY_AMT1	0.07
17_PAY_AMT2	0.07

DATASETS EXPERIMENTS MLI AUTOVIZ HELP PY_CLIENT MOJO2-RUNTIME **MESSAGES[3]** LOGOUT H2OAI

EXPERIMENT SETTINGS



CLASSIFICATION



REPRODUCIBLE



INTERPRETABILITY

SCORER
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR

CPU / MEMORY Notifications Log Trace



ROC



PREC-RECALL

LIFT

GAINS

GPU USAGE

H2O.ai Experiment CreditCard

DRIVERLESS AI 1.4.1 – AI TO DO AI
Licensed to H2O.ai (SN28)

TRAINING DATA

DATASET
CreditCard-train.csv

ROWS
24K

COLUMNS
25

DROPPED COLS
--

VALIDATION DATASET
--

TEST DATASET
Yes
CreditCard-test.csv

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

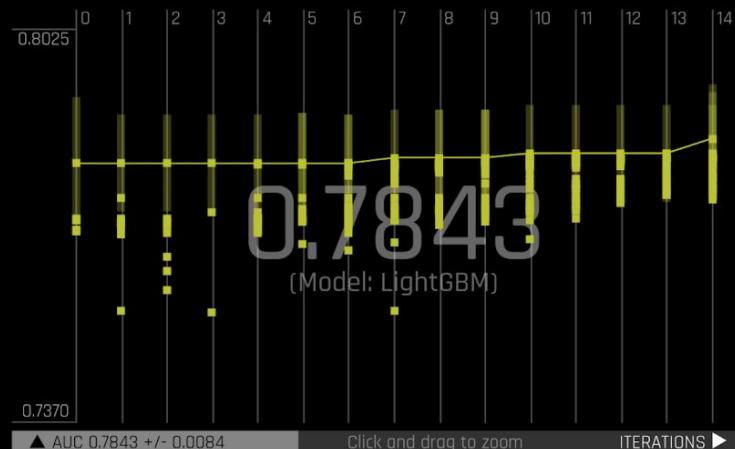
TYPE
bool

COUNT
23999

UNIQUE
2

TARGET FREQ
5369

ITERATION DATA - VALIDATION



STATUS: COMPLETE

- [INTERPRET THIS MODEL](#)
- [SCORE ON ANOTHER DATASET](#)
- [TRANSFORM ANOTHER DATASET...](#)
- [DOWNLOAD \(HOLDOUT\) TRAINING PREDICTIONS](#)

- [DOWNLOAD TEST PREDICTIONS](#)
- [DOWNLOAD PYTHON SCORING PIPELINE](#)
- [DOWNLOAD MOJO SCORING PIPELINE](#)
- [DOWNLOAD EXPERIMENT SUMMARY](#)
- [DOWNLOAD LOGS](#)

EXPERIMENT SETTINGS

6
ACCURACY

2
TIME

6
INTERPRETABILITY

CLASSIFICATION

REPRODUCIBLE

ENABLE GPUs

SCORER
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR

Notifications Log Trace

CPU / MEMORY



MEM

VARIABLE IMPORTANCE

26_InteractionAdd:PAY_0:PAY_2	1.00
30_NumToCatTE:PAY_0:PAY_2.0	0.85
33_ClusterTE:ClusterID81:AGE:BILL_AMT1:EDUCATION:PAY_...	0.36
35_NumToCatWoE:BILL_AMT1:PAY_0.0	0.29
10_PAY_0	0.22
14_PAY_5	0.12
37_NumToCatTE:LIMIT_BAL:PAY_6:PAY_AMT5.0	0.11
24_InteractionSub:PAY_0:PAY_AMT3	0.11
24_InteractionSub:BILL_AMT1:LIMIT_BAL	0.10
32_ClusterTE:ClusterID49:BILL_AMT3:PAY_4:PAY_AMT4.0	0.08
24_InteractionSub:PAY_2:PAY_AMT3	0.08
25_InteractionDiv:BILL_AMT1:BILL_AMT2	0.08
25_InteractionDiv:AGE:PAY_AMT1	0.07
17_PAY_AMT2	0.07

ROC

PREC-RECALL

LIFT

GAINS

SUMMARY

Experiment: vacesiwa, 2018-11-30 12:30, 1.4.1

Settings: 6/2/6, seed=126037824, GPUs enabled

Train data: CreditCard-train.csv (23999, 25)

Validation data: N/A

Test data: CreditCard-test.csv (6000, 24)

Target column: default payment next month (binary, 22.372% target class)

System specs: Docker/Linux, 480 GB, 32 CPU cores, 8/8 GPUs

Max memory usage: 4 GB, 0.791 GB GPU

Recipe: AutoDL (14 iterations, 16 individuals)

Validation scheme: stratified, 1 internal holdout

Feature engineering: 2725 features tested (88 selected)

Timing:

Data preparation: 2.52 secs

Model and feature tuning: 86.74 secs (27 models trained)

Feature evolution: 261.36 secs (149 models trained)

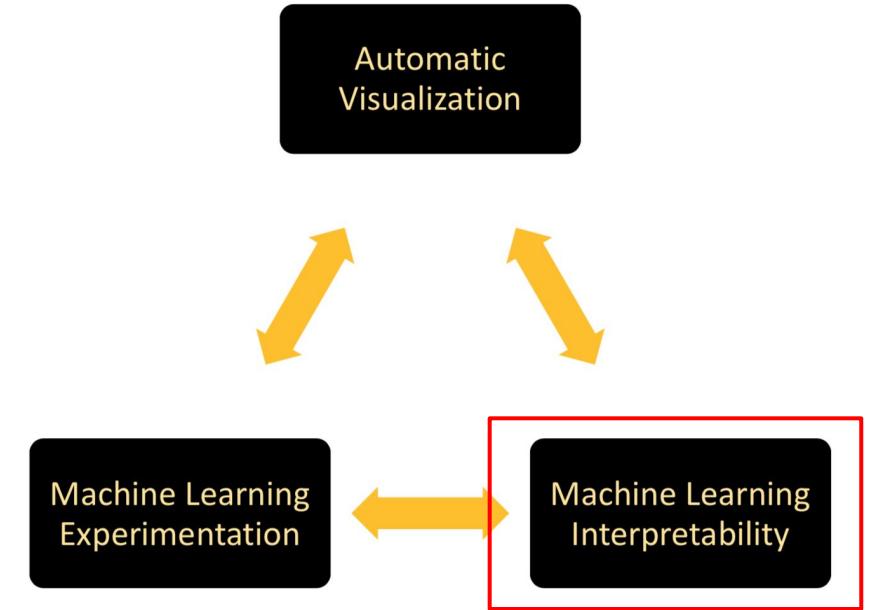
Final pipeline training: 104.05 secs (10 models trained)

Validation score: AUC = 0.78016 +/- 0.0070061 (baseline)

Validation score: AUC = 0.78428 +/- 0.0083514 (final pipeline)

Test score: AUC = 0.80486 +/- 0.0083514 (final pipeline)

Hands-on Experiment: Explain



H2O.ai Experiment CreditCard

DRIVERLESS AI 1.4.1 – AI TO DO AI
Licensed to H2O.ai (SN28)

TRAINING DATA

DATASET
CreditCard-train.csv

ROWS
24K

COLUMNS
25

DROPPED COLS
--

VALIDATION DATASET
--

TEST DATASET
Yes

CreditCard-test.csv

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE
bool

COUNT
23999

UNIQUE
2

TARGET FREQ
5369

ASSISTANT

STATUS: COMPLETE

[INTERPRET THIS MODEL](#)

[SCORE ON ANOTHER DATASET](#)

[TRANSFORM ANOTHER DATASET...](#)

[DOWNLOAD \(HOLDOUT\) TRAINING PREDICTIONS](#)

[DOWNLOAD TEST PREDICTIONS](#)

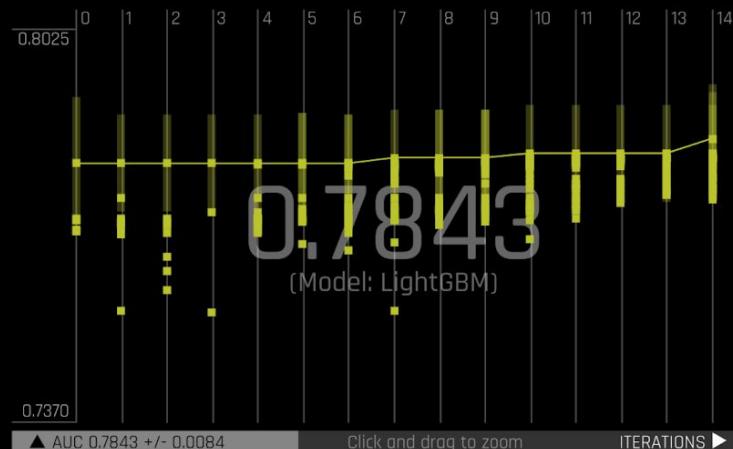
[DOWNLOAD PYTHON SCORING PIPELINE](#)

[DOWNLOAD MOJO SCORING PIPELINE](#)

[DOWNLOAD EXPERIMENT SUMMARY](#)

[DOWNLOAD LOGS](#)

ITERATION DATA - VALIDATION



VARIABLE IMPORTANCE

26_InteractionAdd:PAY_0:PAY_2	1.00
30_NumToCatTE:PAY_0:PAY_2.0	0.85
33_ClusterTE:ClusterID81:AGE:BILL_AMT1:EDUCATION:PAY_...	0.36
35_NumToCatWoE:BILL_AMT1:PAY_0.0	0.29
10_PAY_0	0.22
14_PAY_5	0.12
37_NumToCatTE:LIMIT_BAL:PAY_6:PAY_AMT5.0	0.11
24_InteractionSub:PAY_0:PAY_AMT3	0.11
24_InteractionSub:BILL_AMT1:LIMIT_BAL	0.10
32_ClusterTE:ClusterID49:BILL_AMT3:PAY_4:PAY_AMT4.0	0.08
24_InteractionSub:PAY_2:PAY_AMT3	0.08
25_InteractionDiv:BILL_AMT1:BILL_AMT2	0.08
25_InteractionDiv:AGE:PAY_AMT1	0.07
17_PAY_AMT2	0.07

EXPERIMENT SETTINGS

6
ACCURACY

2
TIME

6
INTERPRETABILITY

CLASSIFICATION

REPRODUCIBLE

ENABLE GPUs

CPU / MEMORY



MEM

Notifications Log Trace

ROC

PREC-RECALL

LIFT

GAINS

SUMMARY

Experiment: vacesiwa, 2018-11-30 12:30, 1.4.1

Settings: 6/2/6, seed=126037824, GPUs enabled

Train data: CreditCard-train.csv (23999, 25)

Validation data: N/A

Test data: CreditCard-test.csv (6000, 24)

Target column: default payment next month (binary, 22.372% target class)

System specs: Docker/Linux, 480 GB, 32 CPU cores, 8/8 GPUs

Max memory usage: 4 GB, 0.791 GB GPU

Recipe: AutoDL (14 iterations, 16 individuals)

Validation scheme: stratified, 1 internal holdout

Feature engineering: 2725 features tested (88 selected)

Timing:

Data preparation: 2.52 secs

Model and feature tuning: 86.74 secs (27 models trained)

Feature evolution: 261.36 secs (149 models trained)

Final pipeline training: 104.05 secs (10 models trained)

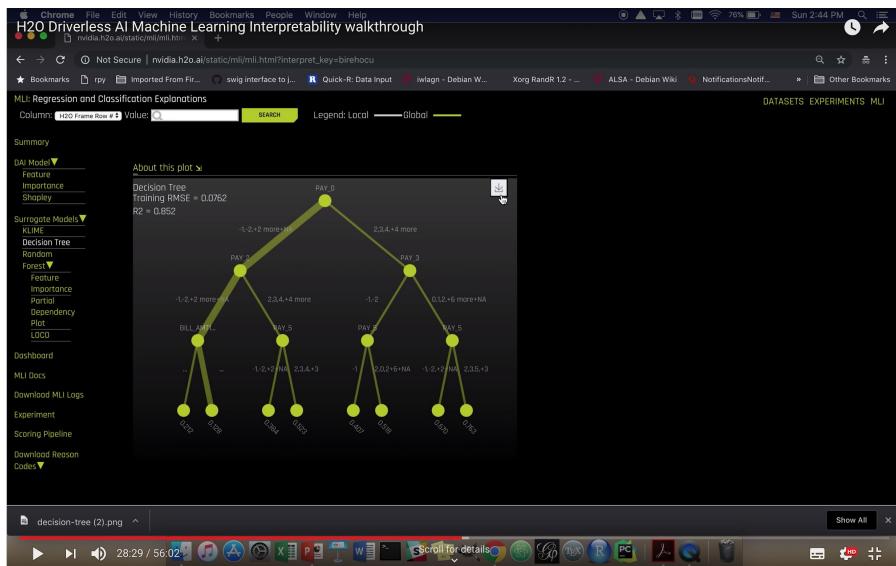
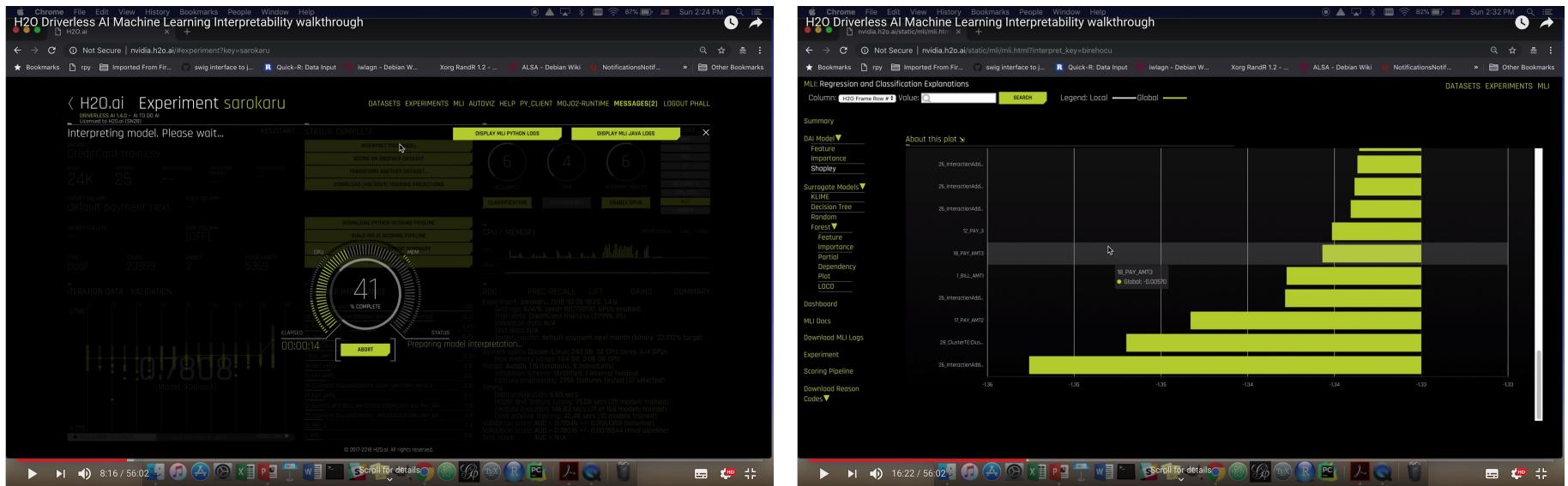
Validation score: AUC = 0.78016 +/- 0.0070061 (baseline)

Validation score: AUC = 0.78428 +/- 0.0083514 (final pipeline)

Test score: AUC = 0.80486 +/- 0.0083514 (final pipeline)

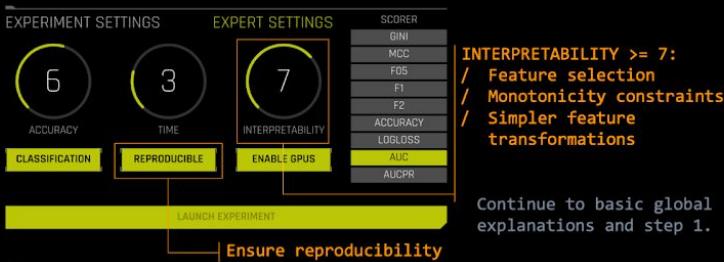
SCORER
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR

Tutorial on YouTube



<https://www.youtube.com/watch?v=5jSU3CUREXY>

0. START HERE: Train a more interpretable model

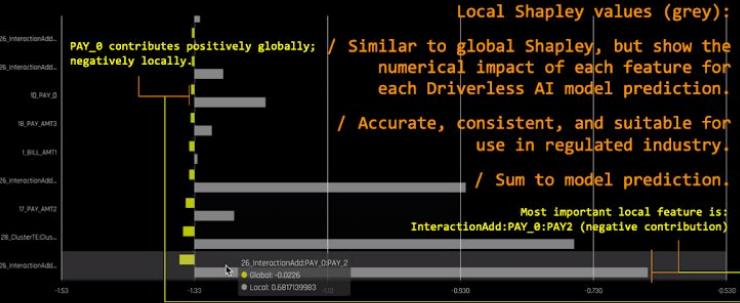


Advanced local explanations

<Understanding model behavior for an individual row>

6. Local Shapley feature importance

<Accurate local feature importance values for each individual row>

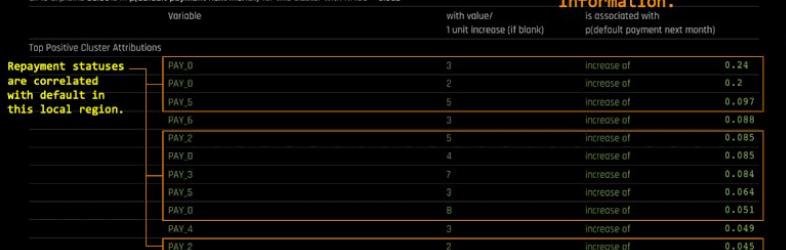


7. Local linear explanations

<Driverless AI model local linear trends around an individual>

Cluster 12 Reason Codes

LIME explains 89.93% in (default payment next month) for this cluster with RMSE = 0.052



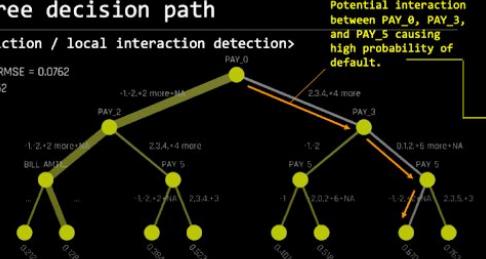
8. Local surrogate tree decision path

<Path of an individual to a prediction / local interaction detection>

Decision path (grey):

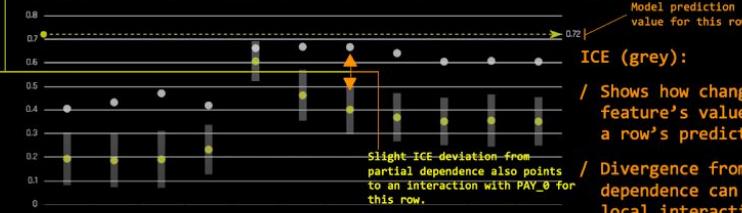
/ Shows approximately how row values impact Driverless AI model predictions.

/ Path can show local interactions.



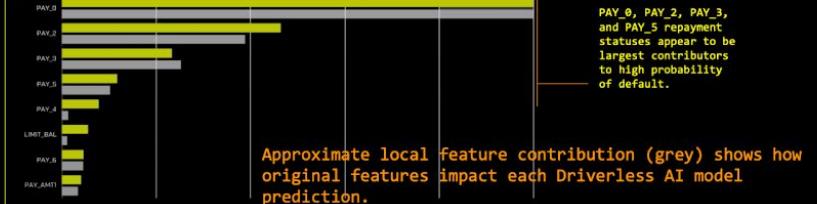
9. Original feature individual conditional expectation (ICE)

<Driverless AI model behavior for similar individuals / local interaction detection>



10. Local original feature importance

<Original features that drive a prediction for an individual>



Basic global explanations

<Understanding overall model behavior>

2. Global original feature importance

<Important original features that drive model behavior>



3. Partial dependence

<Average Driverless AI model prediction for different values of the original variables>



1. Global Shapley feature importance

<Important features created by Driverless AI that drive model behavior>

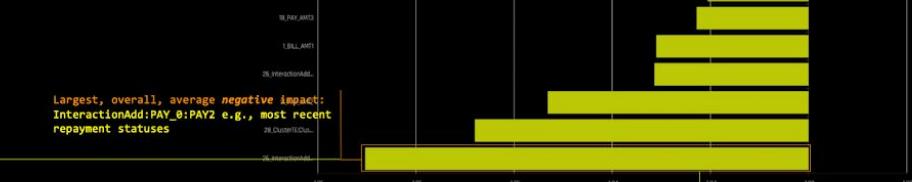
Global Shapley values:

/ Show the average numerical impact of a feature.

/ Positive features push the model's prediction higher on average; negative features push lower on average.

/ Are offsets from the average prediction.

/ Are in the same units as the actuals (e.g., target) for regression; logit units for classification models.



4. Global surrogate decision tree

<Overall flowchart of the Driverless AI model's decision making processes>

Surrogate decision tree models driverless AI predictions:

/ Higher and more frequent features are more important.

/ Shows approximate decision paths to predicted numerical outcomes.

/ Features above or below one-another can indicate an interaction.

/ Thickest edges are most common decision path through tree.

5. Global interpretable model

<Linear model of Driverless AI model predictions>



Interpretable, global linear model (white) or driverless AI predictions (green, increasing, middle):

/ Ranks Driverless AI predictions from lowest (bottom left) to highest (top right).

/ Quantifies nonlinearity of Driverless AI model.

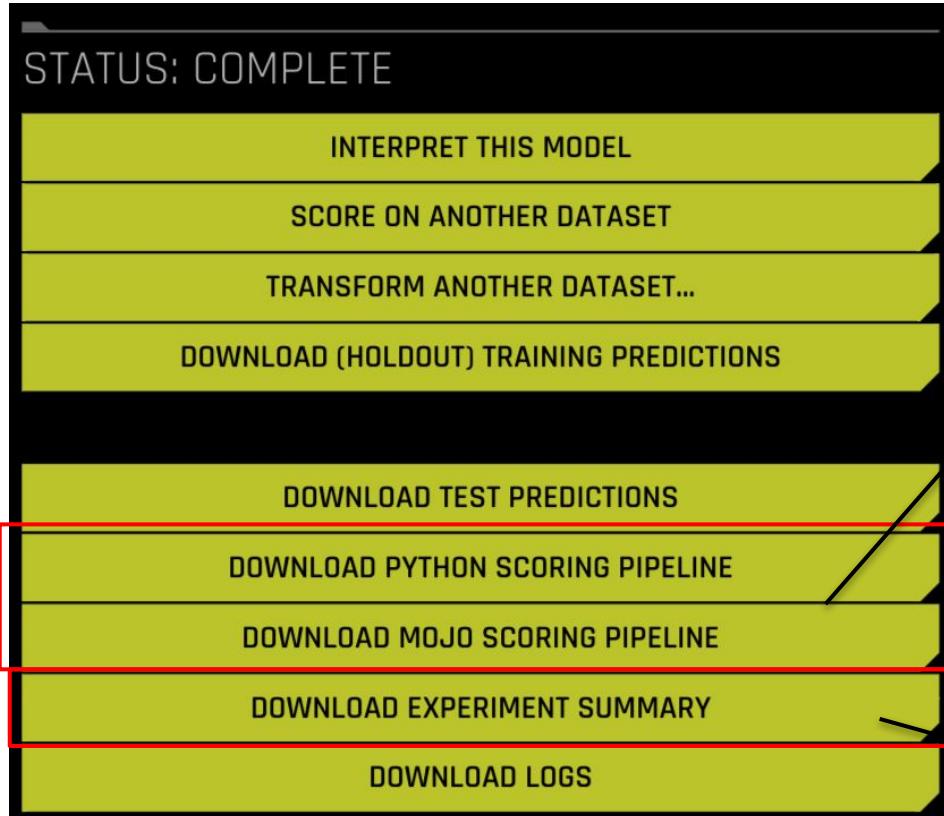
/ Provides basic sanity check of Driverless AI performance by plotting actuals (top, bottom) vs. predicted.



<Confirmatory information between explainability techniques highlighted in yellow.>

Hands-on Experiment: Other Features

Other Features



Auto Report in PDF

Different Deployment Options

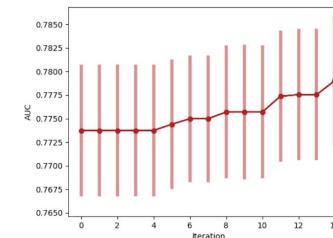
tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
gpu_hist	lossguide	0.000	16.000	0.450	0.400	34	0.770	4.049
gpu_hist	lossguide	0.000	8.000	0.600	0.600	34	0.770	3.577
gpu_hist	lossguide	0.000	128.000	0.500	0.400	174	0.770	11.444
gpu_hist	depthwise	10.000	0.000	0.900	0.700	23	0.767	7.131
gpu_hist	depthwise	10.000	0.000	0.900	0.700	23	0.767	7.186
gpu_hist	depthwise	10.000	0.000	0.900	0.700	23	0.767	7.347
gpu_hist	depthwise	10.000	0.000	0.900	0.700	23	0.767	7.332

tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
gpu_hist	depthwise	6.000	0.000	0.450	0.800	22	0.774	5.272
gpu_hist	depthwise	6.000	0.000	0.450	0.800	22	0.774	5.302
gpu_hist	lossguide	0.000	16.000	0.600	0.700	97	0.772	4.796
gpu_hist	depthwise	8.000	0.000	0.900	1.000	152	0.771	12.262
gpu_hist	depthwise	5.000	0.000	0.600	1.000	63	0.771	2.337
gpu_hist	depthwise	10.000	0.000	0.450	0.800	34	0.768	11.080
gpu_hist	lossguide	0.000	16.000	0.350	0.400	136	0.773	2.692
gpu_hist	lossguide	0.000	128.000	0.450	0.900	156	0.773	12.282

Feature Evolution

During the Model and Feature Tuning Stage, Driverless AI evaluates the effects of different types of algorithms, algorithm parameters, and features. The goal of the Model and Feature Tuning Stage is to determine the best algorithm and parameters to use during the Feature Evolution Stage. In the Feature Evolution Stage, Driverless AI trained xgboost and light gbm models (155 models) where each model evaluated a different set of features. The Feature Evolution Stage uses a genetic algorithm to search the large feature engineering space.

The graph below shows the effect the Model and Feature Tuning Stage and Feature Evolution Stage had on the performance.



Feature Transformation

The result of the Feature Evolution Stage is set of features to use for the final model. Some of these features were automatically created by Driverless AI. The top 14 features used in the final model are shown below, ordered by importance. If no transformer was applied, the feature is an original column.

Feature	Description	Transformer	Relative Importance
26_InteractionAdd: PAY_0: PAY_2	[PAY_0] + [PAY_2]	Interaction	1.000
30_NumToCatTE: PAY_0: PAY_2.0	Out-of-fold mean of the response grouped by [PAY_0, PAY_2] using 5 folds [internal parameters:{10, 3, None}] (numeric columns are bucketed into 250 equally populated bins) [internal parameters:{10, 3, None}]	Numeric to Categorical Target Encoding	0.981
10_PAY_0	PAY_0 (original)	None	0.350
24_InteractionSub: BILL_AMT1: LIMIT_BAL	[BILL_AMT1] - [LIMIT_BAL]	Interaction	0.147
25_InteractionDiv: PAY_2: PAY_AMT2	[PAY_2] / [PAY_AMT2]	Interaction	0.115
25_InteractionDiv: PAY_AMT2	[AGE] / [PAY_AMT2]	Interaction	0.095
17_PAY_AMT2	PAY_AMT2 (original)	None	0.094
12_PAY_3	PAY_3 (original)	None	0.081
25_InteractionDiv: BILL_AMT1: BILL_AMT2	[BILL_AMT1] / [BILL_AMT2]	Interaction	0.079

Other Features

The screenshot shows the H2O.ai interface for managing experiments. At the top, there's a navigation bar with links for DATASETS, EXPERIMENTS, MLI, AUTOVIZ, HELP, PY_CLIENT, MOJO2-RUNTIME, MESSAGES[3], LOGOUT, and H2OAI. Below the navigation is a search bar with placeholder text 'Search for datasets, experiments, or help...'. The main area is titled 'Experiments' and displays a single experiment entry for 'CreditCard'. The experiment details are as follows:

Experiment	Dataset	Target	Validation Score	Test Score	Scorer	Size	Accuracy	Time	Interpretability	Status	Runtime
CreditCard	CreditCard-train.csv	default payme...	0.78428	0.80486	AUC	340MB	6	2	6	Done	00:08:06

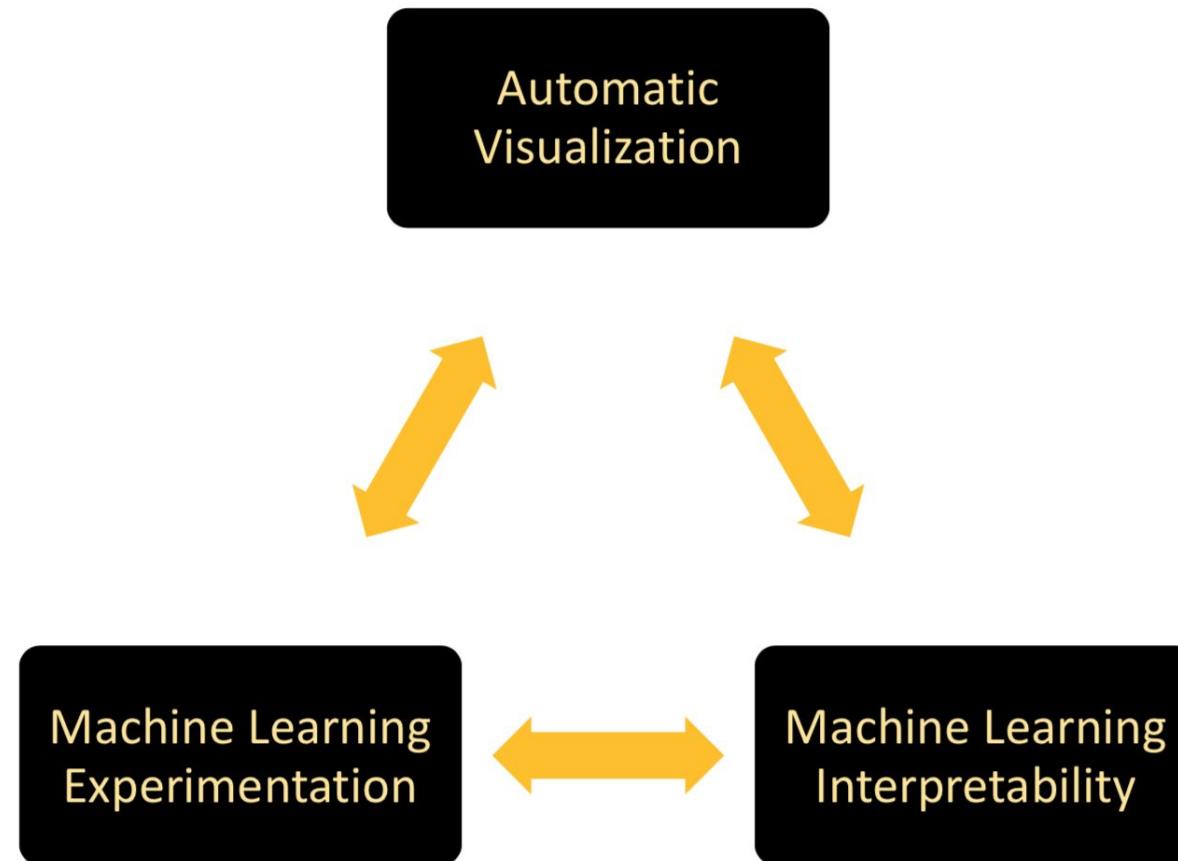
Below the table, there are several buttons and links:

- + NEW EXPERIMENT
- SELECT
- SORT BY ▾
- OPEN
- RESTART FROM LAST CHECKPOINT
- NEW MODEL WITH SAME PARAMS
- DELETE

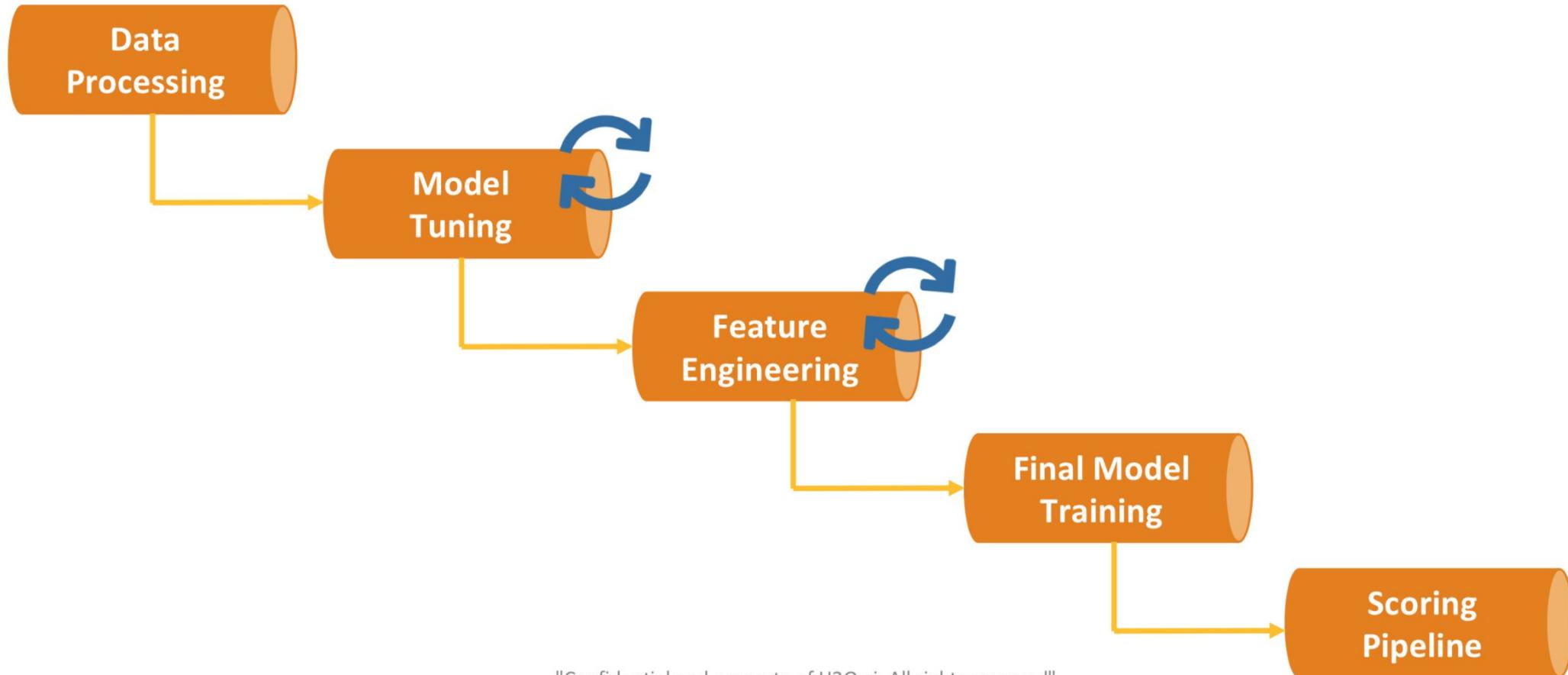
At the bottom of the page, there's a copyright notice: © 2017-2018 H2O.ai. All rights reserved.

Hands-on Experiment: Quick Recap

Driverless AI Components



Driverless AI: Machine Learning Workflow



"Confidential and property of H2O.ai. All rights reserved"

Machine Learning Experimentation



Accuracy



Time



Interpretability

Approximate Reasoning

Local Reason Codes

k-LIME Local Attributions	Variable	with value	is associated with	DEFAULT_PAYMENT_NEXT_MONTH
Top Positive Local Attributions				
	PAY_1	2	increase of	0.34
	PAY_5	2	increase of	0.06
	PAY_3	2	increase of	0.06
Skipped 8 additional attributions, click to view all ...				
Top Negative Local Attributions				
	PAY_AMT3	3000	decrease of	0.01
	BILL_AMT5	24930	decrease of	0.01
	BILL_AMT1	21024	decrease of	0.01

Why will someone Default?

The fact that they haven't paid in 2 months increases their likelihood by 34%

Their Pay Amount is \$3,000 **decreases their likelihood by 1%**

**Hands-on Experiment:
Try Different Settings / Datasets in Data
Folder / Your Own Data
Q & A**

Dive into H2O (Amsterdam)

Part II - H2O-3 Hands-on Training

Václav Belák, Ph.D.

Data Scientist

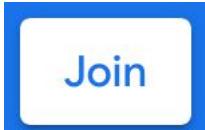
vaclav@h2o.ai

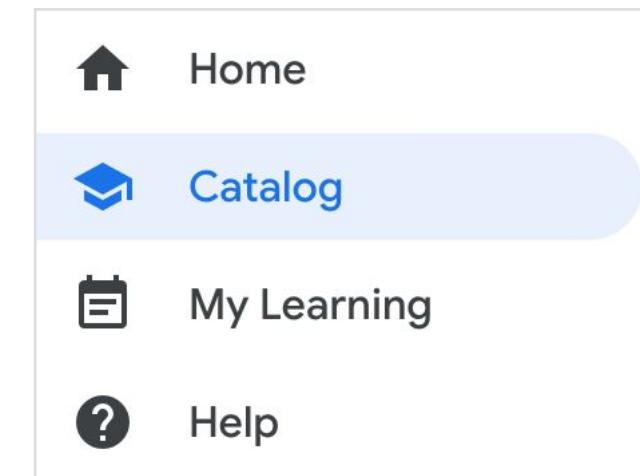
H2O.ai Community Slack Workspace

Online Chat to ask questions, discuss use cases, give feedback and more

- Join the H2O.ai Community Slack Workspace today!
 - <https://tinyurl.com/h2o-community-slack>
- Use emoji to tag messages
 - :question :use_case :mli :get_started :bugs ...
- Reply to message using **threads**
- Check out Community Guide for more info:
 - <https://tinyurl.com/hac-community-guide>

First-time Qwiklab Account Setup

1. Go to <https://h2oai.qwiklab.com/>
2. Click on 
3. Create a new account with a valid email address
4. You will receive a confirmation email
 - Click on the link in the confirmation email
5. Go back to <http://h2oai.qwiklab.com> and log in
6. Go to the *Catalog* on the left bar
7. Choose *Vaclav's Advanced H2O Workshop*
8. and wait for instructions

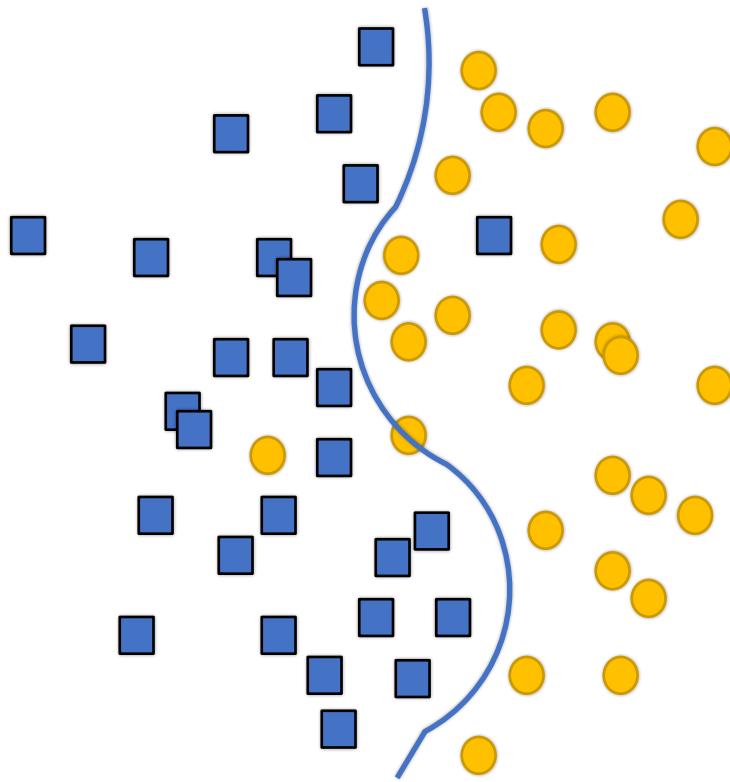


[Start Lab](#)

Agenda

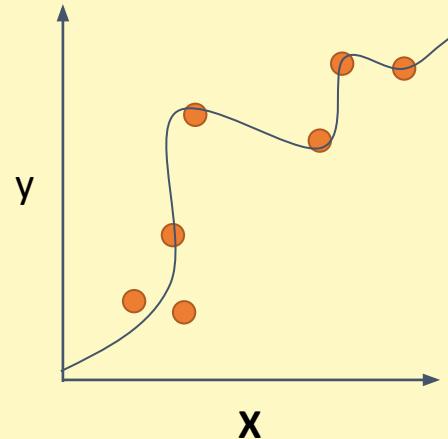
1. Superbrief intro to Machine Learning
2. Introduction to H2O-3
3. Hands On
 - Data Munging
 - Feature Engineering
 - Supervised Learning (GBM and AutoML*)
4. Questions and Answers

What is Machine Learning?



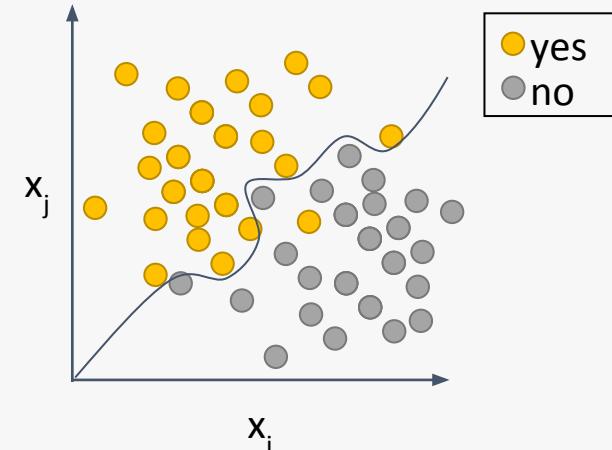
Supervised Learning

Regression:
How much will a customer spend?



H₂O algos:
Penalized Linear Models
Random Forest
Gradient Boosting
XGBoost
Neural Networks
Stacked Ensembles

Classification:
Will a customer churn?



H₂O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
XGBoost
Neural Networks
Stacked Ensembles

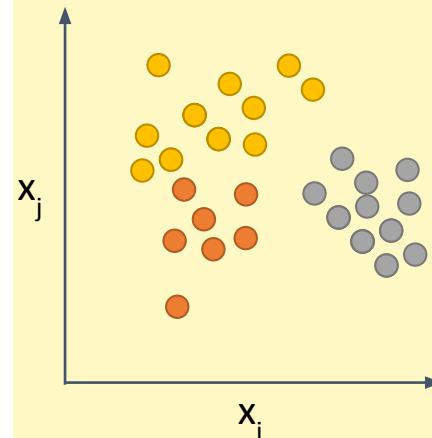
H₂O.ai

H₂O.ai

Unsupervised Learning

Clustering:

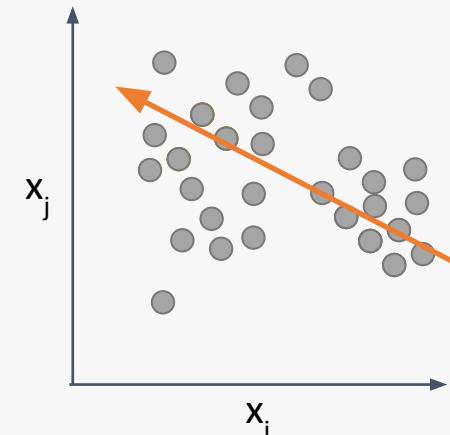
Grouping rows – e.g. creating groups of similar customers



H₂O algos:
k – means

Feature extraction:

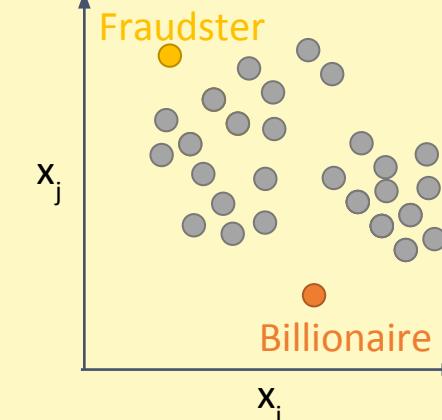
Grouping columns – Create a small number of new representative dimensions



H₂O algos:
Principal components
Generalized low rank models
Autoencoders, Word2Vec

Anomaly detection:

Detecting outlying rows -
Finding high-value or
fraudulent customers

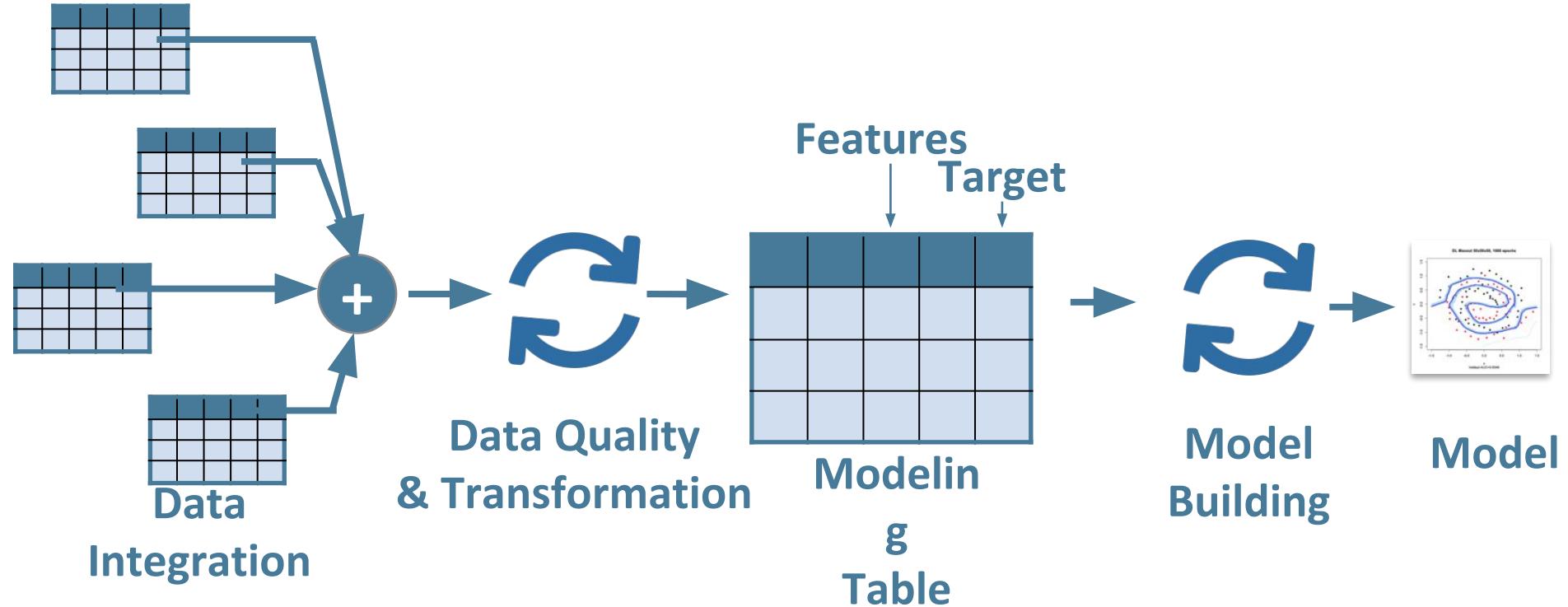


H₂O algos:
Isolation Forest
Generalized low rank models
Autoencoders

H₂O.ai

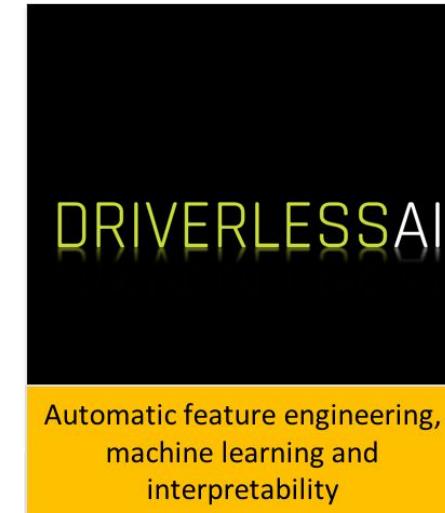
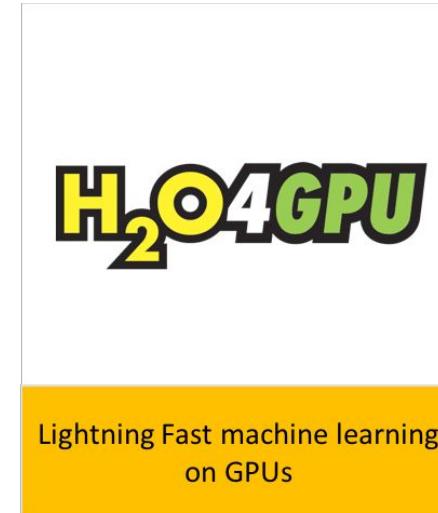
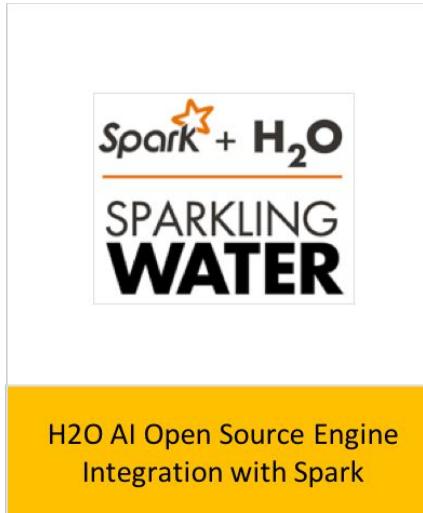
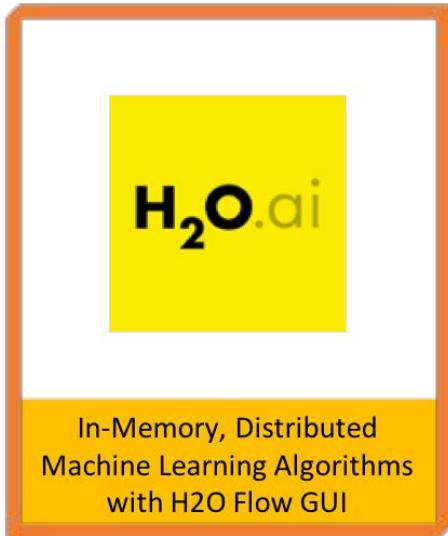
H₂O.ai

Simplified Typical Machine Learning Pipeline



Introduction to H2O

H2O Products



What is H2O?

Math Platform

Open source in-memory AI engine

- Parallelized and distributed algorithms
- GLM, Random Forest, GBM, Deep Learning, etc.

Tech and API

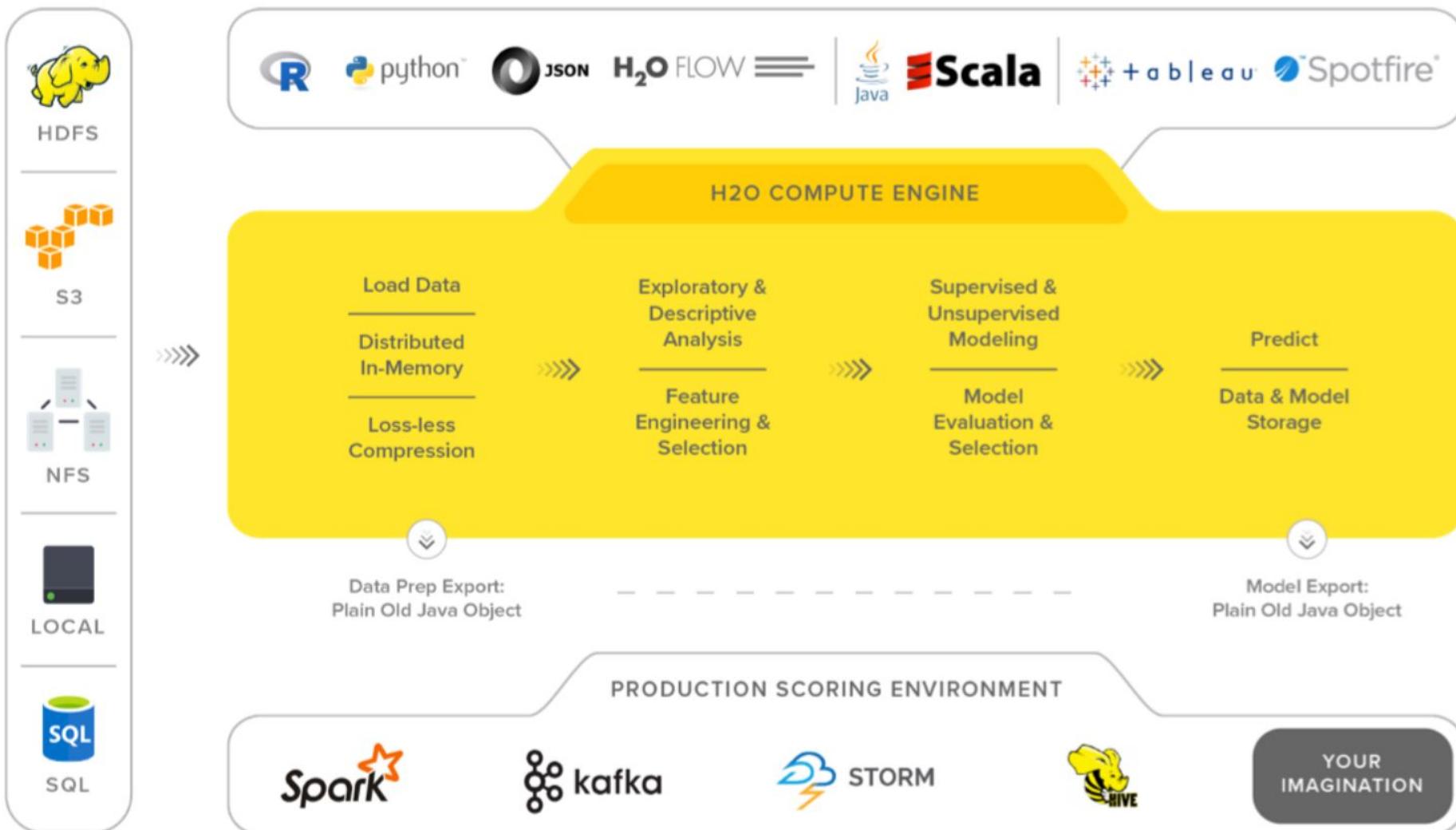
Easy to use and adopt

- Written in Java – perfect for Java Programmers
- Install is lightweight
- REST API (Java) – run H2O from R, Python, WebUI

Big Data

More data? Or better models? BOTH

- Use all of your data – model without sampling
- More Data + Better Models = Better Predictions



R, Python, and Flow

The screenshot shows an RStudio interface with an R script named `all_algos.R`. The script performs variable selection and importance analysis for four different models: GBM, GLM, RF, and Deep Learning. It includes code to load data, train models, and print variable importance. The RStudio environment pane shows the global environment with objects like `my.gbm`, `my.rf`, and `my.glm`. Below the script are four bar charts titled "VI from GBM", "VI from RF", "VI from GLM", and "VI from Deep Learning", each showing the relative importance of various variables.

```
all_algos.R
# Run GBM with variable importance, lambda search and using all factor levels
my.gbm <- h2o.glm(x=mx, y=myv, training_frame=data.hex, family="binomial",standardize=TRUE,lambda_search=T)
# select the best model picked by glm
# best_model <- my.glm$best_model
# Get the normalized coefficients of the best model
en_coeff <- abs(my.glmmodels[[best_model]]$model$normalized_coefficients)
glm.vi <- glmVariableImportanceFromGLM()
print("Variable Importance from GLM")
print(glm.vi)
# Plot variable importance from glm
barplot(glm.vi[,1:20],coefficients, names.arg = glm.vi[,1:20],snames, las=2,main="VI from GLM")
# Run deep learning with variable importance
my.dl <- h2o.deepLearning(x = myx, y = myy, training_frame = data.hex,
                           activation = "Tanh", hidden = c(10,10,10),
                           epochs = 12, variable_importances = TRUE)
# Access variable importance from the built model
print("Variable Importance from Deep Learning")
print(my.dl)
variable_importances;
variable relative_importance scaled_importance percentage
1 duration 13202.887635 1.000000 0.985303
2 euribor3m 194.000000 0.000000 0.000000
3 age 0.000000 0.000000 0.000000
4 job 0.000000 0.000000 0.000000
5 marital 0.000000 0.000000 0.000000
6 education 0.000000 0.000000 0.000000
7 default 0.000000 0.000000 0.000000
8 housing 0.000000 0.000000 0.000000
9 loan 0.000000 0.000000 0.000000
10 contact 0.000000 0.000000 0.000000
11 month 0.000000 0.000000 0.000000
```

The screenshot shows a Jupyter Notebook cell with Python code. The code builds a deep learning model and compares its performance against a GBM model. It also prints a table of results for both models. The output cell shows the AUC values for both models.

```
#deeplearning Model Build Progress: [########################################] 100%
In [7]: # GBM performance on train/test data
train_auc_gbm = data.gbm.model_performance(train).auc()
test_auc_gbm = data.gbm.model_performance(test).auc()

# Deep Learning performance on train/test data
# train_auc_dl = data_dl.model_performance(train).auc()
# test_auc_dl = data_dl.model_performance(test).auc()

# Make a pretty HTML table printout of the results
header = ["Model", "AUC Train", "AUC Test"]
table = [
    ["GBM", train_auc_gbm, test_auc_gbm],
    ["DL ", train_auc_dl, test_auc_dl]
]
h2o.display.H2ODisplay(table, header)

Out[7]:


| Model | AUC Train | AUC Test  |
|-------|-----------|-----------|
| GBM   | 0.9568221 | 0.9307979 |
| DL    | 0.8956065 | 0.8841564 |



In [8]: # Create new H2OFrame of crime observations
examples = [
    {"Date": ["02/08/2015 11:43:58 PM", "02/08/2015 11:00:39 PM"],
     "IUCR": ["1811, 1150"],
     "Primary_Type": ["NARCOTICS", "DECEPTIVE PRACTICE"]},
```

The screenshot shows the H2O Flow interface for a K-Means example. It displays a "Setup Parse" configuration panel where users can define the source URL (`http://s3.amazonaws.com/h2o-public-test-data/smalldata/flow_examples/seeds_dataset.txt`), ID column (`Key_Frame`), parser type (`CSV`), separator (`HT \t (horizontal tab): '09'`), and other options like column headers and single quotes. Below this is an "Edit Column Names and Types" table where columns are mapped to types and values.

Column	Type	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7	Value 8
1	Numeric	15.26	14.88	14.29	13.84	16.14	14.38		
2	Numeric	14.84	14.57	14.09	13.94	14.99	14.21		
3	Numeric	0.871	0.8811	0.905	0.8955	0.9034	0.8951		
4	Numeric	5.763	5.554	5.291	5.324	5.658	5.386		
5	Numeric	3.312	3.333	3.337	3.379	3.562	3.312		
6	Numeric	2.221	1.018	2.699	2.259	1.355	2.462		
7	Numeric	5.22	4.956	4.825	4.805	5.175	4.956		
8	Numeric	1	1	1	1	1	1		

Reading Data with Python

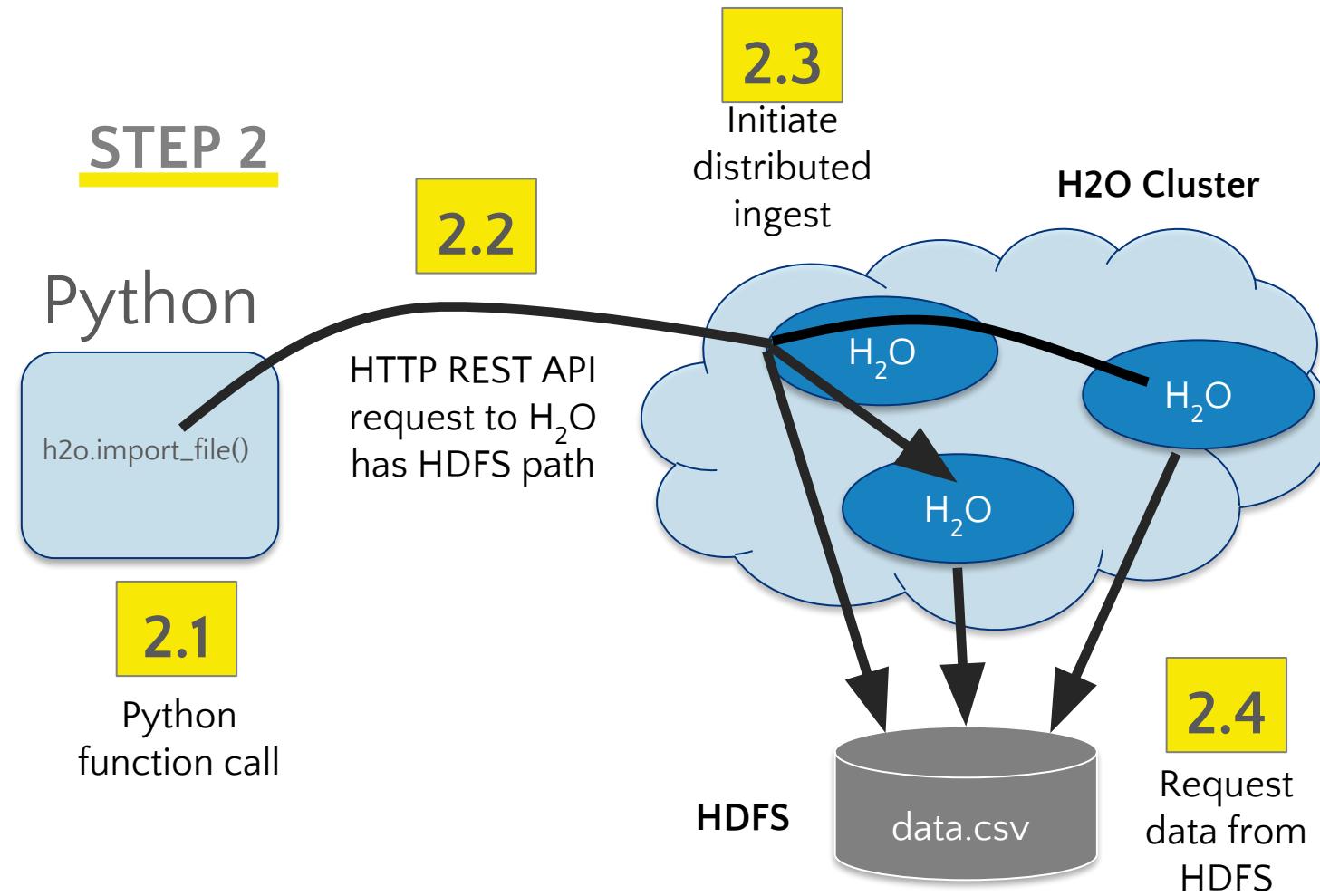
STEP 1



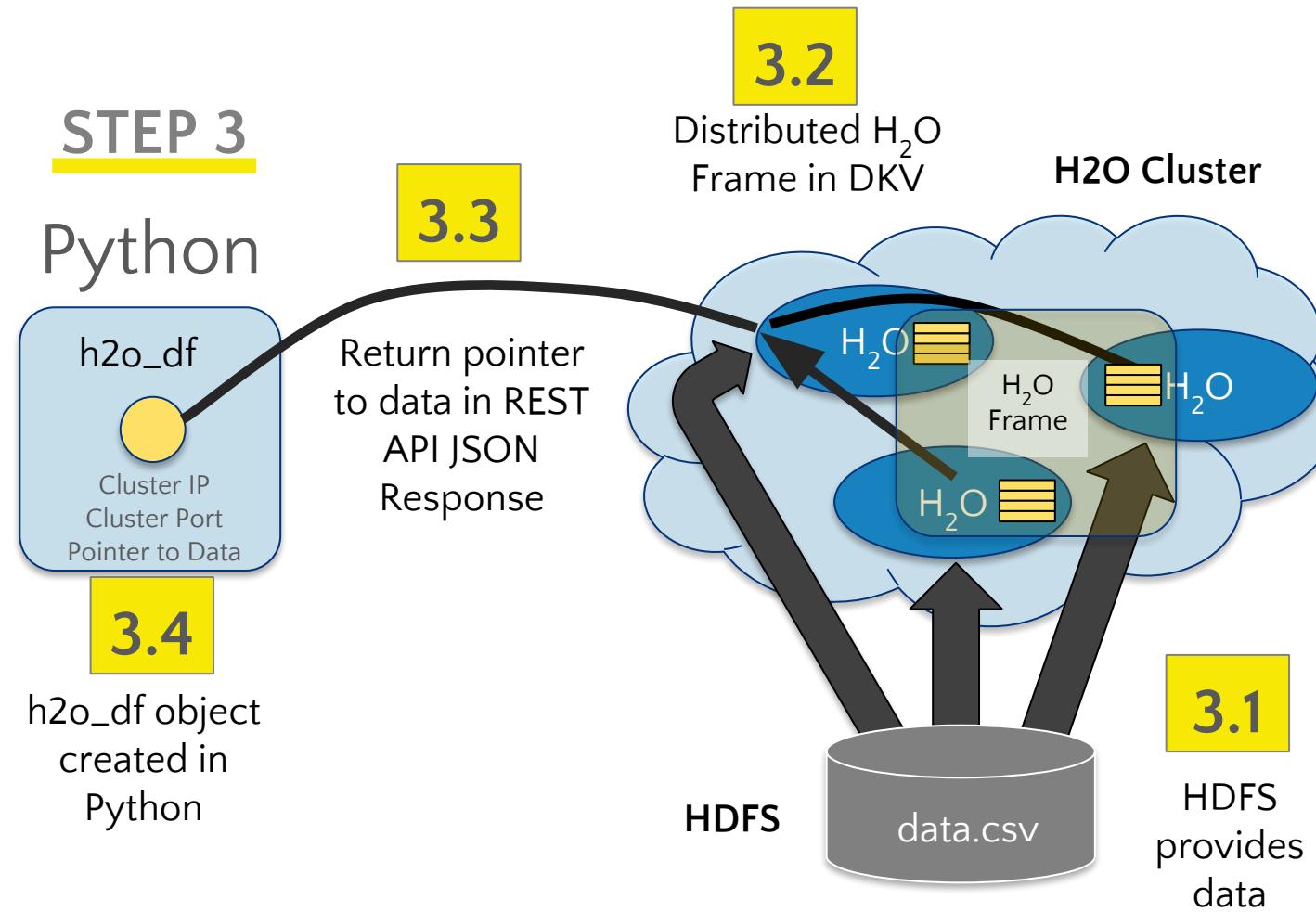
```
h2o_df = h2o.import_file("hdfs://.../data.csv")
```

Python
user

Reading Data with Python



Reading Data with Python



Data Transfer between Python and H2O

`h2o.H2OFrame`

- Turns Pandas DF into an H2OFrame
- Saves the data-frame to CSV and then performs `h2o.upload()` to H2O cluster

`h2o.as_list` OR `h2o_frame.as_data_frame`

- Turns H2OFrame to list or Pandas DF
- Downloads the H2OFrame locally and will parse with Pandas if `use_pandas` is set to True

The size of H2OFrame may exceed the memory of your client workstation

Hands On

Case Study: Lending Club Dataset

- Loan data from 2007 up until 2015 including rejected applications and accepted application
- **Goal:** Predict the loan delinquency based on the information supplied when applying for a loan
- Data:

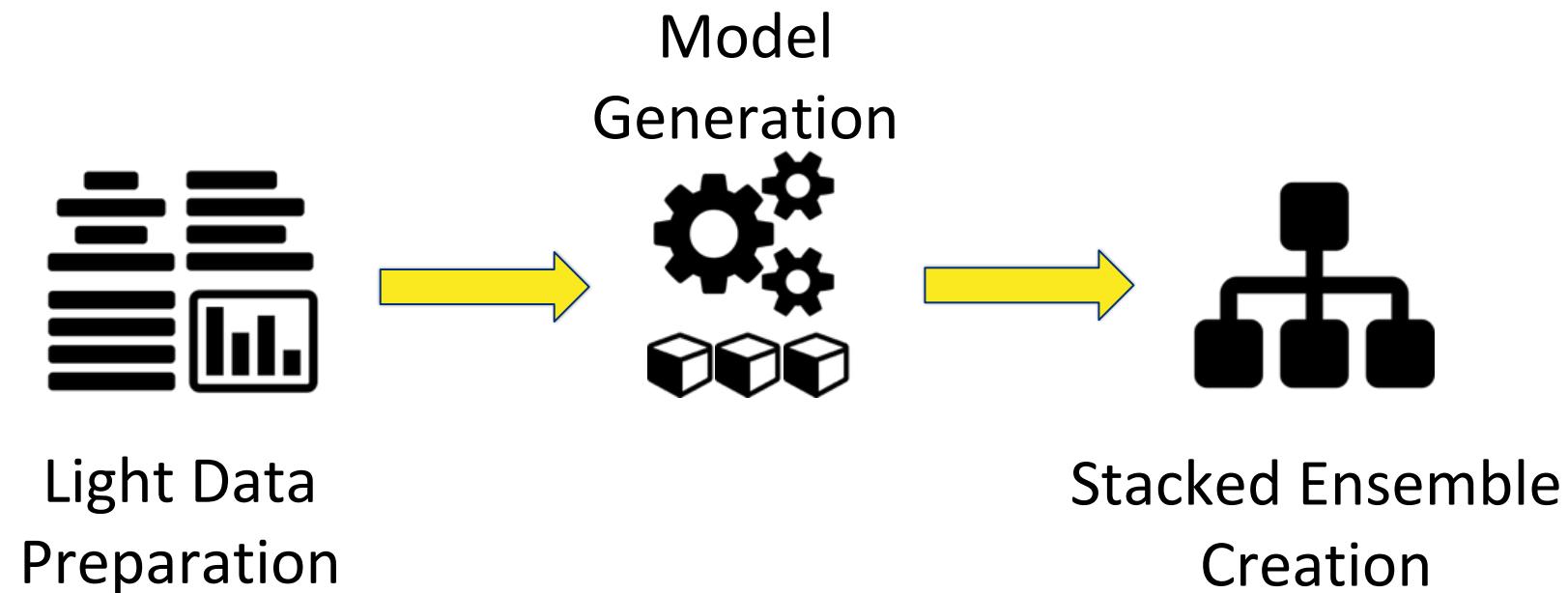
Type	Features
Demographic Information	annual_inc, home_ownership, emp_length
Loan Information	purpose, term, desc, int_rate
Response Column	bad_loan

Workflow

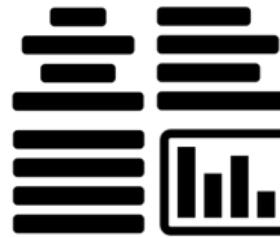
1. Filter data to relevant observations
2. Create response column
3. Clean columns
4. Add new features
5. Train GBM model
6. Evaluate performance
7. Interpret GBM model
8. Save & reuse model
9. AutoML
10. End the lab

AutoML

H2O Auto ML Pipeline

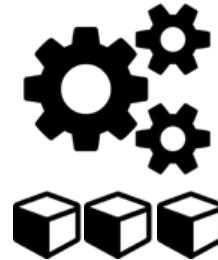


H2O AutoML Light Data Preparation



- Impute missing data
- Standardize numeric features
- One-hot encode categorical features

H2O AutoML Model Generation



- Explore multiple algorithms
 - GLMs, Random Forests, GBMs, Deep Neural Networks
- Perform random grid search over hyper-parameter space
- Use early stopping rules for models and grids

H2O AutoML Stacked Ensemble Creation



- Stacked ensemble of all the models
- Stacked ensemble of the best models for each of the algorithms

Productionalizing Models

Predicting on New Data / Productionalization

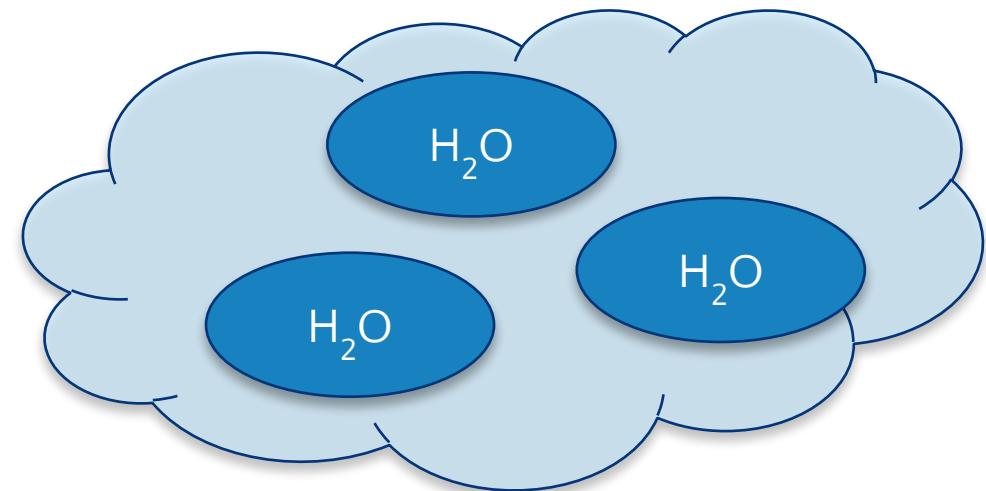
Standalone / RealTime

- Build a MOJO or POJO
- Pure Java and self-contained
- Doesn't need H2O cluster



H2O Cluster / Batch

- Save & load models directly in H2O
- Batch scoring using predict()
- Needs H2O cluster



Q & A

[Extra Slides] Driverless AI Training: Appendix