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

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**

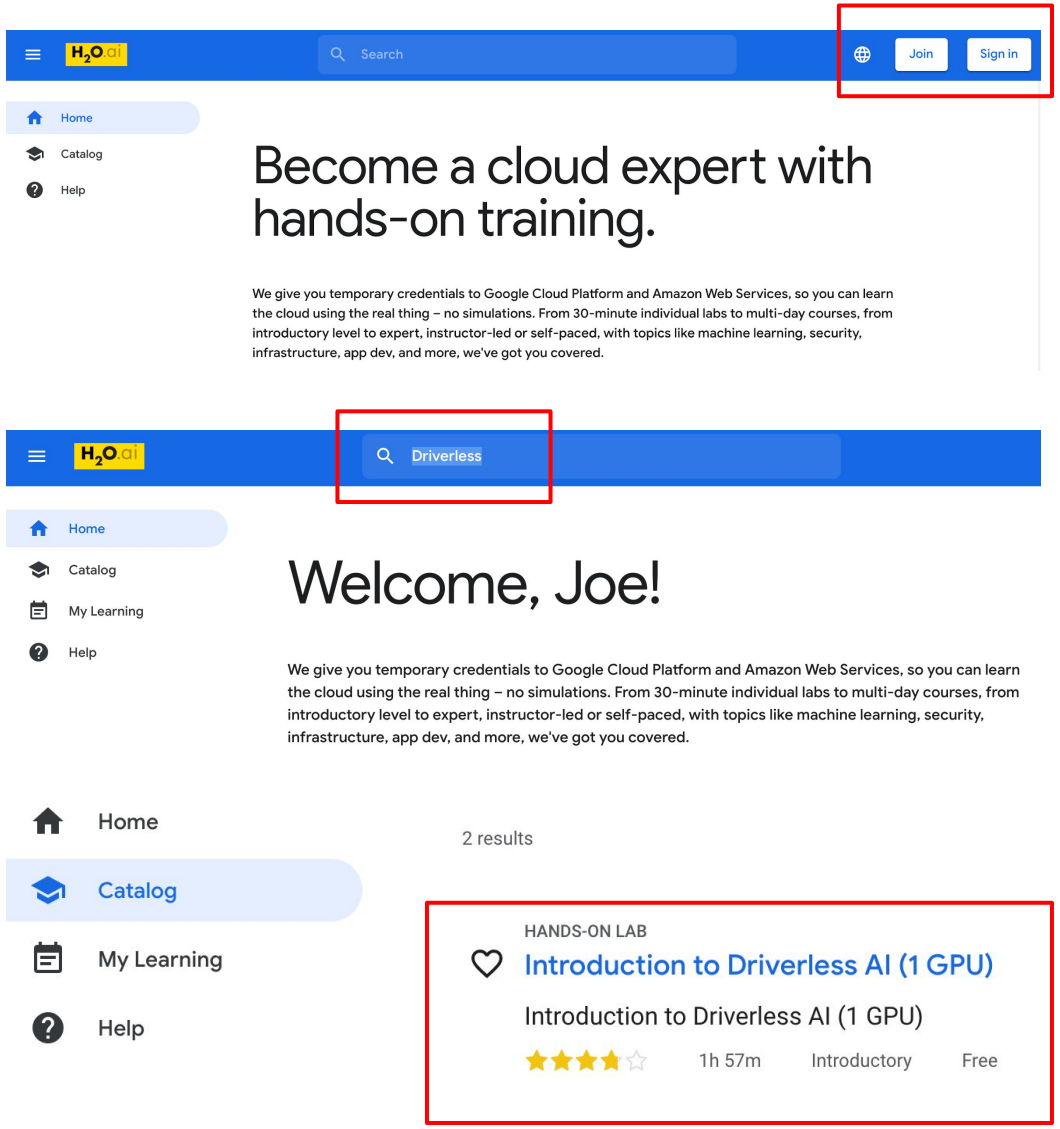
Driverless AI Training (9am - 12pm) Agenda

- **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

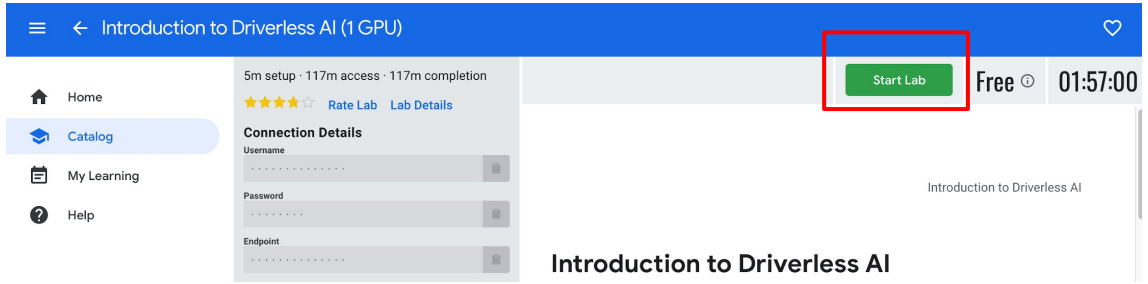
Setting Up Qwiklabs

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)**”

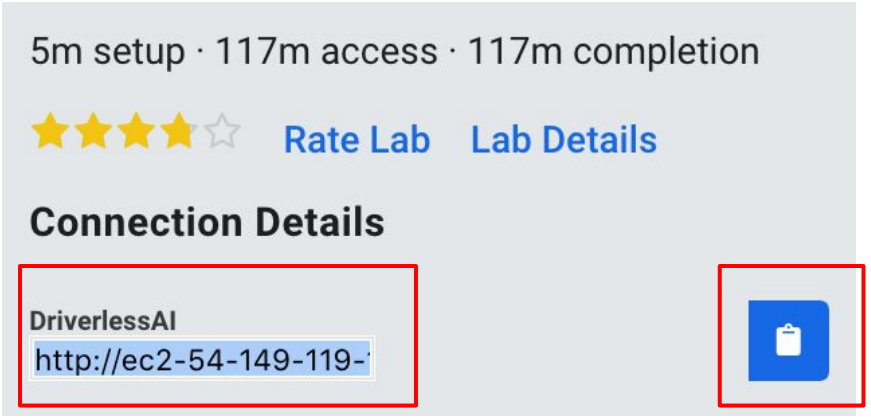


Setting Up Qwiklabs

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**.

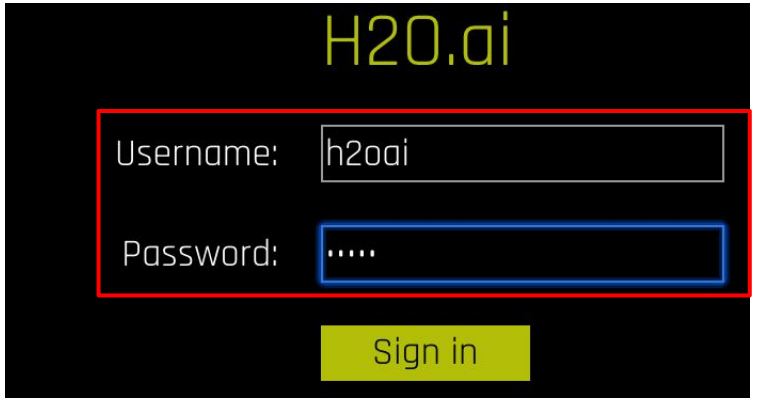


7. You may see a “502 Bad Gateway” error message. It is harmless. Wait for another minute and try again.

H2O.ai

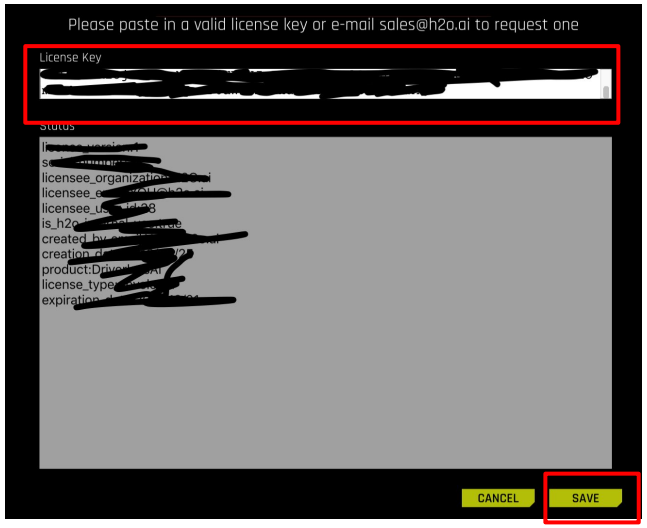
Setting Up Qwiklabs

8. Scroll to down and click “**I agree to these terms**”.
9. Enter Username: **h2oai**
10. Enter Password: **h2oai**



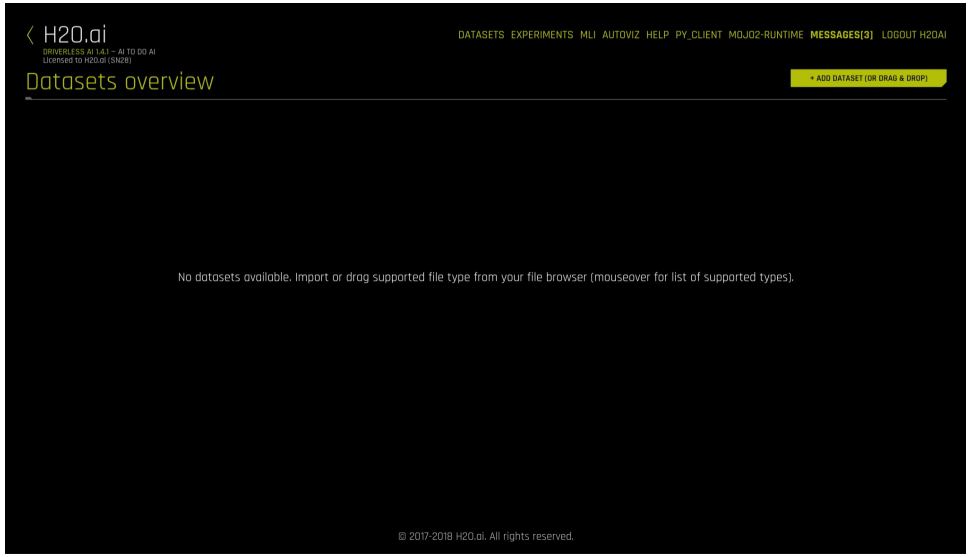
Setting Up Qwiklabs

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



K8O11BsW2uJwDlvZxScRC95k_2rsI2jyZ0z_5XaCsqDtbdClTGRphzGYa9Z2bRwen1jaPIYY8vOVdm3_4zkMqNZARi9VeBvgyiuldKTeVnAFD2bxfYPPX5rACwRPX_QFsgI7GJ-B6KYOALUph3wUKslmtvwIrjQM_aAz09JWKi-UbaUhE2XI67rw03oYu284uTAS3tgTjvLJGpRLxn1aNWxpJrYoEo2wy4aHMNoBT53q7uRkgW6J5qkdppZnhEKjvxLPUWr4mFefY3RPEHU6_QL31rcMGWkjMh2iKAZo3kbpSs7J0odKnQqPTy7HtAL1Qc-Nc3xyNMTYOh5wmf0W8GxpY2Vuc2VfdmVyc2lvbjoxCnNlcmlhbF9udW1iZXI6Mjg4ODYKbGljZW5zZWVfb3JnYW5pemF0aW9uOkgyTyBRd2lrbGFiCmxpY2Vuc2VlX2VtYWlsOnRvbWsrbWVldHVwcXdpa2xhYkBoMm8uYWkKbGljZW5zZWVfdXNlcl9pZDoyODg4Ngppc19oMm9faW50ZXJuYWxfdXNlOmZhbHNlCmNyZWF0ZWRfYnlfZW1haWw6cGF0cmljay5tb3JhbkBoMm8uYWkKY3JlYXRpb25fZGF0ZToyMDE4LzExLzMwCnByb2R1Y3Q6RHJpdmVybGVzc0FJCmxpY2Vuc2VfdHlwZTpwaWxvdApleHBpcmF0aW9uX2RhdGU6MjAxOC8xMi8wNgo=

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	• 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
2307
MAKERS GONNA MAKE
MAKER
THIRSTY?


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

Automatic feature engineering, machine learning and interpretability

Secure multi-tenant H2O clusters

Worldwide Recognition in the H2O.ai Community

Open source community

222 OF FORTUNE 500 ♥ H_2O

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

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

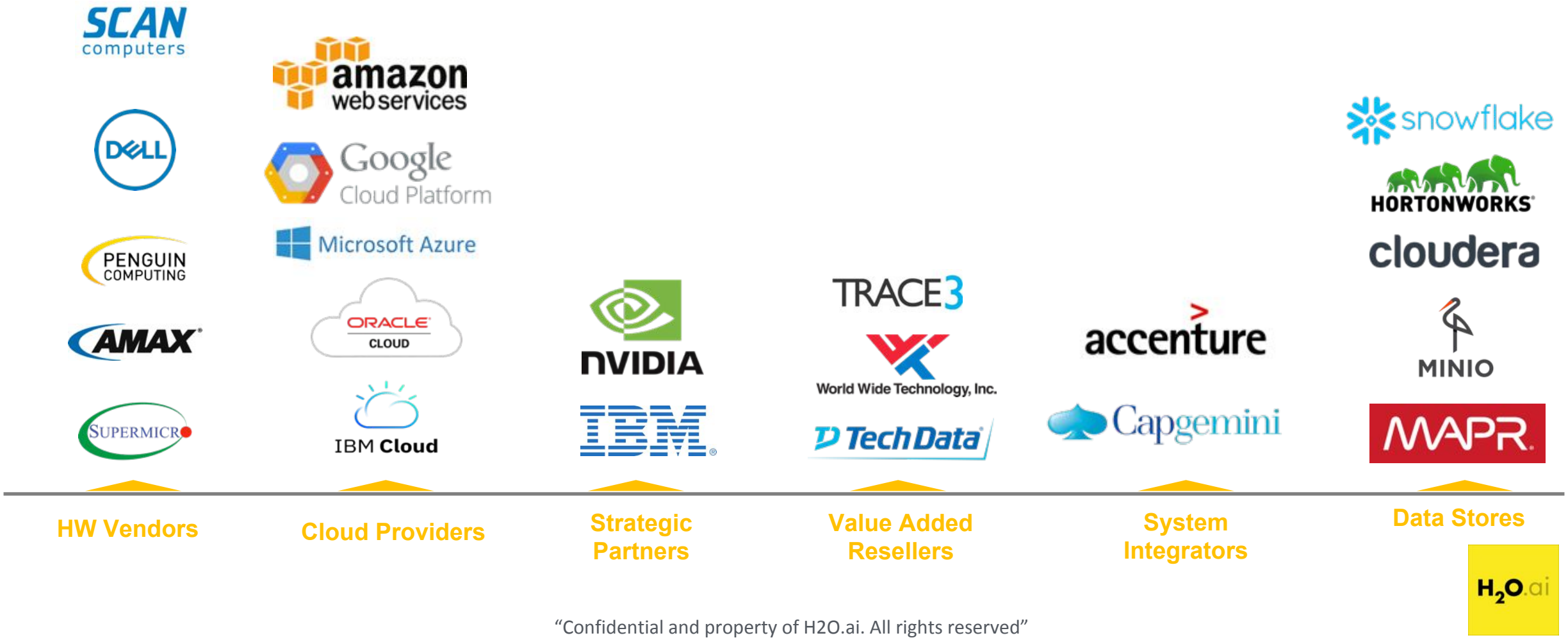
- **H2O.ai customers gave the highest overall score** among all the vendors for sales relationship and account management, customer support (onboarding, troubleshooting, etc.) and overall service and support

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

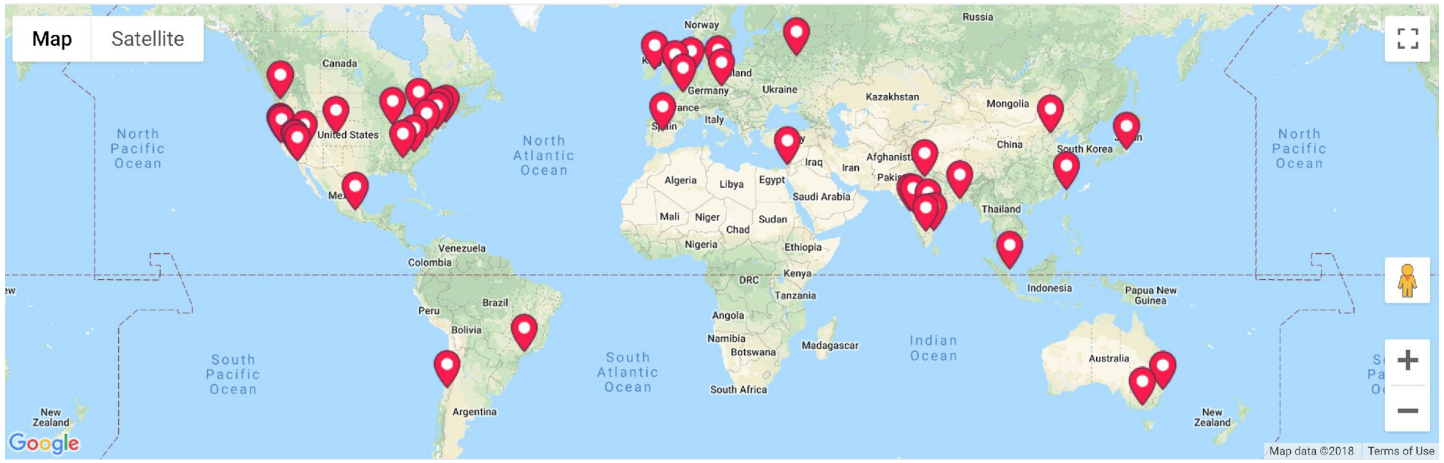
Source: Gartner (February 2018)

Get the Gartner Magic Quadrant here

Partner Ecosystem




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	Groups	Countries
102,646	42	20

London Artificial Intelligence & Deep Learning

London, United Kingdom · 7,410 members ·
Public group

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 H_2O.ai

H_2O.ai

#360Selfie Time!

Meetup Tomorrow (Dec 5th)
Hilton Amsterdam Centraal Station (6-9pm)

5
DEC

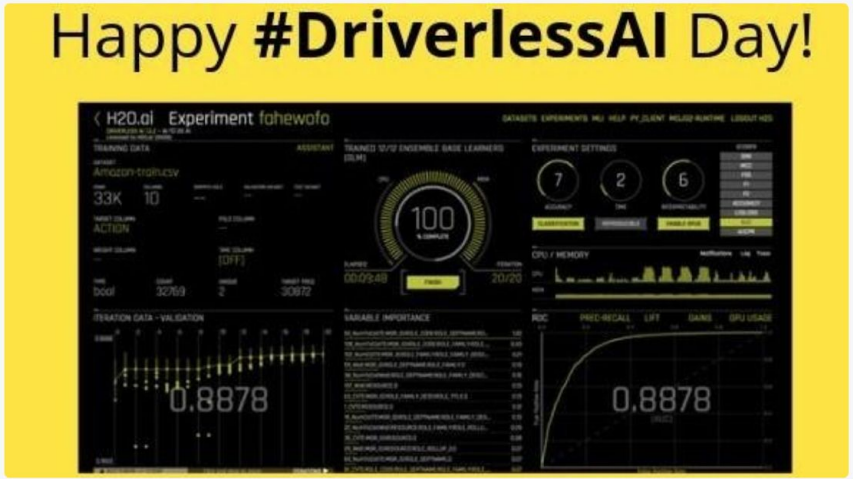
Wednesday, December 5, 2018

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

Share:



Organizer tools

Wednesday, December 5, 2018
6:00 PM to 9:00 PM
Add to calendar

DOUBLETREE BY HILTON AMSTERDAM CENTRAAL STATION
Oosterdoksstraat 4 · Amsterdam

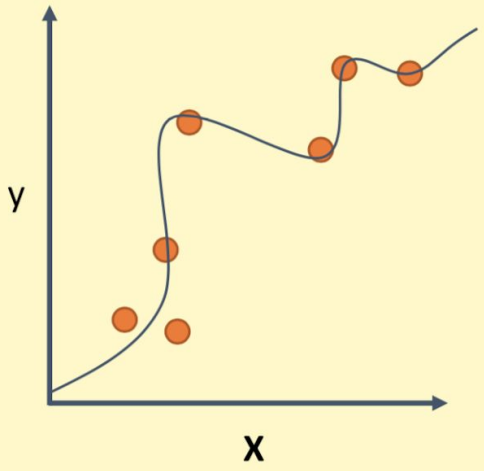
How to find us
Look for rooms Manchester 1 + 2 & Birmingham 1 + 2

H2O.ai

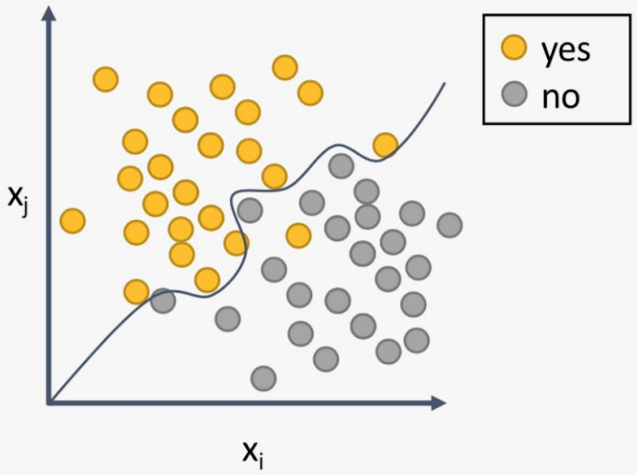
Driverless AI Overview

Supervised Learning

Regression:
How much will a customer spend?



Classification:
Will a customer churn?

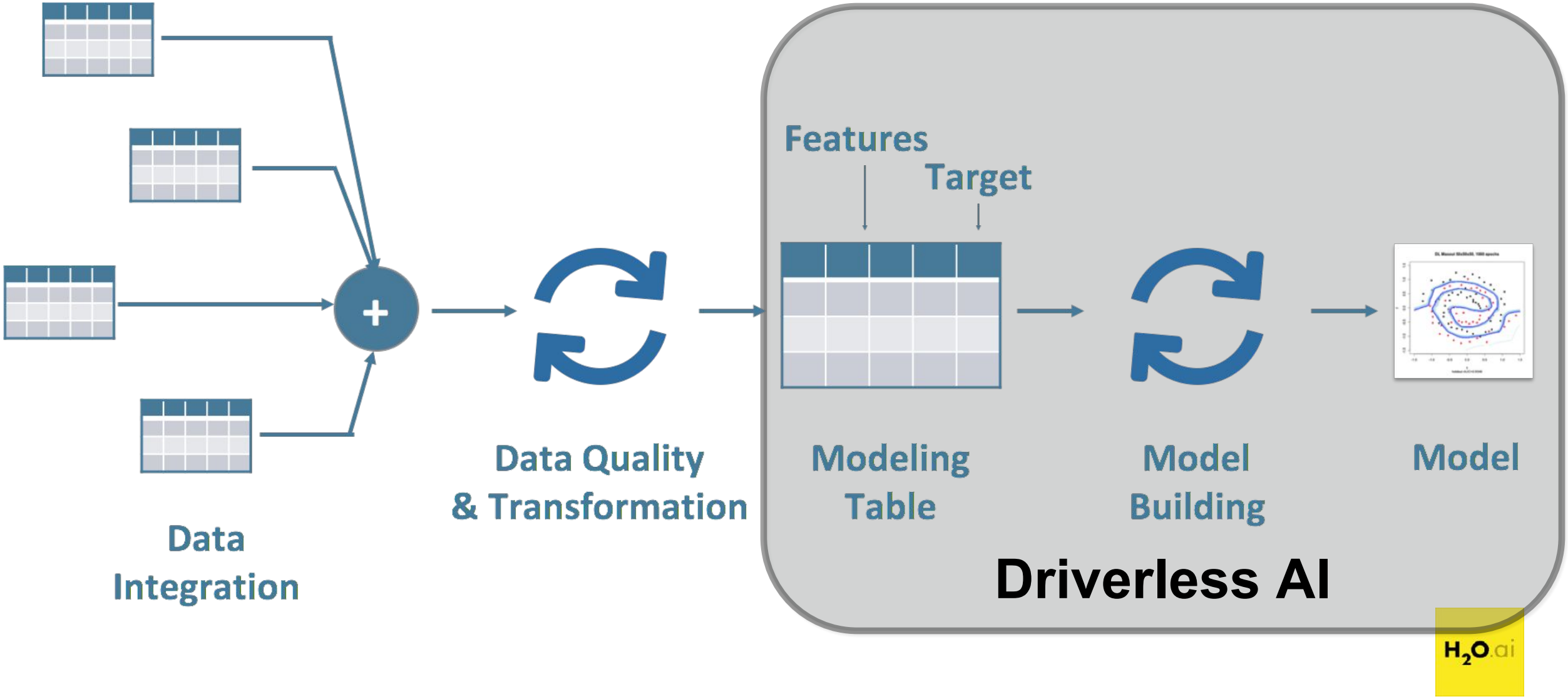


H2O.ai

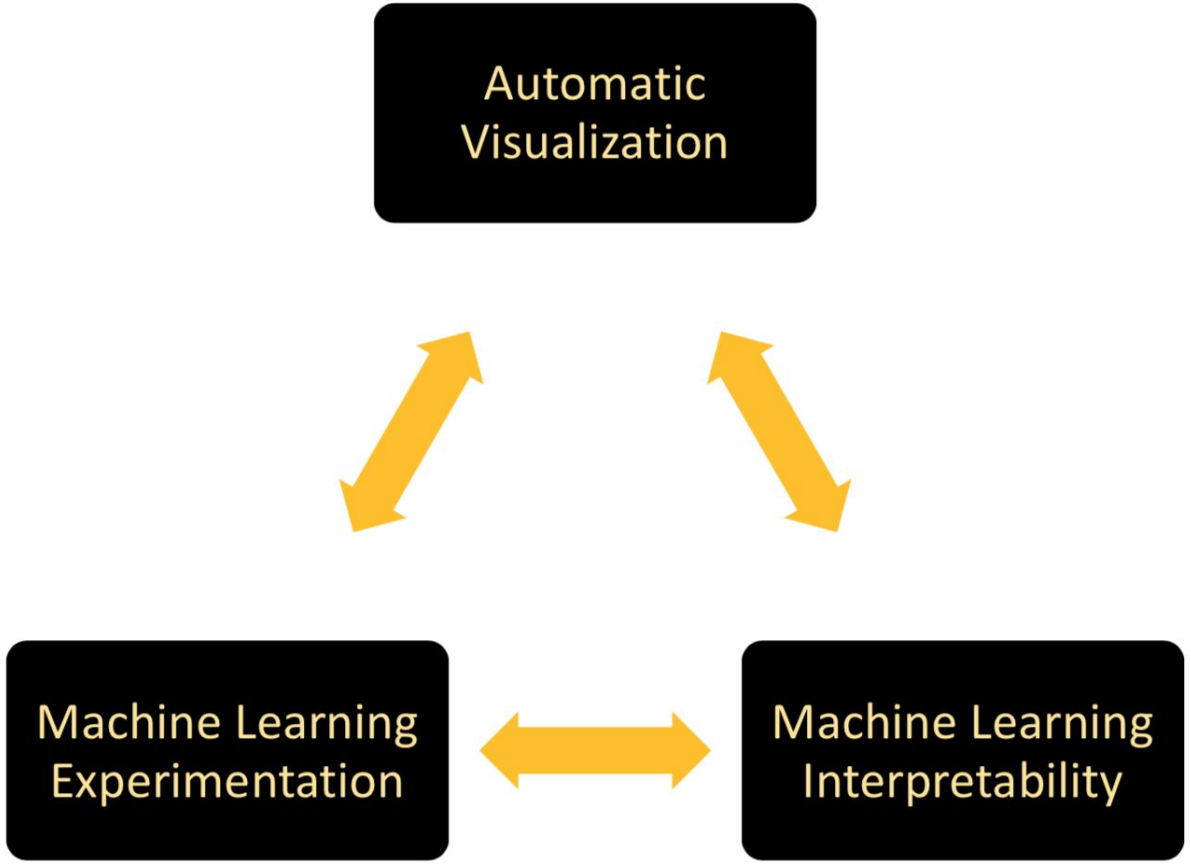
Problems Addressed by Driverless AI

- 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

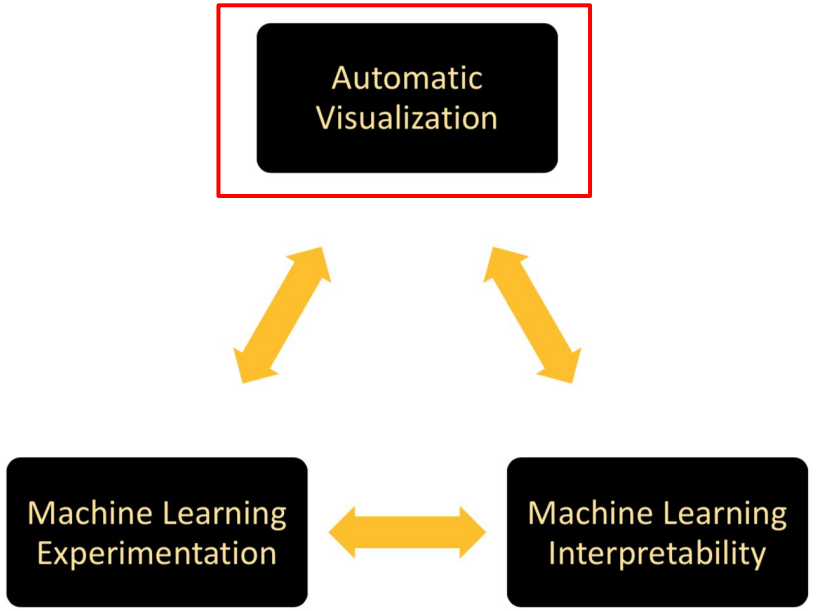
Driverless AI: Automates Data Science and ML Workflows

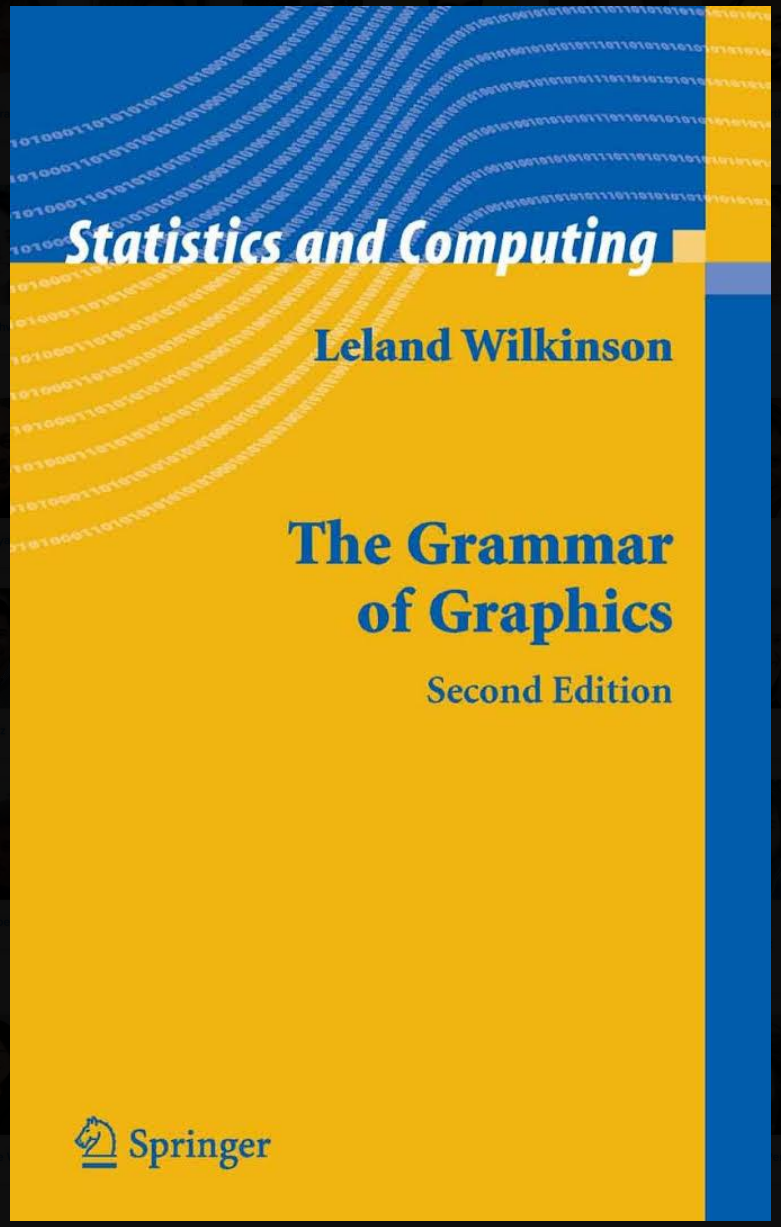


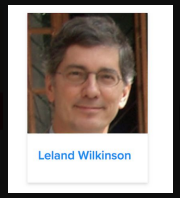
Driverless AI Components



Driverless AI:
Automatic Visualization



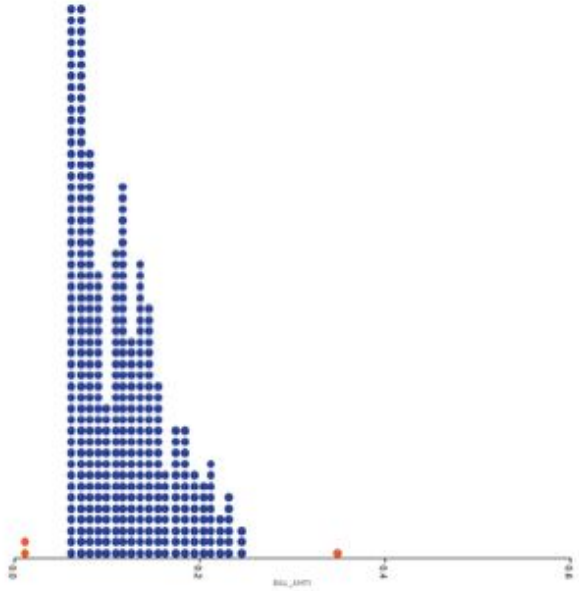


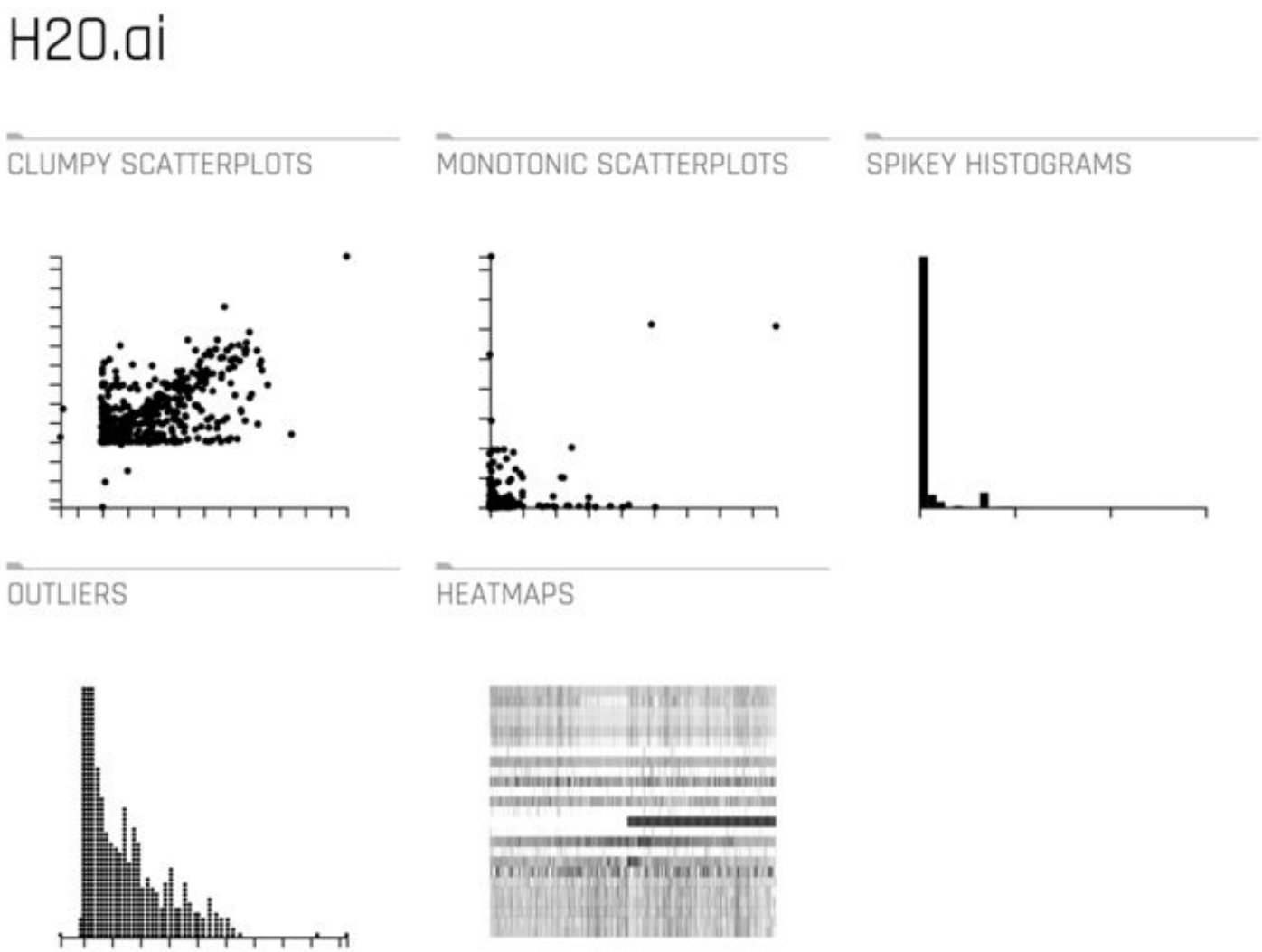


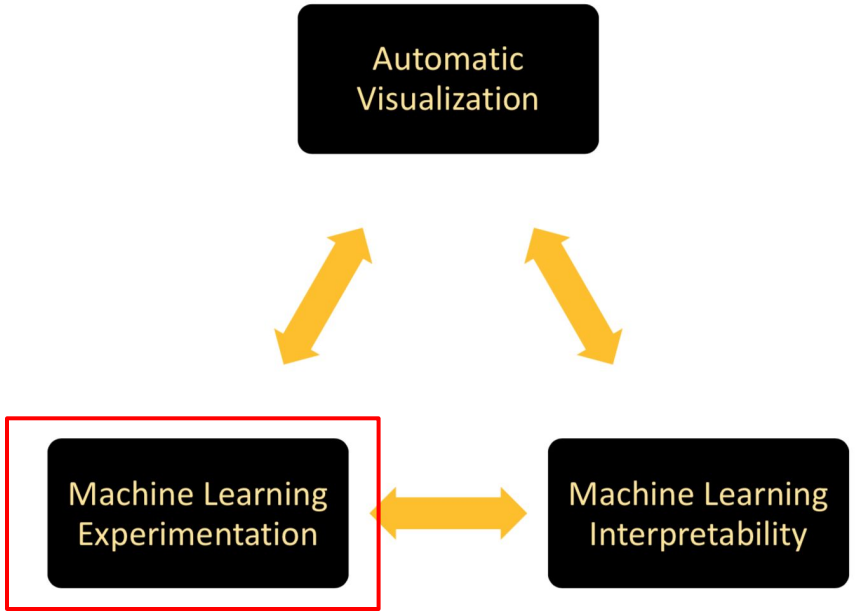
Origin of R Package `ggplot2`

Automatic Visualization

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

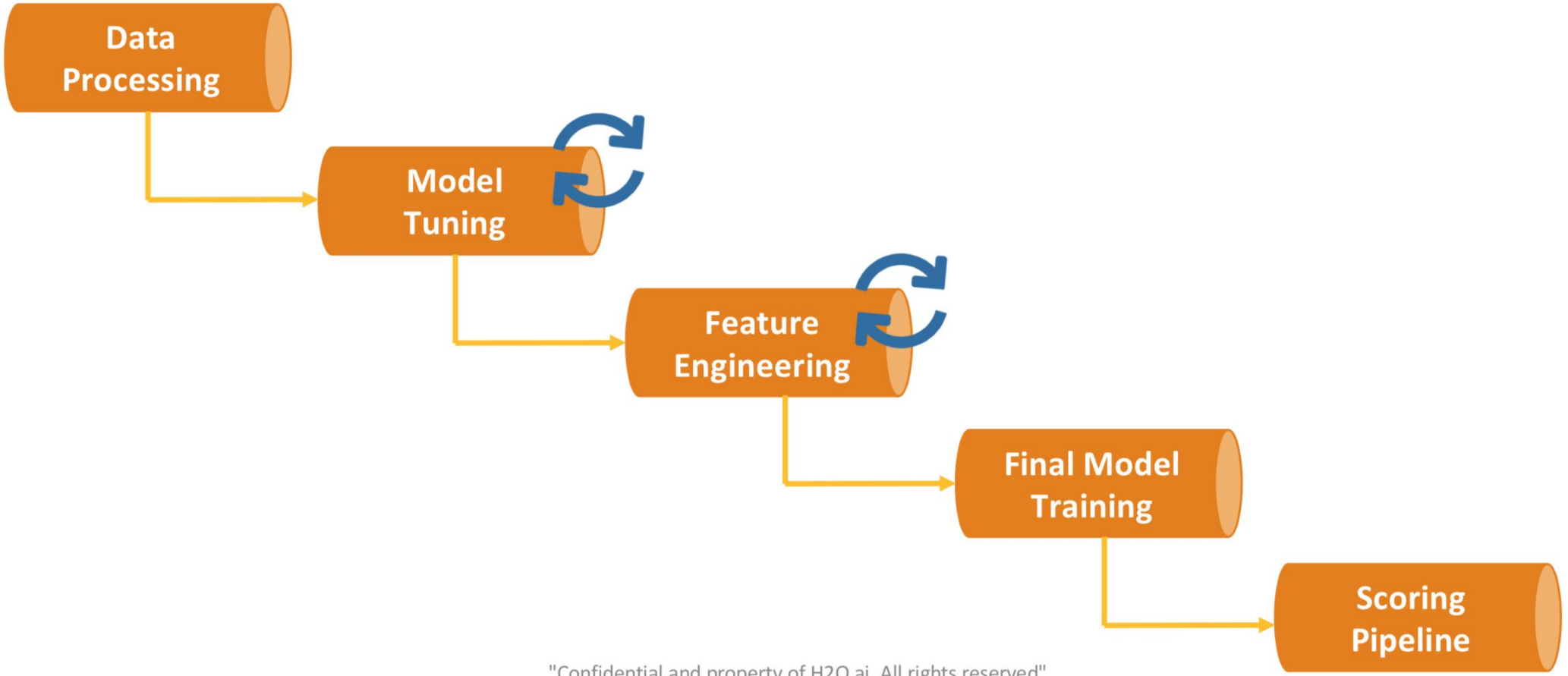






Driverless AI:
Machine Learning Experimentation

Driverless AI: Machine Learning Workflow



H2O.ai

Statistical Learning vs Deep Learning - We Do Both!

What do these settings mean?

ACCURACY
- Training data size: **1,000,000 rows, 8 cols**
- Feature evolution: **GBM, 5-fold CV**
- Final pipeline: **Ensemble (5xGBM, 2xTensorFlow), 5-fold CV**

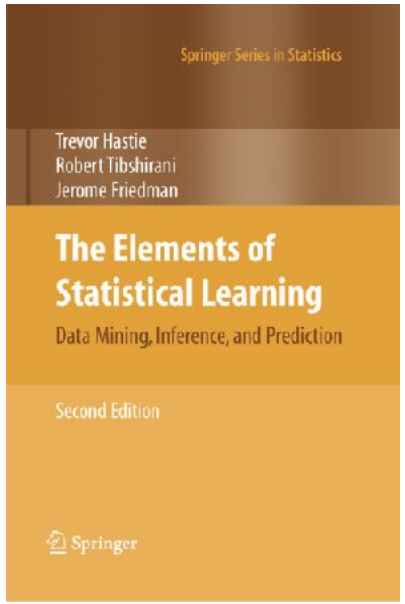
TIME
- Feature evolution: **8 individuals**, up to **47 iterations**
- Early stopping: After **5** iterations of no improvement

INTERPRETABILITY
- Feature pre-pruning strategy: None
- Monotonicity constraints: disabled
- Feature engineering search space (where applicable):
['Clustering', 'Date', 'FrequencyEncoding', 'Identity', 'Interactions', 'NumEncoding', 'TargetEncoding', 'Text', 'TruncatedSVD', 'WeightOfEvidence']

GBM models to train:
- Model and feature tuning: **165**
- Feature evolution: **620**
- Final pipeline: **35**

Estimated max. total memory usage:
- Feature engineering: **2.6GB**
- GPU XGBoost: **656.0MB**

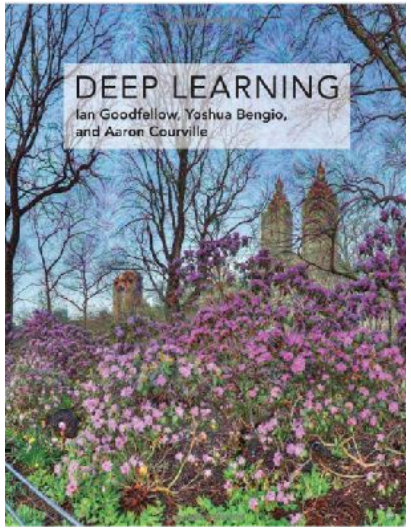
Estimated runtime: **10 minutes**



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

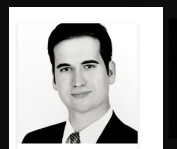
H2O.ai

48th

Branden Murray

Dmitry Larko

25th

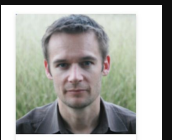


Mathias Müller

4th

Marios Michailidis

1st



Mark Landry

33rd

Kaggle Grandmasters (and their Highest Rank)

113 Grandmasters	980 Masters	3,339 Experts	46,135 Contributors	33,242 Novices

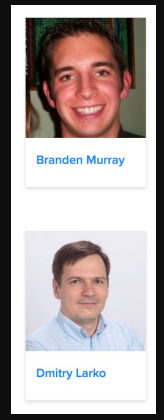
About 80,000 Kagglers

13th

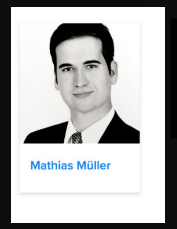
H_2O Team

H_2O.ai

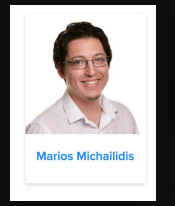
48th



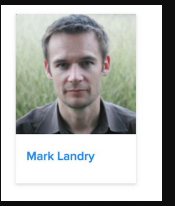
25th



4th



1st



33rd

181st

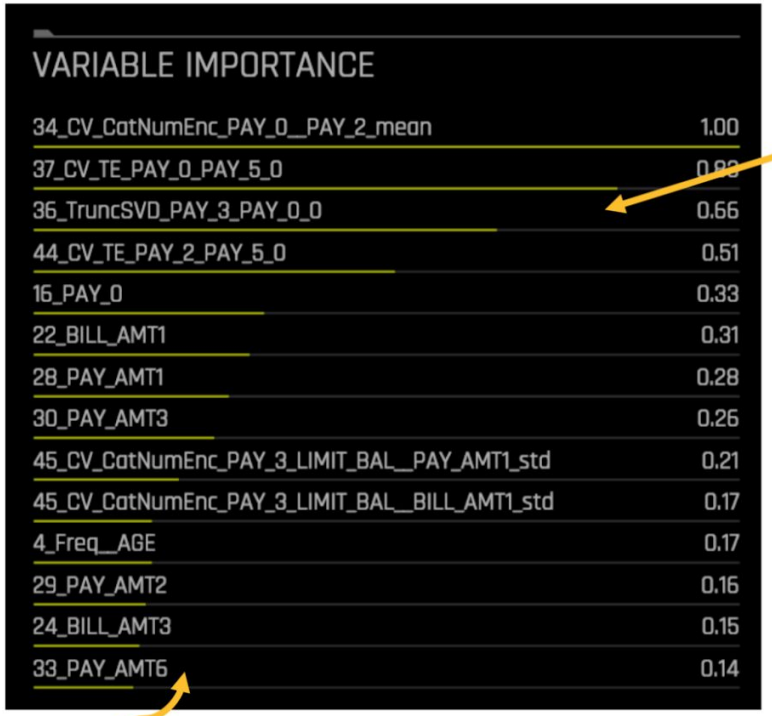
Hoping to get closer to them at some point …

13th

H2O Team

H2O.ai

Automatic Feature Engineering: Kaggle Grandmasters' Recipes



Generated Features

Original Features

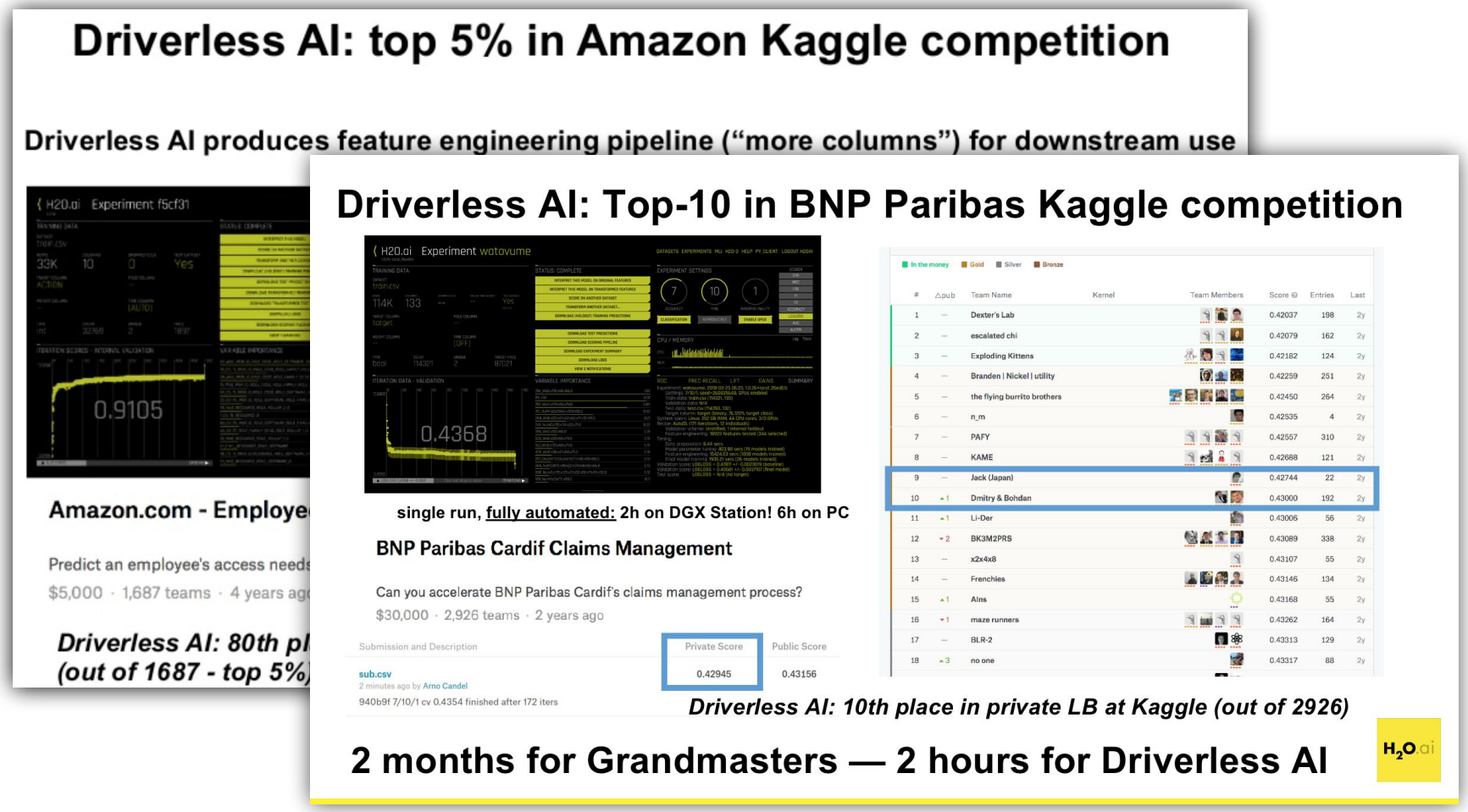
Feature Transformations

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

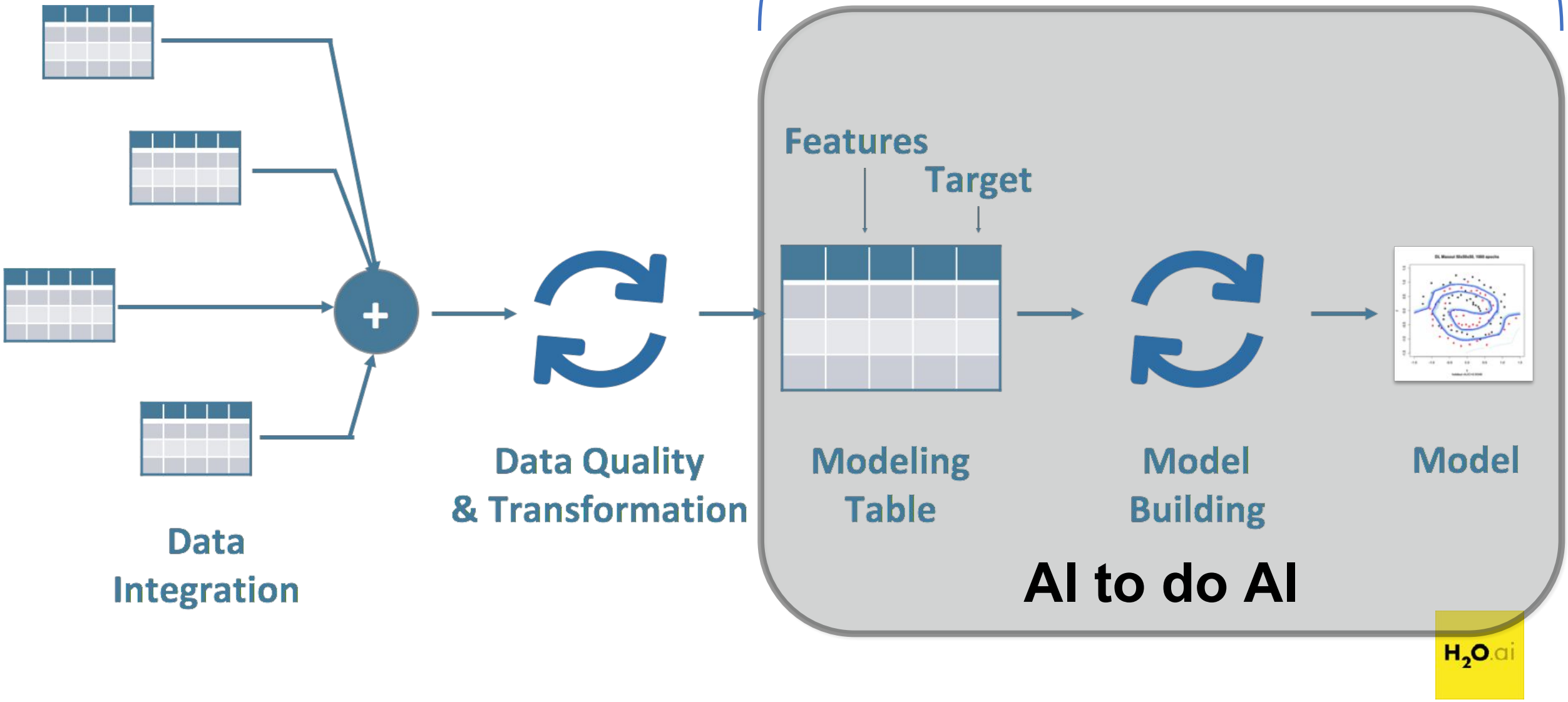
H2O.ai

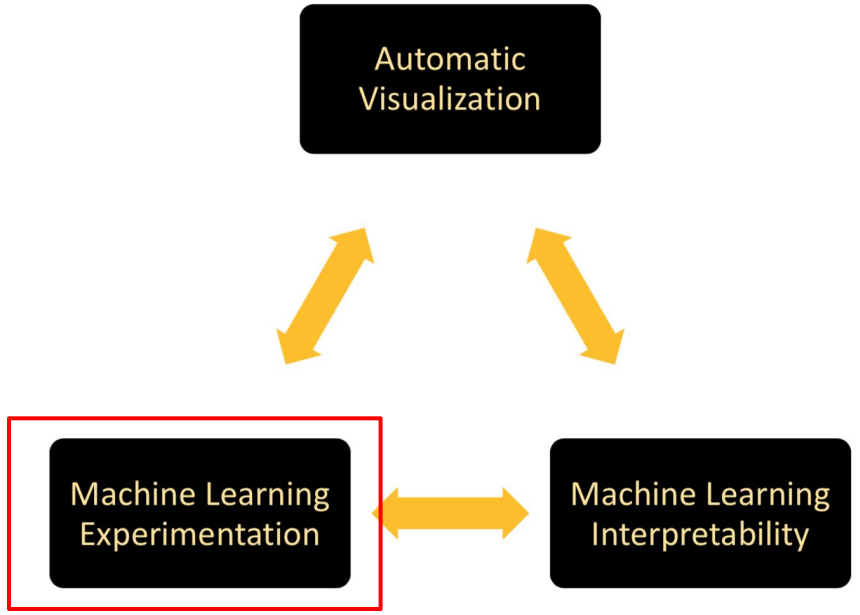
Accuracy

- 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



Deployment: Auto Generated Pipelines





Driverless AI:
Machine Learning Interpretability

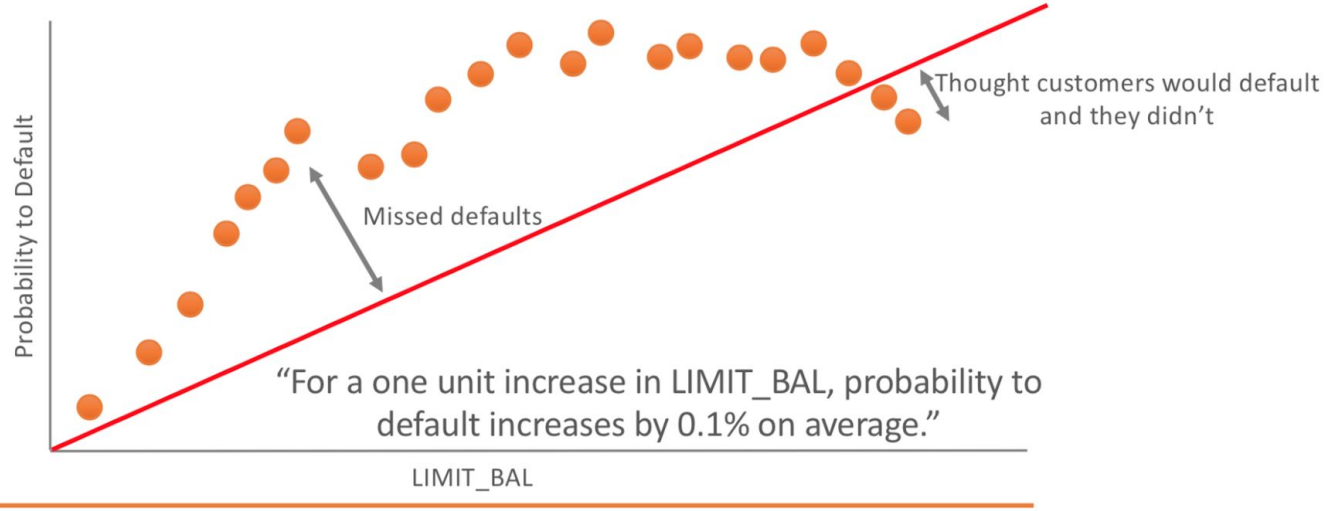
Interpretability

- 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

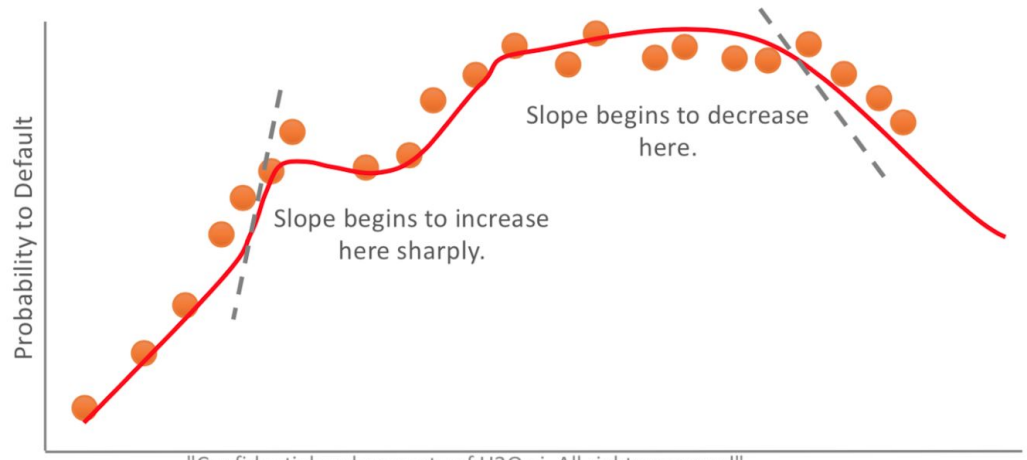
Linear Models

Exact explanations for approximate models.



Machine Learning

Approximate explanations for exact models.



H2O.ai

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%**

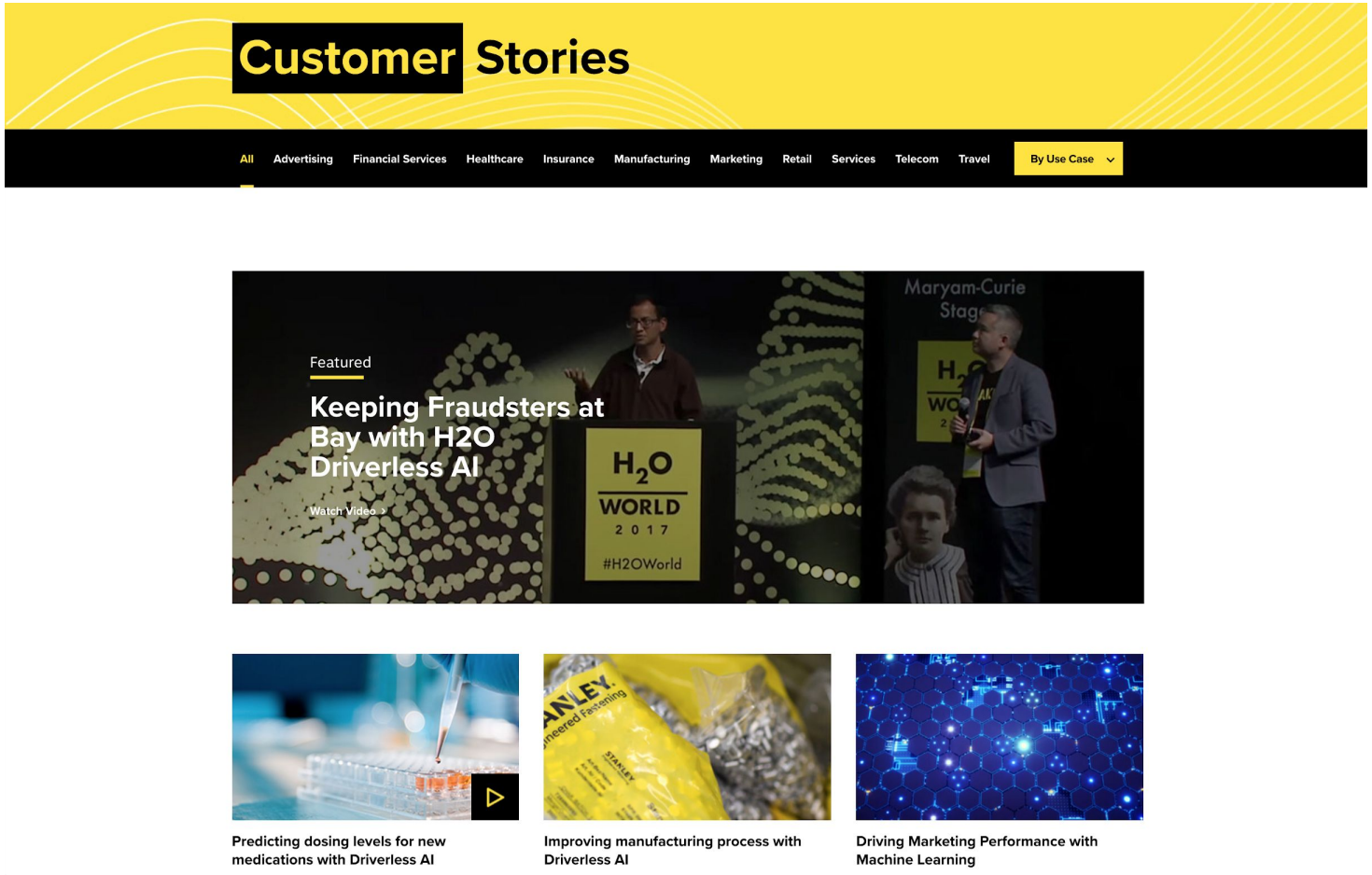
Their Pay Amount is $3,000 **decreases their likelihood by 1%**

Driverless AI:
Real-World Use Cases

H2O Driverless AI Delivers Value in Every Industry

Financial Services	Healthcare	Marketing	Manufacturing
+6% Accuracy	Near perfect scores	2.5X Performance	1 month savings
Matched 10 years of Machine Learning Expertise	Increased customer satisfaction	Outperforms alternative digital marketing	Accurately predicting supply chain

www.h2o.ai/customer-stories/



H2O.ai

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.

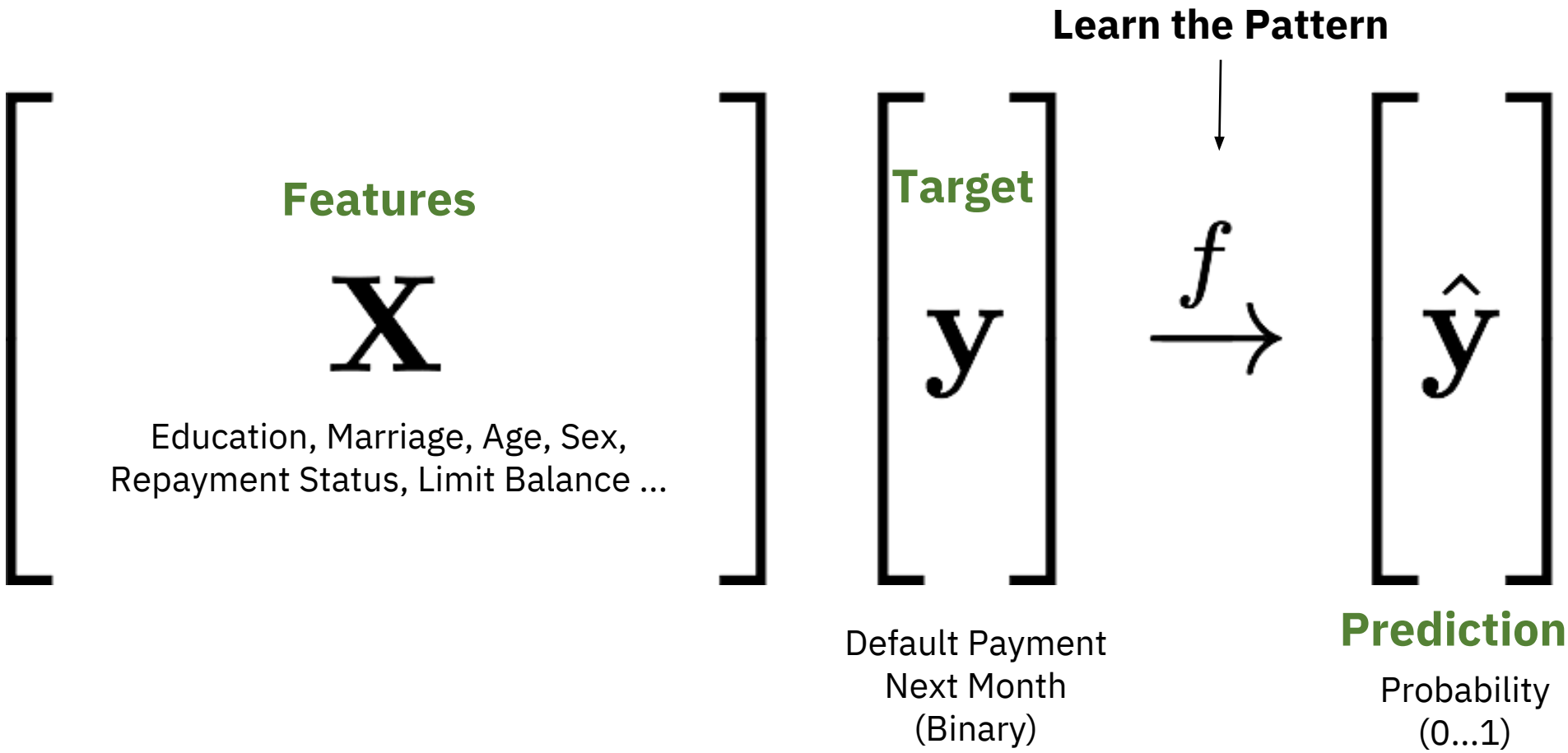
Credit Card Example

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**

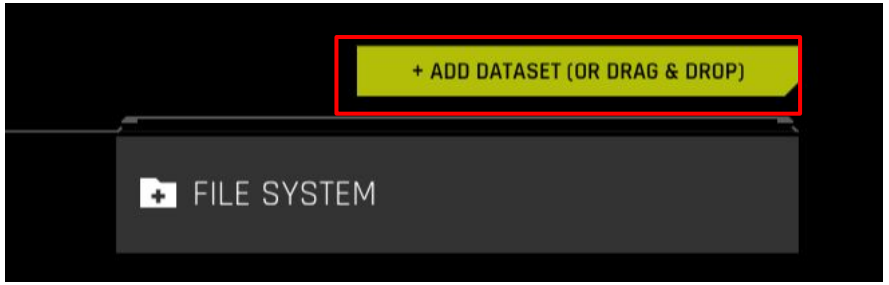
Learning from Credit Card Data

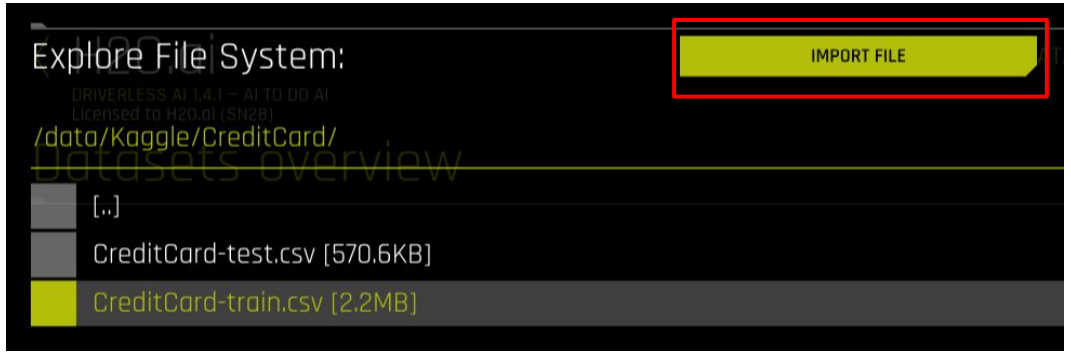


H2O.ai

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

Credit Card Data





Credit Card Data

2. Click on **CreditCard-train.csv**. You will see four available options:
   a. Details
   b. Visualize
   c. Predict
   d. Delete

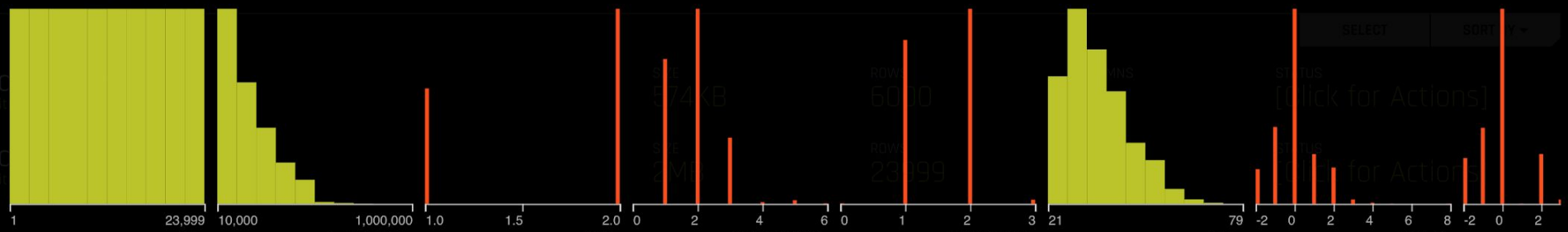
3. Let's start with **Details**

Dataset Details: CreditCard-train.csv

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

DATASET ROWS

Start typing to filter out columns



Name	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0
Type	int	int	int	int	int	int	int
Count	23999	23999	23999	23999	23999	23999	23999
Mean	12000	165498.7158	1.6283	1.8475	1.5572	35.3809	-0.0031
Min	1	10000	1	0	0	21	-2
Max	23999	1000000	2	6	3	79	8
StdDev	6928.059	129130.743	0.483	0.78	0.522	9.271	1.123
Freq	1	2740	15078	11360	12876	1284	11738
Unique	23999	79	2	7	4	55	11

Dataset Details: CreditCard-train.csv

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

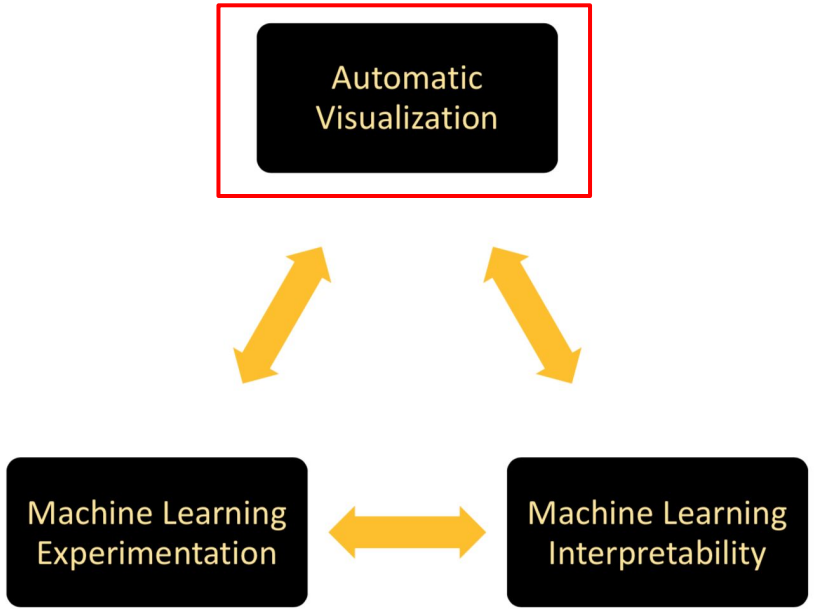
DATASET OVERVIEW

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	0	689	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	8617	5670	35835	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	6500	2870	1000	6500	6500	6500
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59695	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	0
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20515	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	316	0	632
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601

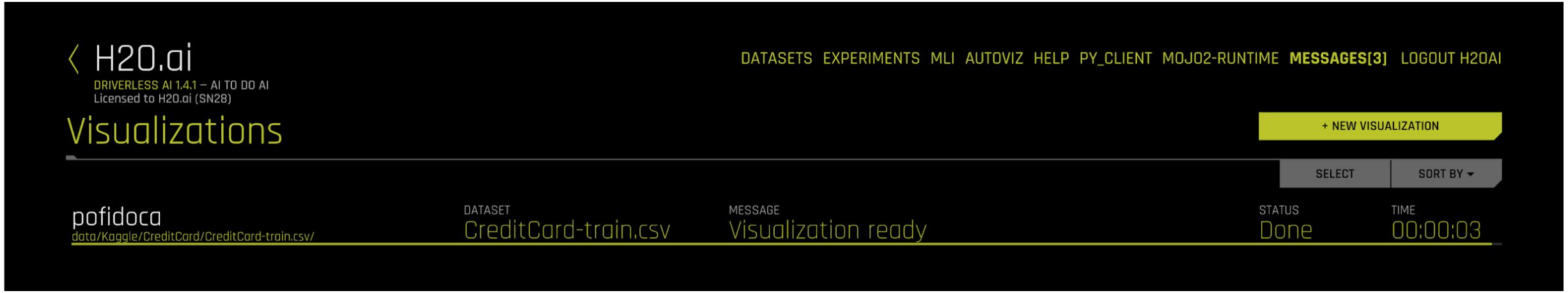
1 2 3 4 5 6 7 8 9 10 ... 1044 Next Page

Rows 1-23 of 23999 total

Hands-on Experiment: **Visualize**



Datasets → CreditCard-train.csv → Visualize



H2O.ai

H2O.ai

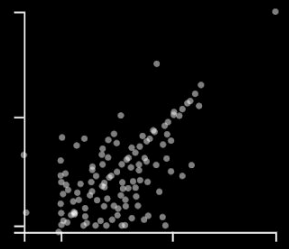
DRIVERLESS AI 1.4.1 – AI TO DO AI
Licensed to H2O.ai (SN28)

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

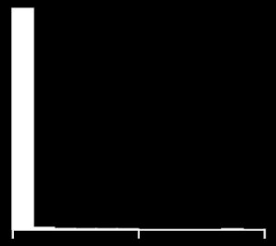
Visualizations for: CreditCard-train.csv

DISPLAY LOGS

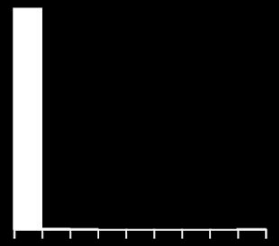
CORRELATED SCATTERPLOTS



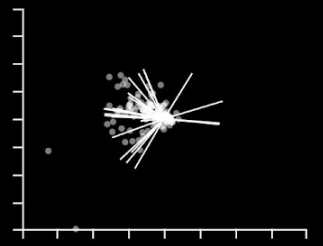
SKEWED HISTOGRAMS



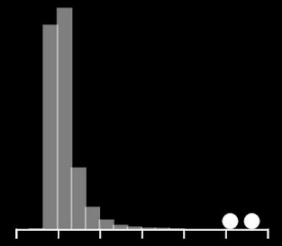
GAPS HISTOGRAMS



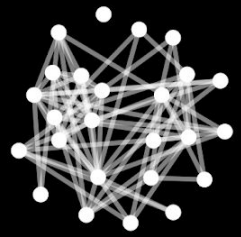
BIPLOTS



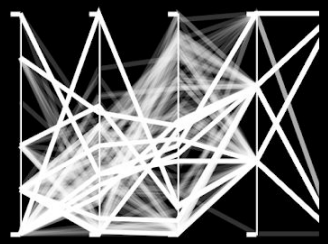
OUTLIERS



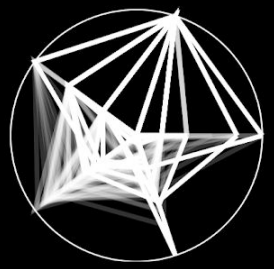
CORRELATION GRAPH



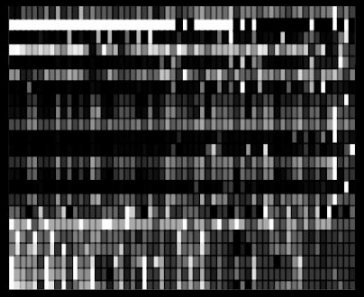
PARALLEL COORDINATES PLOT

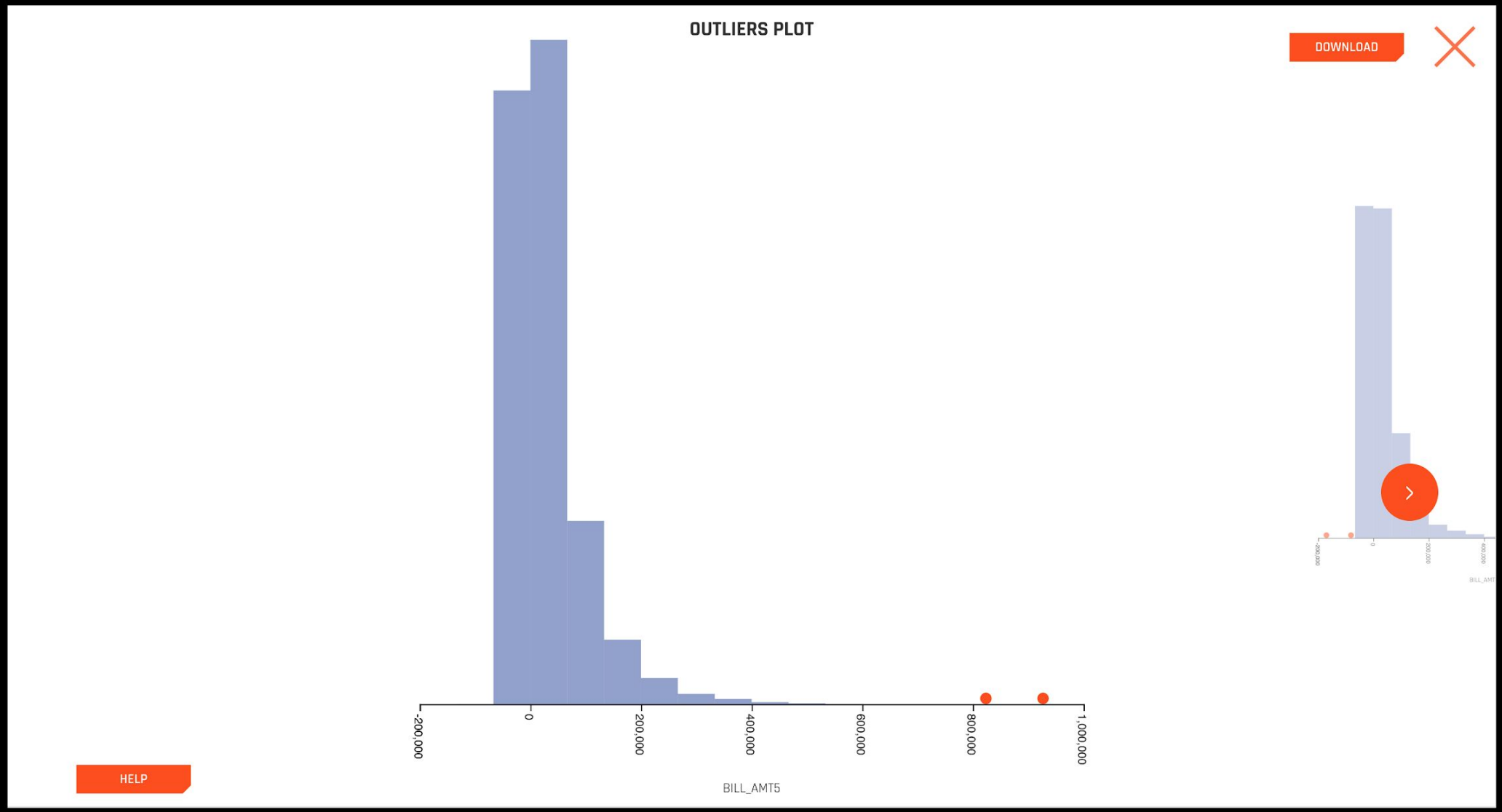


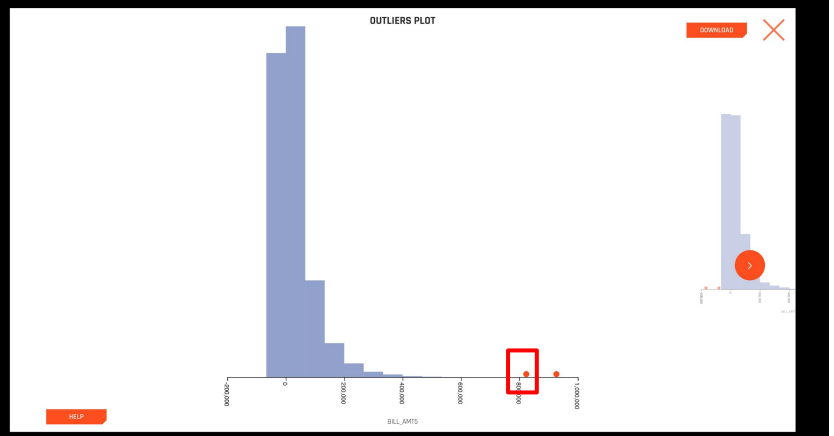
RADAR PLOT



DATA HEATMAP

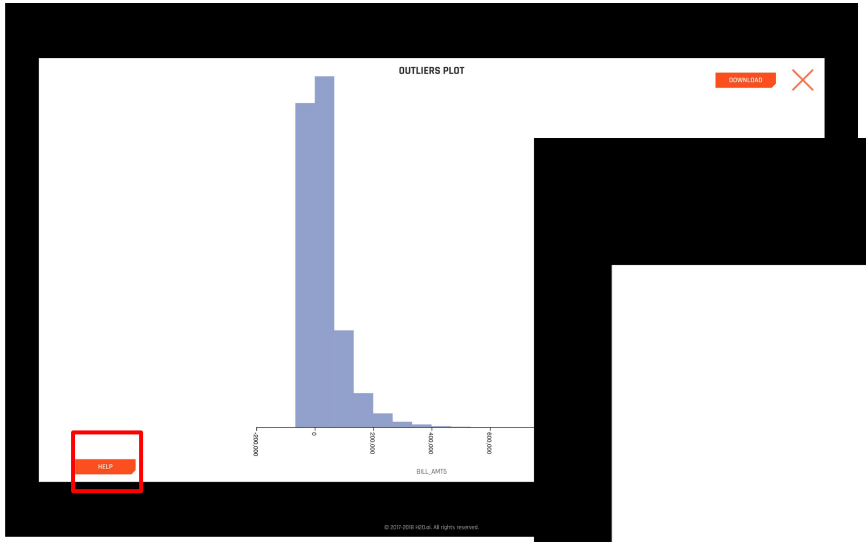






ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	P/
20893	550000	1	1	2	35	2	2	2	2	2	0	539092	552234	565550	572805	823540	501370	

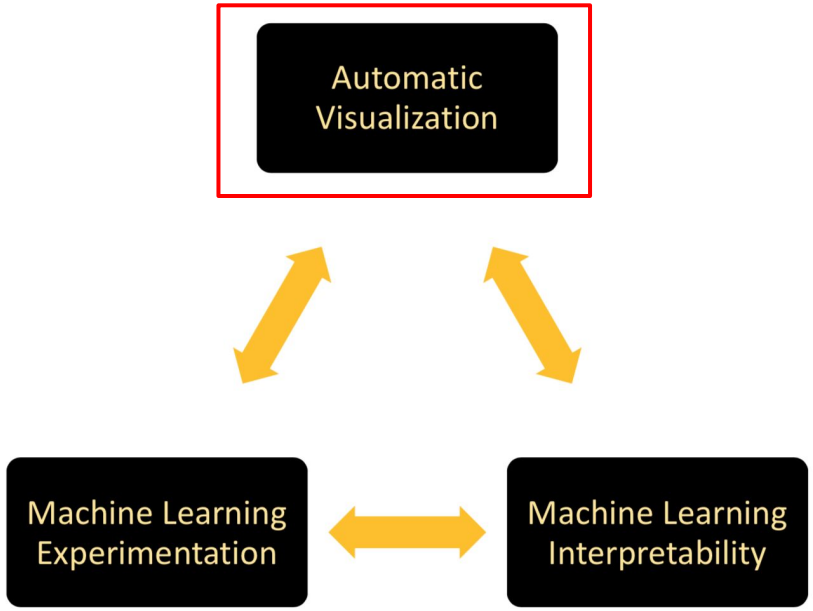
Rows 1-1 of 1 total



OUTLIERS PLOT

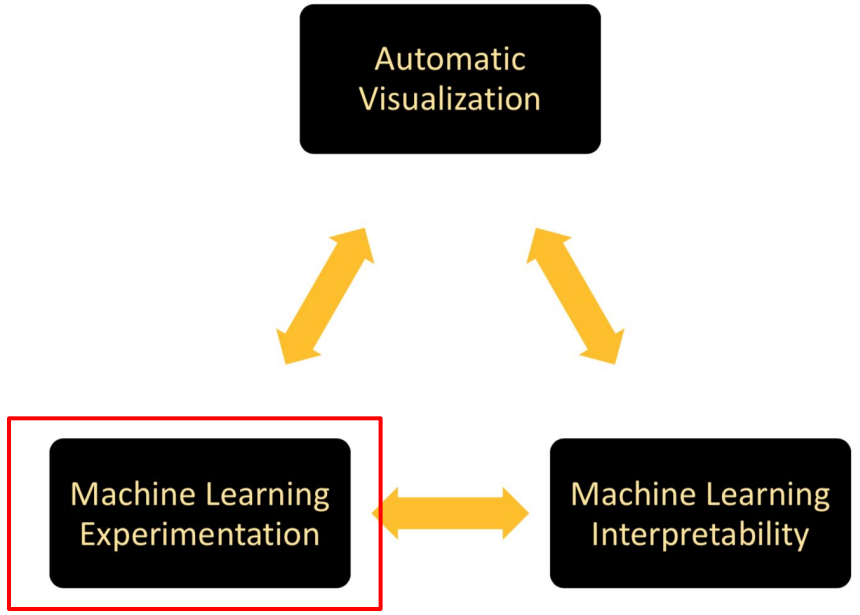
DOWNLOAD

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.

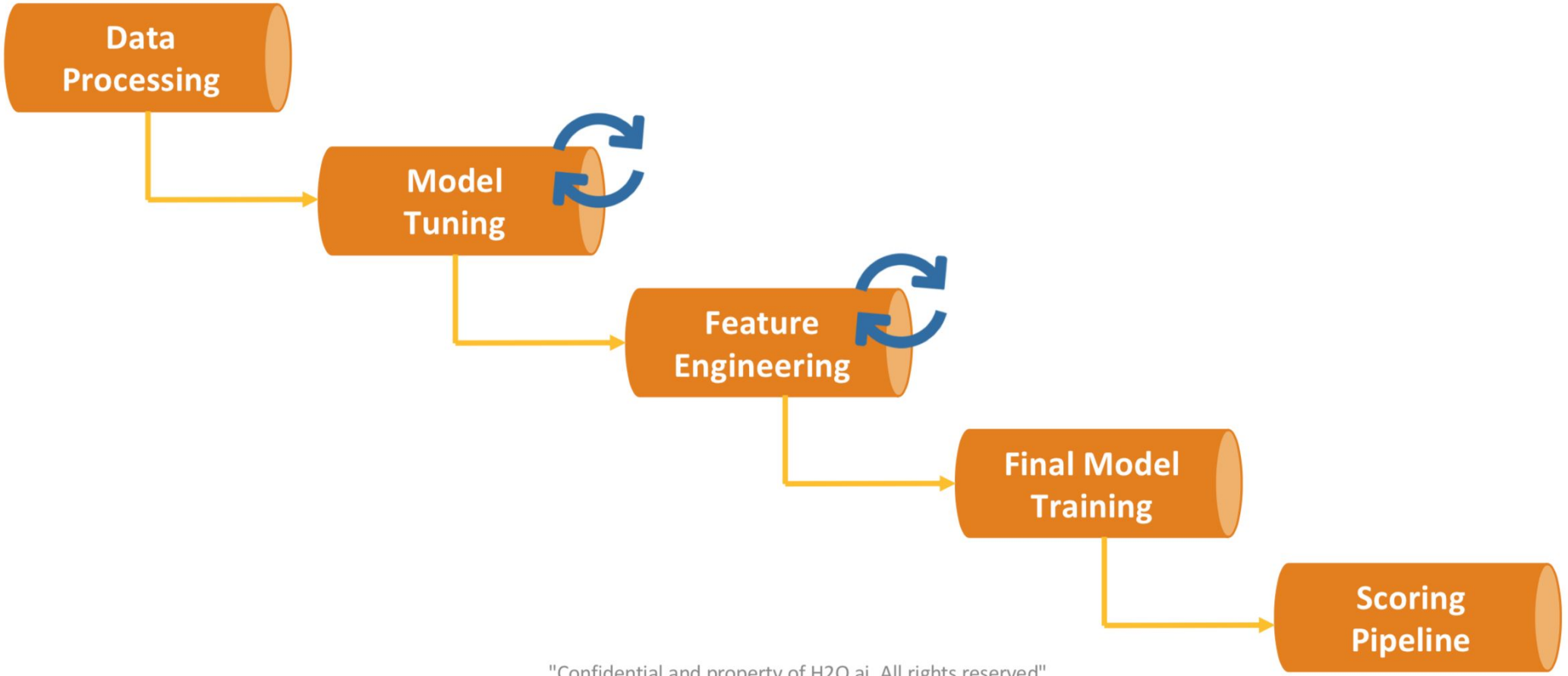


Hands-on Experiment:
Visualize - Explore and Play with the Graphs

Hands-on Experiment: **Predict**

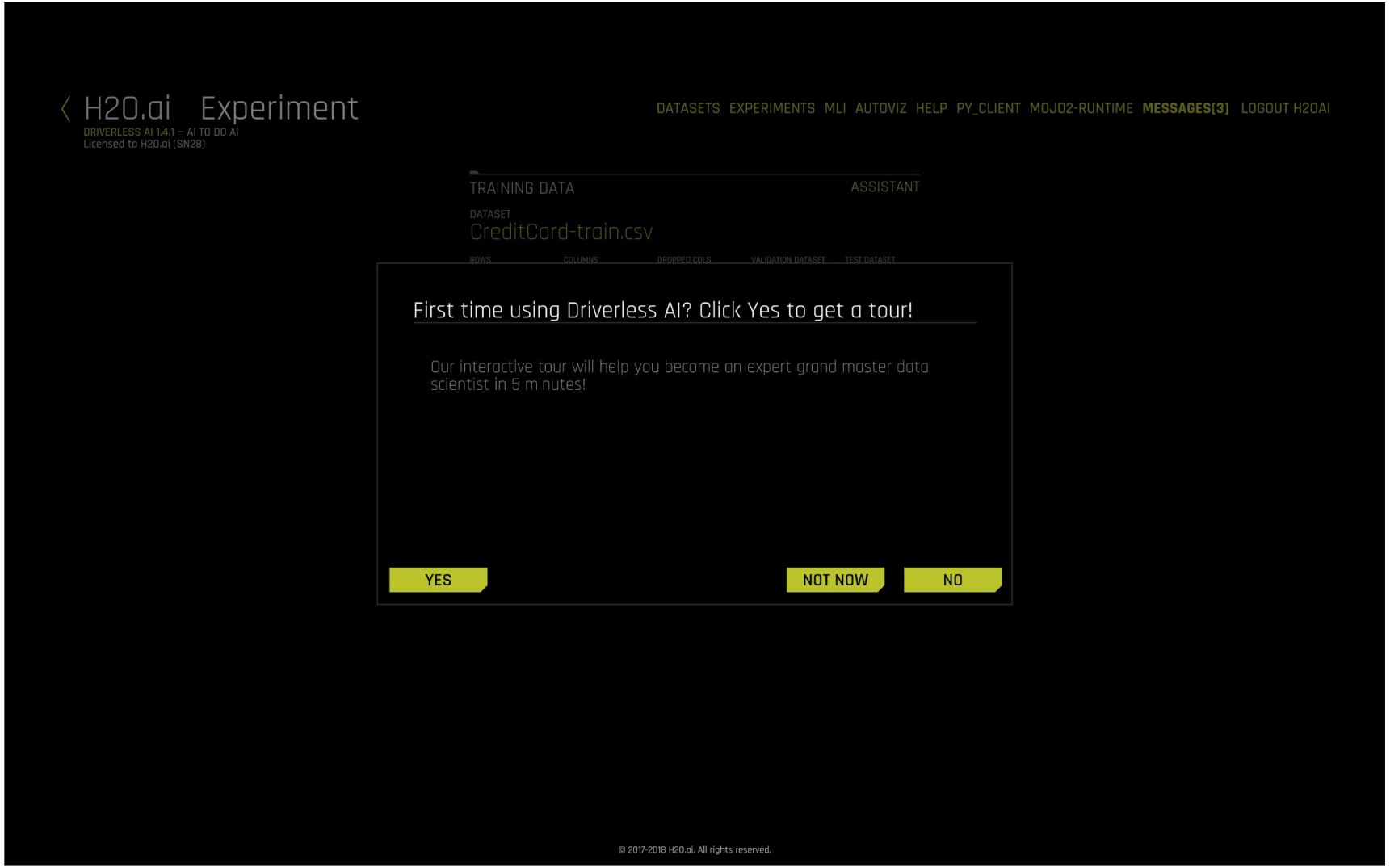


Driverless AI: Machine Learning Workflow

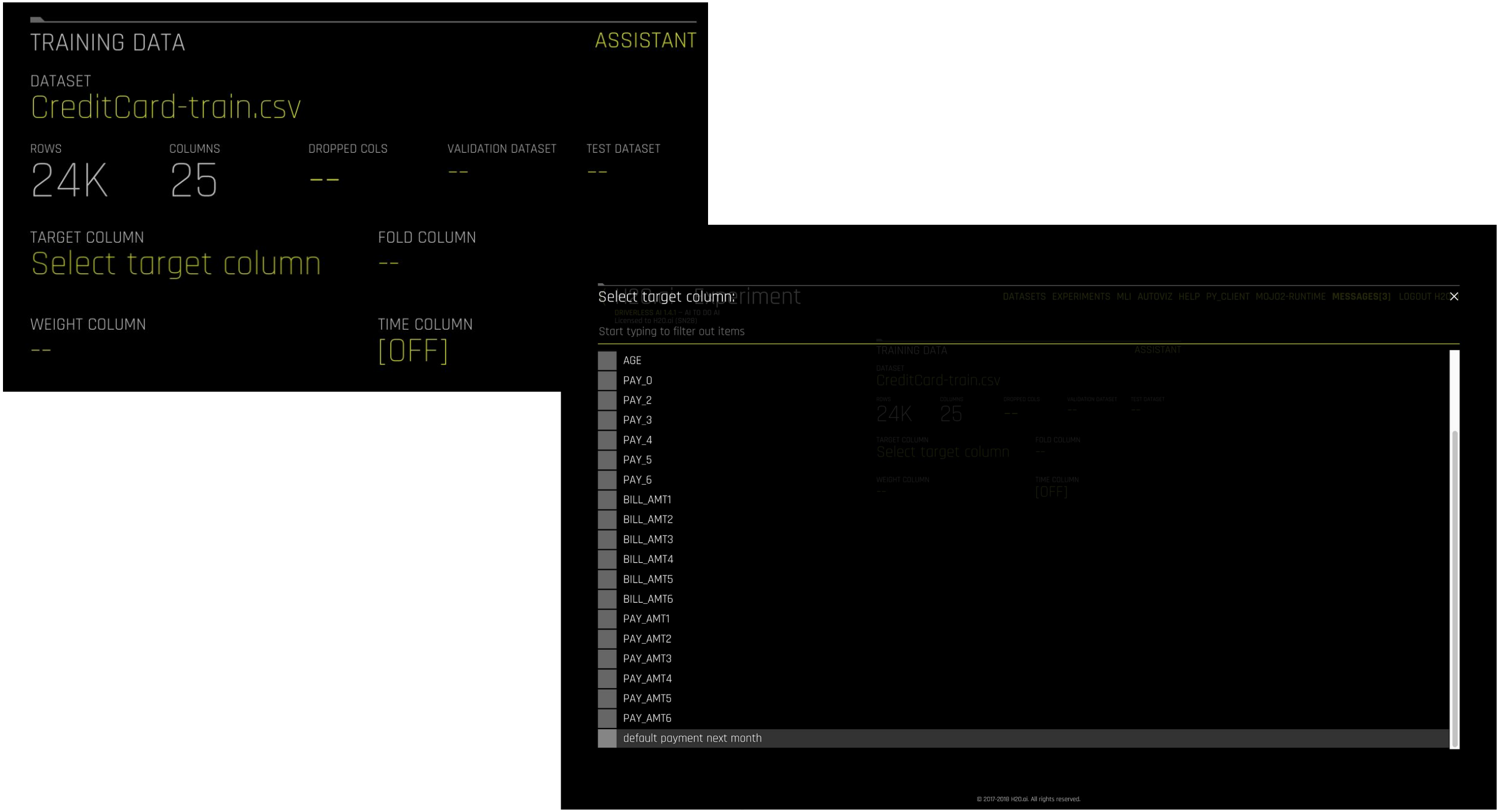


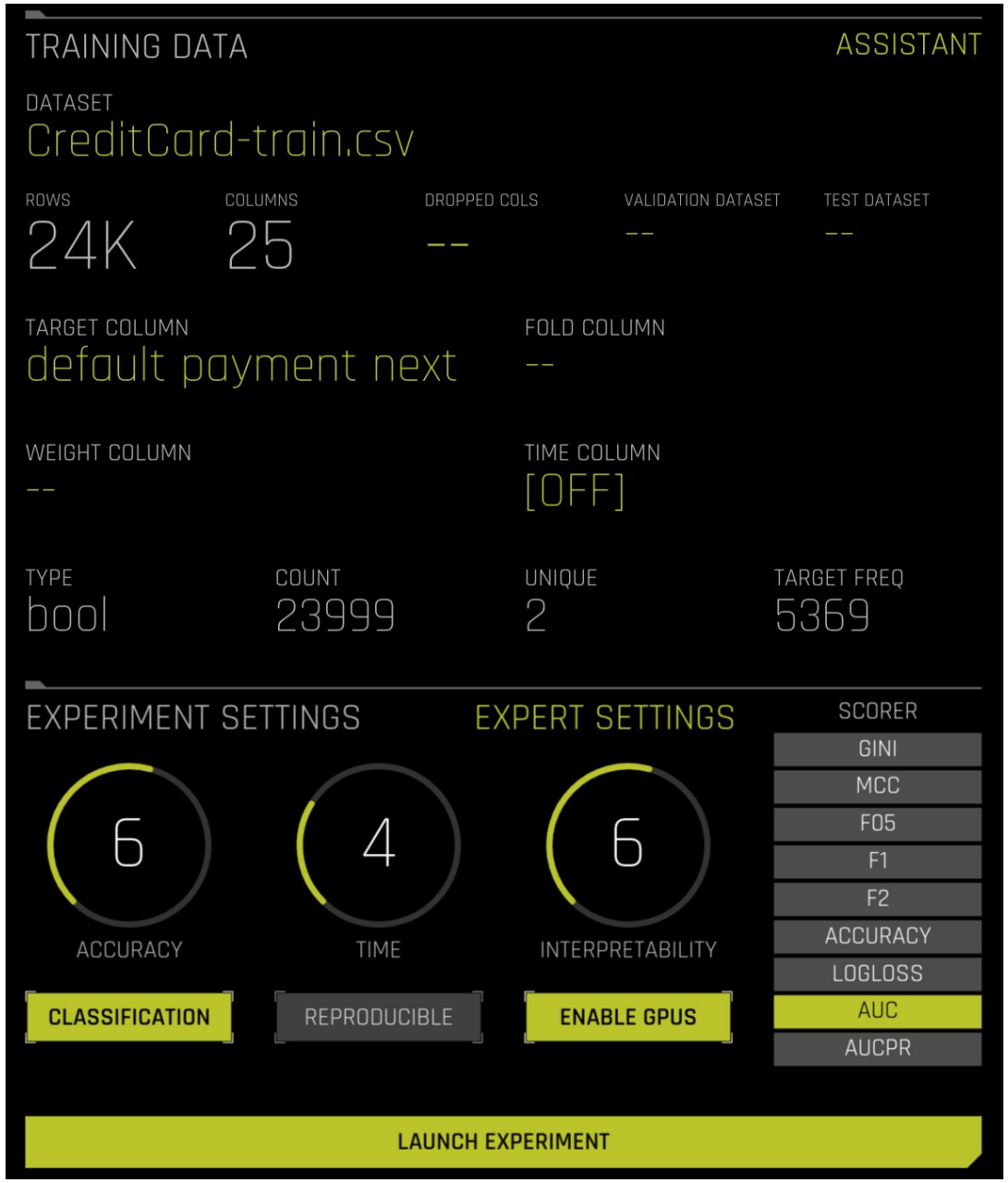
H2O.ai

Datasets → CreditCard-train.csv → Predict



H2O.ai


TRAINING DATA
ASSISTANT
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]
Select target column:
DATASETS EXPERIMENTS MLI AUTOVIZ HELP PY_CLIENT MOJO2-RUNTIME MESSAGES[3] LOGOUT H2
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
© 2017-2018 H2O.ai. All rights reserved.


TRAINING DATA
ASSISTANT
DATASET
CreditCard-train.csv
ROWS
24K
COLUMNS
25
DROPPED COLS
--
VALIDATION DATASET
--
TEST DATASET
--
TARGET COLUMN
default payment next
FOLD COLUMN
--
WEIGHT COLUMN
--
TIME COLUMN
[OFF]
TYPE
bool
COUNT
23999
UNIQUE
2
TARGET FREQ
5369
EXPERIMENT SETTINGS
EXPERT SETTINGS
SCORER
6
ACCURACY
4
TIME
6
INTERPRETABILITY
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR
CLASSIFICATION
REPRODUCIBLE
ENABLE GPUS
LAUNCH EXPERIMENT

H2O.ai

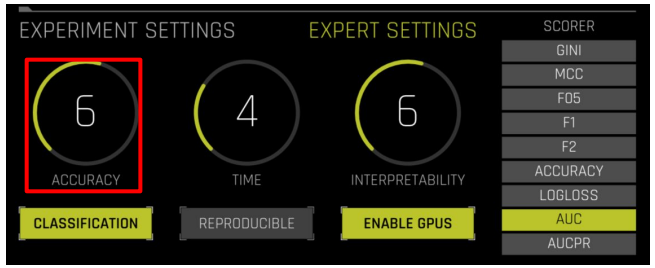
Machine Learning Experimentation

Accuracy

Time

Interpretability

Accuracy



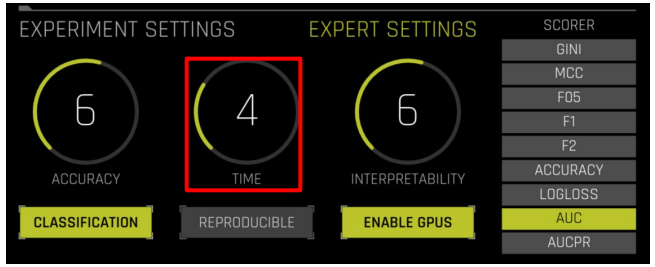
Accuracy	Max Rows x Cols	Ensemble Level	Target Transformation	Parameter Tuning Level	Num Folds	Only First Fold Model	Distribution Check
1	100K	0	False	0	3	True	No
2	1M	0	False	0	3	True	No
3	50M	0	True	1	3	True	No
4	100M	0	True	1	3-4	True	No
5	200M	1	True	1	3-4	True	Yes
6	500M	2	True	1	3-5	True	Yes
7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

H2O.ai

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.

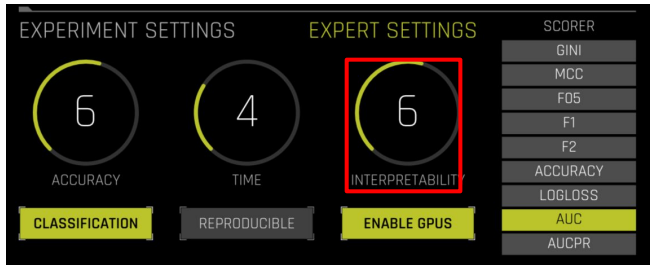


Time	Iterations	Early Stopping Rounds
1	1-5	None
2	10	5
3	30	5
4	40	5
5	50	10
6	100	10
7	150	15
8	200	20
9	300	30
10	500	50

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

H2O.ai

Interpretability

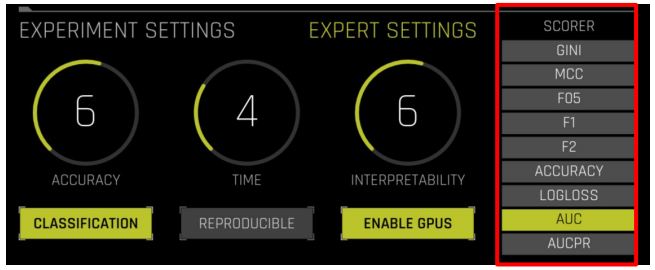


Interpretability	Ensemble Level	Target Transformation	Feature Engineering	Feature Pre-Pruning	Monotonicity Constraints
1 - 3	<= 3			None	Disabled
4	<= 3	Inverse		None	Disabled
5	<= 3	Anscombe	Clustering (ID, distance) Truncated SVD	None	Disabled
6	<= 2	Logit Sigmoid		Feature selection	Disabled
7	<= 2		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

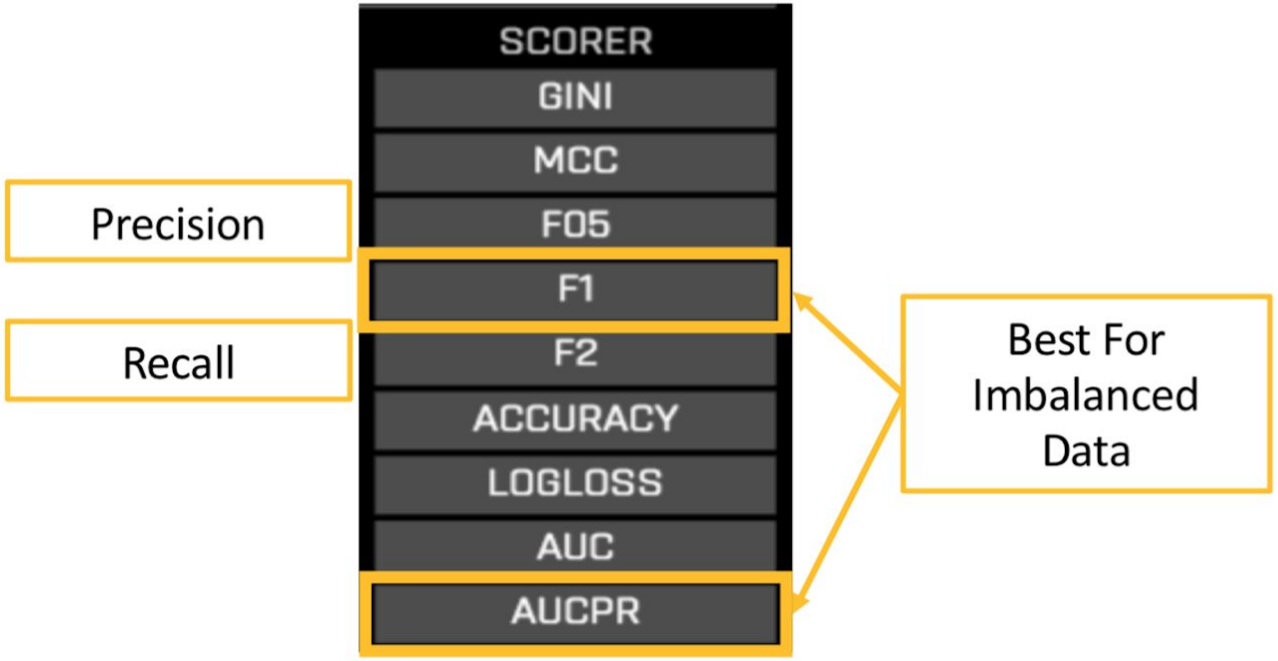
http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/running-experiment.html#interpretability

H2O.ai

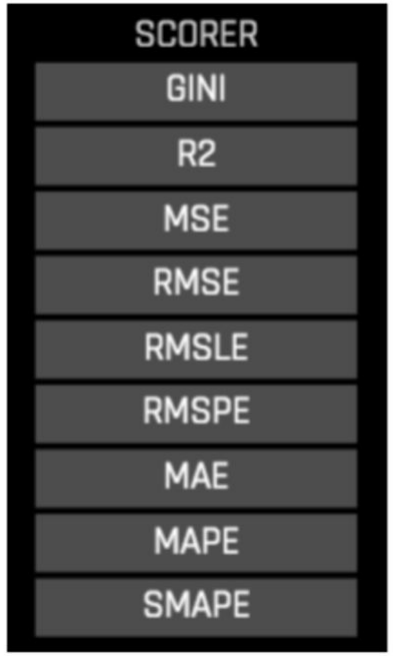
Scorer (Metrics)



Classification



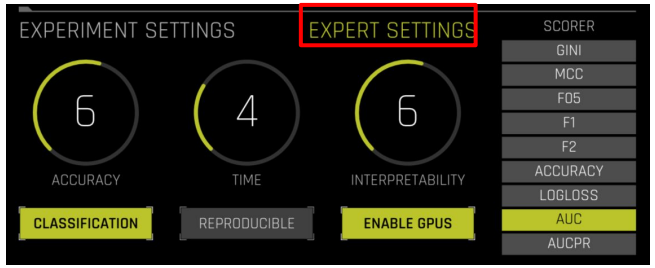
Regression

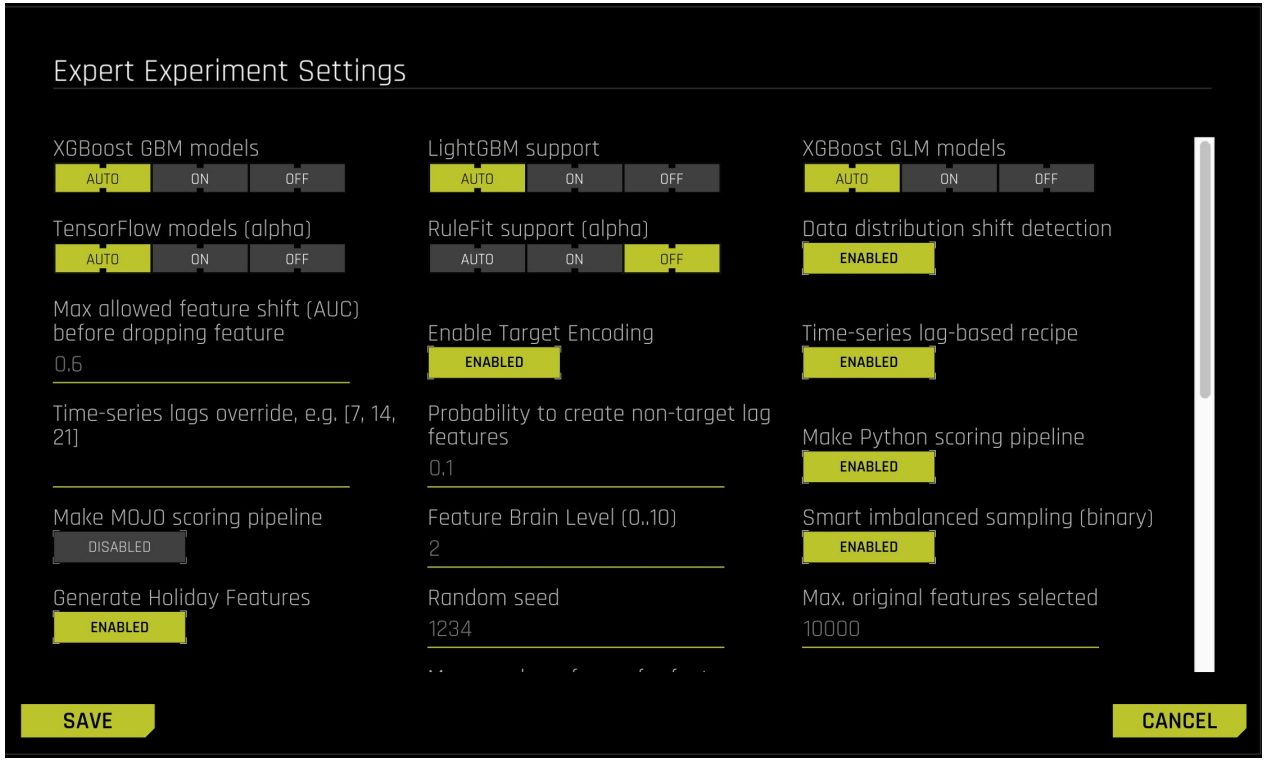


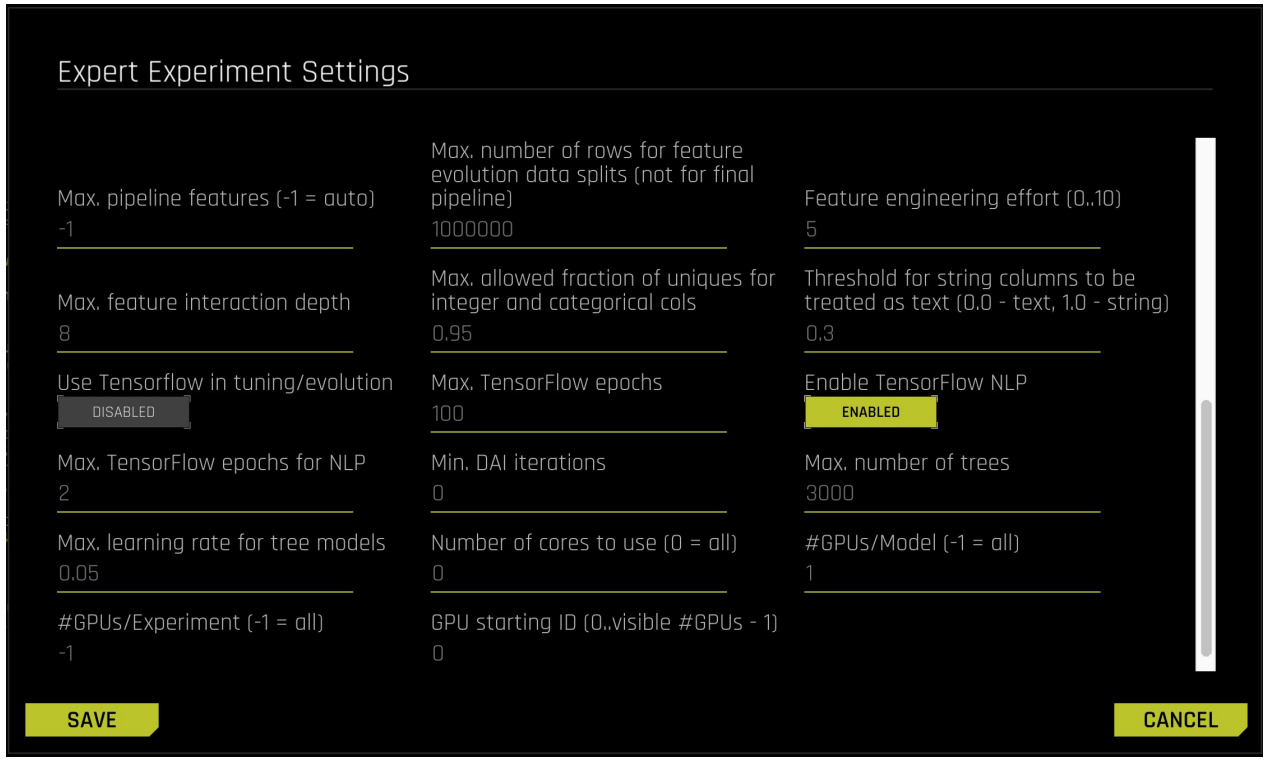
http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/scorers.html

H2O.ai

Expert Settings

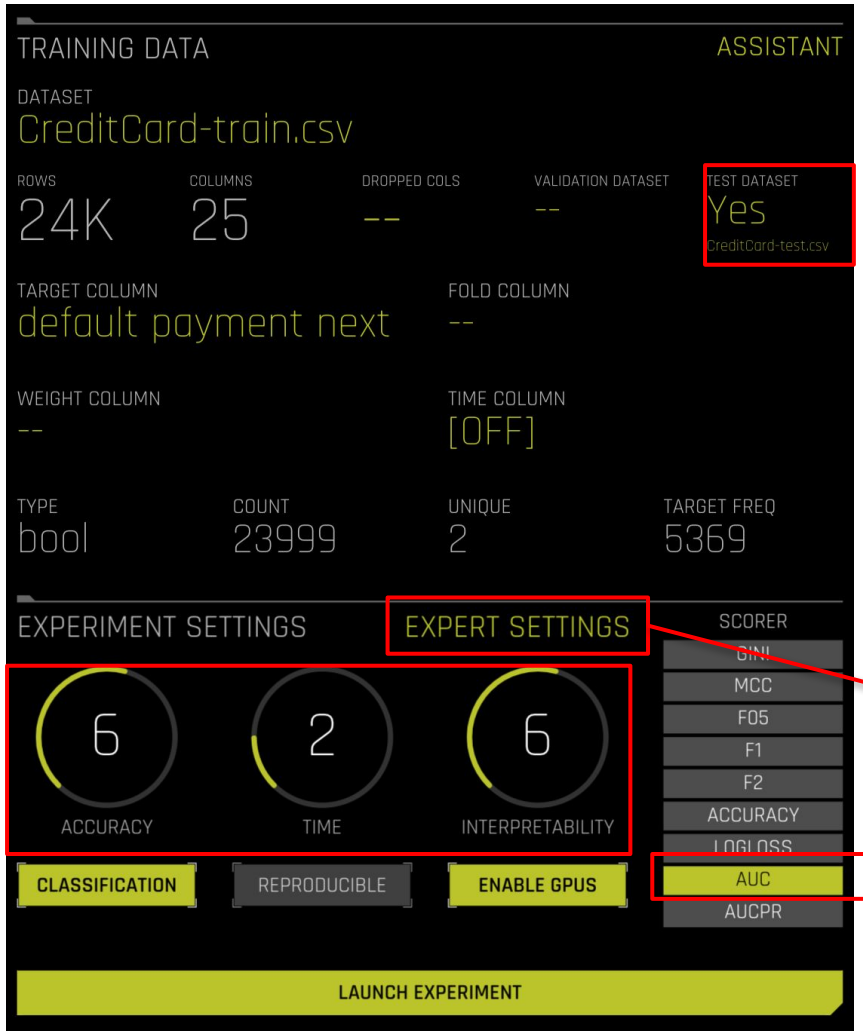




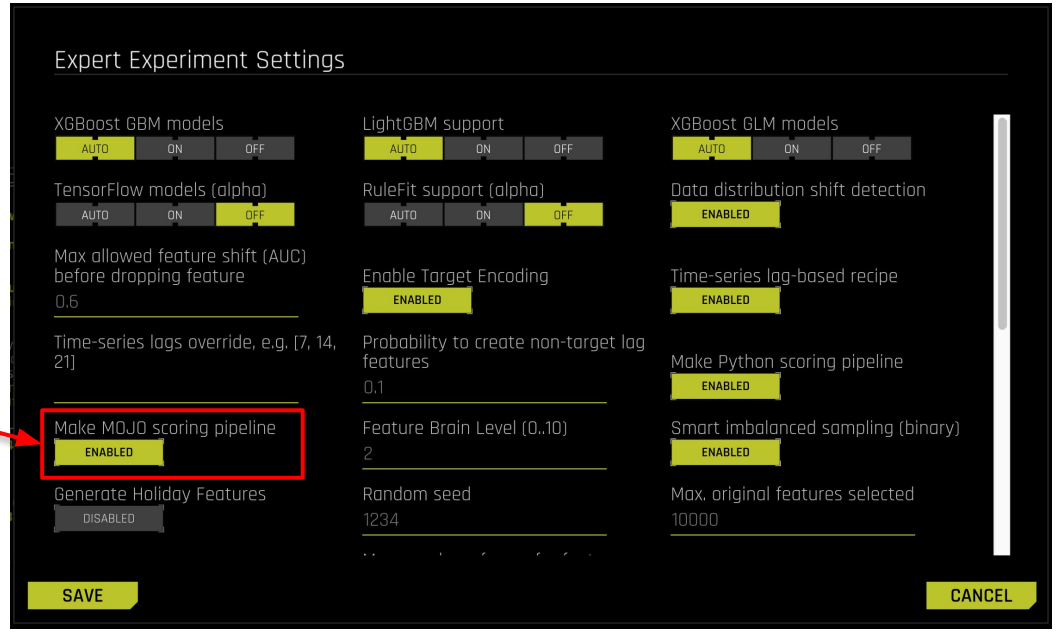


H2O.ai

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



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



H2O.ai Experiment CreditCard

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

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

TRAINING DATA

ASSISTANT

DATASET
CreditCard-train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
24K	25	--	--	Yes
				CreditCard-test.csv

TARGET COLUMN
default payment next

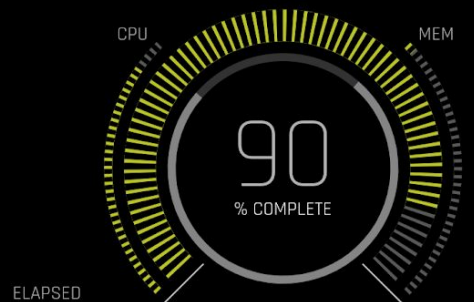
FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE	COUNT	UNIQUE	TARGET FREQ
bool	23999	2	5369

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



ELAPSED
00:04:09

FINISH

ITERATION
10/14

EXPERIMENT SETTINGS

EXPERT SETTINGS

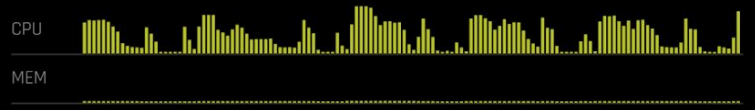
ACCURACY	TIME	INTERPRETABILITY
6	2	6

CLASSIFICATION | REPRODUCIBLE | ENABLE GPUS

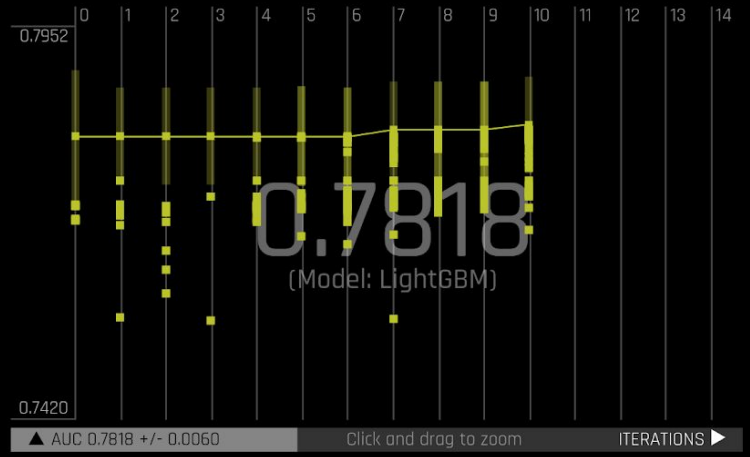
SCORER
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR

CPU / MEMORY

Notifications Log Trace



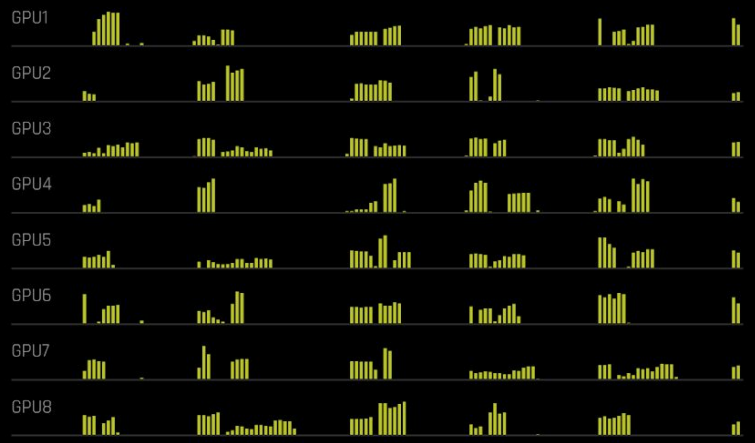
ITERATION DATA - VALIDATION



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

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)

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

TRAINING DATA

ASSISTANT

DATASET
CreditCard-train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
24K	25	--	--	Yes
				CreditCard-test.csv

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE	COUNT	UNIQUE	TARGET FREQ
bool	23999	2	5369

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

EXPERT SETTINGS

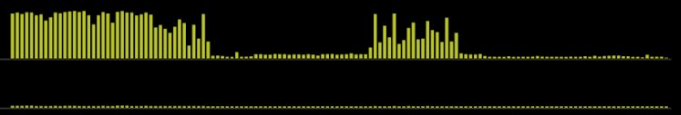
ACCURACY	TIME	INTERPRETABILITY
6	2	6

CLASSIFICATION | REPRODUCIBLE | ENABLE GPUS

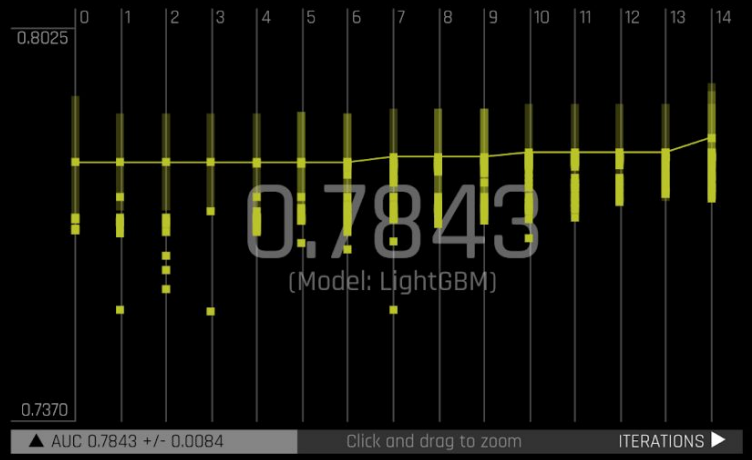
SCORER: GINI, MCC, F05, F1, F2, ACCURACY, LOGLOSS, **AUC**, AUCPR

CPU / MEMORY

Notifications Log Trace



ITERATION DATA - VALIDATION



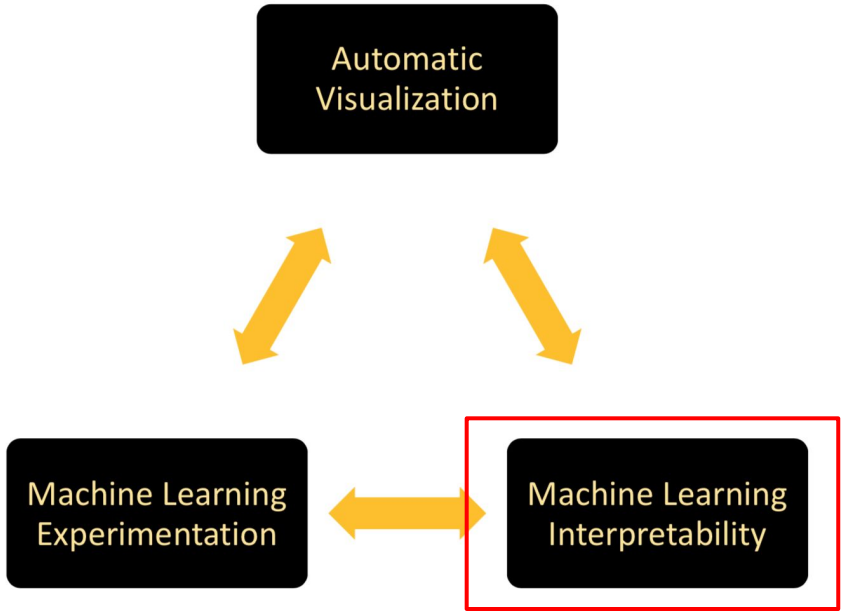
VARIABLE IMPORTANCE

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

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)

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

TRAINING DATA

ASSISTANT

DATASET
CreditCard-train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
24K	25	--	--	Yes
				CreditCard-test.csv

TARGET COLUMN
default payment next

FOLD COLUMN
--

WEIGHT COLUMN
--

TIME COLUMN
[OFF]

TYPE	COUNT	UNIQUE	TARGET FREQ
bool	23999	2	5369

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

EXPERT SETTINGS

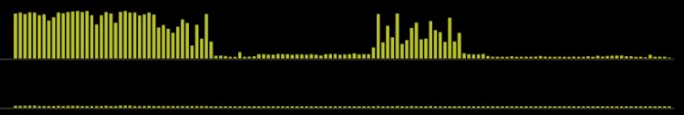
ACCURACY	TIME	INTERPRETABILITY
6	2	6

CLASSIFICATION | REPRODUCIBLE | ENABLE GPUS

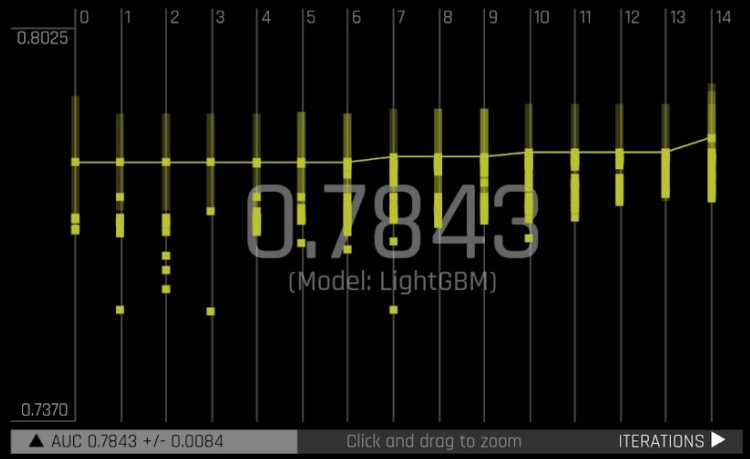
SCORER: GINI, MCC, F05, F1, F2, ACCURACY, LOGLOSS, **AUC**, AUCPR

CPU / MEMORY

Notifications Log Trace



ITERATION DATA - VALIDATION



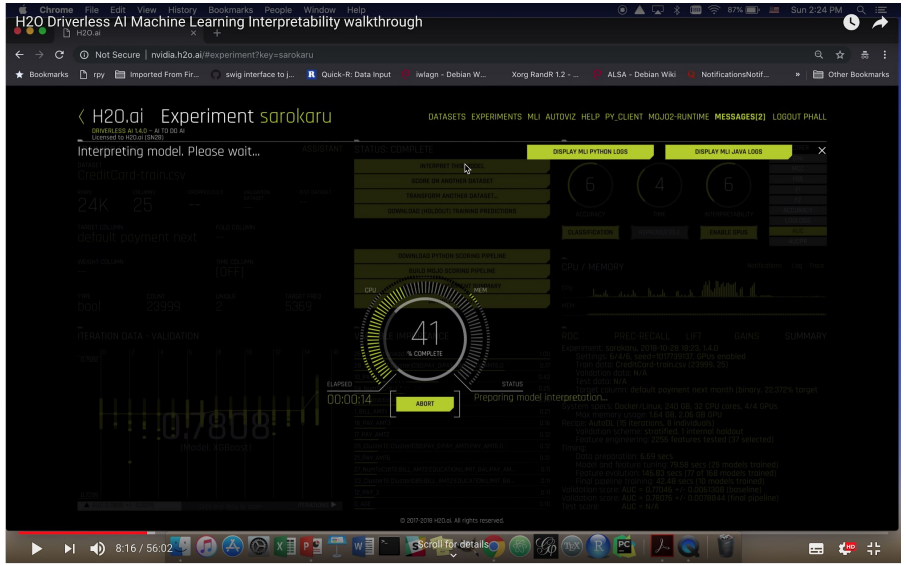
VARIABLE IMPORTANCE

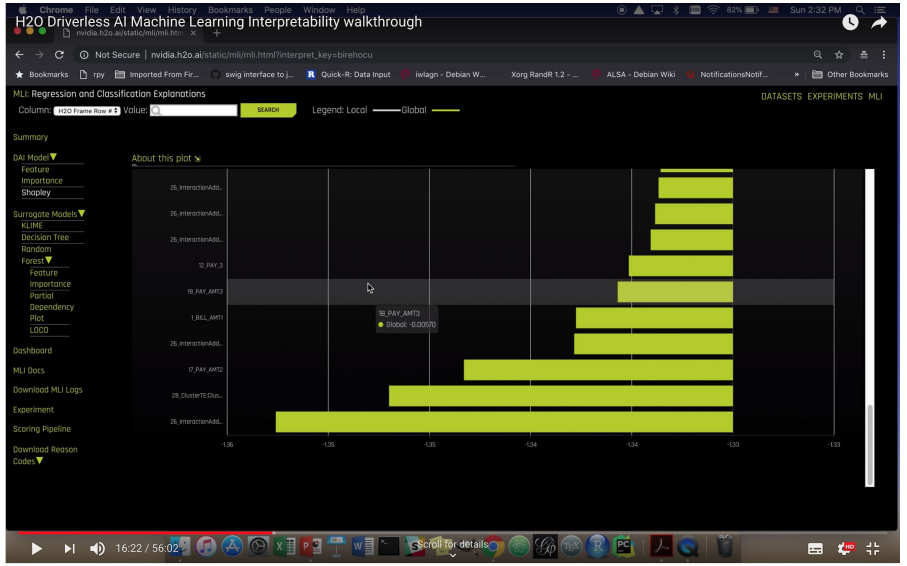
26_InteractionAdd:PAY_0:PAY_2	1.00
30_NumToCatTE:PAY_0:PAY_2.0	0.85
33_ClusterTE:ClusterID81:AGE:BILL_AMT1:EDUCATION:PAY_...	0.36
35_NumToCatWoE:BILL_AMT1:PAY_0.0	0.29
10_PAY_0	0.22
14_PAY_5	0.12
37_NumToCatTE:LIMIT_BAL:PAY_6:PAY_AMT5.0	0.11
24_InteractionSub:PAY_0:PAY_AMT3	0.11
24_InteractionSub:BILL_AMT1:LIMIT_BAL	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

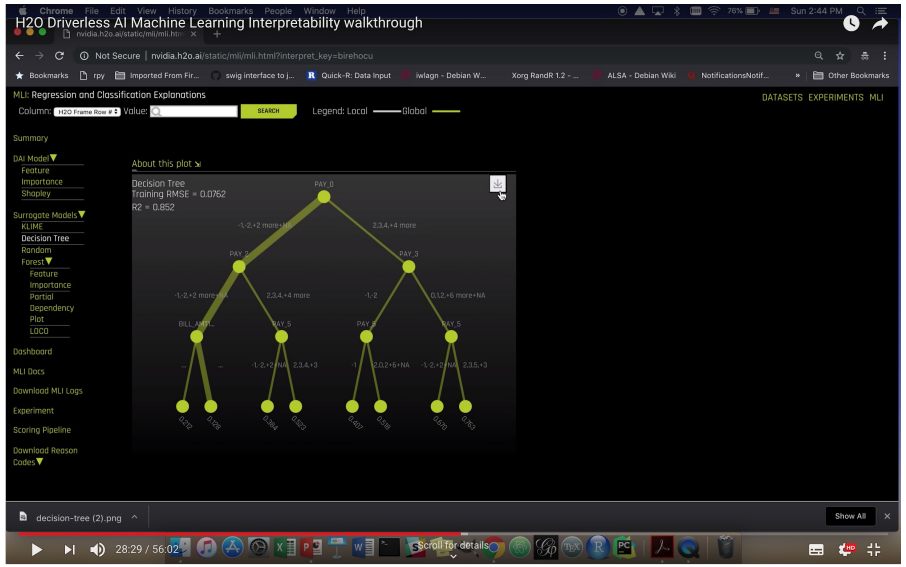
ROC PREC-RECALL LIFT GAINS SUMMARY

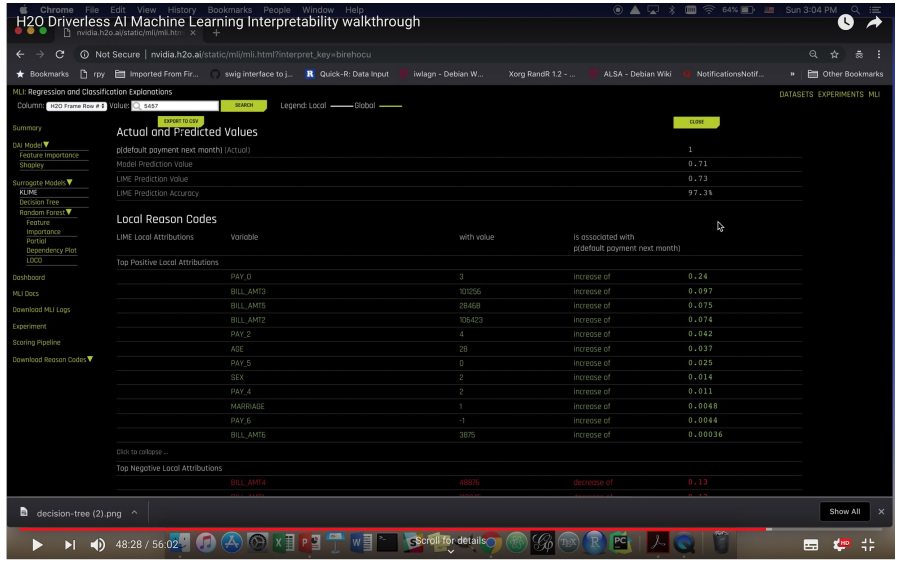
Experiment: vacesiwa, 2018-11-30 12:30, 1.4.1
Settings: 6/2/6, seed=126037824, GPUs enabled
Train data: CreditCard-train.csv (23999, 25)
Validation data: N/A
Test data: CreditCard-test.csv (6000, 24)
Target column: default payment next month (binary, 22.372% target class)
System specs: Docker/Linux, 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)

Tutorial on YouTube









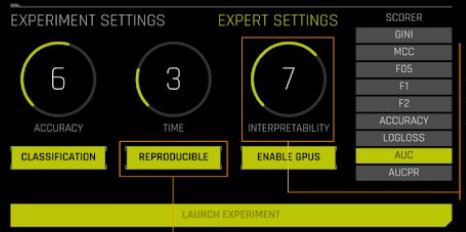
https://www.youtube.com/watch?v=5jSU3CUReXY

How to explain a model with H2O Driverless AI

<and the Kaggle credit card data>

<Confirmatory information between explainability techniques highlighted in yellow.>

0. START HERE: Train a more interpretable model



INTERPRETABILITY >= 7:
/ Feature selection
/ Monotonicity constraints
/ Simpler feature transformations

Ensure reproducibility

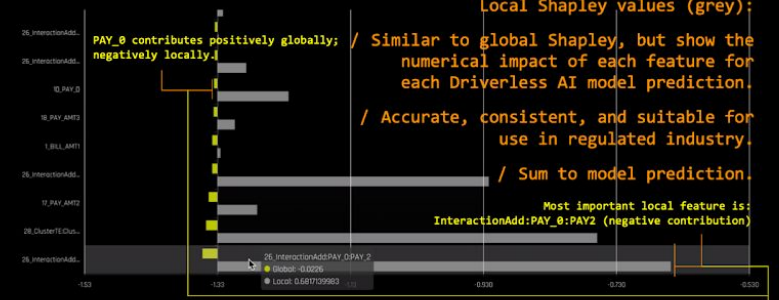
Continue to basic global explanations and step 1.

Advanced local explanations

<Understanding model behavior for an individual row>

6. Local Shapley feature importance

<Accurate local feature importance values for each individual row>



Local Shapley values (grey):

/ Similar to global Shapley, but show the numerical impact of each feature for each Driverless AI model prediction.

/ Accurate, consistent, and suitable for use in regulated industry.

/ Sum to model prediction.

7. Local linear explanations

<Driverless AI model local linear trends around an individual>

LIME-derived coefficients compliment local Shapley values with local trend information.

Cluster 12 Reason Codes

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

Variable	with value/ 1 unit increase (if blank)	is associated with p(default payment next month)	
Top Positive Cluster Attributions			
PAY_0	3	increase of	0.24
PAY_0	2	increase of	0.2
PAY_5	5	increase of	0.097
PAY_6	3	increase of	0.088
PAY_2	5	increase of	0.085
PAY_0	4	increase of	0.085
PAY_3	7	increase of	0.084
PAY_5	3	increase of	0.064
PAY_0	8	increase of	0.051
PAY_4	3	increase of	0.049
PAY_2	2	increase of	0.045

Repayment statuses are correlated with default in this local region.

8. Local surrogate tree decision path

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

Potential interaction between PAY_0, PAY_3, and PAY_5 causing high probability of default.

Decision path (grey):

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

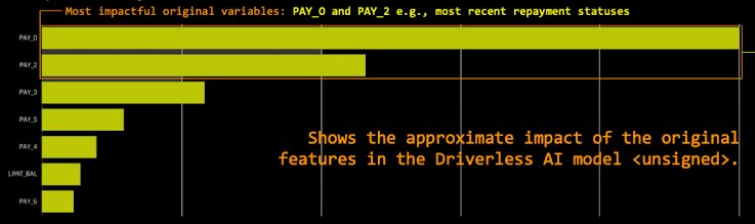
/ Path can show local interactions.

Basic global explanations

<Understanding overall model behavior>

2. Global original feature importance

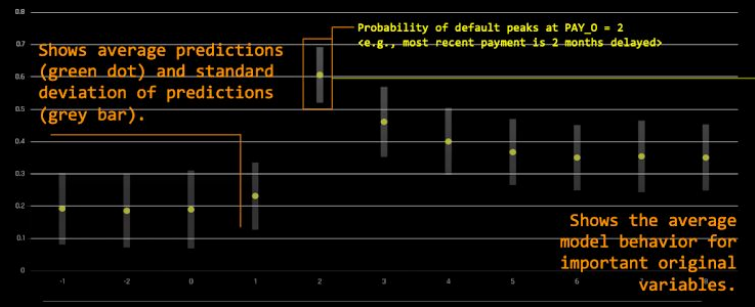
<Important original features that drive model behavior>



Shows the approximate impact of the original features in the Driverless AI model <unsigned>.

3. Partial dependence

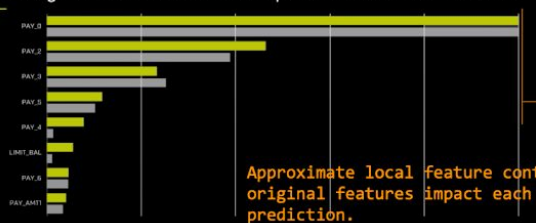
<Average Driverless AI model prediction for different values of the original variables>



Shows the average model behavior for important original variables.

10. Local original feature importance

<Original features that drive a prediction for an individual>

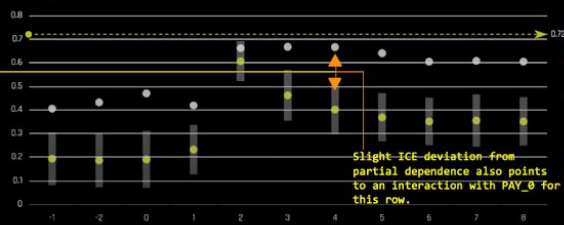


PAY_0, PAY_2, PAY_3, and PAY_5 repayment statuses appear to be largest contributors to high probability of default.

Approximate local feature contribution (grey) shows how original features impact each Driverless AI model prediction.

9. Original feature individual conditional expectation (ICE)

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



ICE (grey):

/ Shows how changing a feature's value changes a row's prediction.

/ Divergence from partial dependence can show local interactions.

1. Global Shapley feature importance

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

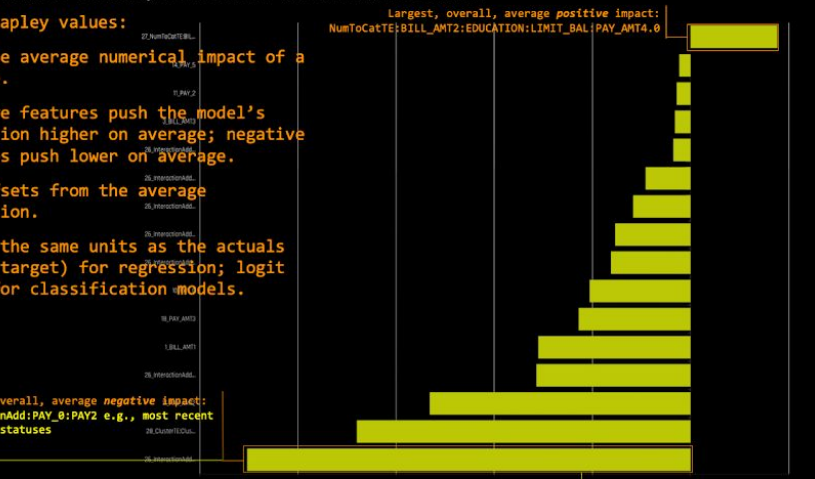
Global Shapley values:

/ 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>

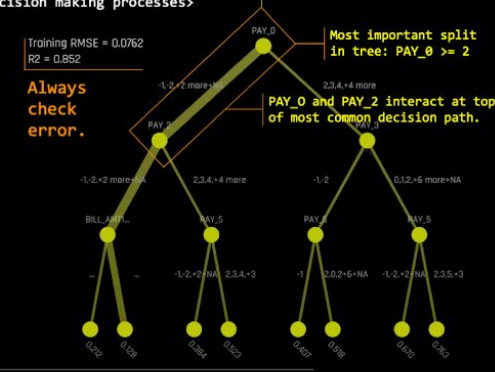
Surrogate decision tree models driverless AI predictions:

/ 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.

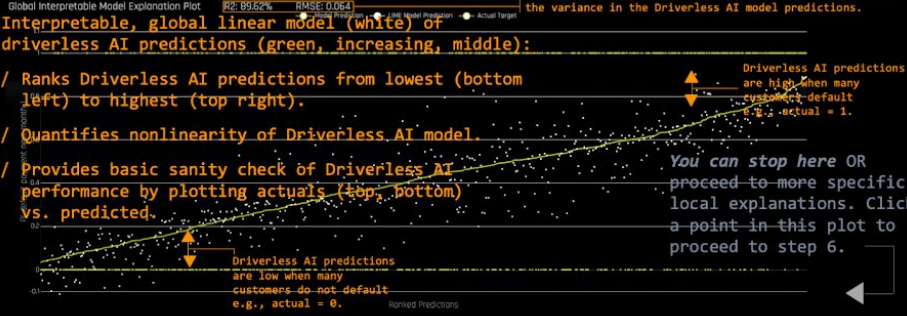
/ Thickest edges are most common decision path through tree.



Always check error.

5. Global interpretable model

<Linear model of Driverless AI model predictions>



High R2 and low RMSE show that this Driverless AI model is nearly linear e.g., a linear model explains nearly 90% of the variance in the Driverless AI model predictions.

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

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

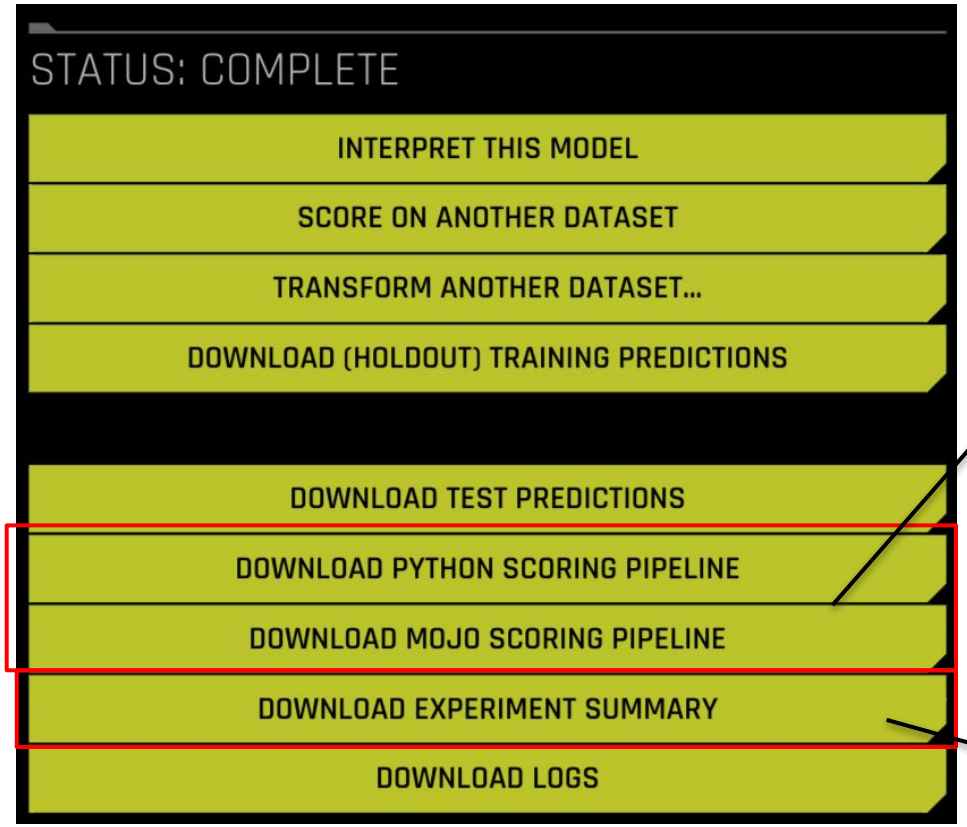
/ Quantifies nonlinearity of Driverless AI model.

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

You can stop here OR proceed to more specific local explanations. Click a point in this plot to proceed to step 6.

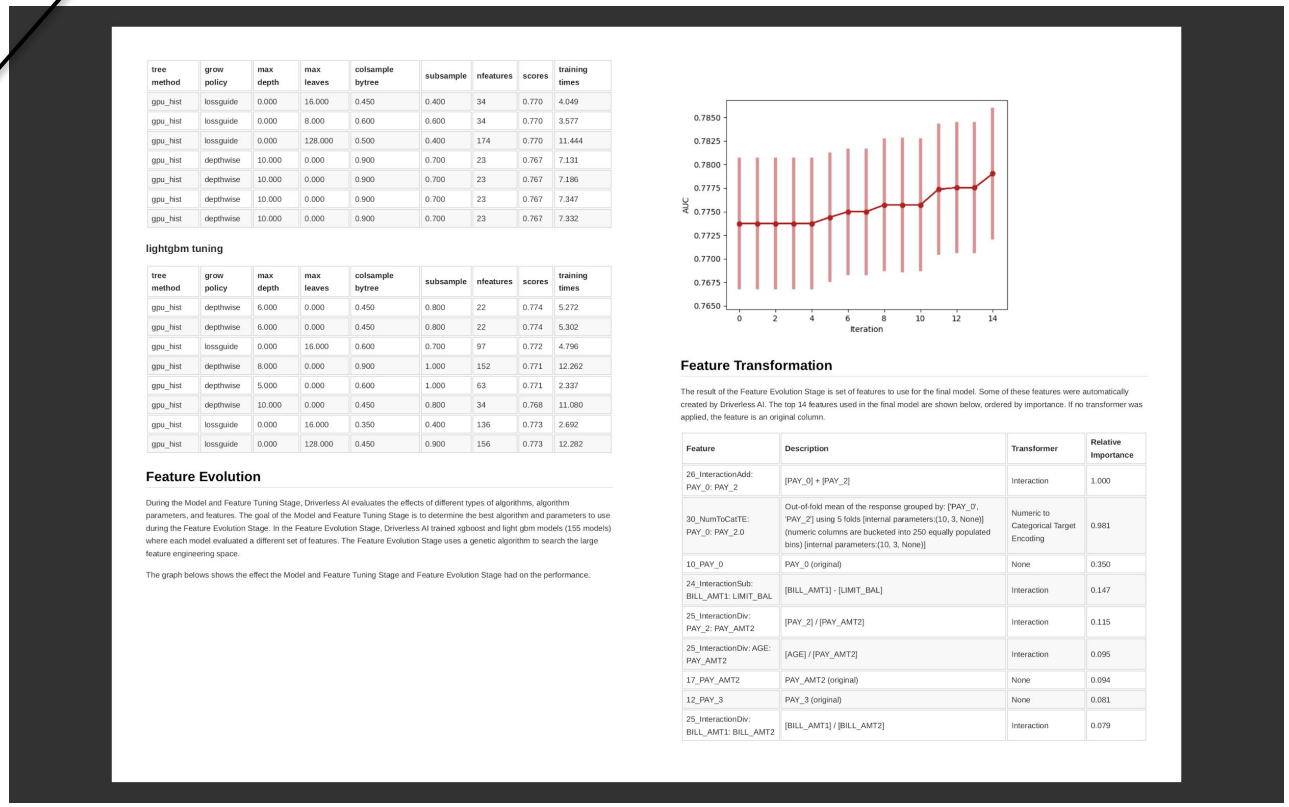
Hands-on Experiment:
Other Features

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

lightgbm tuning

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 belows shows the effect the Model and Feature Tuning Stage and Feature Evolution Stage had on the performance.

Feature Transformation

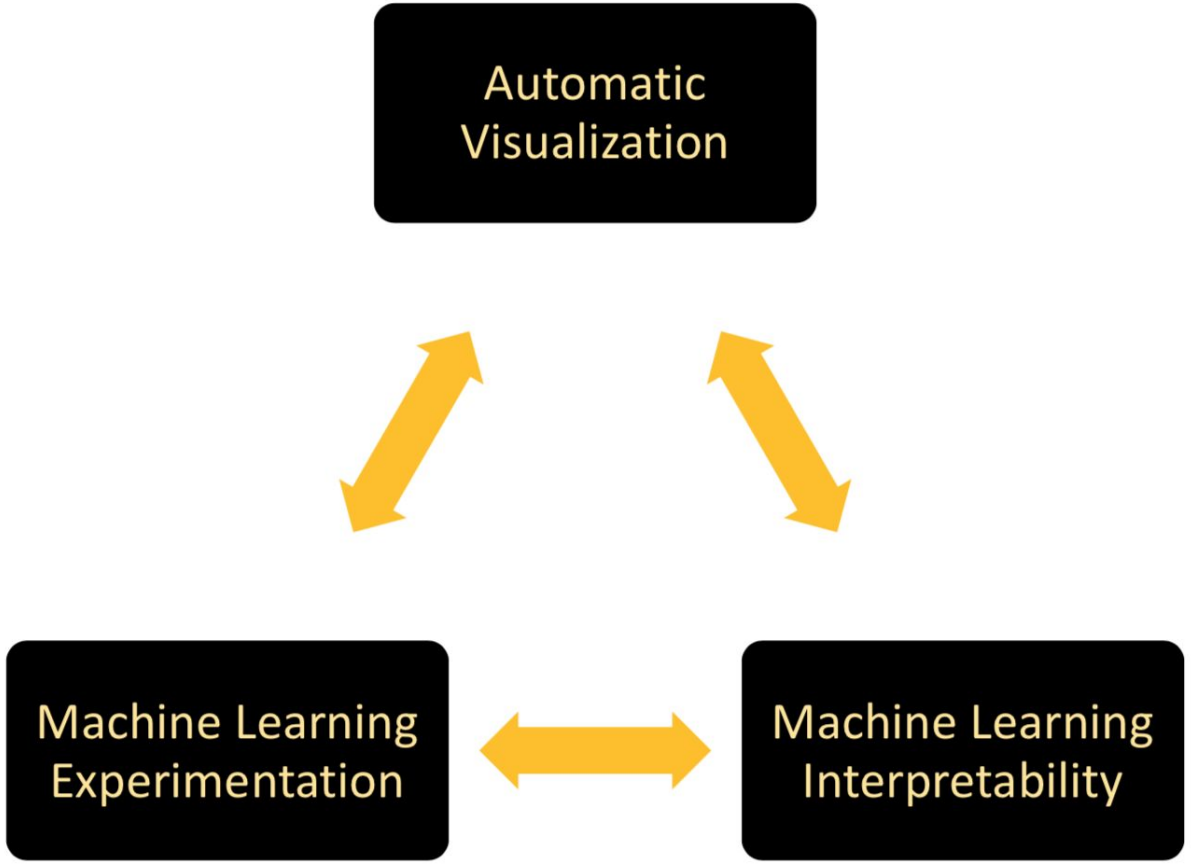
The result of the Feature Evolution Stage is set of features to use for the final model. Some of these features were automatically created by Driverless AI. The top 14 features used in the final model are shown below, ordered by importance. If no transformer was applied, the feature is an original column.

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: AGE: 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

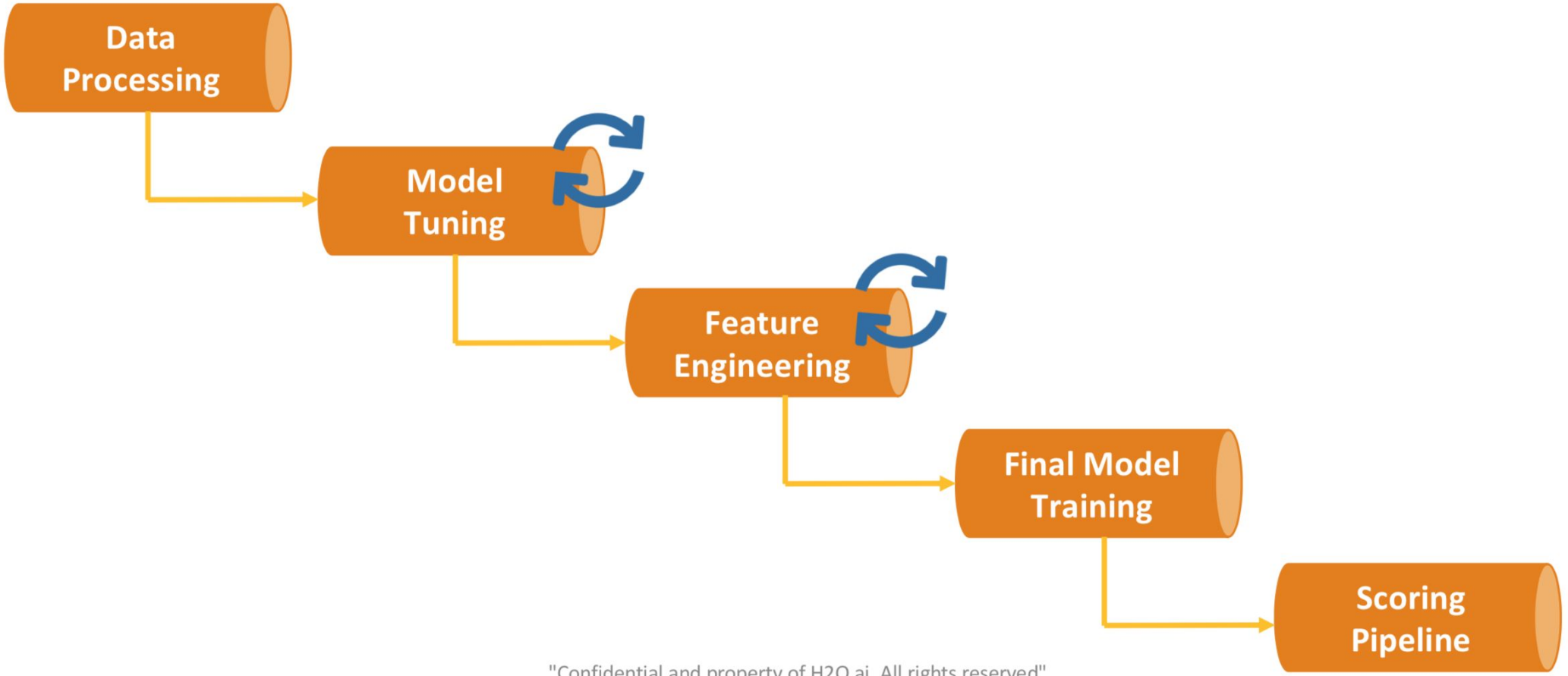
Other Features

Hands-on Experiment:
Quick Recap

Driverless AI Components



Driverless AI: Machine Learning Workflow



H2O.ai

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

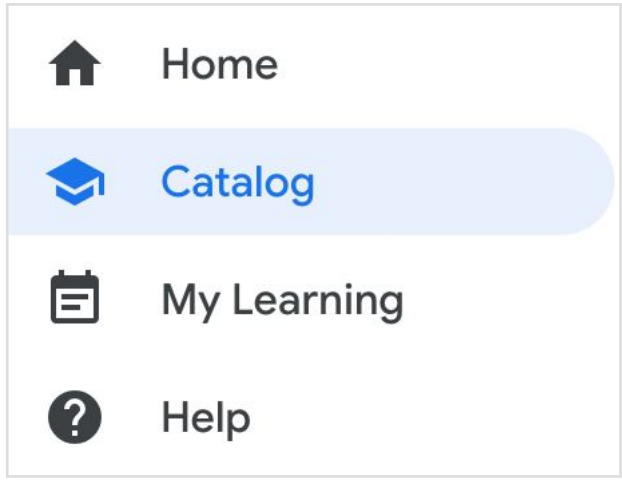
- Join the H2O.ai Community Slack Workspace today!
  - https://tinyurl.com/h2o-community-slack
- Use **emoji** to tag messages
  - :question :use_case :mli :get_started :bugs …
- Reply to message using **threads**
- Check out Community Guide for more info:
  - https://tinyurl.com/hac-community-guide

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. 

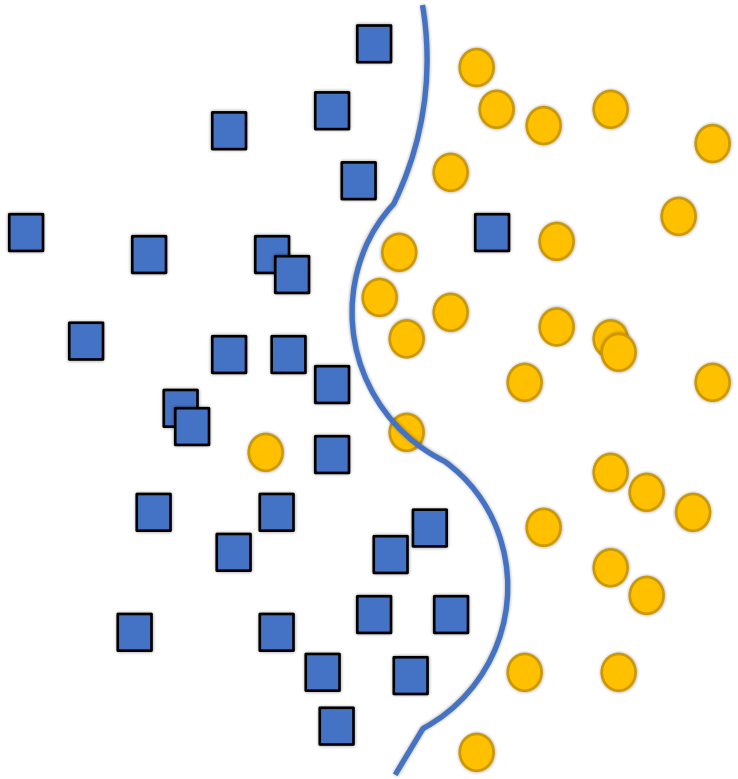
and wait for instructions



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

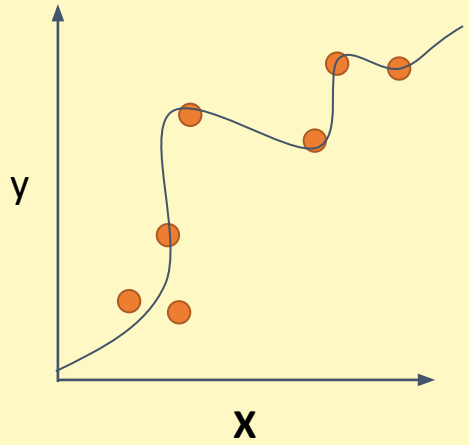
What is Machine Learning?



H2O.ai

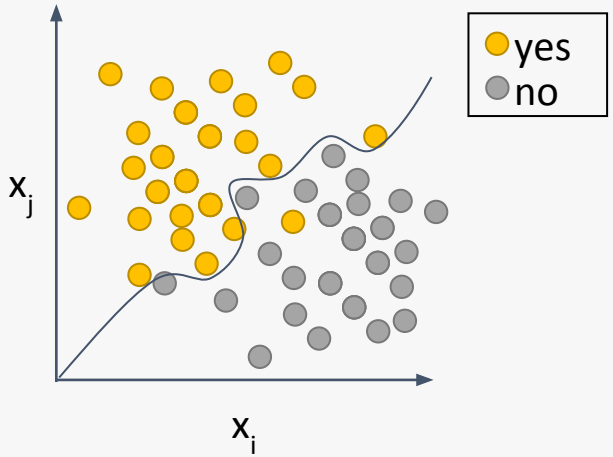
Supervised Learning

Regression:
How much will a customer spend?



H_2O algos:
Penalized Linear Models
Random Forest
Gradient Boosting
XGBoost
Neural Networks
Stacked Ensembles

Classification:
Will a customer churn?



H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
XGBoost
Neural Networks
Stacked Ensembles

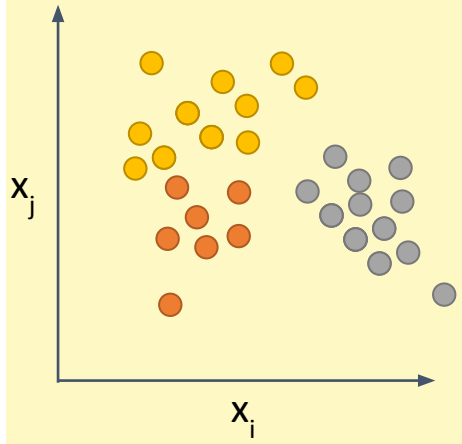
H_2O.ai

H_2O.ai

Unsupervised Learning

Clustering:

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

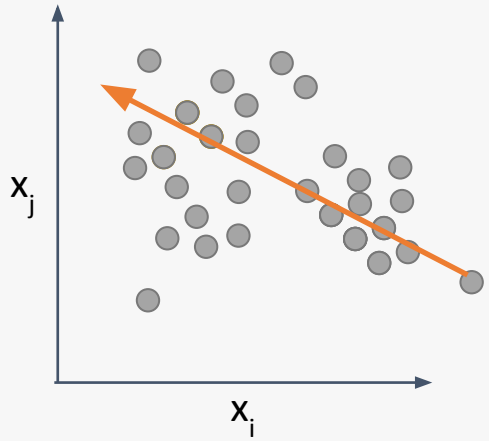


H_2O algos:

k – means

Feature extraction:

Grouping columns – Create a small number of new representative dimensions



H_2O algos:

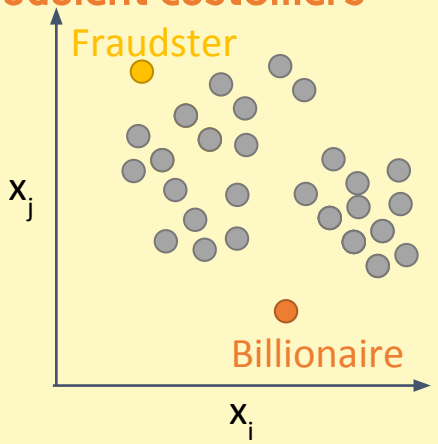
Principal components

Generalized low rank models

Autoencoders, Word2Vec

Anomaly detection:

Detecting outlying rows - Finding high-value or fraudulent customers



H_2O algos:

Isolation Forest

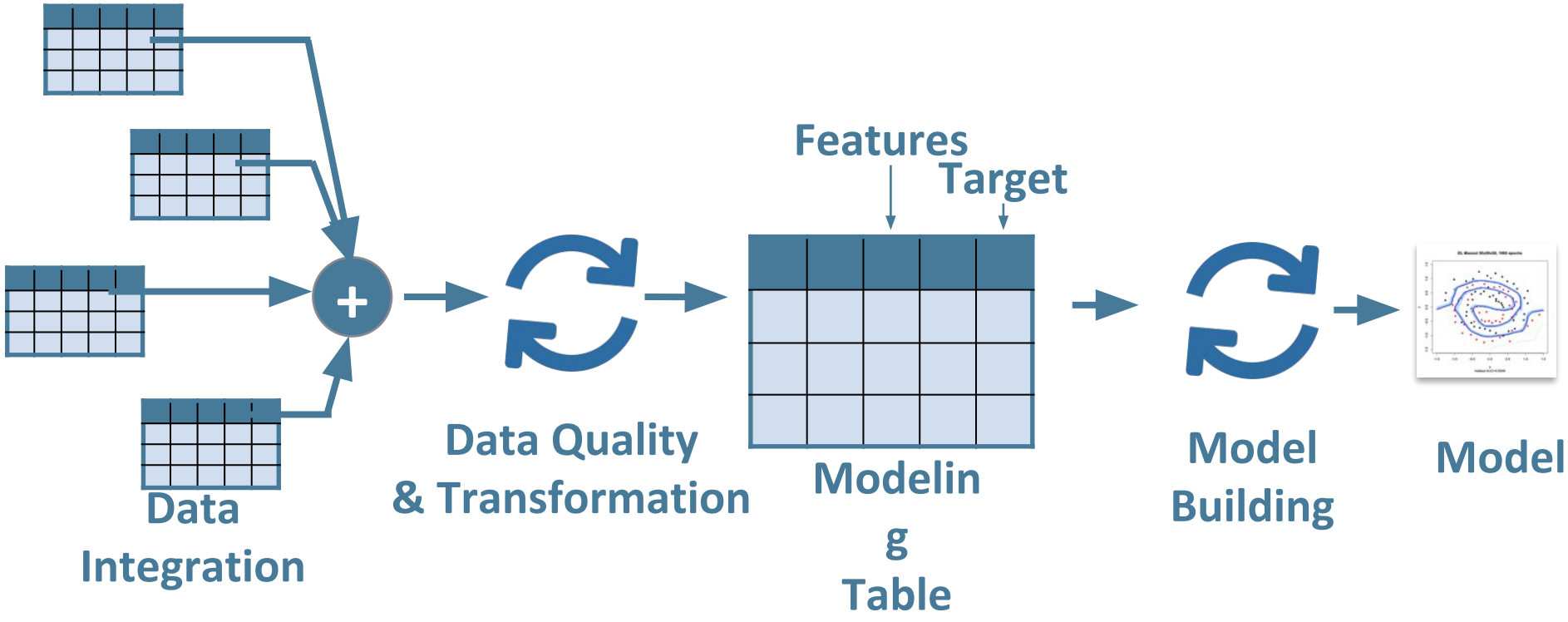
Generalized low rank models

Autoencoders

H_2O.ai

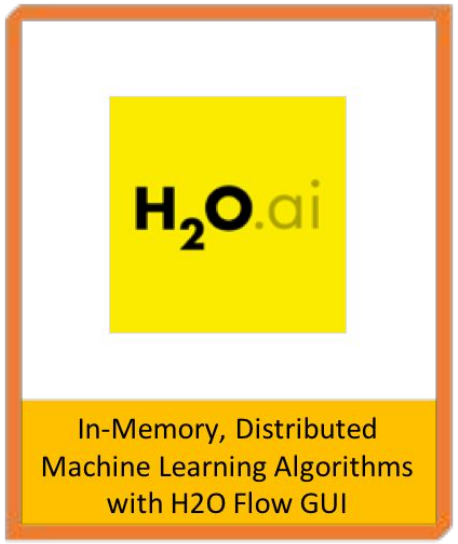
H_2O.ai

Simplified Typical Machine Learning Pipeline



Introduction to H2O

H2O Products



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

H2O AI Open Source Engine Integration with Spark

Lightning Fast machine learning on GPUs

Automatic feature engineering, machine learning and interpretability

Secure multi-tenant H2O clusters

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

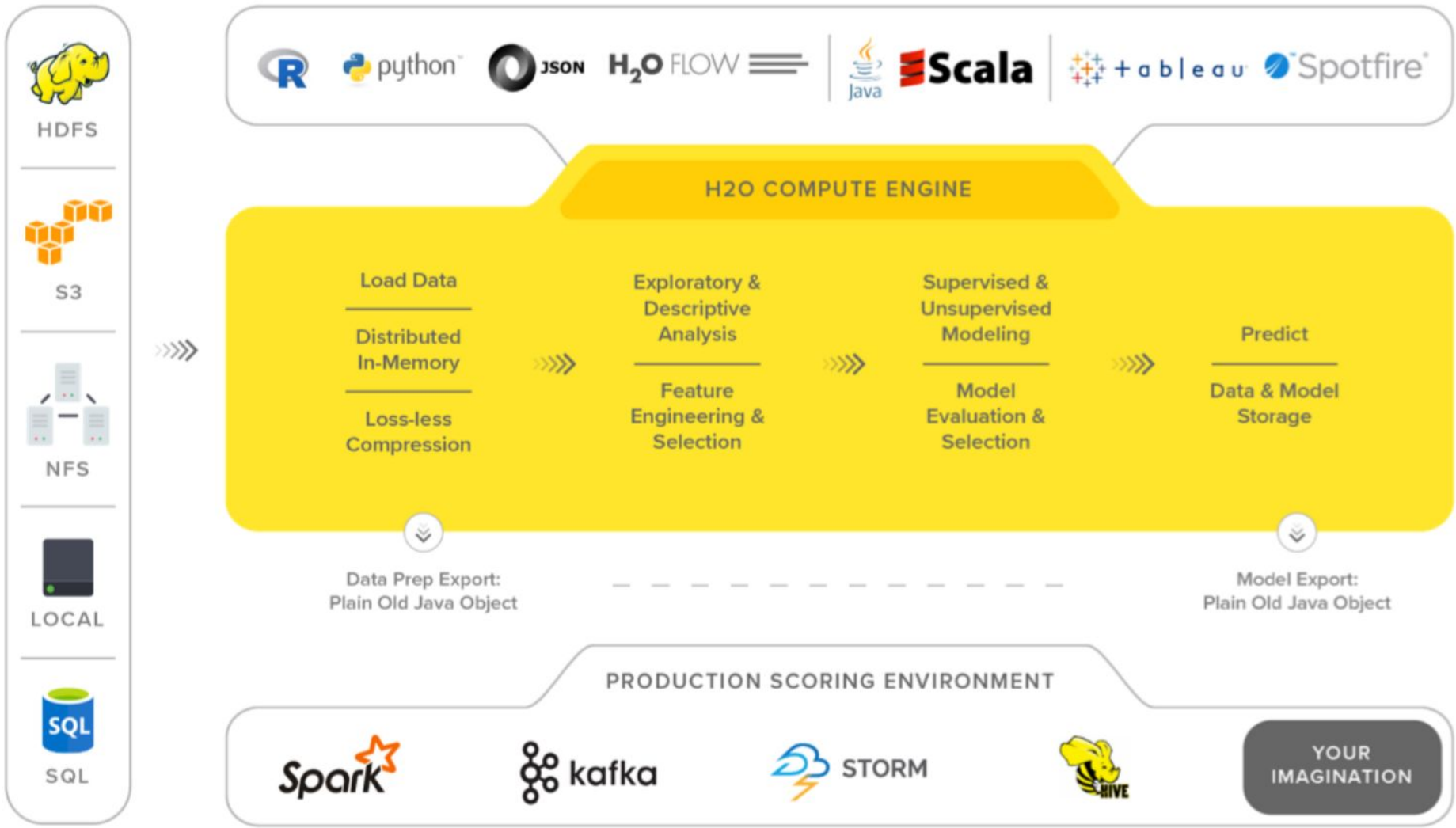
- Written in Java – perfect for Java Programmers
- Install is lightweight
- REST API (Java) – run H2O from R, Python, WebUI

Big Data

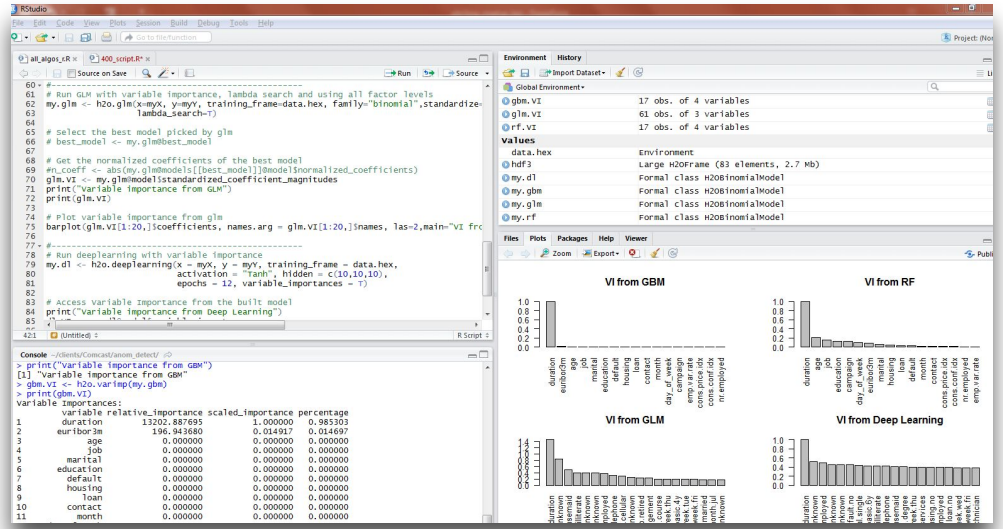
More data? Or better models? BOTH

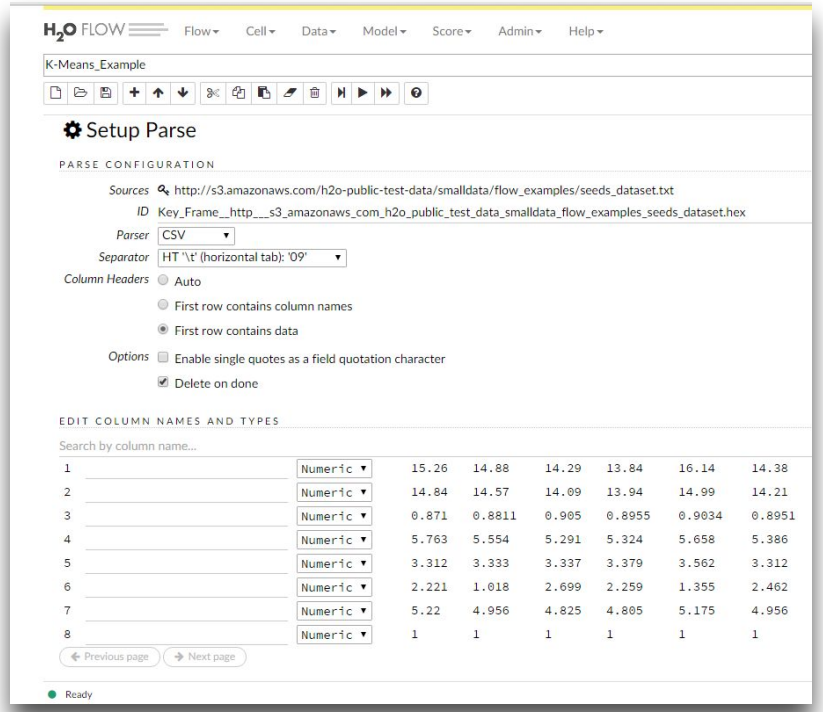
- Use all of your data – model without sampling
- More Data + Better Models = Better Predictions

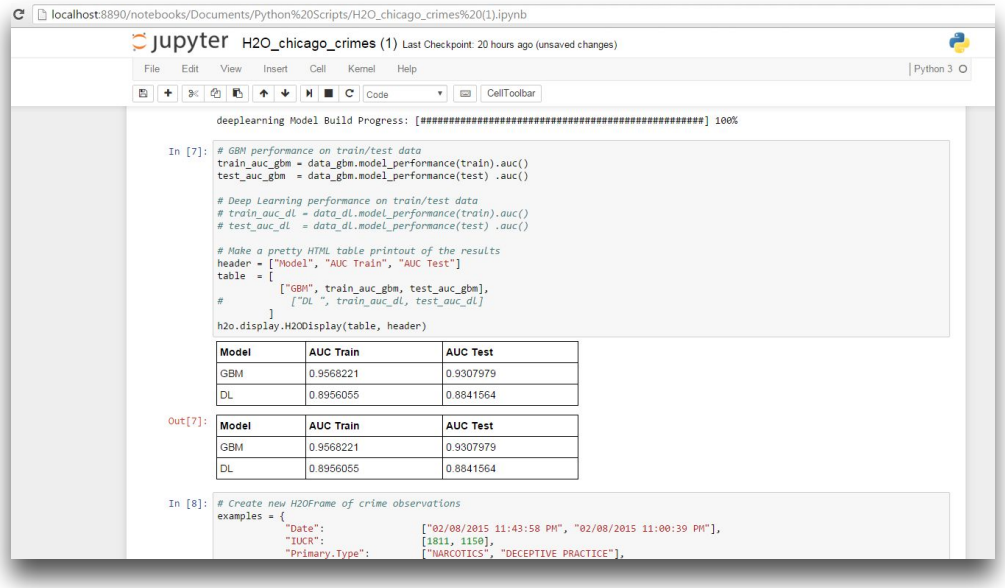
H2O.ai


HDFS
S3
NFS
LOCAL
SQL
SQL
R
python
JSON
H2O FLOW
Java
Scala
+ableau
Spotfire
H2O COMPUTE ENGINE
Load Data
Distributed In-Memory
Loss-less Compression
Exploratory & Descriptive Analysis
Feature Engineering & Selection
Supervised & Unsupervised Modeling
Model Evaluation & Selection
Predict
Data & Model Storage
Data Prep Export: Plain Old Java Object
Model Export: Plain Old Java Object
PRODUCTION SCORING ENVIRONMENT
Spark
kafka
STORM
HIVE
YOUR IMAGINATION

R, Python, and Flow







H2O.ai

Reading Data with Python

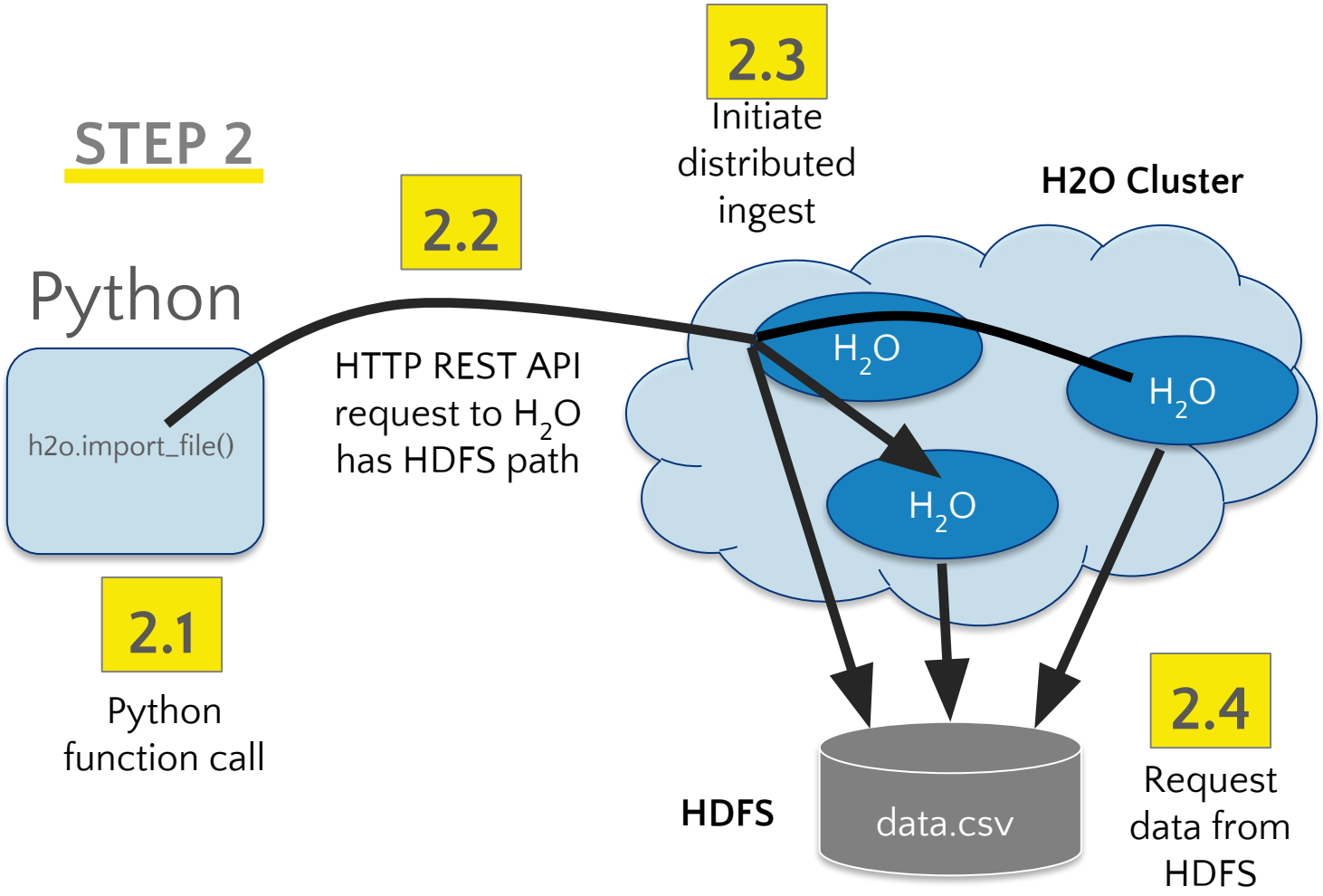
STEP 1

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

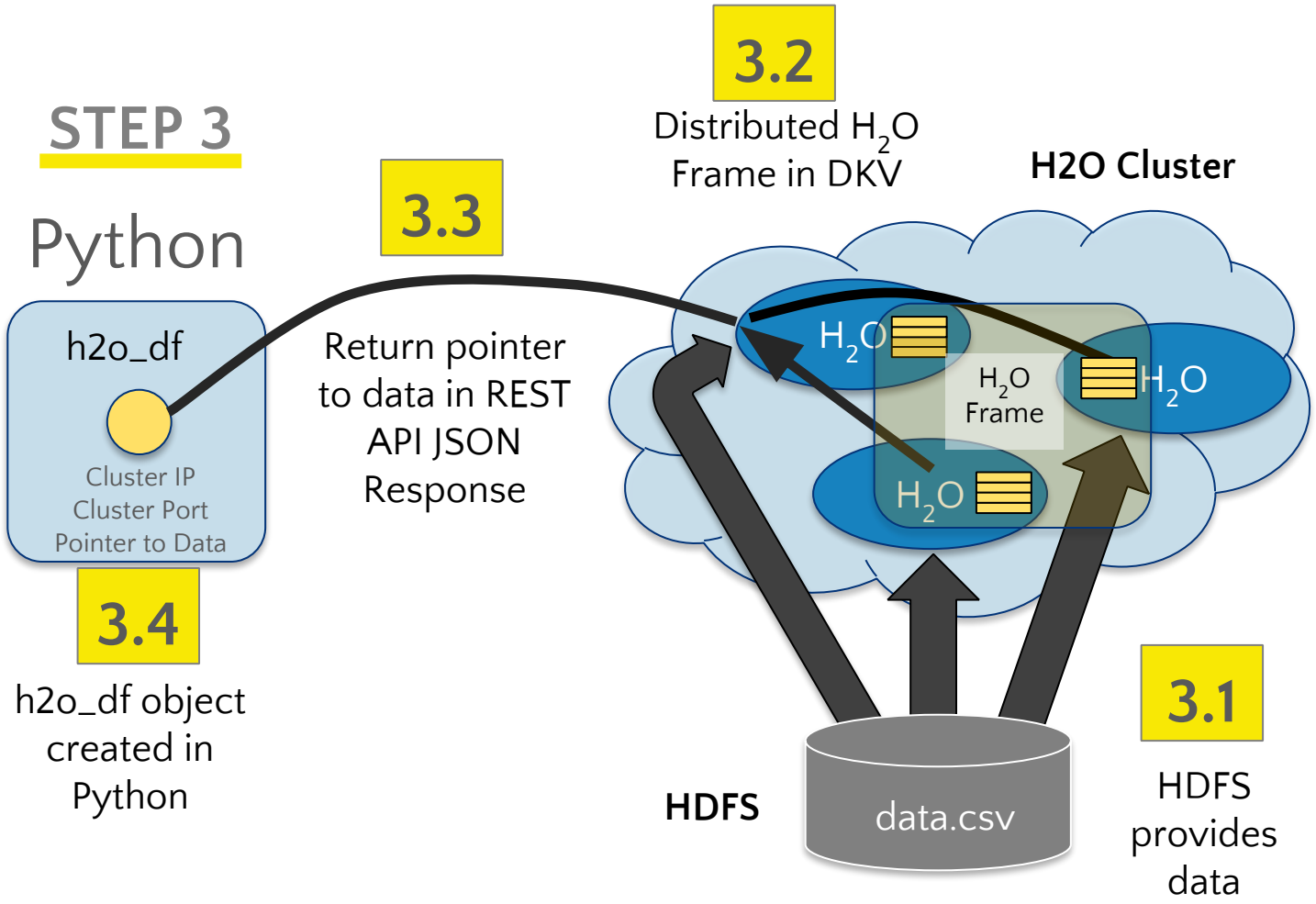
Python user

H2O.ai

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

H2O.ai

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:

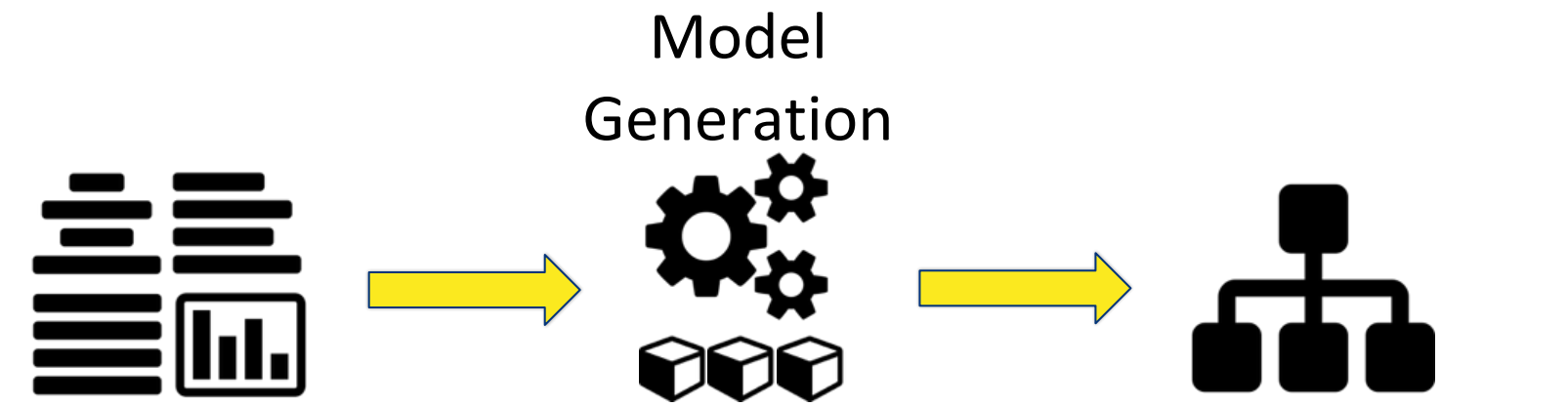
Type	Features
Demographic Information	annual_inc, home_ownership, emp_length
Loan Information	purpose, term, desc, int_rate
Response Column	**bad_loan**

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

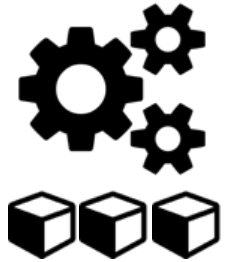
H2O Auto ML Pipeline



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

H2O.ai

Productionalizing Models

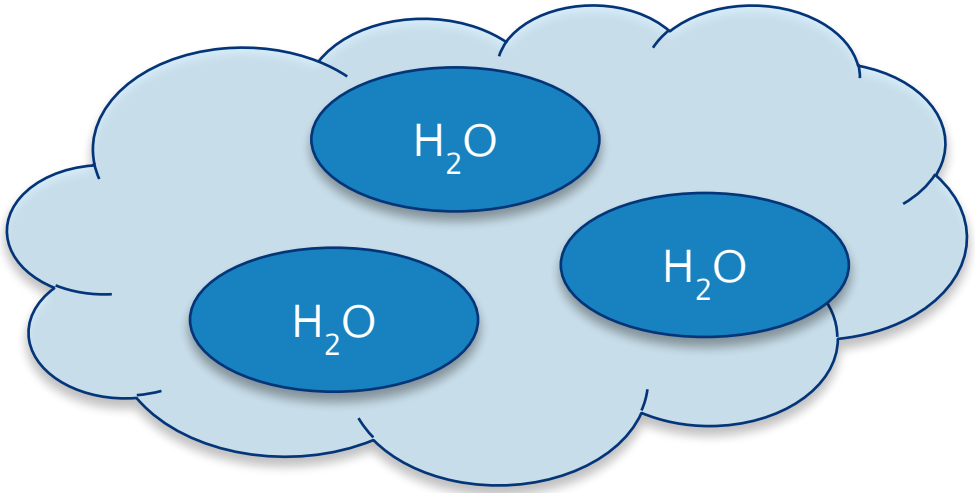
Predicting on New Data / Productionalization

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**