QCon London

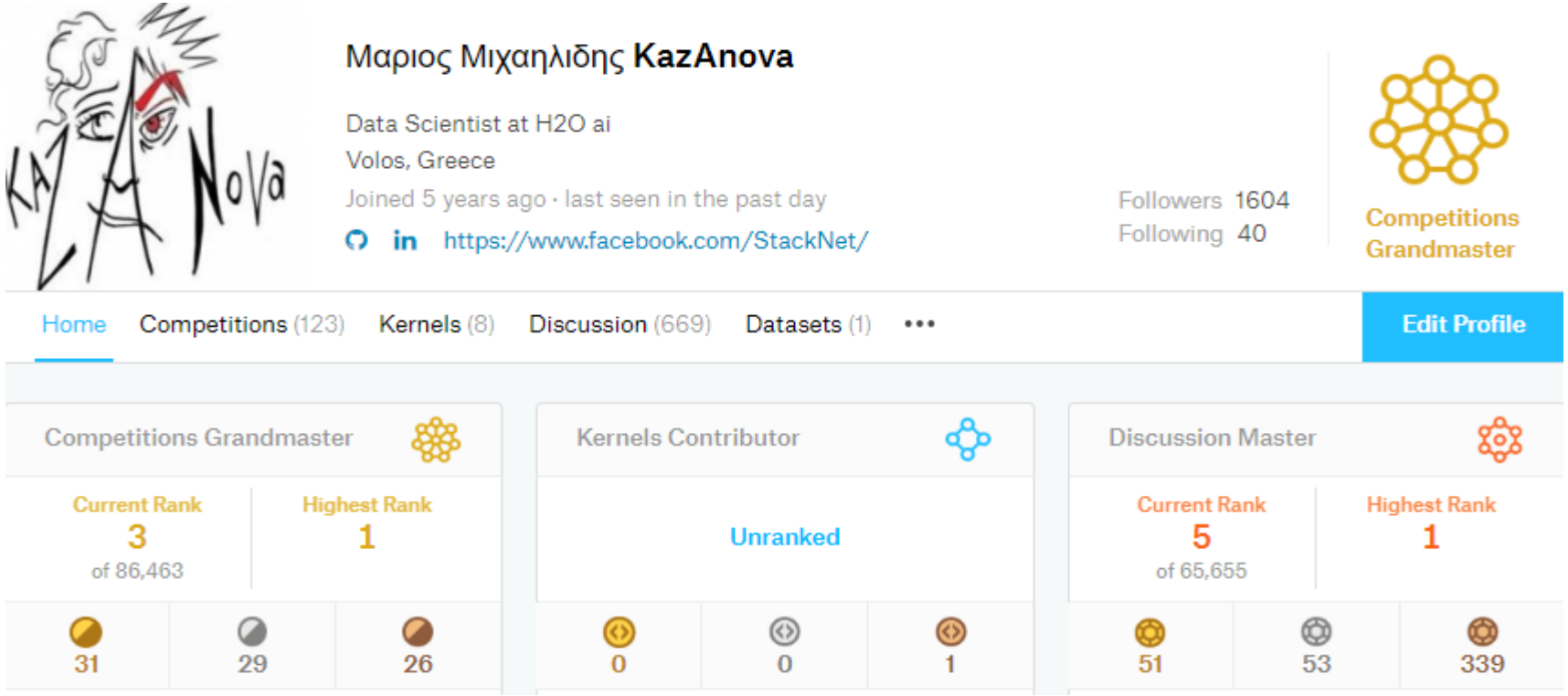
H2O Driverless AI

An AI that creates AI!

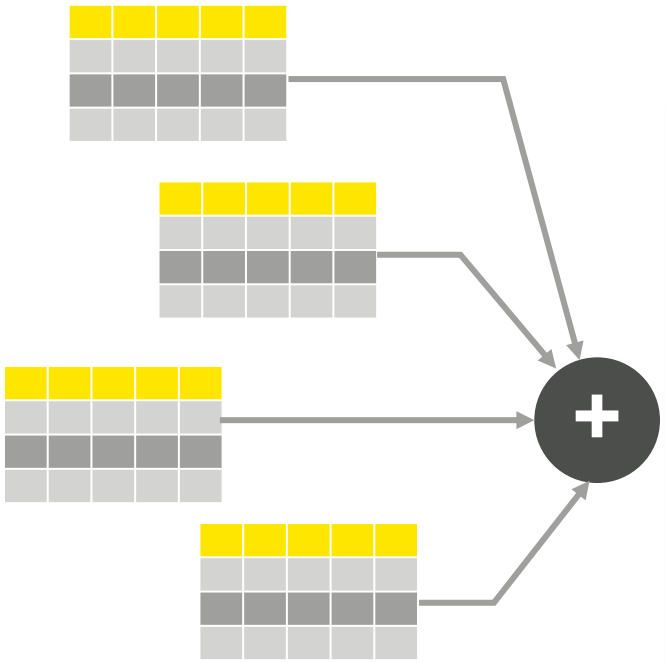
Marios Michailidis

Background

- Competitive data scientist at **H2O**.ai
- PhD in ensemble methods at UCL
- Former kaggle #1 – over 120+ competitions



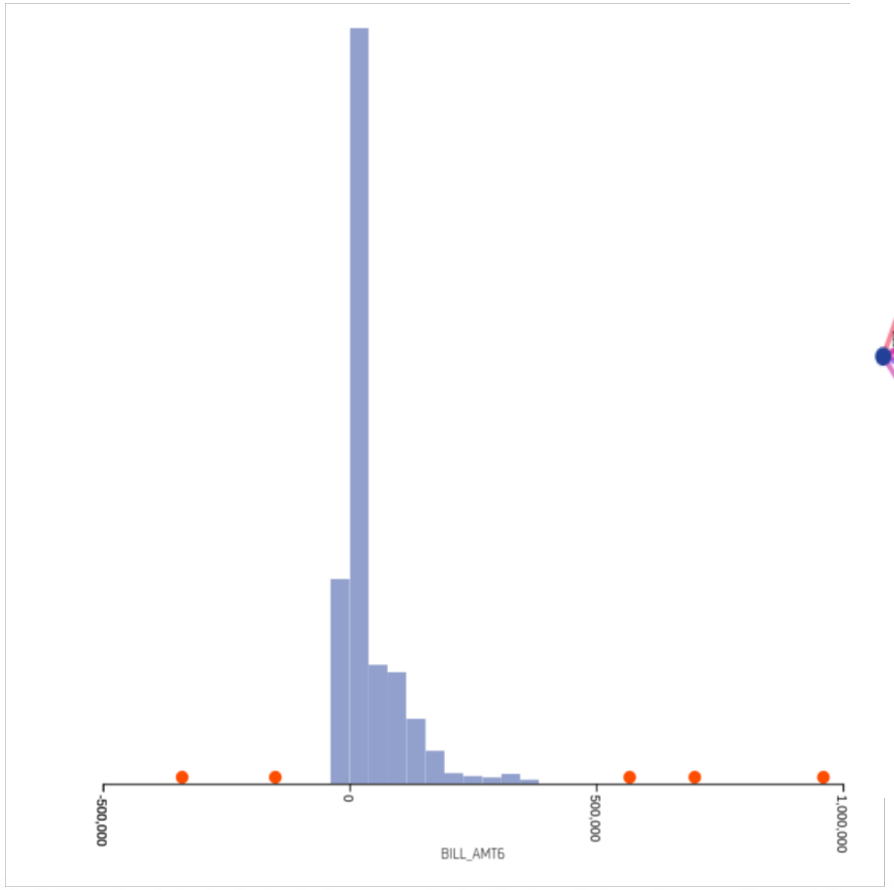
Challenges in the Machine Learning Workflow



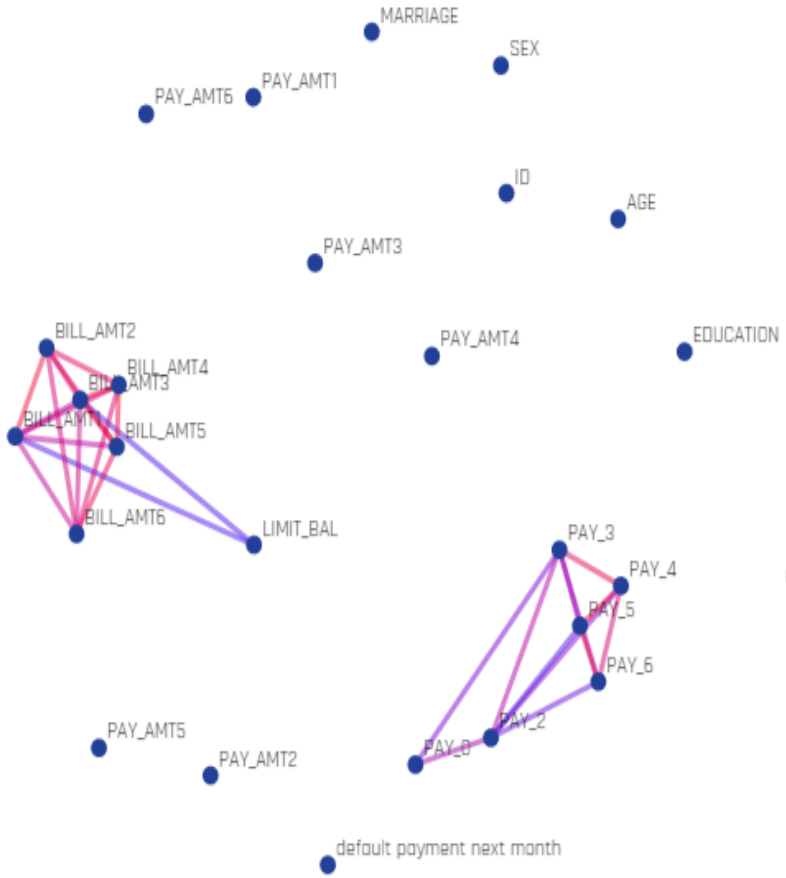
Data Integration

- Insight – Visualization
- Cross Validation
- Feature Engineering
- Model Selection
- Hyper Parameter Optimization
- Feature Selection
- Ensemble
- Understanding/Interpreting the results
- Deploy/Productionize

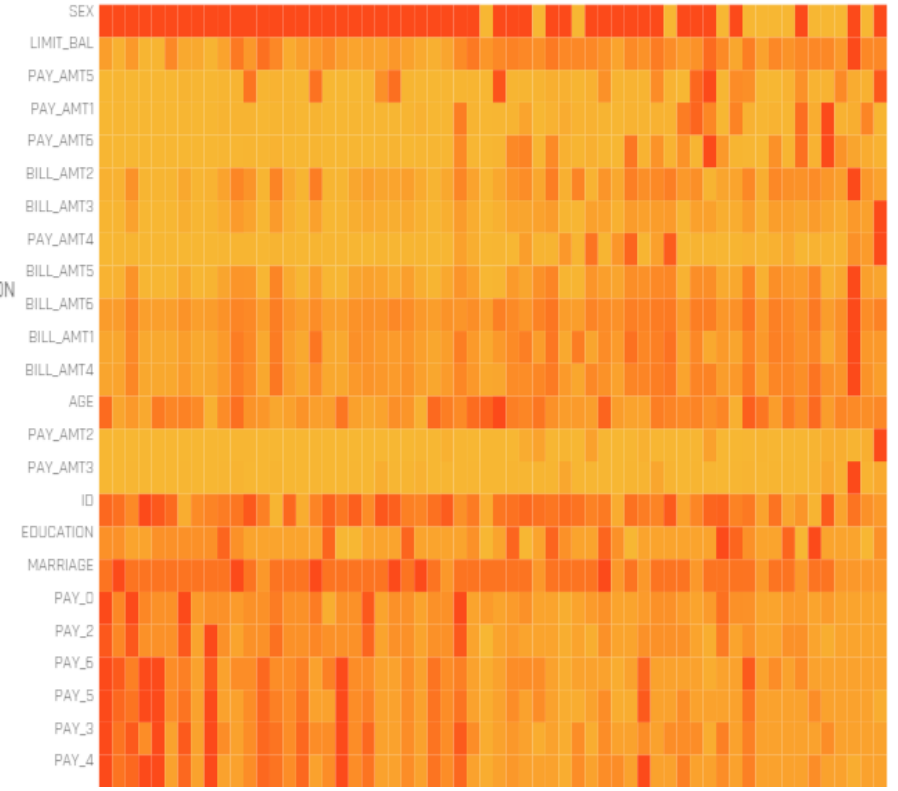
Visualizations



outlier detection



Correlation graph



Heat map

Feature engineering: Categorical features

L.Encoding
2
2
1
3
1
2
3
1
1
3

Frequency
3
3
4
3
4
3
3
4
4
3

Animal	Cost
Dog	108.33
Dog	108.33
Cat	293.75
Fish	38.33
Cat	293.75
Dog	108.33
Fish	38.33
Cat	293.75
Cat	293.75
Fish	38.33

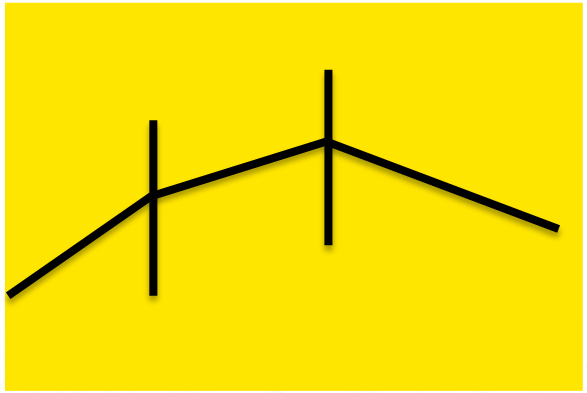
Is_Dog?	Is_Cat?	Is_Fish?
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	1

Animal	Avg.cost
Cat	293.75
Dog	108.333
Fish	38.3333

Feature engineering: Numerical features

Age
40
45
45
45
49
51
56
61
62
71
73
74
77
78
78

Age
40-51
52-56
57-77
78+



Age
40-51
40-51
40-51
40-51
40-51
40-51
52-56
57-77
57-77
57-77
57-77
57-77
78+
78+

Age
40
45
45
45
missing
51
56
61
62
71
73
74
77
78
78

Mean= 61.14

Age
40-51
52-56
57-77
78+
missing

Age
40
45
45
45
61.14
51
56
61
62
71
73
74
77
78
78

Age
40-51
40-51
40-51
40-51
missing
40-51
52-56
57-77
57-77
57-77
57-77
57-77
78+
78+

Feature engineering: Interactions

Age	income
40	40
45	14
45	100
45	33
49	20
51	44
56	65
61	80
62	55
71	15
73	18
74	33
77	32
78	48
78	41

*	+
1600	80
630	59
4500	145
1485	78
980	69
2244	95
3640	121
4880	141
3410	117
1065	86
1314	91
2442	107
2464	109
3744	126
3198	119

Animal	color
Dog	brown
Dog	white
Cat	black
Fish	gold
Cat	mixed
Dog	black
Fish	white
Cat	white
Cat	black
Fish	brown

Animal_color
Dog_brown
Dog_white
Cat_black
Fish_gold
Cat_mixed
Dog_black
Fish_white
Cat_white
Cat_black
Fish_brown

Age	Animal
13	Dog
12	Dog
15	Cat
3	Fish
16	Cat
10	Dog
6	Fish
20	Cat
13	Cat
2	Fish

Animal	Age
Dog	11.66
Cat	16
Fish	3.66

derived
11.66
11.66
16
3.66
16
11.66
3.66
16
16
3.66

Feature engineering: Text

This dog is cute ! I love dogs

dog	this	is	jedi	love	cute	empero r	bingo	!	I
2	1	1	0	1	1	0	0	1	1

- SVD
- Word2vec

word	f1	f2	f3	f4	f5
dog	0.8	0.2	0.1	-0.1	-0.4
this	-0.2	0	0.1	0.8	0.3
love	0.1	0.2	0.7	-0.5	0.1
cute	0.2	0.2	0.8	-0.4	0

King – Man = Queen

$$X = U\Sigma V^T$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

X U Σ V

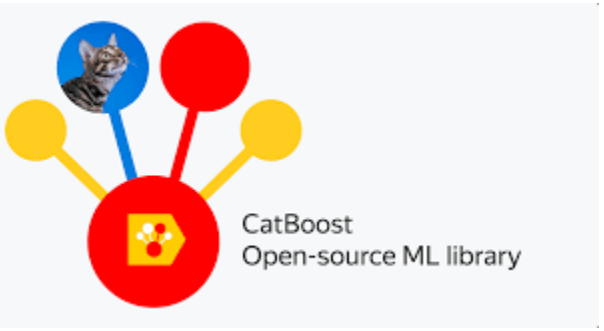
Feature engineering: Time series

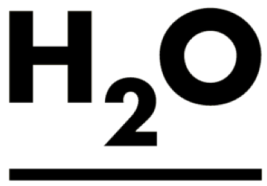
Date	Sales	Day	Lag1	Lag2	Year	MA2	weeknum
1/1/2018	100	-	-	-	-	-	-
2/1/2018	150	1	100	-	2018	100	1
3/1/2018	160	2	150	100	2018	125	1
4/1/2018	200	3	160	150	2018	155	1
5/1/2018	210	4	200	160	2018	180	1
6/1/2018	150	5	210	200	2018	205	1
7/1/2018	160	6	150	210	2018	180	1
8/1/2018	120	7	160	150	2018	155	2
9/1/2018	80	8	120	160	2018	140	2
10/1/2018	70	9	80	120	2018	100	2
		10		1	2018	4	2

Common open source packages used in ML

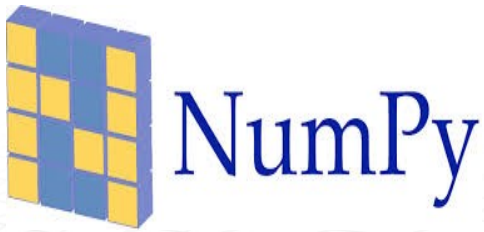
Microsoft LightGBM

dmlc XGBoost





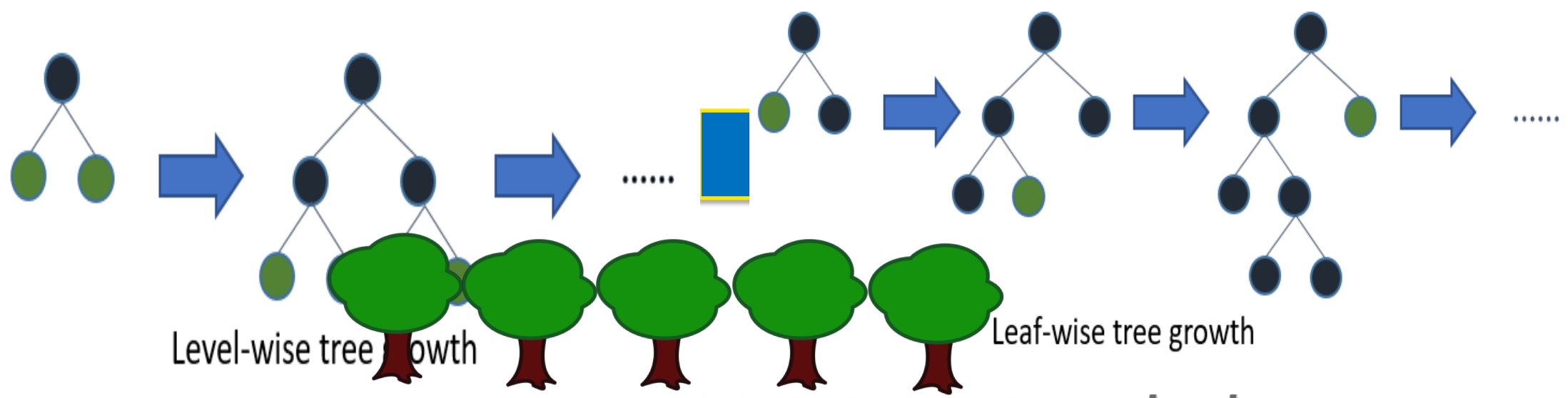
PYTORCH



LIBLINEAR
LIBFFM
LIBSVM

Hyper parameter tuning

- Optimizing maximum depth
- Tree expansion (loss vs depth)
-
-



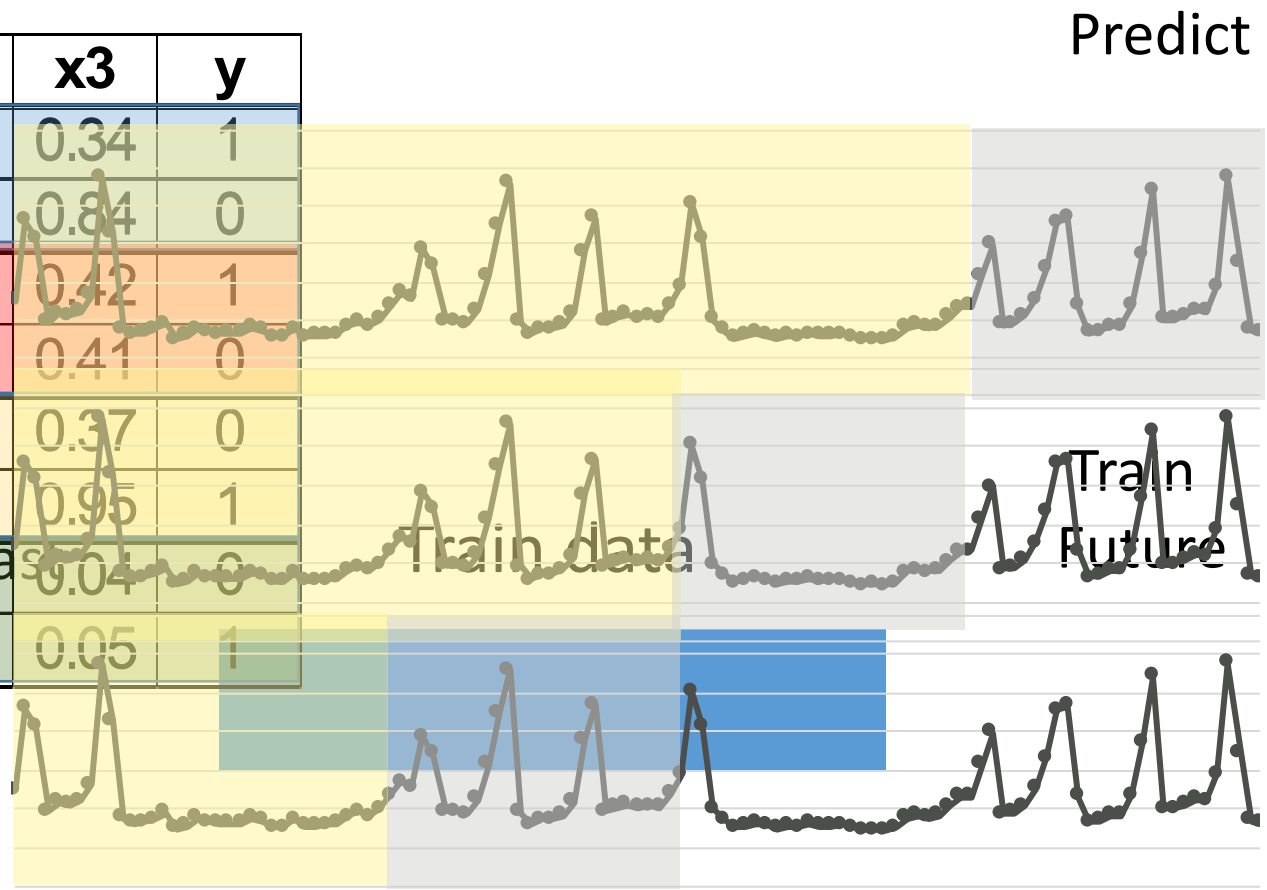
dmlc XGBoost

Validation approach

Fold : 1 2 3 4

K=4

x0	x1	x2	x3	y
0.94	0.27	0.80	0.34	1
0.02	0.22	0.17	0.84	0
0.83	0.11	0.23	0.42	1
0.74	0.26	0.03	0.41	0
0.08	0.29	0.76	0.37	0
0.71	0.76	0.43	0.95	1
0.08	0.72	0.97	0.04	0
0.84	0.79	0.89	0.05	1

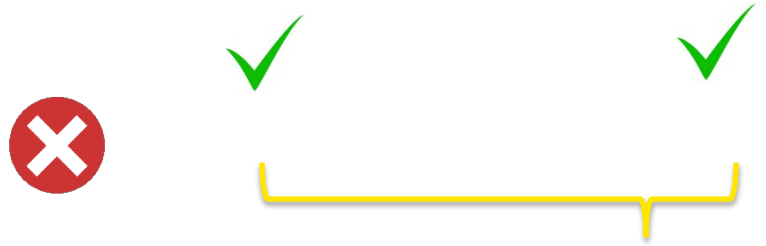


pred
0.96
0.03
0.90
0.02
0.03
0.77
0.08
0.90

Genetic algorithm approach

x1	x2	x3	x4	y
0.14	0.69	0.01	0.71	1
0.22	0.44	0.45	0.69	1
0.12	0.35	0.51	0.23	0
0.22	0.42	0.79	0.60	1
0.93	0.82	0.72	0.50	1
0.32	0.58	0.28	0.22	0
0.95	0.59	0.68	0.09	1
0.34	0.58	0.35	0.81	0
0.05	0.80	0.28	0.86	1
0.23	0.49	0.63	0.03	0
0.05	0.34	0.53	0.73	1
0.74	0.02	0.33	0.56	0

Iteration 1/10



XGBoost

x2 + x4

+ Random Transformation

sqrt(x2)

Tuned XGBoost

X% accuracy

Feature	Importance
x2*x4	1
x4	0.5
sqrt(x2)	0.3
x3	0.02
x2	0.05

Feature ranking based on permutations

Train

0.5	0.52	0.69	0.2	1
0.18	0.94	0.57	0.68	0
0.67	0.6	0.76	0.91	1
0.25	0.68	0.57	0.07	1
0.83	0.07	0.76	0.56	0
0.49	0.92	0.64	0.02	0
0.41	0.17	0.7	0.57	1
0.56	0.96	0.75	0.59	1
0.21	0.96	0.14	0.09	0
0.61	0.61	0.38	0.07	0
0.26	0.07	0.13	0.87	1
0.02	0.66	0.27	0.48	1

Repeat for all features

0.02
0.56
0.21
0.41
0.61
0.26

fit LightGBM

score

Validation 80% accuracy

70% accuracy (drop)

The 10% difference is how important that feature is

Stacking

A

X0	x1	x2	xn	y
0.17	0.25	0.93	0.79	1
0.35	0.61	0.93	0.57	0
0.44	0.59	0.56	0.46	0
0.37	0.43	0.74	0.28	1
0.96	0.07	0.57	0.01	1

B

X0	x1	x2	xn	y
0.89	0.72	0.50	0.66	0
0.58	0.71	0.92	0.27	1
0.10	0.35	0.27	0.37	0
0.47	0.68	0.30	0.98	0
0.39	0.53	0.59	0.18	1

C

X0	x1	x2	xn	y
0.29	0.77	0.05	0.09	?
0.38	0.66	0.42	0.91	?
0.72	0.66	0.92	0.11	?
0.70	0.37	0.91	0.17	?
0.59	0.98	0.93	0.65	?

Train algorithm **0** on A and make predictions for B and C and save to **B1**, **C1**

Train algorithm **1** on A and make predictions for B and C and save to **B1**, **C1**

Train algorithm **2** on A and make predictions for B and C and save to **B1**, **C1**

Consider datasets A,B,C. Target variable (y) is known for A,B

Preds3
0.45
0.23
0.99
0.34
0.05

Train algorithm **3** on B1 and make predictions for C1

Machine Learning Interpretability?

"The ability to explain or to present in understandable terms to a human."

- Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint. 2017. https://arxiv.org/pdf/1702.08608.pdf

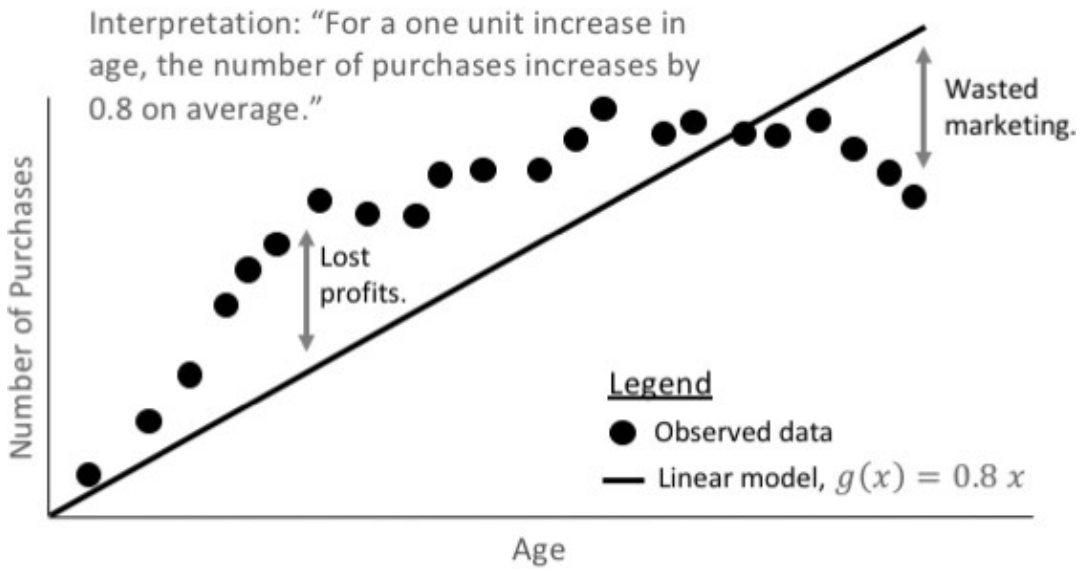
FAT*: https://www.fatml.org/resources/principles-for-accountable-algorithms

XAI: https://www.darpa.mil/program/explainable-artificial-intelligence

Interpretability and accuracy trade-off

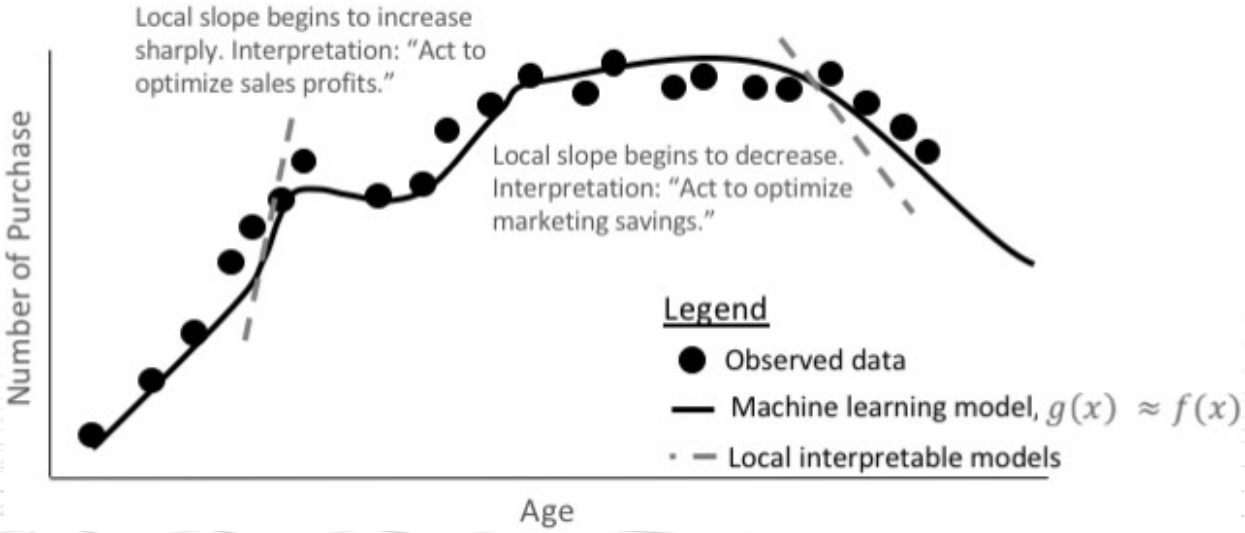
Linear Models

Exact explanations for ***approximate*** models.



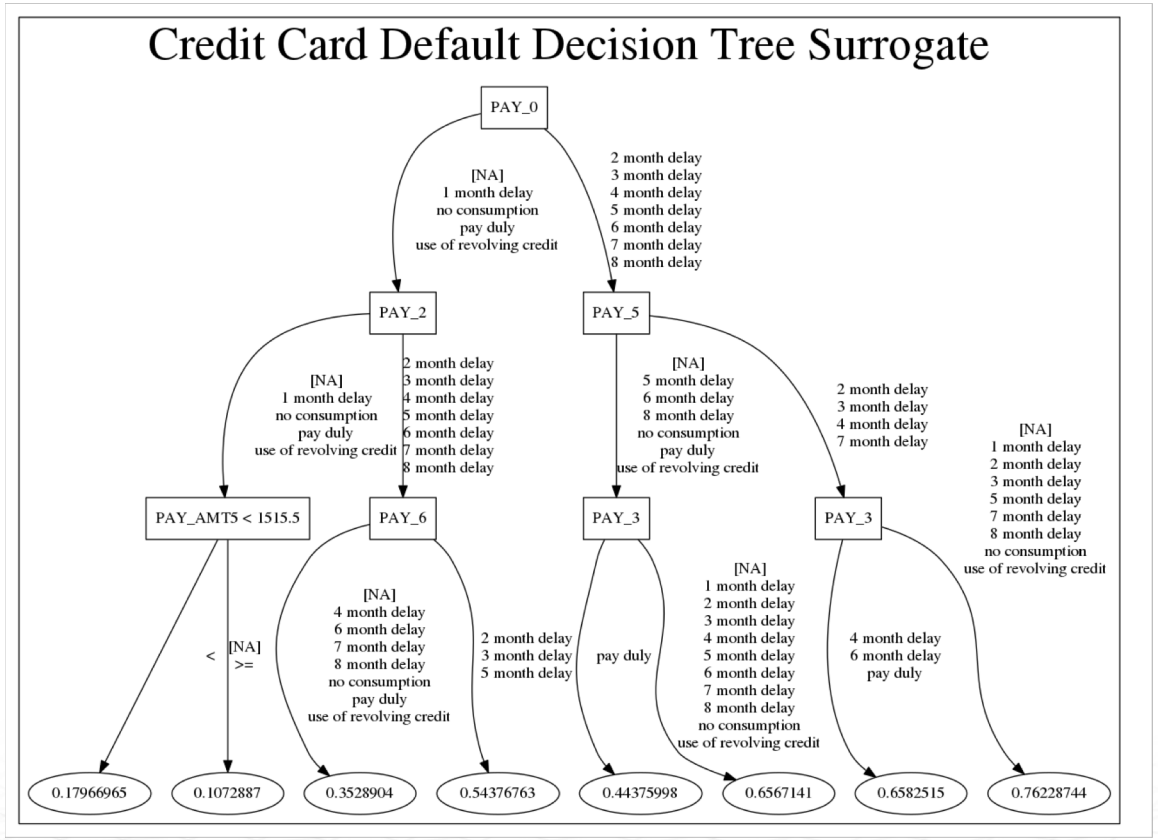
Machine Learning

Approximate explanations for ***exact*** models.

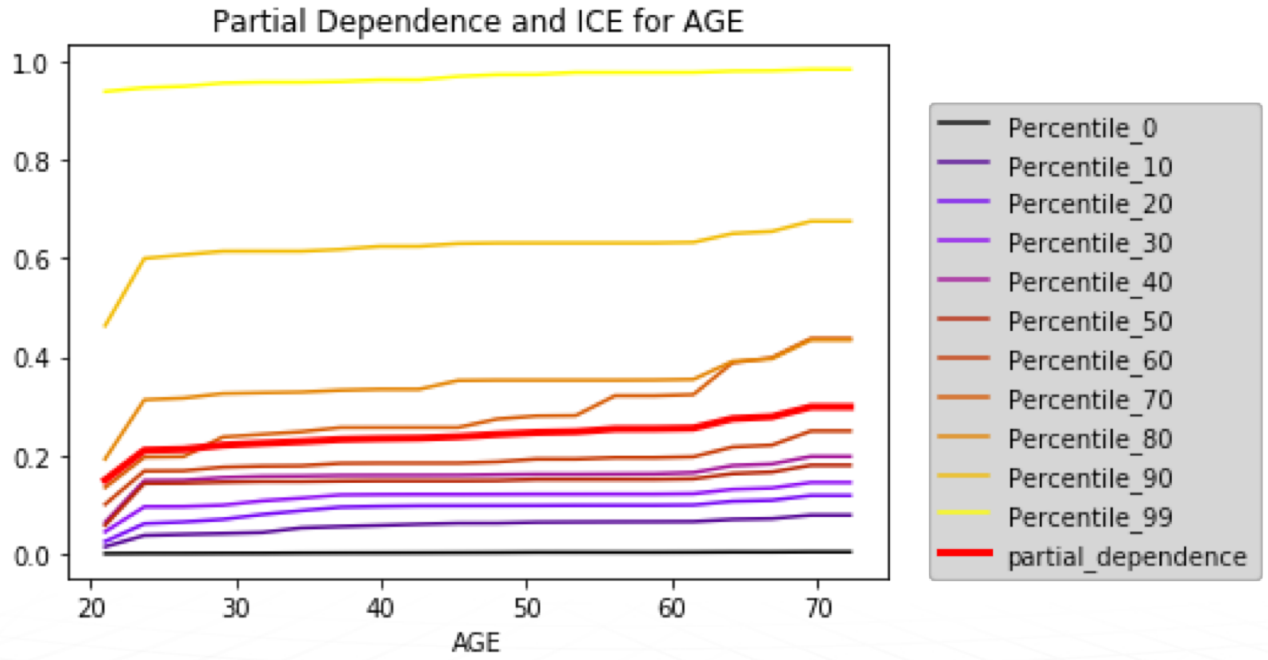


Machine Learning Interpretability examples

Decision Tree Surrogate Models



Partial Dependence and Individual Conditional Expectation



https://github.com/jphall663/interpretable_machine_learning_with_python

H2O.ai Product Suite

Open Source

In-memory, distributed machine learning algorithms with H2O Flow GUI

H2O AI open source engine integration with Spark

Lightning fast machine learning on GPUs

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise support subscriptions

Automatic feature engineering, machine learning and interpretability

- Enterprise software
- Built for domain users, analysts and data scientists – GUI-based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment

H2O is the open source leader in **AI**.

Democratize AI for **Everyone**.

AI for good.

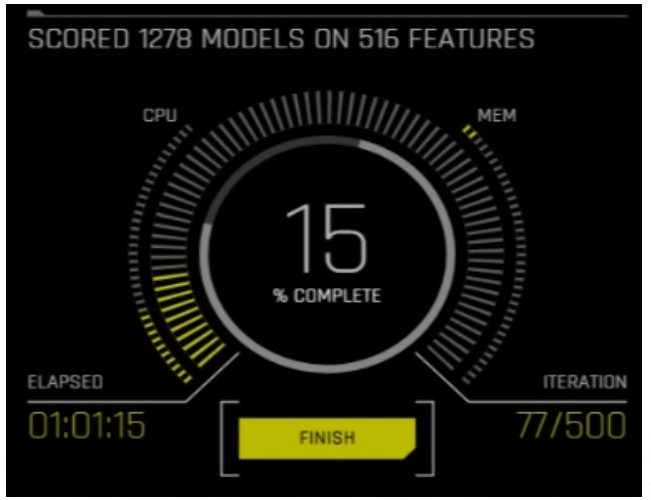
H2O Driverless AI process

- Some input data
- A target variable
- An objective (or a success metric)
- Some allocated resources

x1	x2	x3	x4	y
0.14	0.69	0.01	0.71	1
0.22	0.44	0.45	0.69	1
0.12	0.35	0.51	0.23	0
0.22	0.42	0.79	0.60	1
0.93	0.82	0.72	0.50	1
0.32	0.58	0.28	0.22	0
0.95	0.59	0.68	0.09	1
0.34	0.58	0.35	0.81	0
0.05	0.80	0.28	0.86	1
0.23	0.49	0.63	0.03	0
0.05	0.34	0.53	0.73	1
0.74	0.02	0.33	0.56	0

AUC, Accuracy, Precision…

Will there be a default?



Time, hardware

Insight-visualization

Feature Engineering

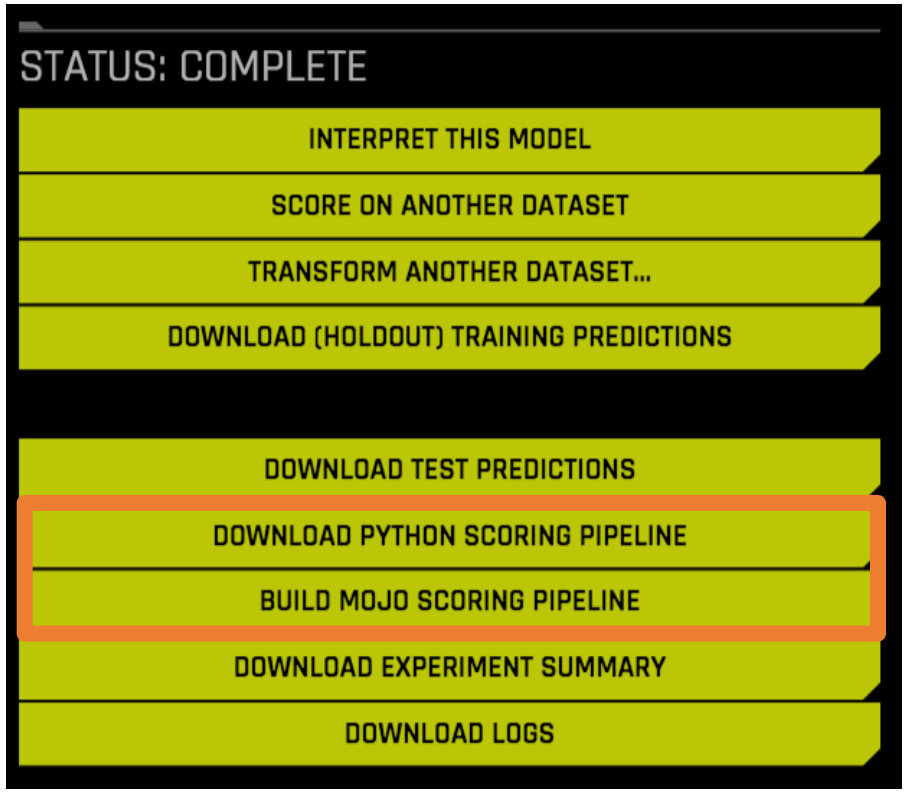
Predictions

Model Interpretability

Scoring Pipeline

- Python

- MOJO (Java)



Demo

Thank you

marios@h2o.ai
in/mariosmichailidis/
@StackNet_
https://www.facebook.com/StackNet/