

Automatic Feature Engineering in H₂O Driverless AI



Jo-fai (Joe) Chow
Data Science Evangelist /
Community Manager

joe@h2o.ai
@matlabulous

More Info → [https://bit.ly/
h2o_meetups](https://bit.ly/h2o_meetups)

H2O.ai HQ Mountain View





H2O.ai Prague

About Me

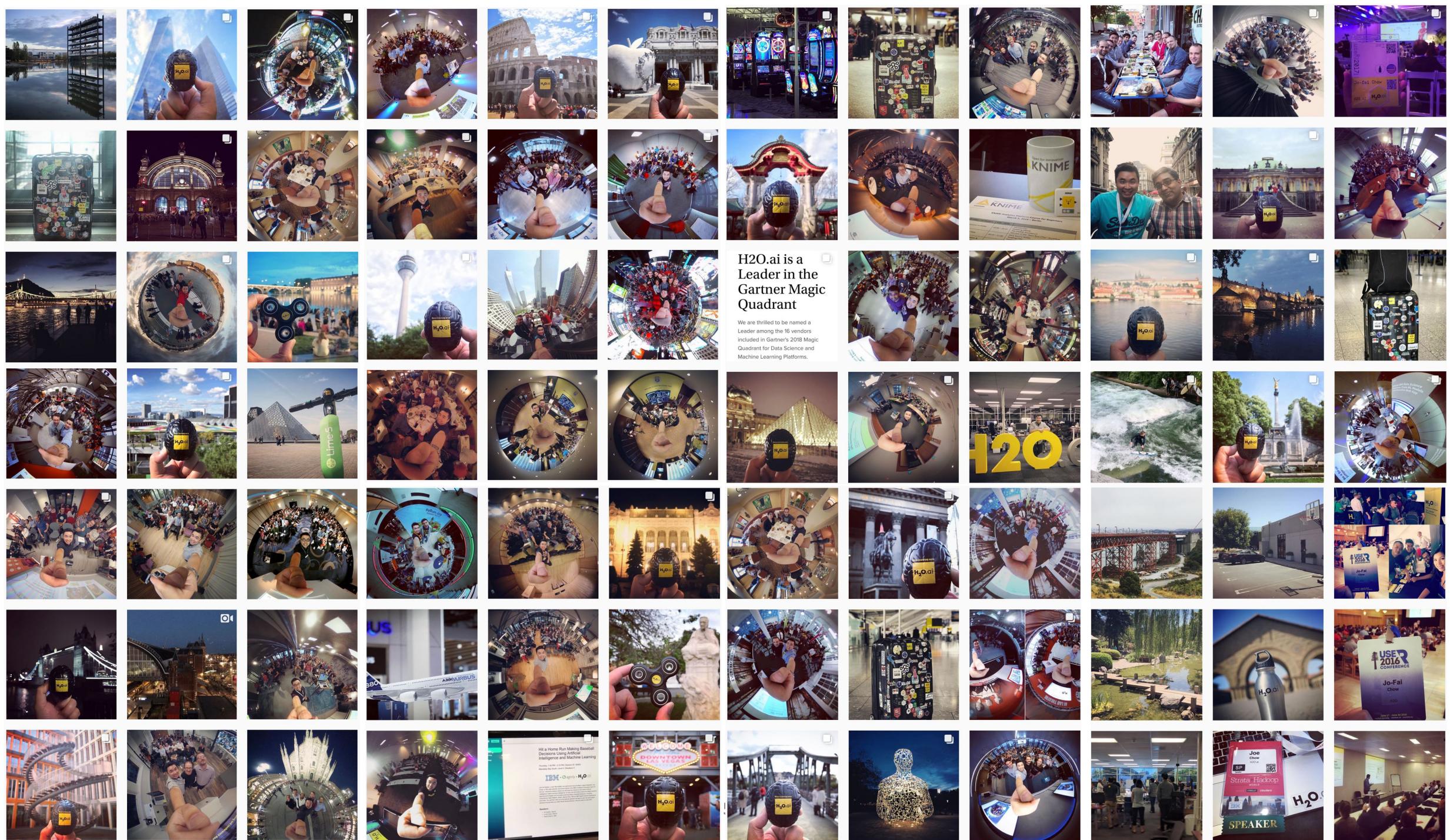


Jo-fai (Joe) Chow

Data Science Evangelist & Community Manager

joe@h2o.ai

- Before H₂O²
 - Water Engineer / EngD Researcher / Matlab Fan Boy
(wonder why  @matlabulous?)
 - Discovered R, Python, H₂O ... never look back again
 - Data Scientist at Virgin Media (UK), Domino Data Lab (US)
 - At H₂O ...
 - Data Scientist / Evangelist /
 - Sales Engineer / Solution Architect /
 - Event Organiser
 - Photographer
 - ... The harsh reality of startup life ...



H2O.ai is a
Leader in the
Gartner Magic
Quadrant

We are thrilled to be named a
Leader among the 16 vendors
included in Gartner's 2018 Magic
Quadrant for Data Science and
Machine Learning Platforms.

H2O

USER
CONFERENCE
Jo-Fai Chow

SP
Strata Hadoop
world
SPEAKER
H2O

Driverless AI Delivers “Expert Data Scientist in a Box”

- Created and supported by world renowned AI experts
- Empowers companies to accomplish AI and ML with a single platform
- Performs the function of an expert data scientist and adds more power to both novice and expert teams
- Details and highlights insights and interpretability with easy to understand results and visualizations



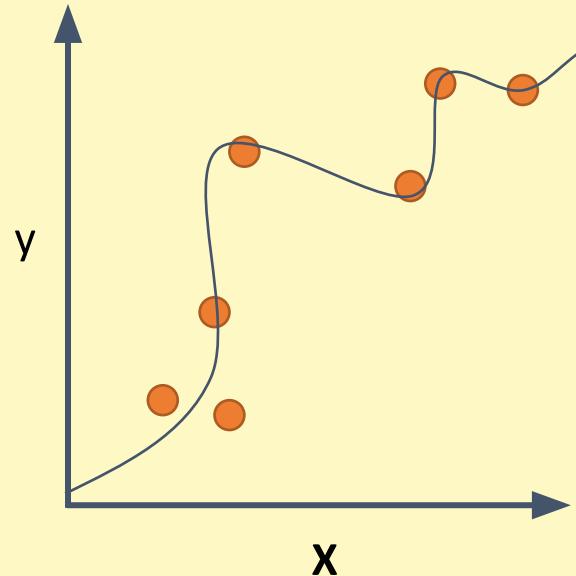
21 day free trial for [Driverless AI](#)

H₂O.ai

Supervised Learning

Regression:

How much will a customer spend?

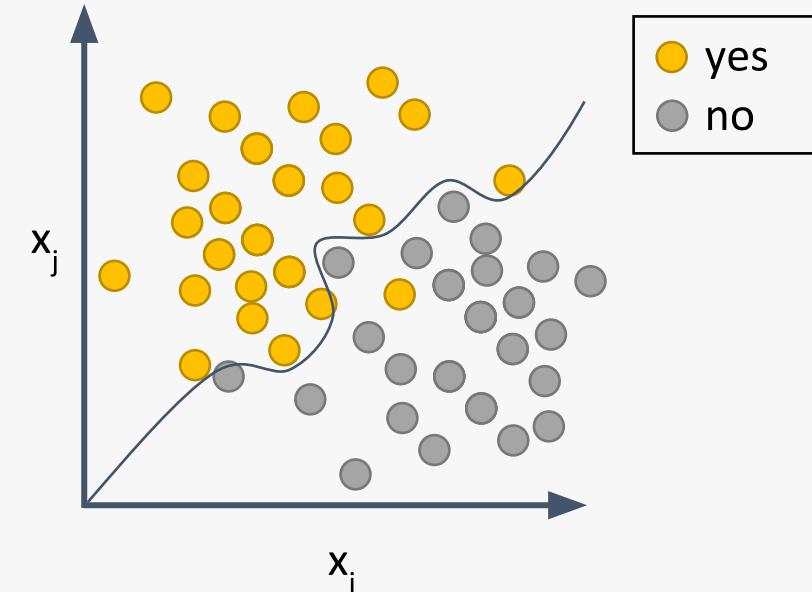


H₂O algos:

Penalized Linear Models
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

Classification:

Will a customer churn?

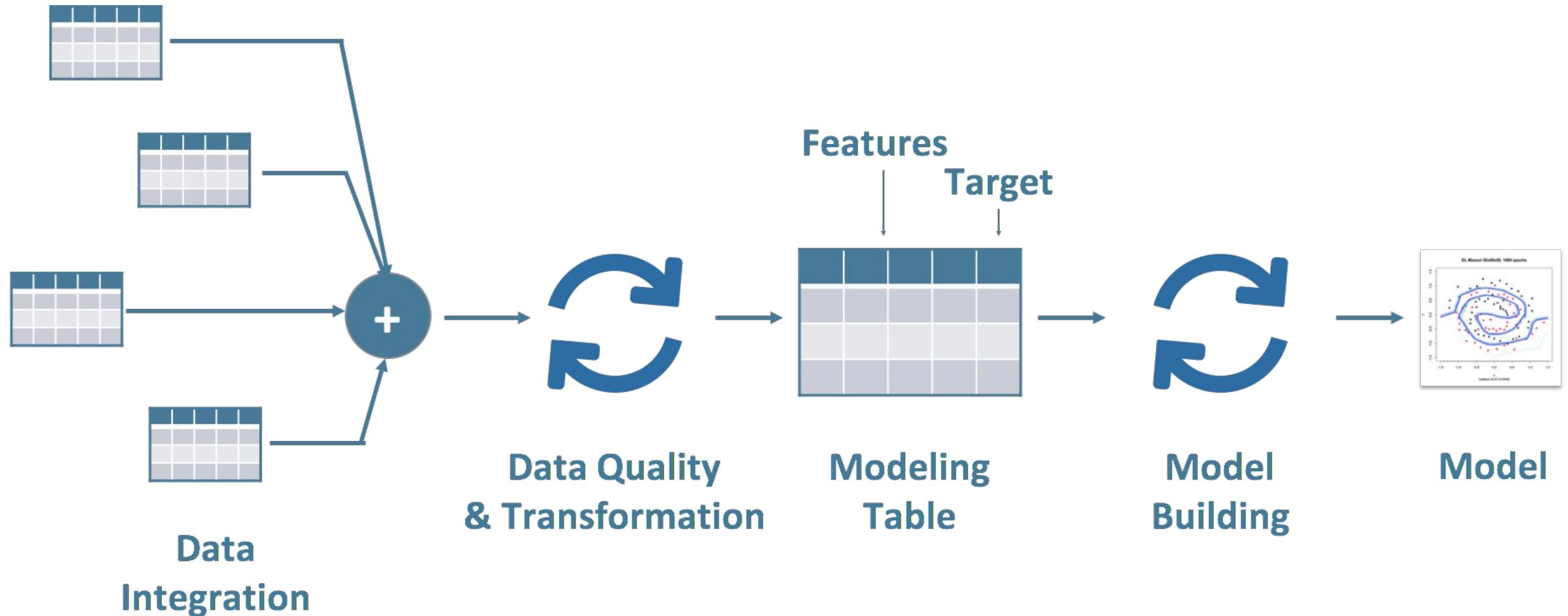


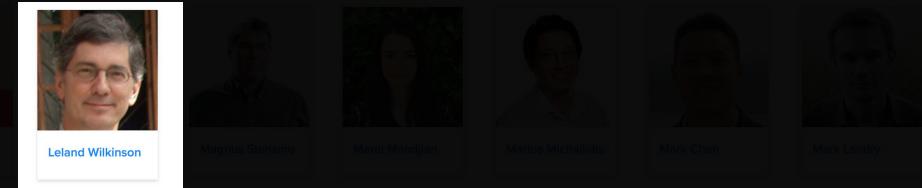
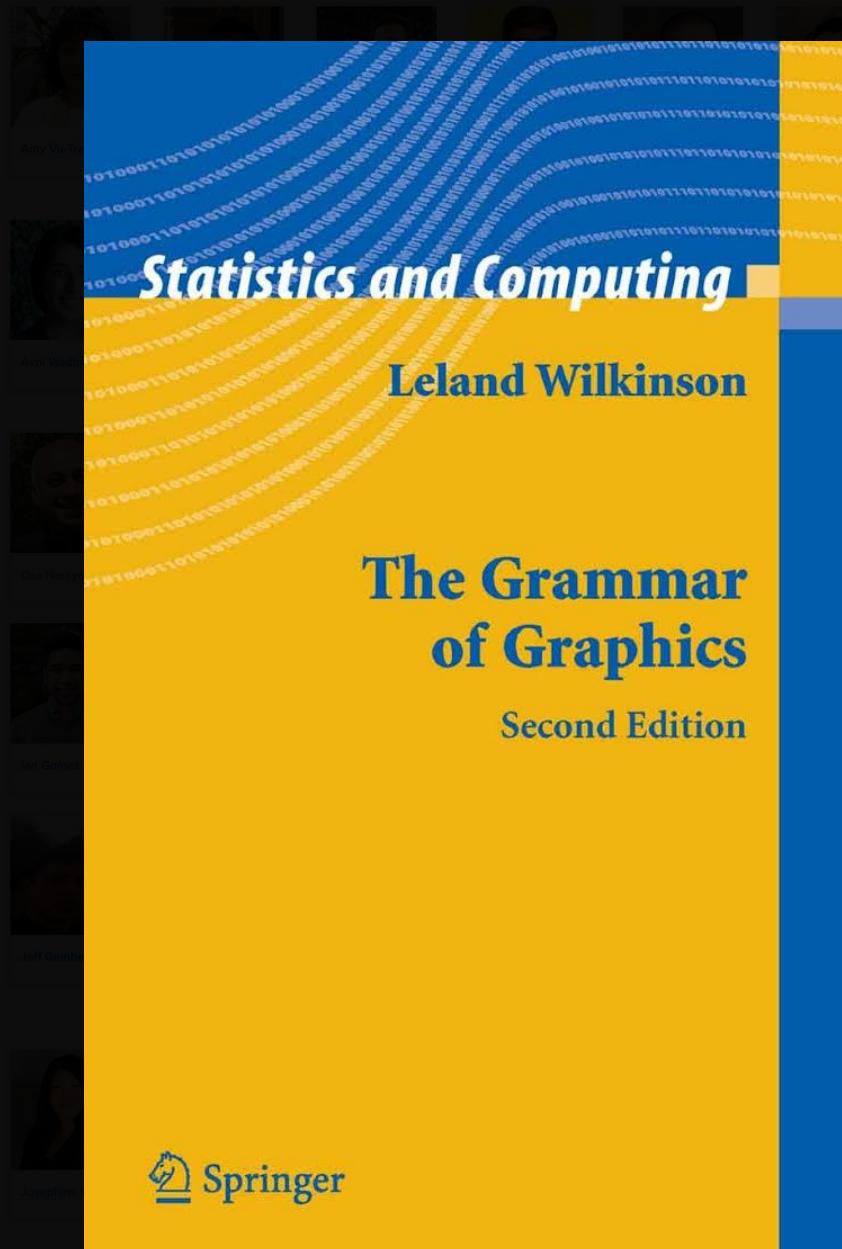
H₂O algos:

Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

H₂O.ai

Driverless AI: Automates Data Science and ML Workflows





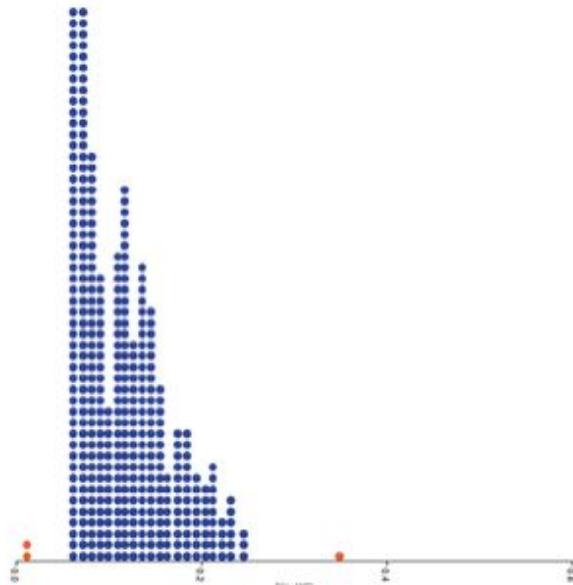
Origin of R Package `ggplot2`



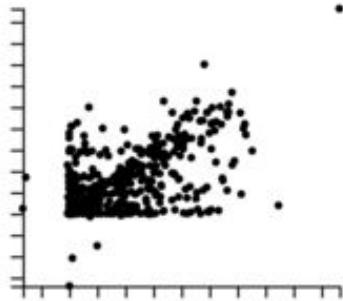
Automatic Visualization

H2O.ai

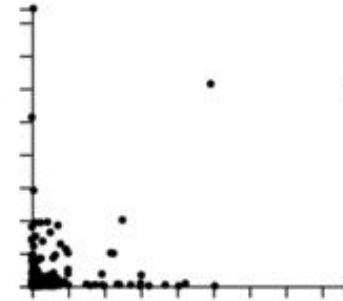
Automatic Scagnostics and other visualizations to generate the most relevant visualizations for each dataset



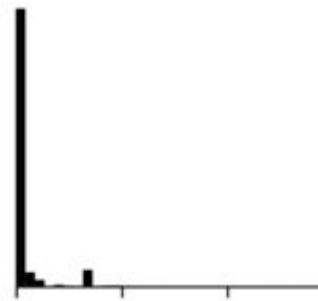
CLUMPY SCATTERPLOTS



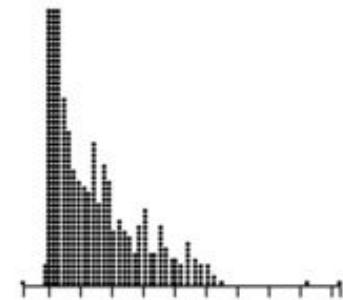
MONOTONIC SCATTERPLOTS



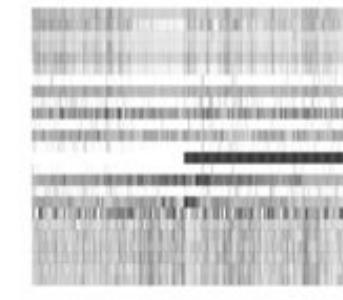
SPIKEY HISTOGRAMS



OUTLIERS



HEATMAPS



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

H₂O.ai



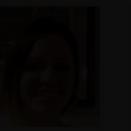
Kaggle Grand Masters (and their Highest Rank)



About 80,000 Kagglers



Amy Vu-Tran



Angela Barz

48th



Parvathy Chaudhary



Arno Candel



Ashwin Banbur



Leland Wilkinson



Magnus Stensmo



Maral Mandjarian



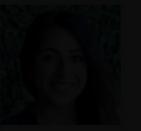
Marios Michailidis



Mark Chan



Mark Landry



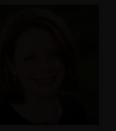
Avni Wadhwa



Benjamin Campbell



Branden Murray



Carl Andrews



Chandan Manocha



Chen Poff



Mateusz Dymczyk



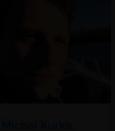
Mathias Müller



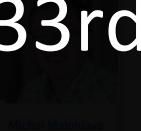
Matt Dowd



Megan Kurka



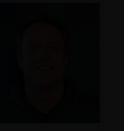
Michael Kurka



Michal Matolcova



Das Narayanan



David Crawford



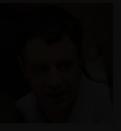
Dmitry Larko



Eric Quiggin



Erin LaBell



Gregory Kanhevsky



Michal Raksa



Monika Mullerova



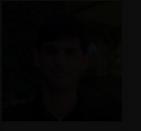
Navdeep Gill



Nishi Mehta



Nisha Shukhar



Nishant Kalonia



Ian Gomez



Jacqueline Scott



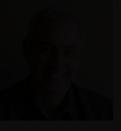
Jia Bining



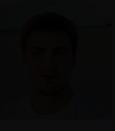
Jinhui Han



Jan Gamec



Jeff Follt



Ondrej Blazek



Pasha Slastenov



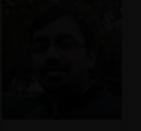
Patrick Abeyoum



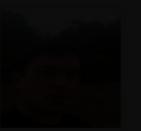
Patrick Hall



Patrick Rice



Prithviraj Babu



Jeff Gambara



Jo-Fai Chow



Josephine Wang



Justin Loyola



Karen Heyeropham



Kathy Lee



Kristina Arabo



Lauren DiPerna



Srikar Ambati



Terry Tang



Tom Kraljevic



Venkat Swaminathan

Venkatesh Yadav



Vinod Iyengar

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

181st



13th



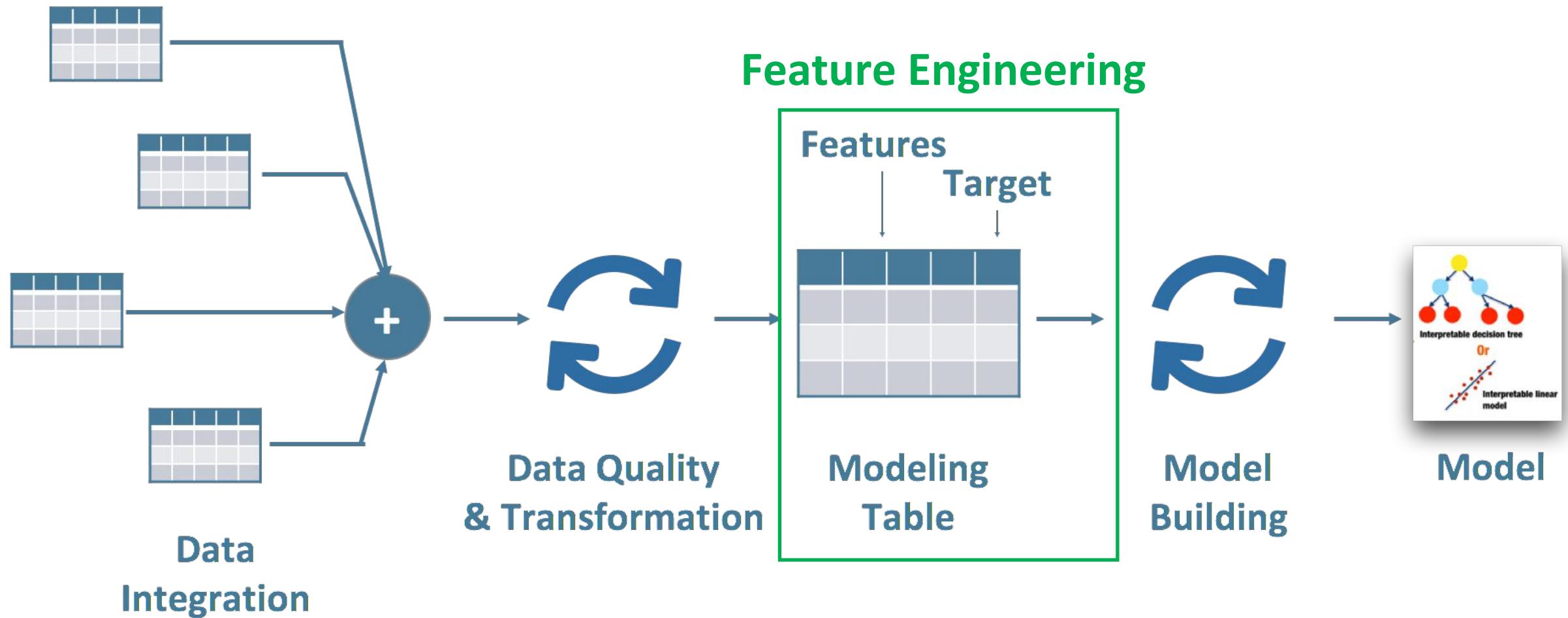
Wan Pham



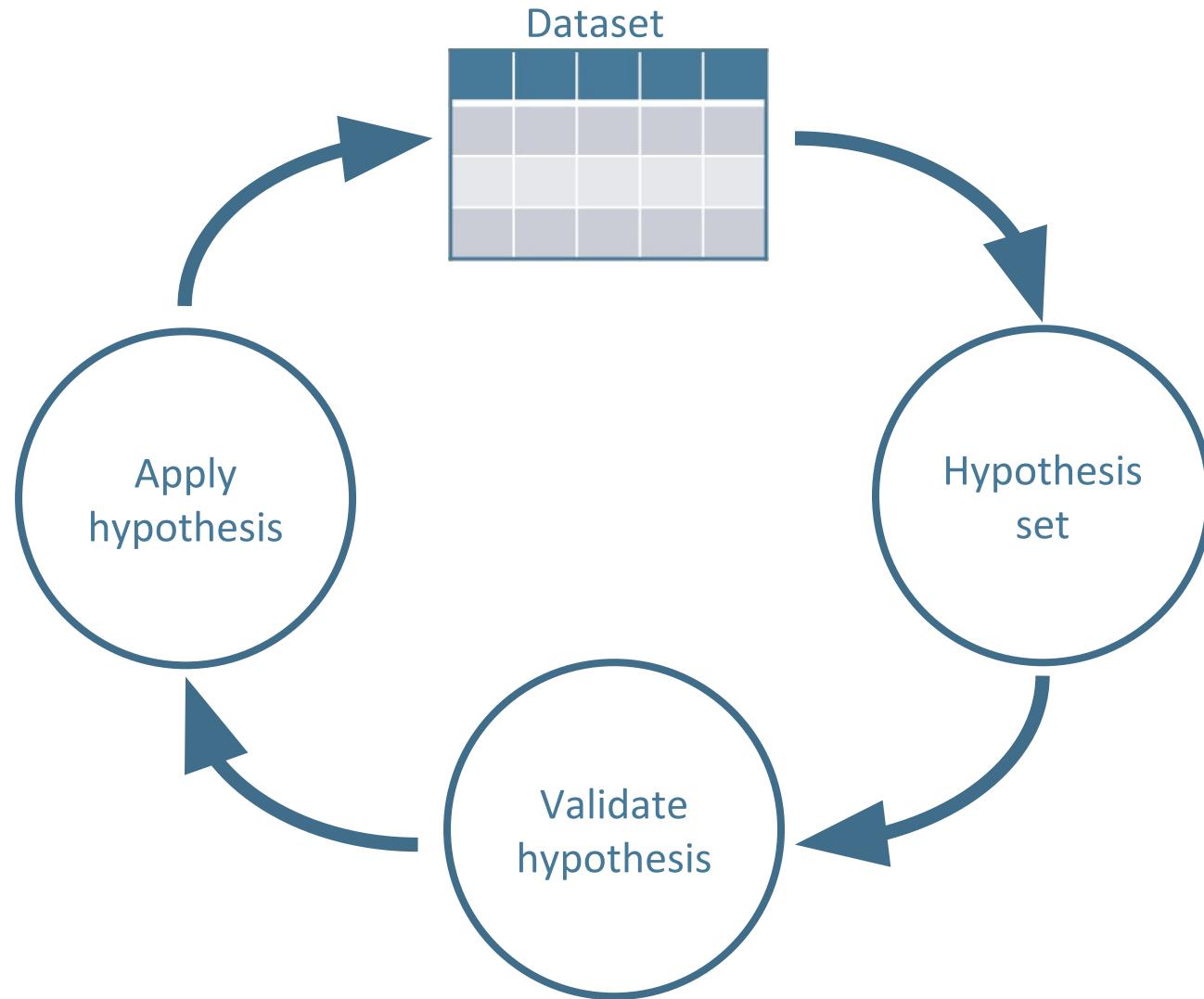
Wendy Wong

H₂O TeamH₂O.ai

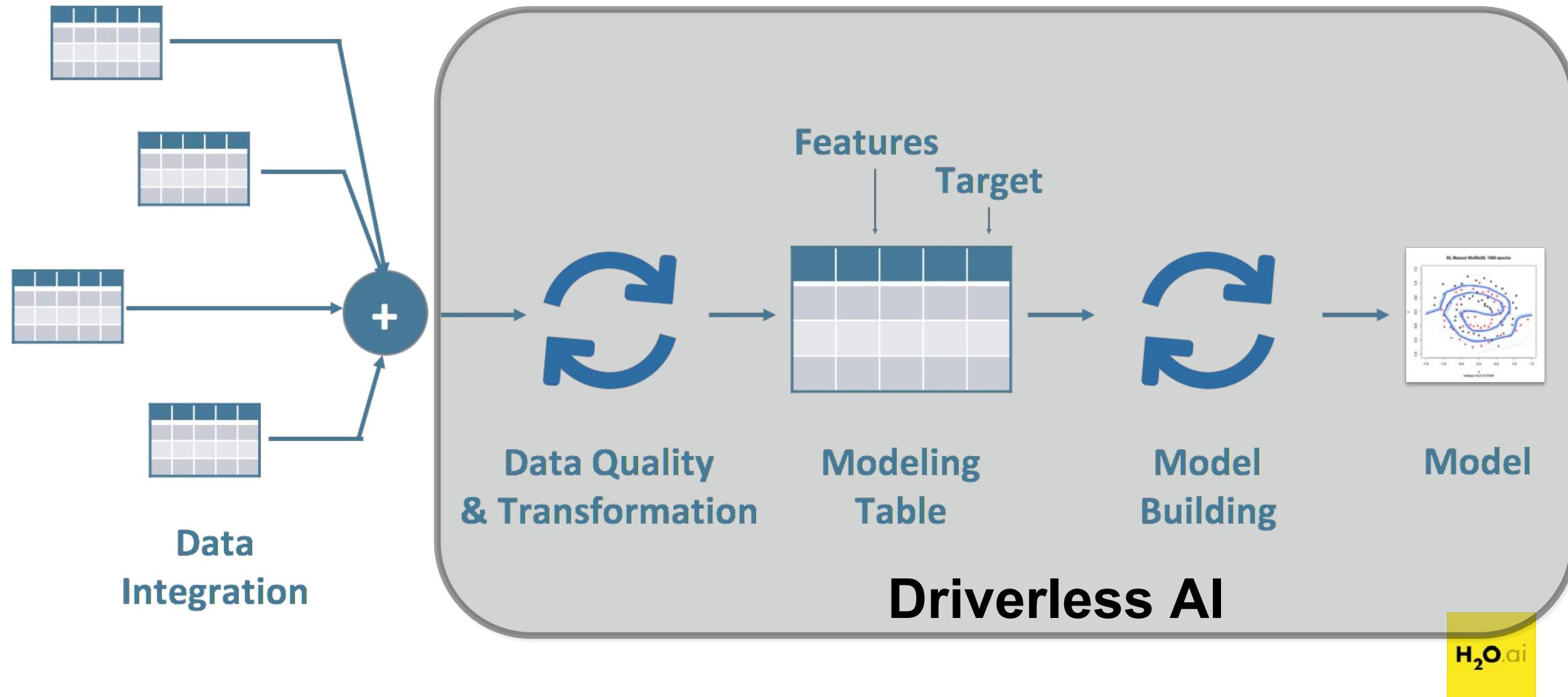
Typical Enterprise Machine Learning Workflow



Feature Engineering cycle



Driverless AI: Automates Data Science and ML Workflows



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

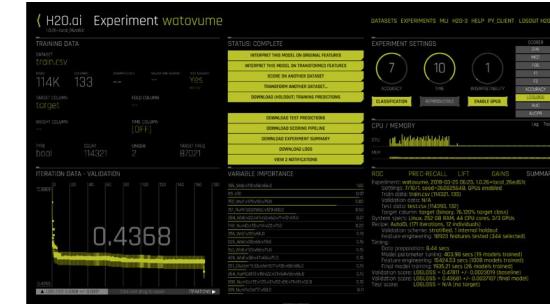


Amazon.com - Employee Access Needs

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

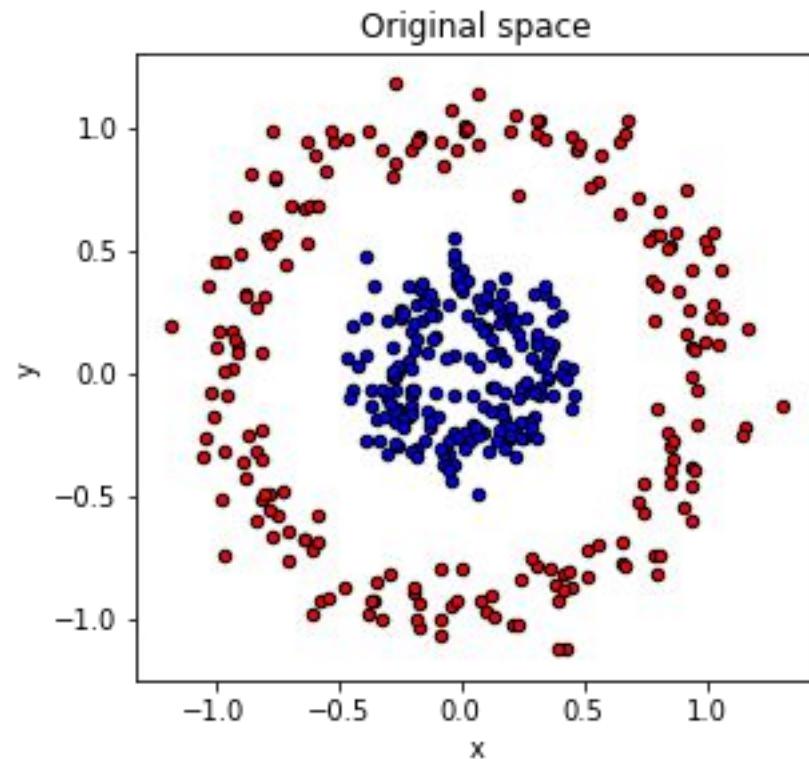
Interpretability

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



What is Feature Engineering?

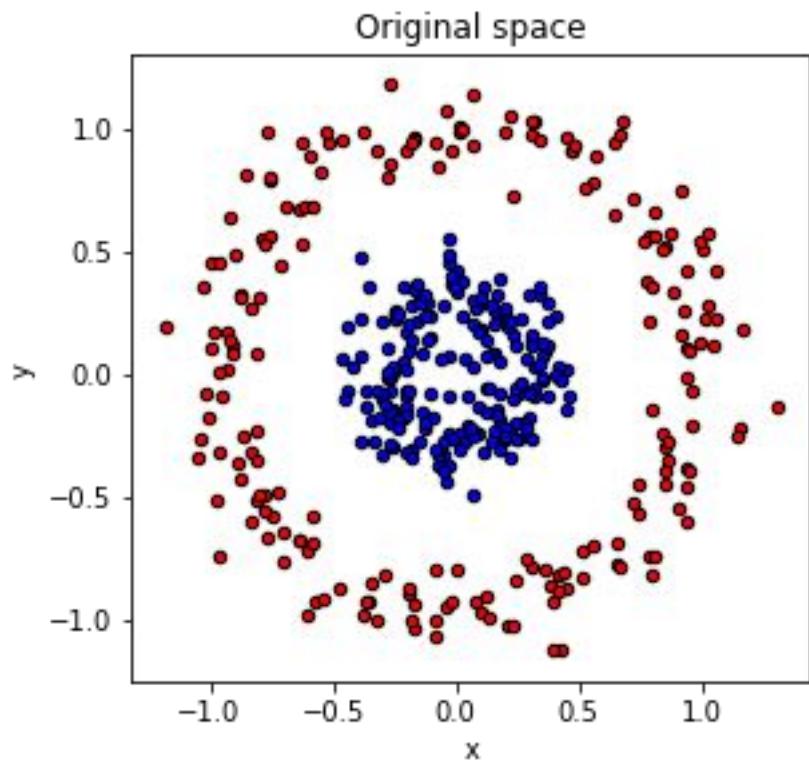
What is feature engineering



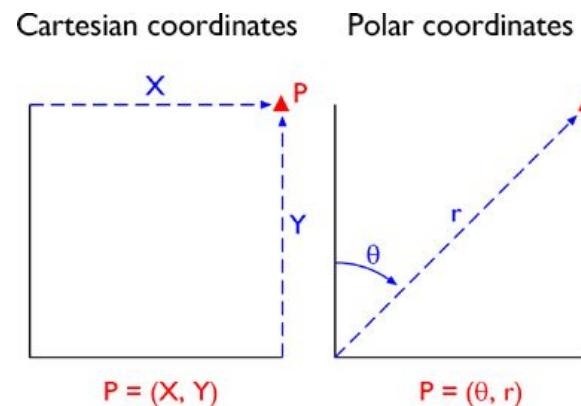
Not possible to separate using linear classifier



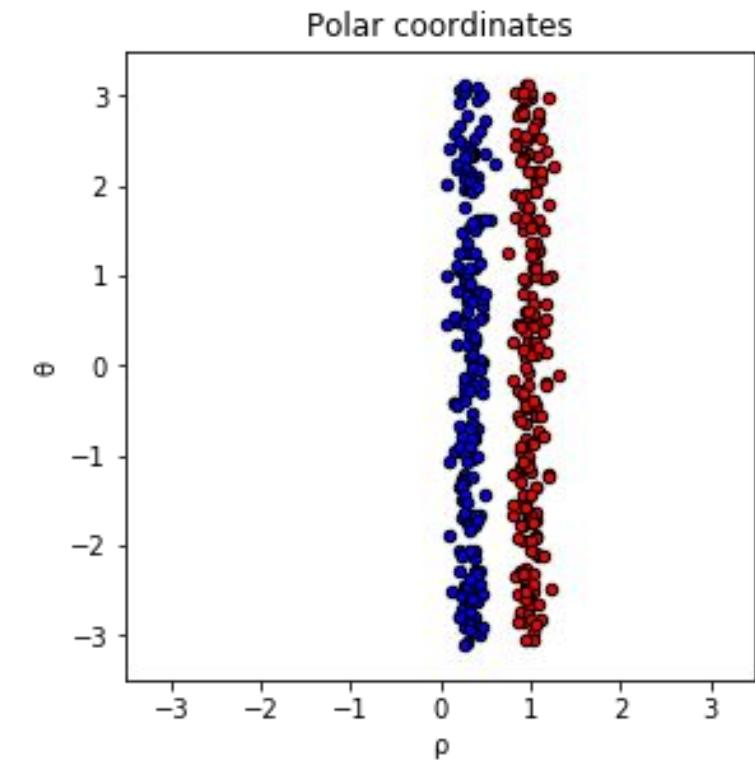
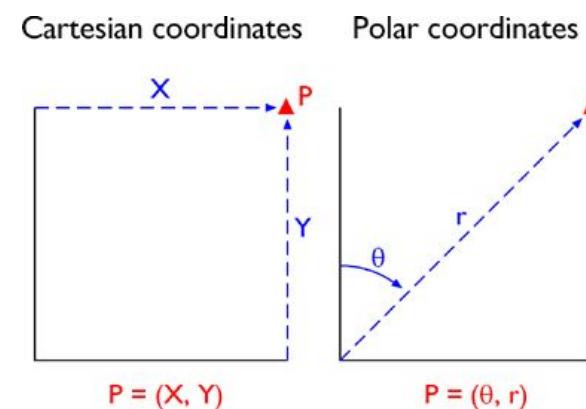
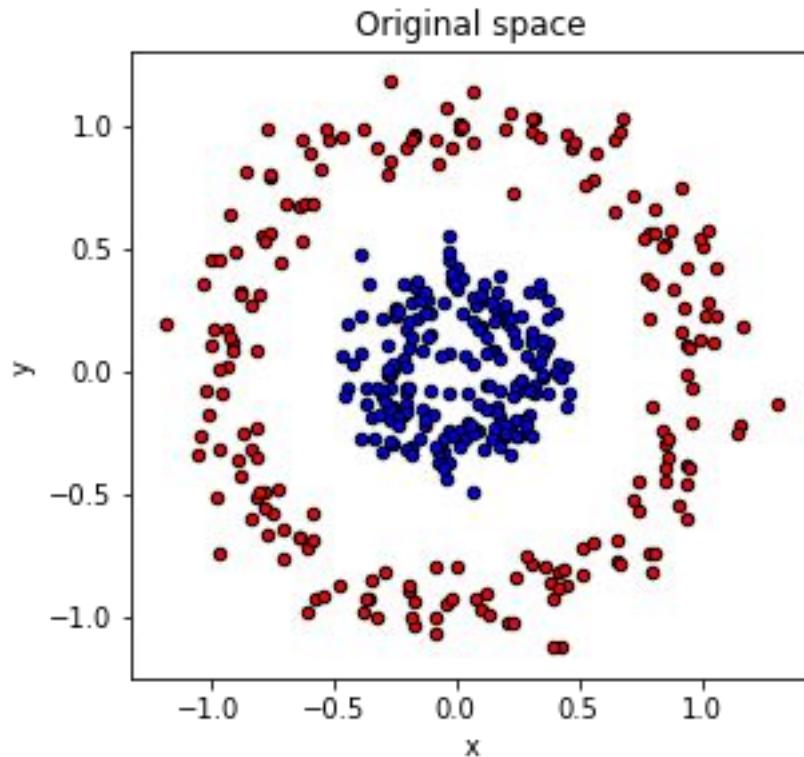
What is feature engineering



What if we use polar coordinates instead?



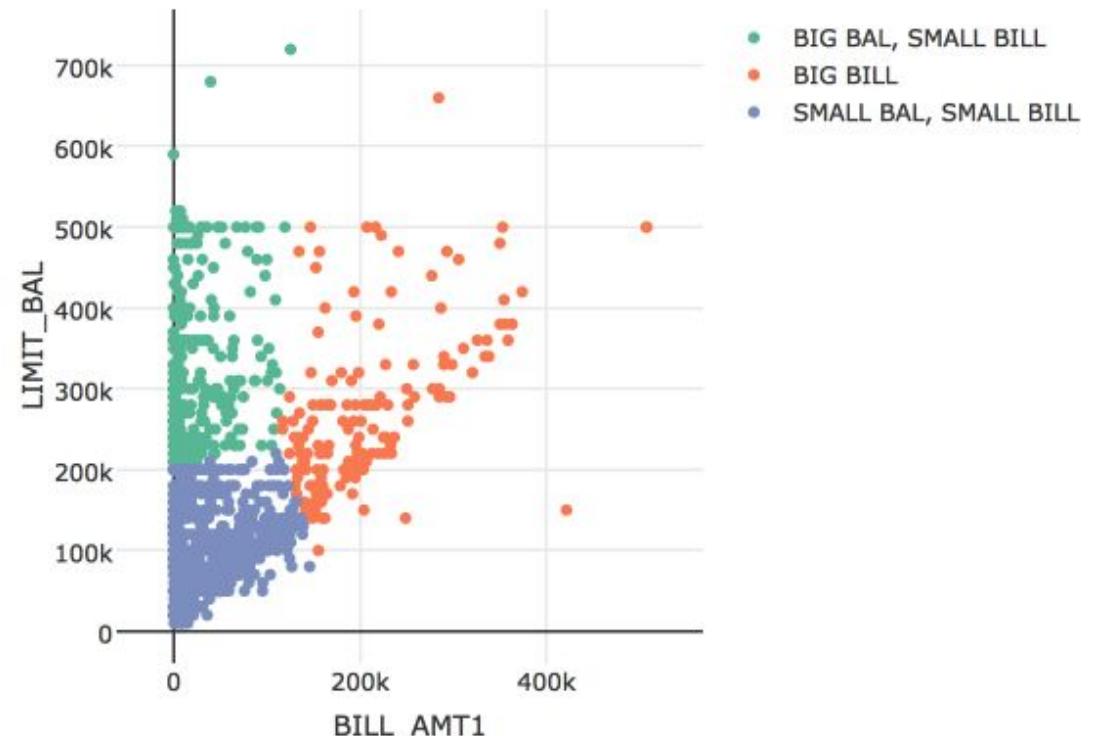
What is feature engineering



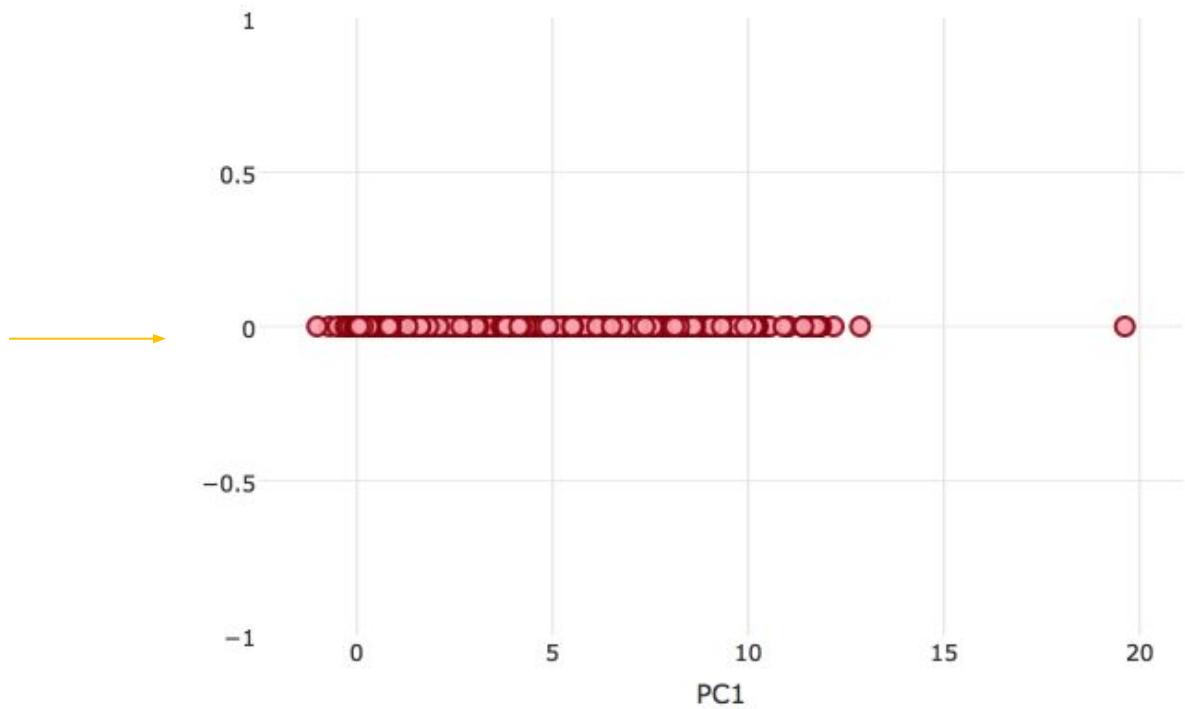
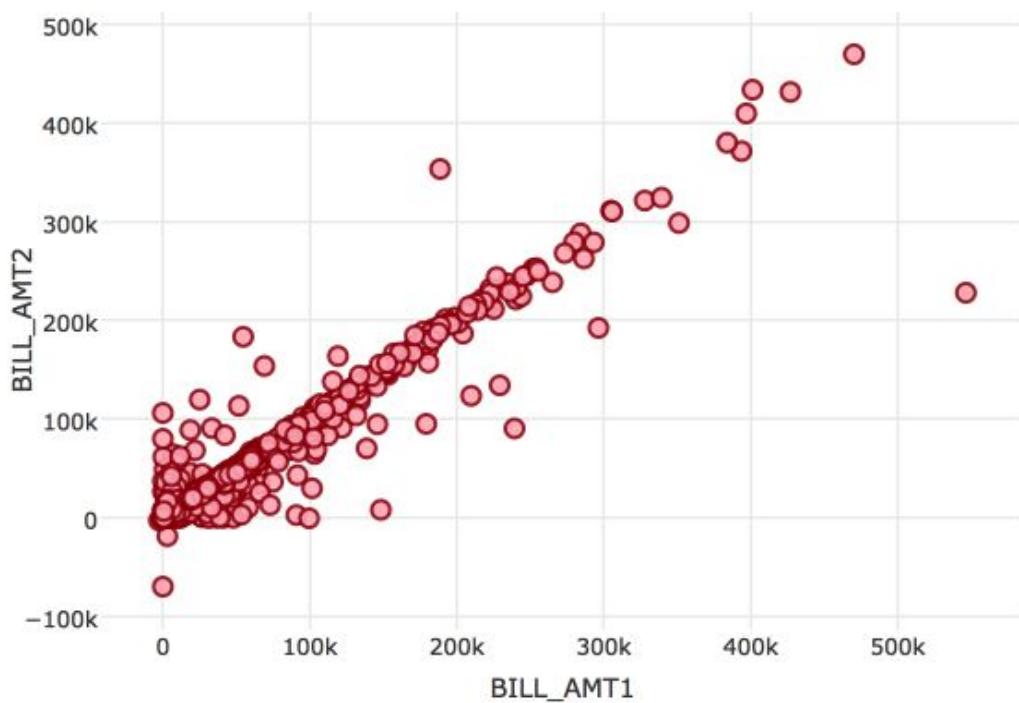
Clustering

Cluster Transformations

- Distance to a specific cluster
- Cross Validation Target Encoding by Cluster ID



Truncated SVD



Text Features

9_TxtTE:Description.0	1.00
27_WoE:HelpfulnessNumerator:Summary.0	0.31
20_WoE:HelpfulnessDenominator:Summary:UserId.0	0.20
4_CVTE:Summary.0	0.18
24_InteractionSub:HelpfulnessDenominator:Helpfu...	0.17
28_ClusterTE:ClusterID70:HelpfulnessDenominator:...	0.15
10_Txt:Description.22	0.05
10_Txt:Description.3	0.05
2_CVTE:ProductId.0	0.04
10_Txt:Description.5	0.03
10_Txt:Description.8	0.03
6_HelpfulnessDenominator	0.03
10_Txt:Description.18	0.03
10_Txt:Description.11	0.03

TxtTE – Train a linear model on the text components from TF-IDF

Txt – Components from a TF-IDF Matrix

There are Many More Tricks! Kaggle Grand Master Out of the Box

VARIABLE IMPORTANCE	
34_CV_CatNumEnc_PAY_0_PAY_2_mean	1.00
37_CV_TE_PAY_0_PAY_5_0	0.83
36_TruncSVD_PAY_3_PAY_0_0	0.66
44_CV_TE_PAY_2_PAY_5_0	0.51
16_PAY_0	0.33
22_BILL_AMT1	0.31
28_PAY_AMT1	0.28
30_PAY_AMT3	0.26
45_CV_CatNumEnc_PAY_3_LIMIT_BAL_PAY_AMT1_std	0.21
45_CV_CatNumEnc_PAY_3_LIMIT_BAL_BILL_AMT1_std	0.17
4_Freq_AGE	0.17
29_PAY_AMT2	0.16
24_BILL_AMT3	0.15
33_PAY_AMT6	0.14

Generated Features

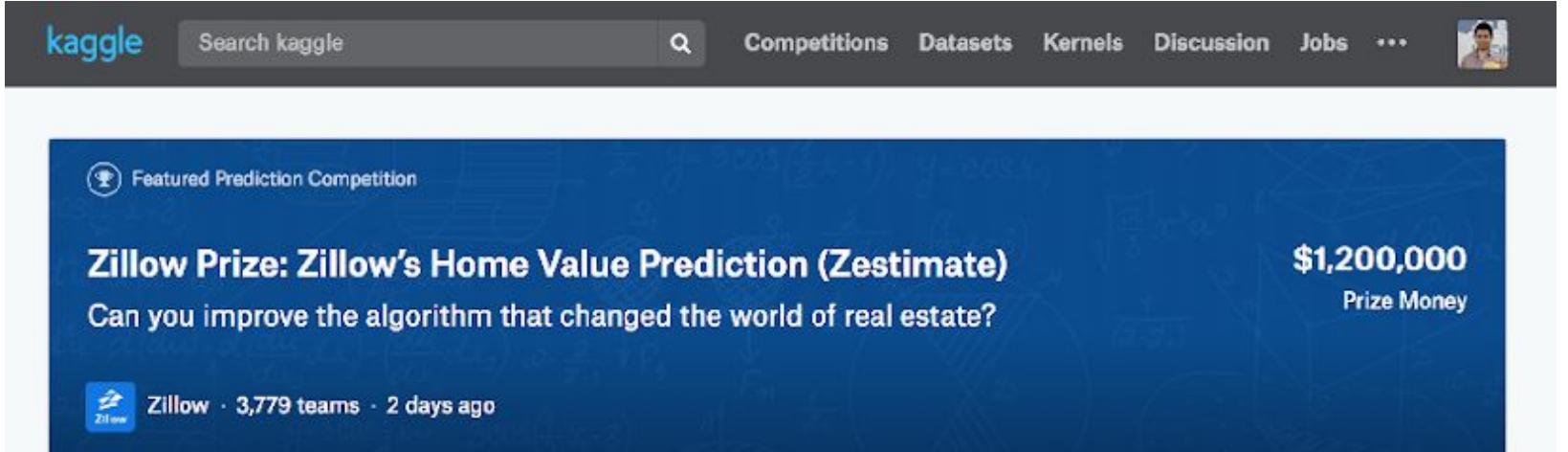
- Automatic Text Handling
- Frequency Encoding
- Cross Validation Target Encoding
- Truncated SVD
- Clustering and more

Original Features

Feature Transformations

How about Auto Feature Engineering + Marios' StackNet?

Driverless AI + StackNet (9 months ago)

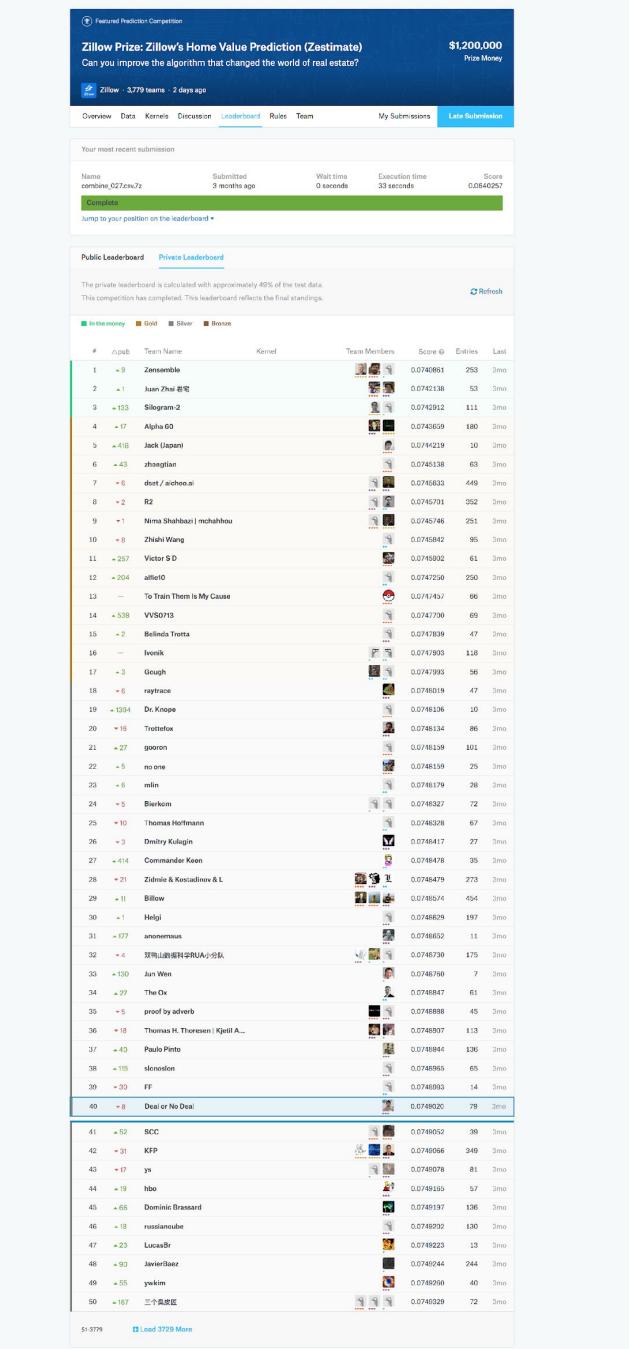


The screenshot shows the Kaggle homepage with the navigation bar: kaggle, Search kaggle, Competitions, Datasets, Kernels, Discussion, Jobs, and a user profile icon. Below the navigation is a banner for the "Featured Prediction Competition" titled "Zillow Prize: Zillow's Home Value Prediction (Zestimate)". The banner features a large "\$1,200,000 Prize Money" and the text "Can you improve the algorithm that changed the world of real estate?". It also includes the Zillow logo and the text "Zillow · 3,779 teams · 2 days ago".

Competition Round One (Top 100 to Next Round)

Rank	Team Name	Score	Entries	Last
40	Deal or No Deal	0.0749020	79	3mo
41	SCC	0.0749052	39	3mo
42	KFP	0.0749066	349	3mo

Finished above my H2O Kaggle Grandmasters colleagues



The screenshot shows the competition page for the "Zillow Prize: Zillow's Home Value Prediction (Zestimate)". At the top, it says "Zillow Prize: Zillow's Home Value Prediction (Zestimate)" and "Can you improve the algorithm that changed the world of real estate?". It shows "3,779 teams · 2 days ago" and a "Leaderboard" tab. The leaderboards section has tabs for "Public Leaderboard" and "Private Leaderboard". A note says "The private leaderboard is calculated with approximately 40% of the test data. This competition has completed. This leaderboard reflects the final standings." Below the tabs is a table with columns: Name, Score, Submitted, Wait time, Execution time, and Score.

Name	Score	Submitted	Wait time	Execution time	Score
combine_027cuz2	0.0749020	3 days ago	0 seconds	33 seconds	0.0640257

Public Leaderboard Private Leaderboard

The private leaderboard is calculated with approximately 40% of the test data. This competition has completed. This leaderboard reflects the final standings.

In the money Gold Silver Bronze

#	Rank	Team Name	Kernel	Team Members	Score	Entries	Last
1	49	Zensimble			0.0749051	283	3mo
2	1	Juan Zhai			0.0749038	53	3mo
3	153	Silograph-2			0.0749012	111	3mo
4	47	AlphaGO			0.0749009	180	3mo
5	418	Jack (Japan)			0.0749003	10	3mo
6	43	zhengqian			0.0749000	63	3mo
7	6	dset / achoo.ai			0.0748633	449	3mo
8	2	R2			0.0748791	262	3mo
9	1	Nirna Shahzai mchahzou			0.0748746	251	3mo
10	8	Zhiyi Wang			0.0748642	95	3mo
11	257	Victor S D			0.0748602	61	3mo
12	200	alife0			0.0747598	250	3mo
13	—	To Train Them Is My Cause			0.0747457	66	3mo
14	338	VVS0713			0.0747700	69	3mo
15	2	Belinda Trotta			0.0747839	47	3mo
16	—	Ivanik			0.0747903	118	3mo
17	6	Grouph			0.0747893	56	3mo
18	6	raytrace			0.0748019	47	3mo
19	139	Dr.Knoope			0.0748136	10	3mo
20	18	Trottelot			0.0748134	88	3mo
21	27	gooron			0.0748139	101	3mo
22	5	no one			0.0748179	29	3mo
23	6	mlin			0.0748327	72	3mo
24	5	Bierkorn			0.0748328	67	3mo
25	10	Thomas Hoffmann			0.0748417	27	3mo
26	3	Dmitry Kulagin			0.0748478	35	3mo
27	414	Commander Keen			0.0748479	273	3mo
28	21	Zidanie & Kosadine & L			0.0748474	494	3mo
29	11	Zillow			0.0748629	197	3mo
30	1	Heiji			0.0748652	11	3mo
31	177	anonymouse			0.0748730	175	3mo
32	4	汉山山海经HanshanHaike			0.0748760	7	3mo
33	130	Juri Winn			0.0748847	61	3mo
34	27	The Ox			0.0748888	45	3mo
35	5	prod by adverb			0.0748897	113	3mo
36	18	Thomas H. Thoresen Kjetil A...			0.0748944	136	3mo
37	43	Paulo Pinto			0.0748955	69	3mo
38	115	sknedlon			0.0748993	14	3mo
39	30	FF			0.0749020	79	3mo
40	8	Deal or No Deal			0.0749020	79	3mo
41	52	SCC			0.0749052	39	3mo
42	31	KFP			0.0749066	349	3mo
43	17	ys			0.0749078	81	3mo
44	19	hbo			0.0749165	57	3mo
45	65	Dominic Brassard			0.0749197	136	3mo
46	18	russiancoupe			0.0749202	130	3mo
47	23	LucasBr			0.0749223	19	3mo
48	93	JavierBaez			0.0749244	244	3mo
49	55	ywkim			0.0749298	40	3mo
50	167	三个臭皮匠			0.0749329	72	3mo

Live Demo


BNP PARIBAS CARDIF

BNP Paribas Cardif Claims Management

Can you accelerate BNP Paribas Cardif's claims management process?
\$30,000 · 2,926 teams · 2 years ago

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Overview

Description	As a global specialist in personal insurance, BNP Paribas Cardif serves 90 million clients in 36 countries across Europe, Asia and Latin America.
Evaluation	In a world shaped by the emergence of new uses and lifestyles, everything is going faster and faster. When facing unexpected events, customers expect their insurer to support them as soon as possible. However, claims management may require different levels of check before a claim can be approved and a payment can be made. With the new practices and behaviors generated by the digital economy, this process needs adaptation thanks to data science to meet the new needs and expectations of customers.
Prizes	
Timeline	
About Bnp Paribas Cardif	

Description

As a global specialist in personal insurance, **BNP Paribas Cardif** serves 90 million clients in 36 countries across Europe, Asia and Latin America.

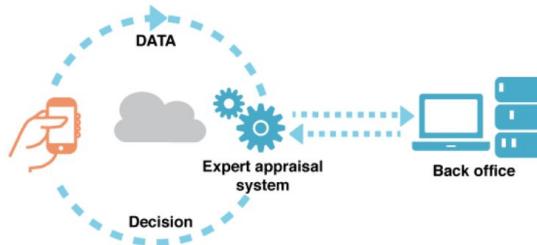
Evaluation

In a world shaped by the emergence of new uses and lifestyles, everything is going faster and faster. When facing unexpected events, customers expect their insurer to support them as soon as possible. However, claims management may require different levels of check before a claim can be approved and a payment can be made. With the new practices and behaviors generated by the digital economy, this process needs adaptation thanks to data science to meet the new needs and expectations of customers.

Prizes

Timeline

About Bnp Paribas Cardif



The diagram illustrates the data flow in the claims management process. It starts with a mobile phone icon on the left, connected by a dashed blue arrow labeled "DATA" to a central cloud icon. From the cloud, another dashed blue arrow labeled "Decision" points to a gear icon labeled "Expert appraisal system". Finally, a dotted blue arrow points from the appraisal system to a server icon labeled "Back office".

In this challenge, BNP Paribas Cardif is providing an anonymized database with two categories of claims:

1. claims for which approval could be accelerated leading to faster payments
2. claims for which additional information is required before approval

Kagglers are challenged to predict the category of a claim based on features available early in the process, helping BNP Paribas Cardif accelerate its claims process and therefore provide a better service to its customers.

LINK: <https://techdayhq.com/london/register#attend>

REGISTRATION CODE: **LDNFREE**





Turner & Townsend

H₂O.ai

- More Info, Code, and Slides
 - [bit.ly/
h2o_meetups](https://bit.ly/h2o_meetups)
- Contact
 - joe@h2o.ai
 - [@matlabulous](https://twitter.com/matlabulous)
 - github.com/woobe

Appendix

Target Mean Encoding

Pay 1	Default Payment
Up To Date	0
Up To Date	0
Up To Date	0
Missed 1 Mo	1
Missed 1 Mo	0
Missed 1 Mo	0
Missed 5 Mo	1

Target Mean Encoding

Pay 1	Default Payment	Mean Target Encoding
Up To Date	0	0
Up To Date	0	0
Up To Date	0	0
Missed 1 Mo	1	0.33
Missed 1 Mo	0	0.33
Missed 1 Mo	0	0.33
Missed 5 Mo	1	1

Target Mean Encoding

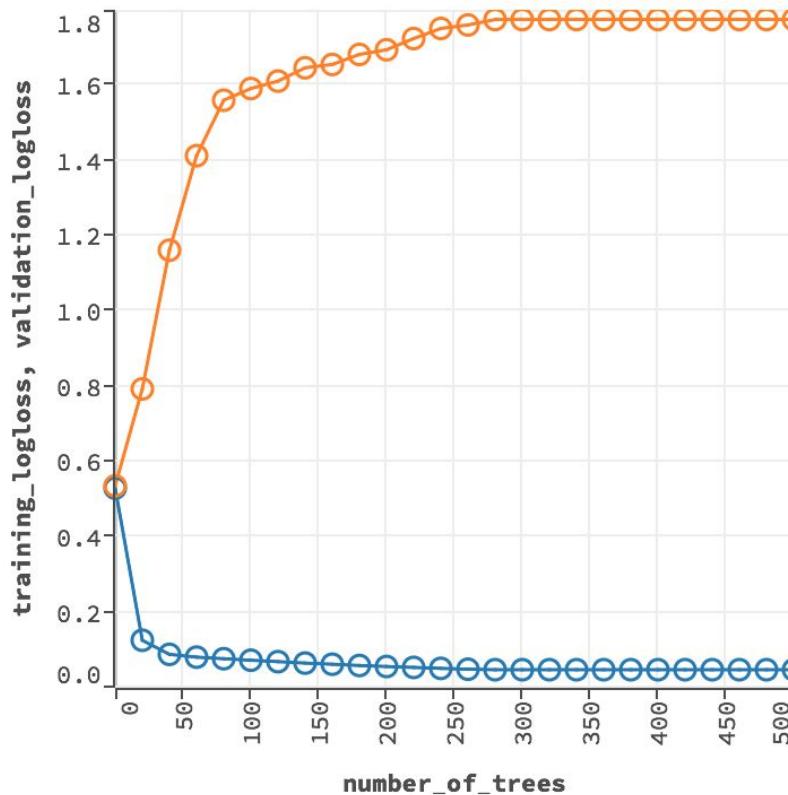
- Mean Target Encoding is based on the response column of the rows
- The lower the number of rows in the group, the more it reveals the response column value

Pay 1	Default Payment	Mean Target Encoding
Missed 5 Mo	1	1

Worst Case Scenario: Response Column = Mean Target Encoding

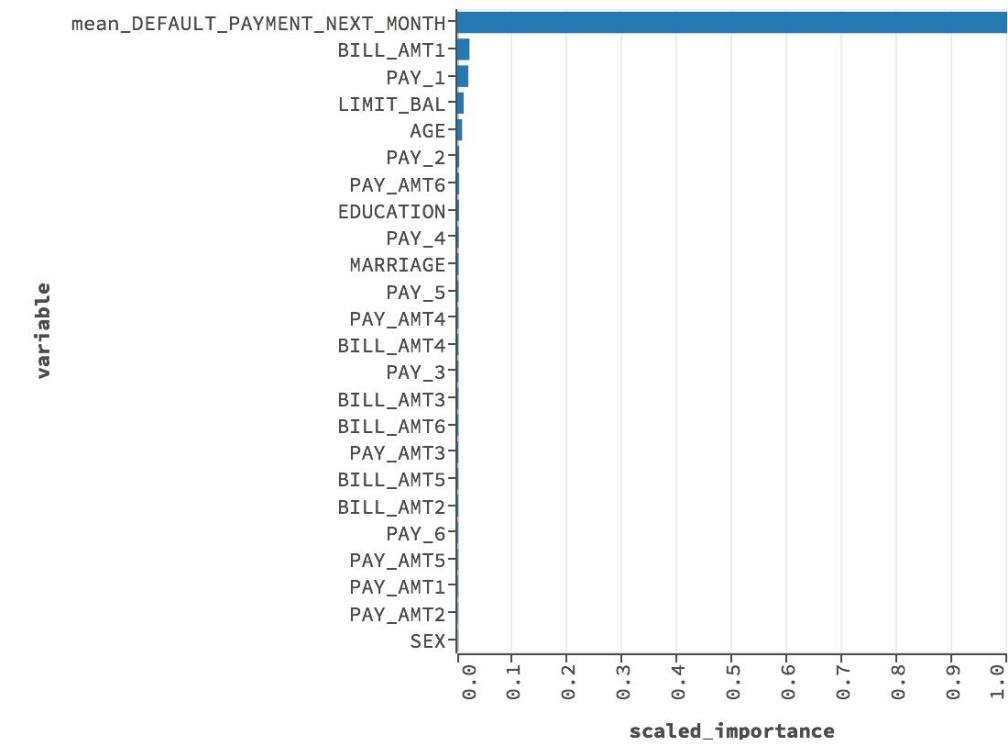
Effects of Data Leakage

▼ SCORING HISTORY - LOGLOSS



Scoring History: Training vs Testing

▼ VARIABLE IMPORTANCES



Data Leakage Feature is the only important feature

Cross Validation Target Encoding

Fold	Pay 1	Default Payment
1	Up To Date	0
2	Up To Date	0
3	Up To Date	0
2	Missed 1 Mo	1
1	Missed 1 Mo	0
3	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment
2	Up To Date	0
3	Up To Date	0
2	Missed 1 Mo	1
3	Missed 1 Mo	0

Fold	Pay 1	Default Payment
1	Up To Date	0
1	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment
1	Up To Date	0
1	Missed 1 Mo	0
1	Missed 5 Mo	1

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	
1	Missed 5 Mo	1	

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	

Cross Validation Target Encoding

Fold	Pay 1	Default Payment	Mean Target Encoding
2	Up To Date	0	0
3	Up To Date	0	0
2	Missed 1 Mo	1	0.5
3	Missed 1 Mo	0	0.5

Fold	Pay 1	Default Payment	CV Target Encoding
1	Up To Date	0	0
1	Missed 1 Mo	0	0.5
1	Missed 5 Mo	1	NA