Automatic Feature Engineering
in H_2O Driverless AI

Jo-fai (Joe) Chow

Data Science Evangelist / Community Manager

joe@h2o.ai

@matlabulous

More Info → https://bit.ly/

h2o_meetups


H2O.ai HQ
Mountain View
2307
MAKERS GONNA MAKE
MAKER
THIRSTY?

H2O.ai Prague

About Me

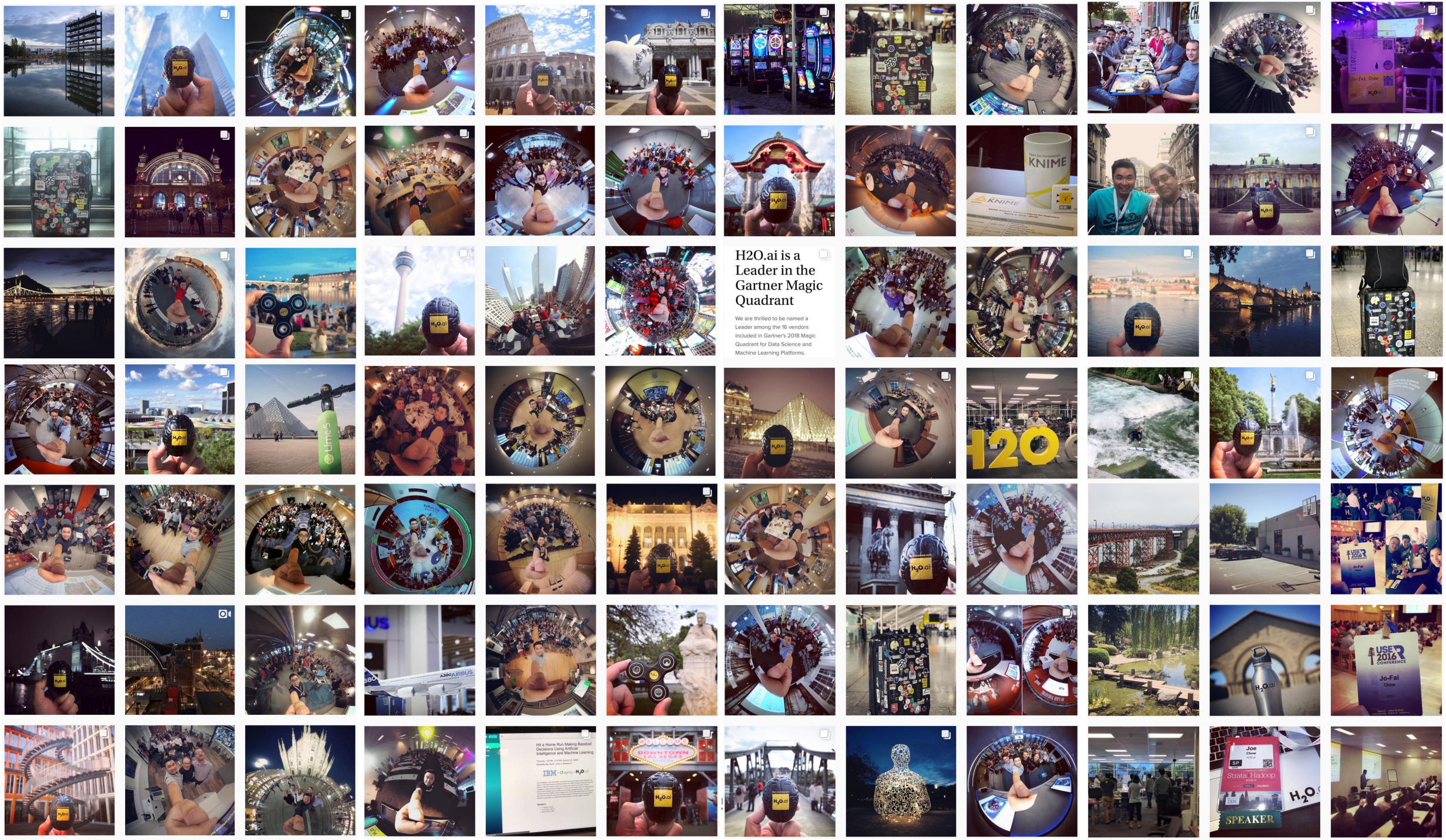
Jo-fai (Joe) Chow

Data Science Evangelist & Community Manager

joe@h2o.ai

- **Before H_2O**
  - Water Engineer / EngD Researcher / Matlab Fan Boy
    (wonder why @matlabulous?)
  - Discovered R, Python, H_2O … never look back again
  - Data Scientist at Virgin Media (UK), Domino Data Lab (US)
- **At H_2O …**
  - 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.
KNIME
H2O
USER 2016 CONFERENCE
Jo-Fai Chow
Strata Hadoop
SPEAKER


H2O.ai
#360Selfie Time!

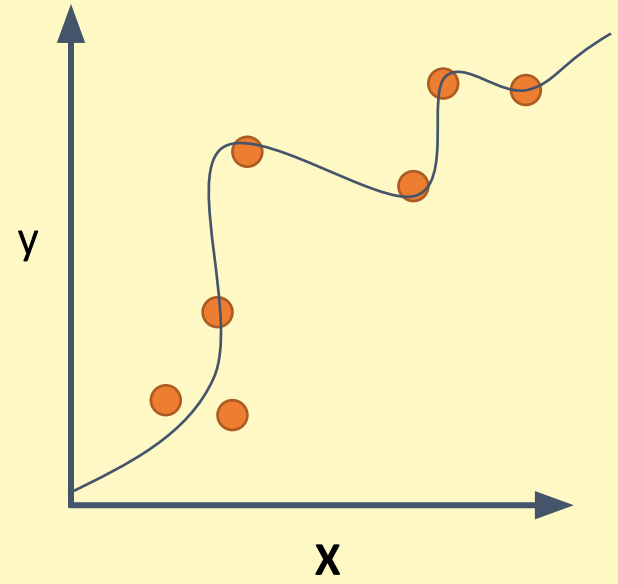
Driverless AI Delivers "Expert Data Scientist in a Box"

- 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]

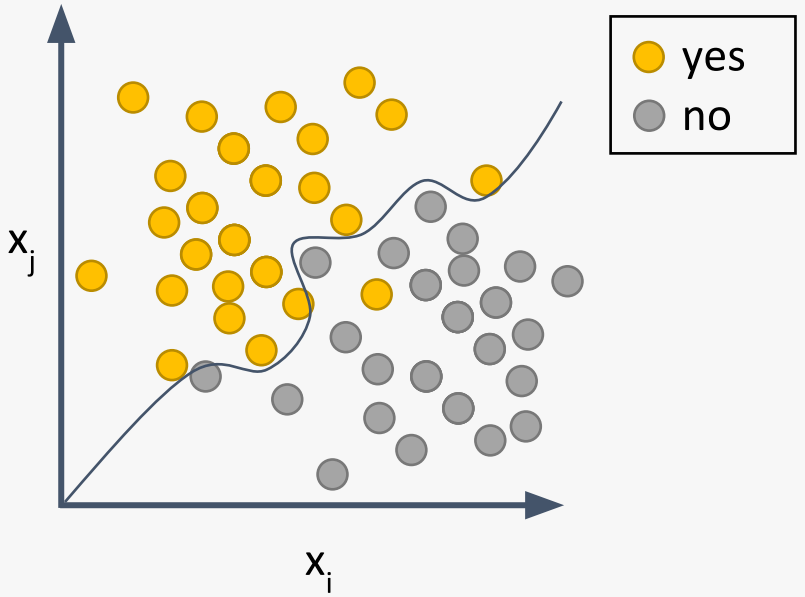
Supervised Learning

Regression:
How much will a customer spend?



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

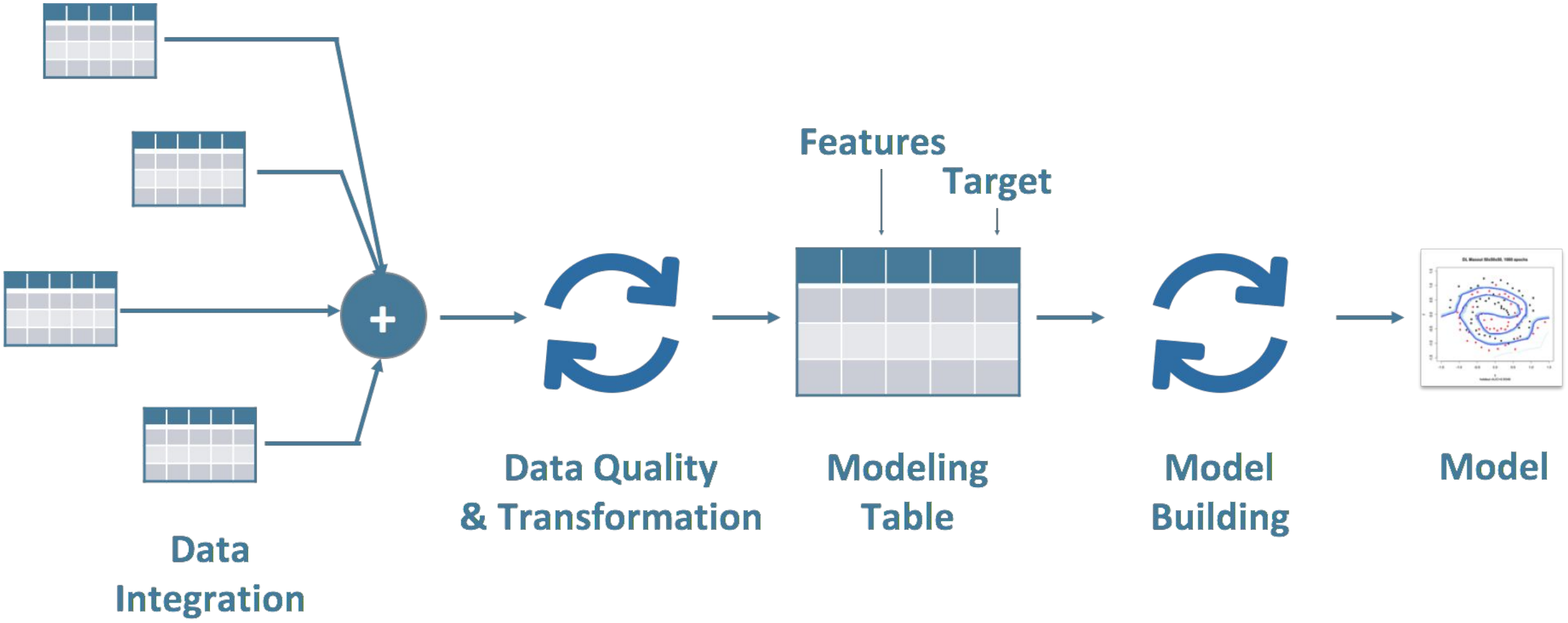
Classification:
Will a customer churn?

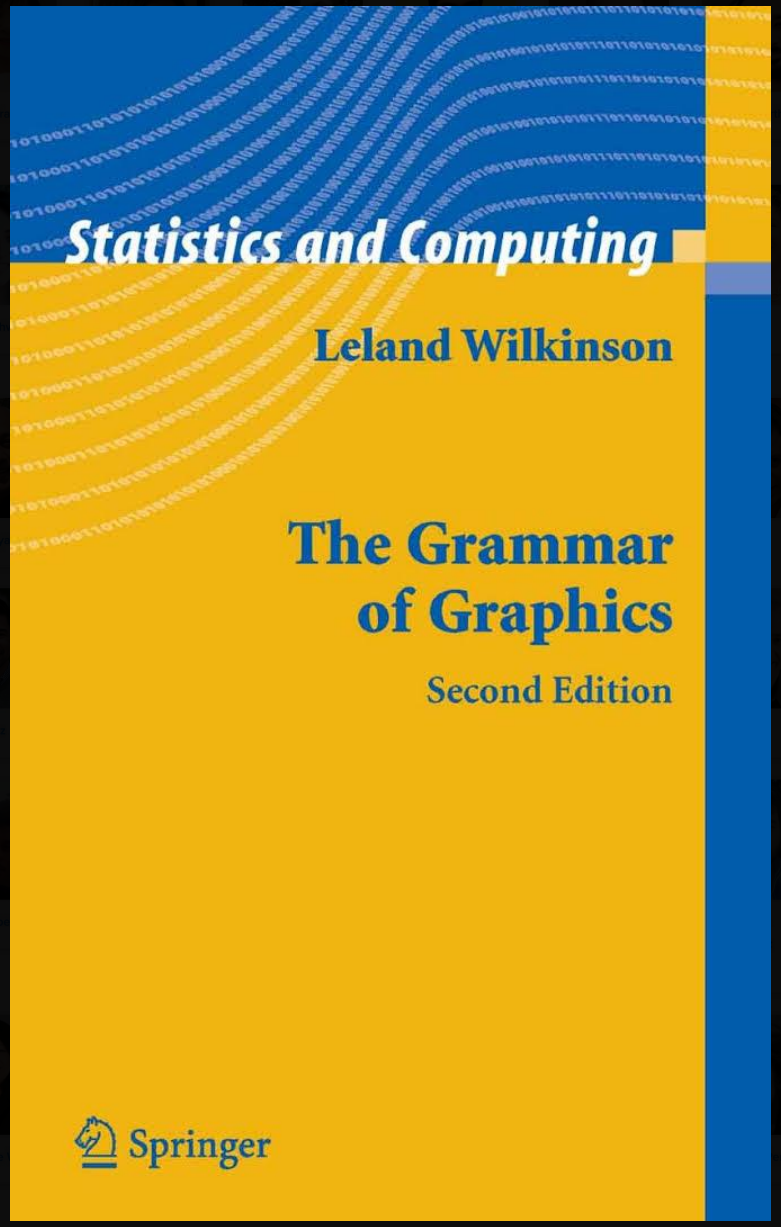


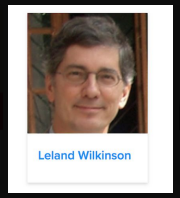
H_2O algos:
Penalized Linear Models
Naïve Bayes
Random Forest
Gradient Boosting
Neural Networks
Stacked Ensembles

H_2O.ai

Driverless AI: Automates Data Science and ML Workflows



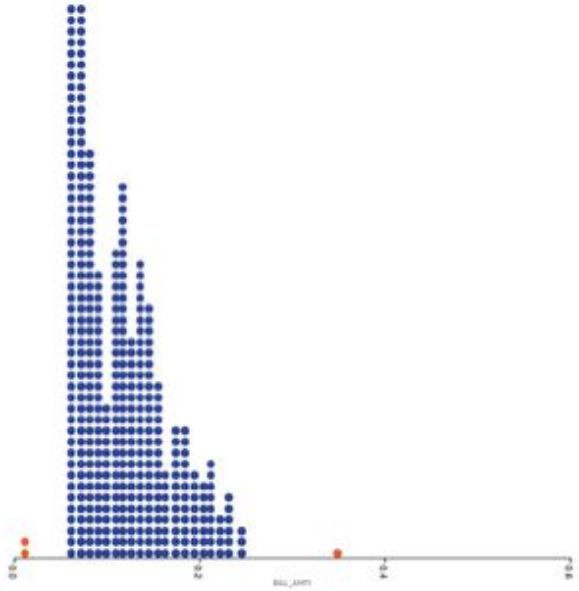


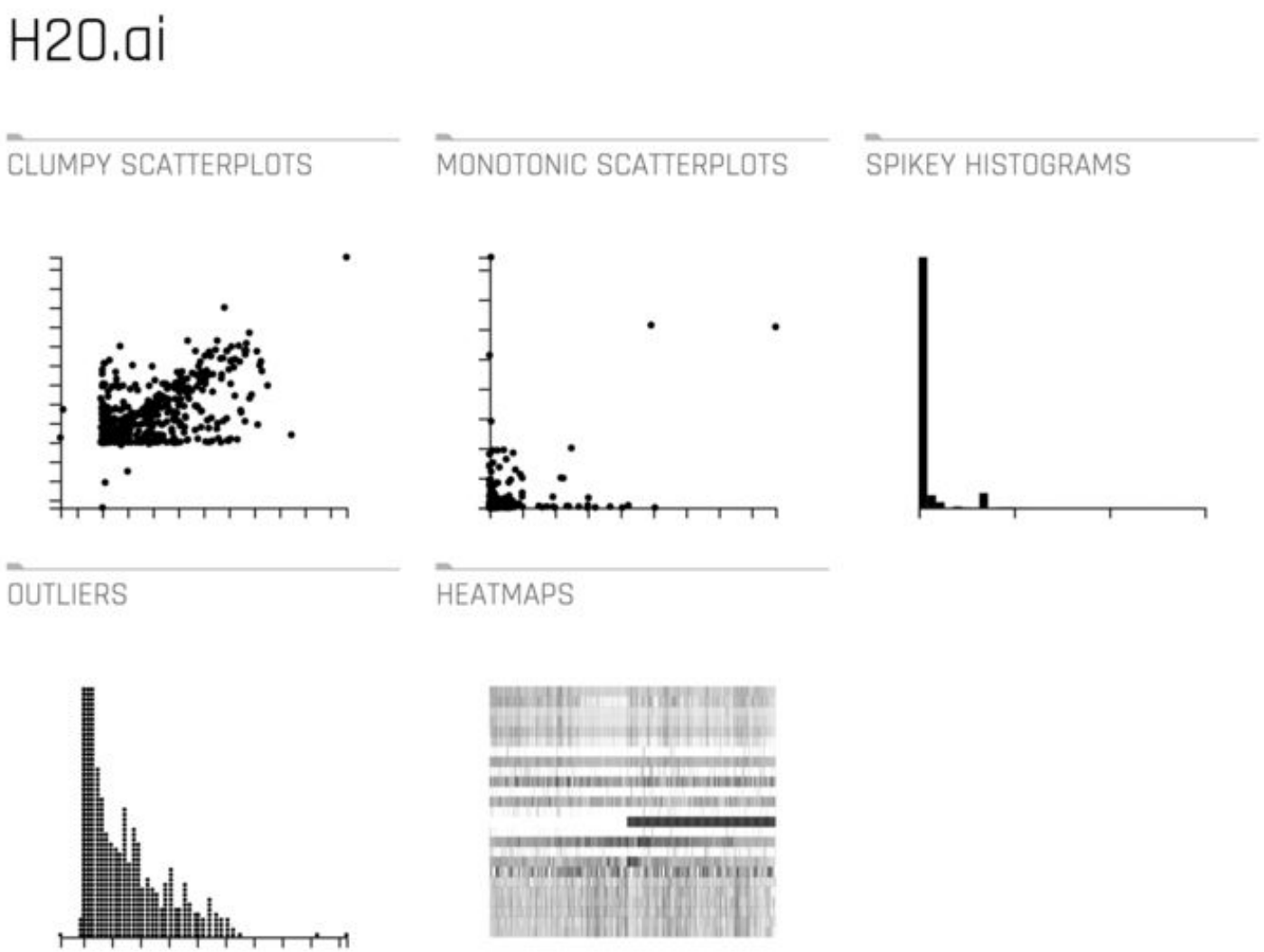


Origin of R Package `ggplot2`

Automatic Visualization

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



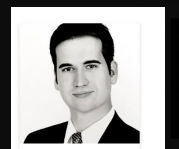


48th

Branden Murray

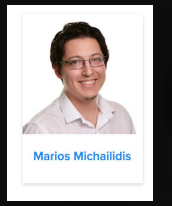
Dmitry Larko

25th



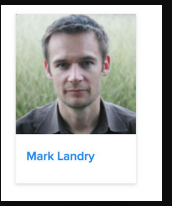
Mathias Müller

4th



Marios Michailidis

1st



Mark Landry

33rd

Kaggle Grand Masters (and their Highest Rank)

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

About 80,000 Kagglers

13th

H$_2$O Team

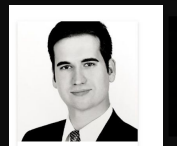
H$_2$O.ai

48th

Branden Murray

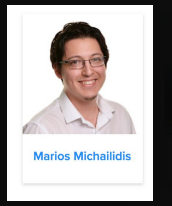
Dmitry Larko

25th



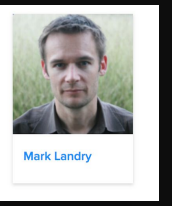
Mathias Müller

4th



Marios Michailidis

1st



Mark Landry

33rd

Jo-Fai Chow

181st

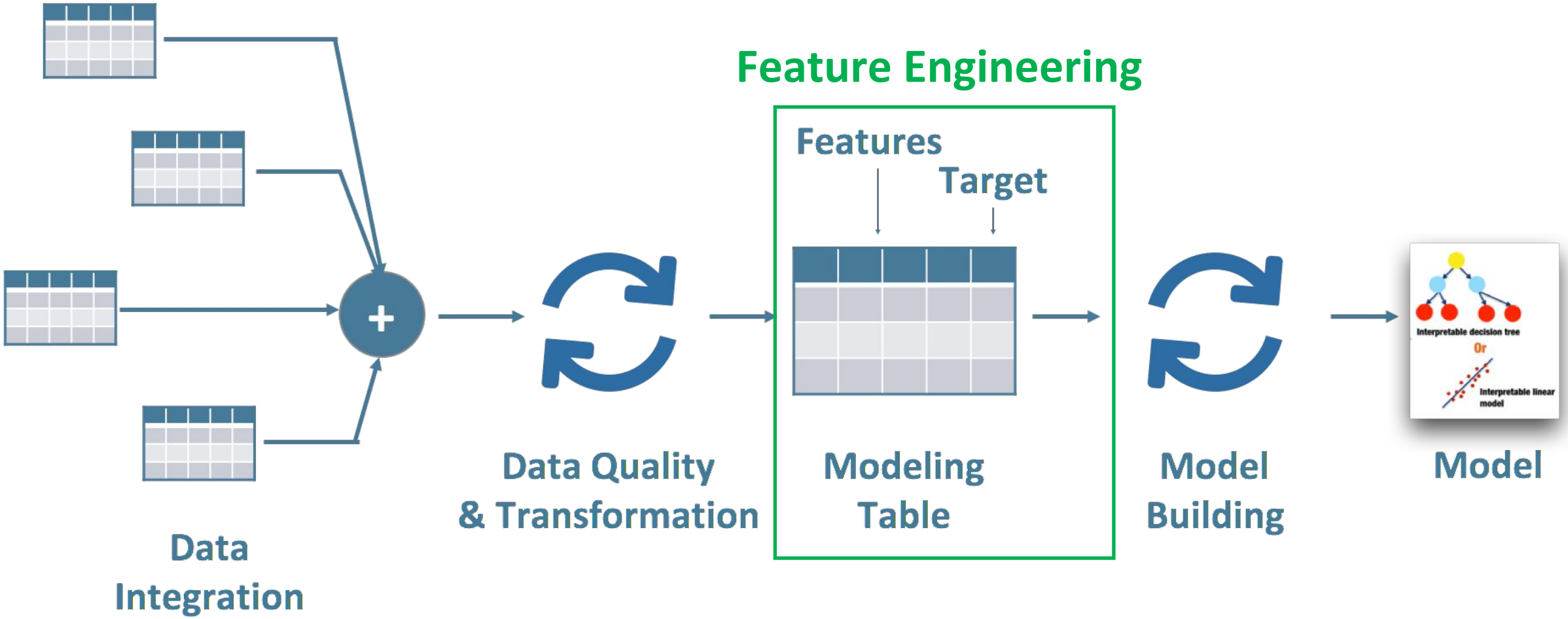
Hoping to get closer to them at some point …

13th

H2O Team

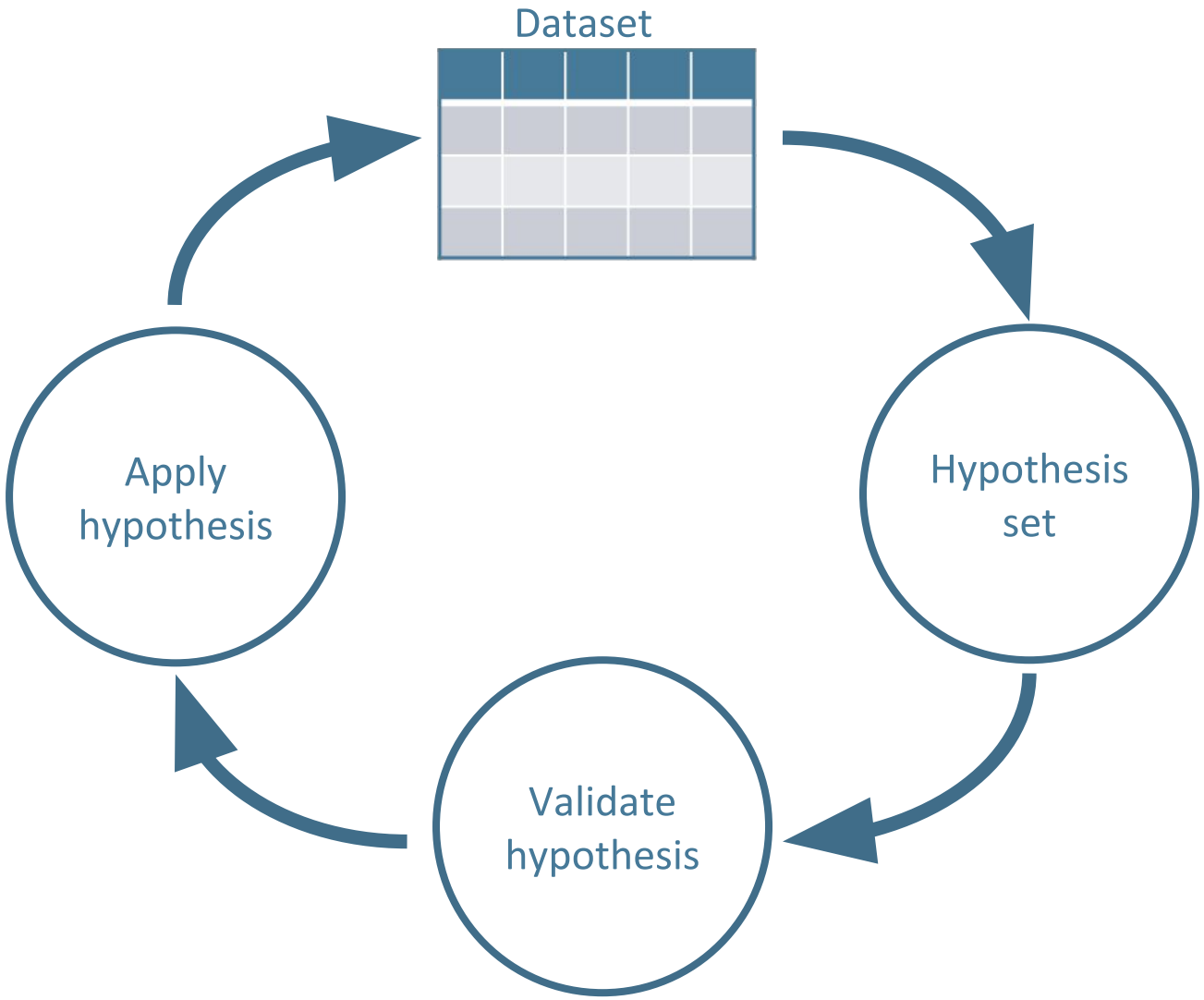
H2O.ai

Typical Enterprise Machine Learning Workflow



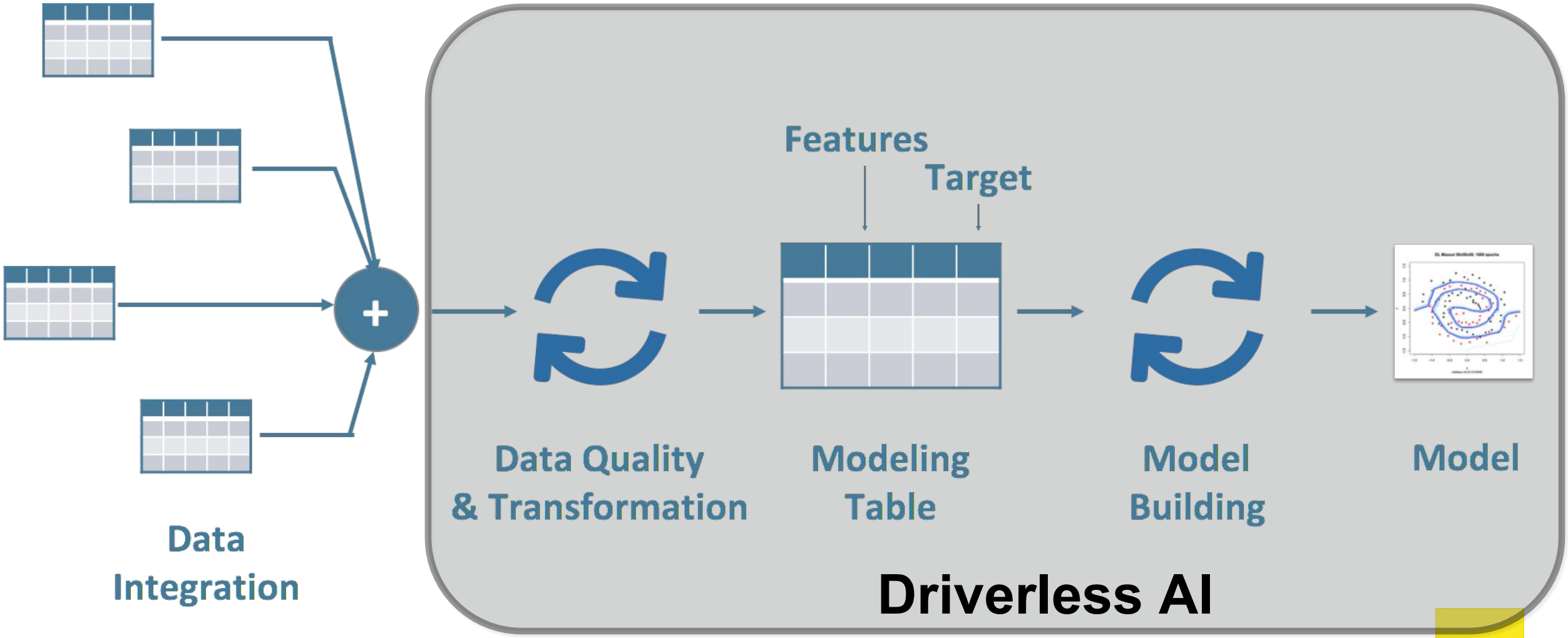
H2O.ai

Feature Engineering cycle



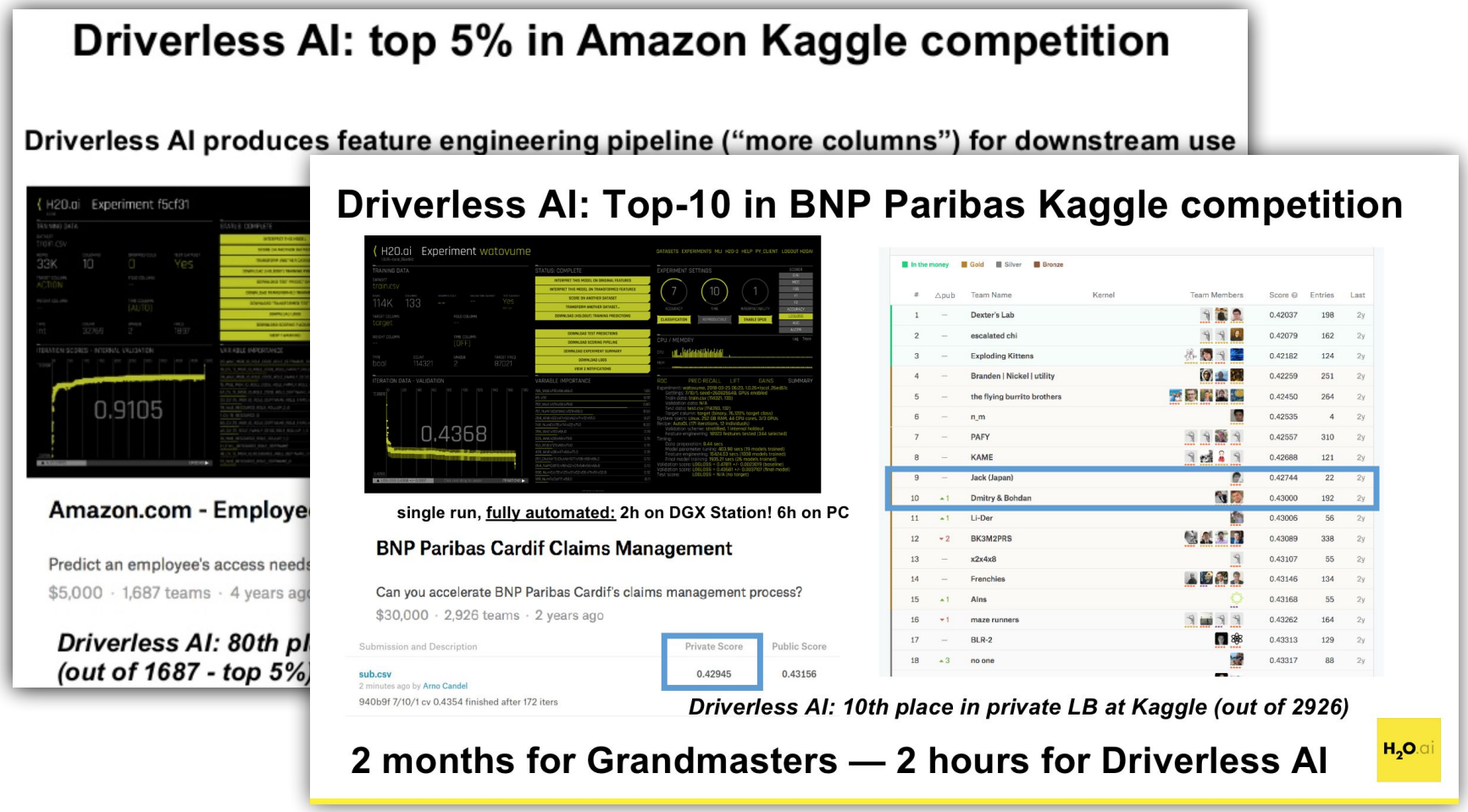
H2O.ai

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

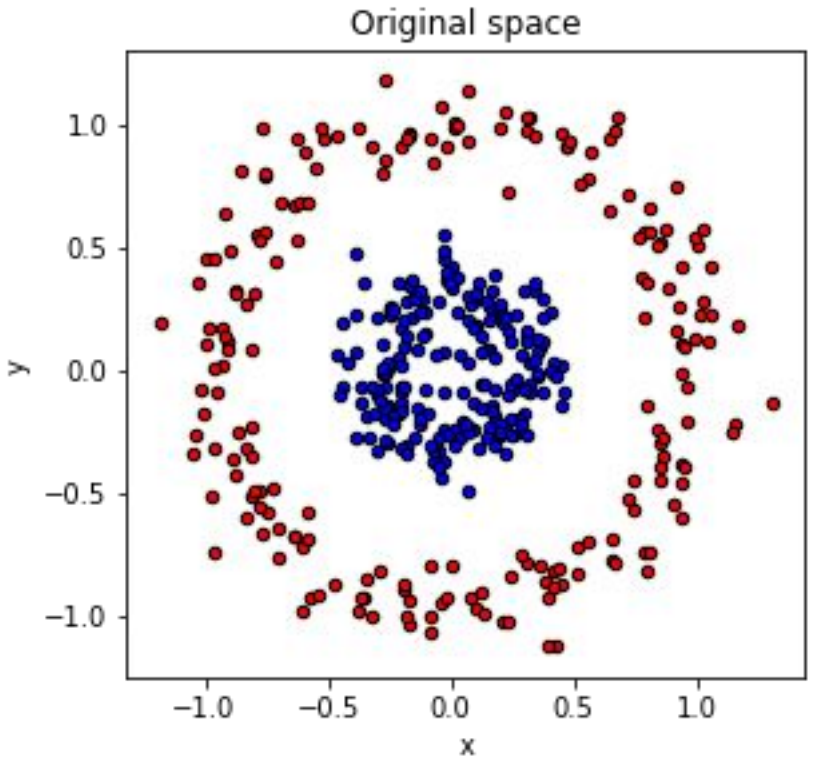


Interpretability

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

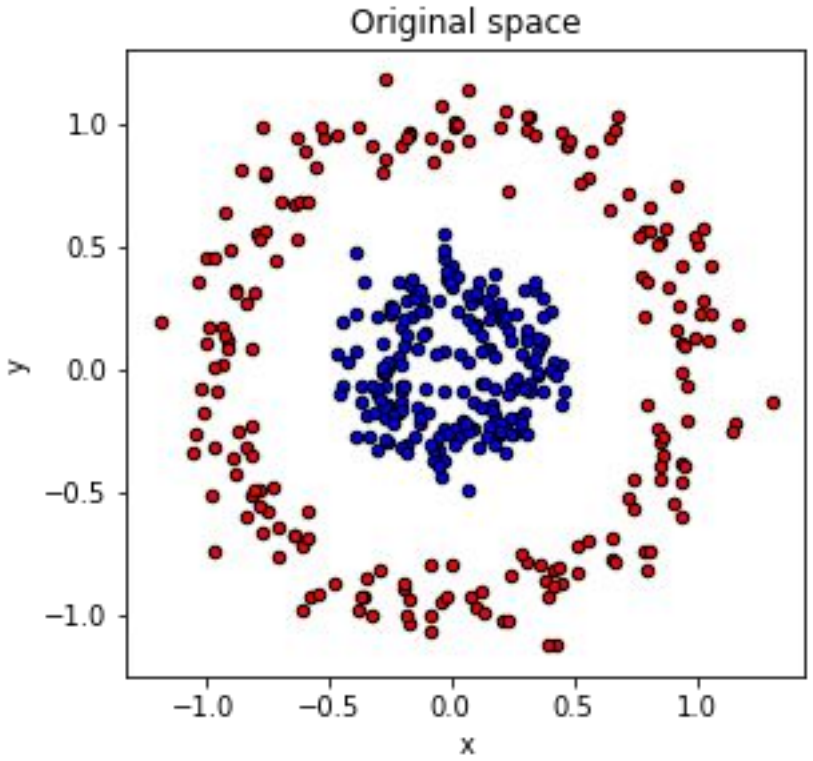
What is Feature Engineering?

What is feature engineering

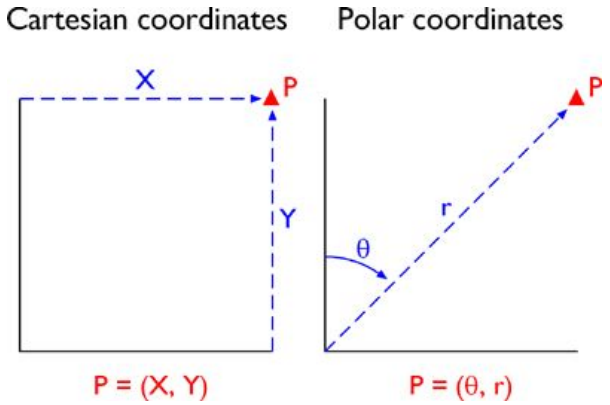


Not possible to separate using linear classifier

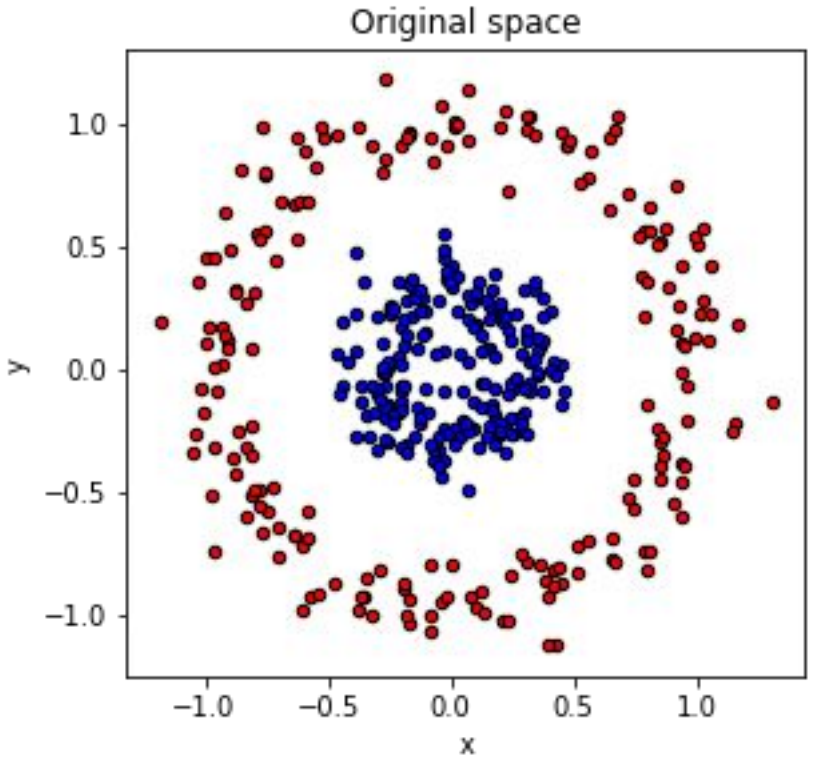
What is feature engineering

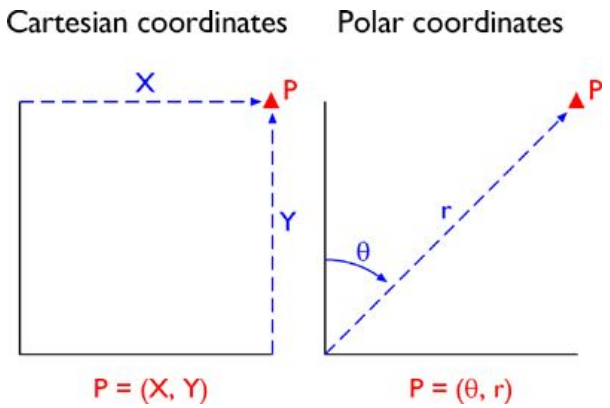


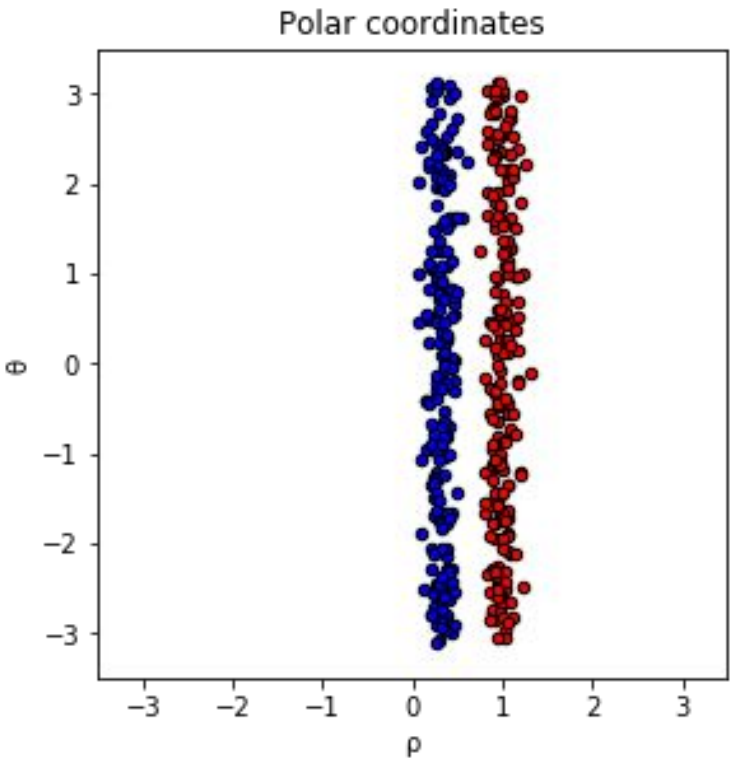
What if we use polar coordinates instead?



What is feature engineering



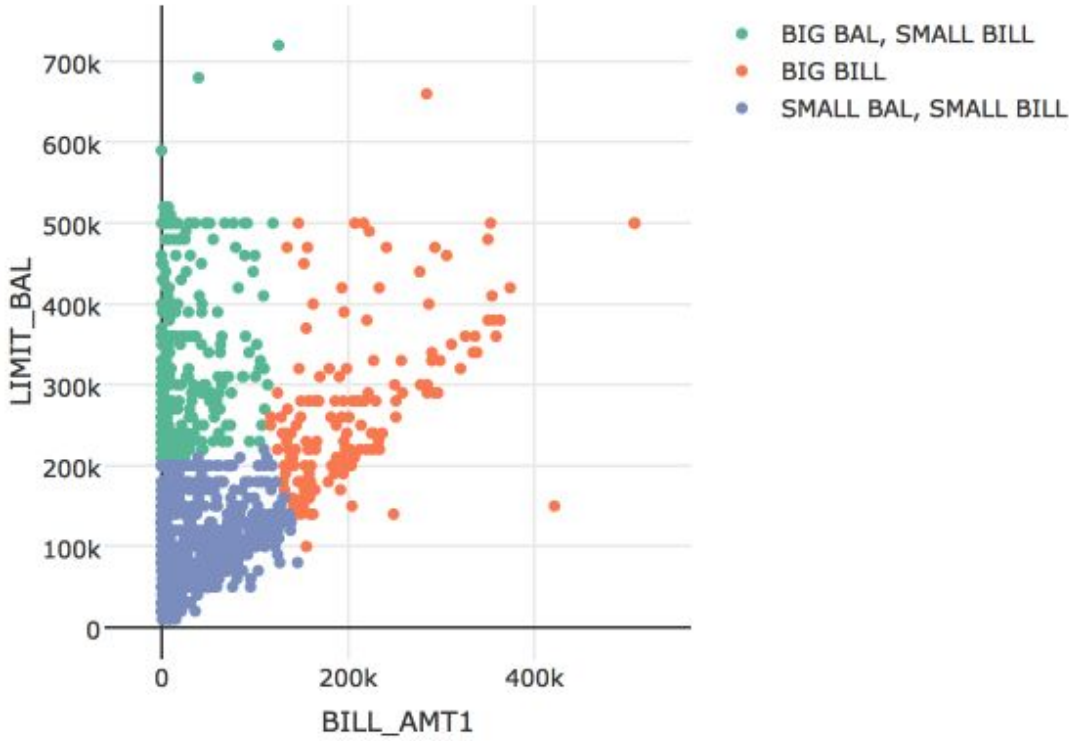




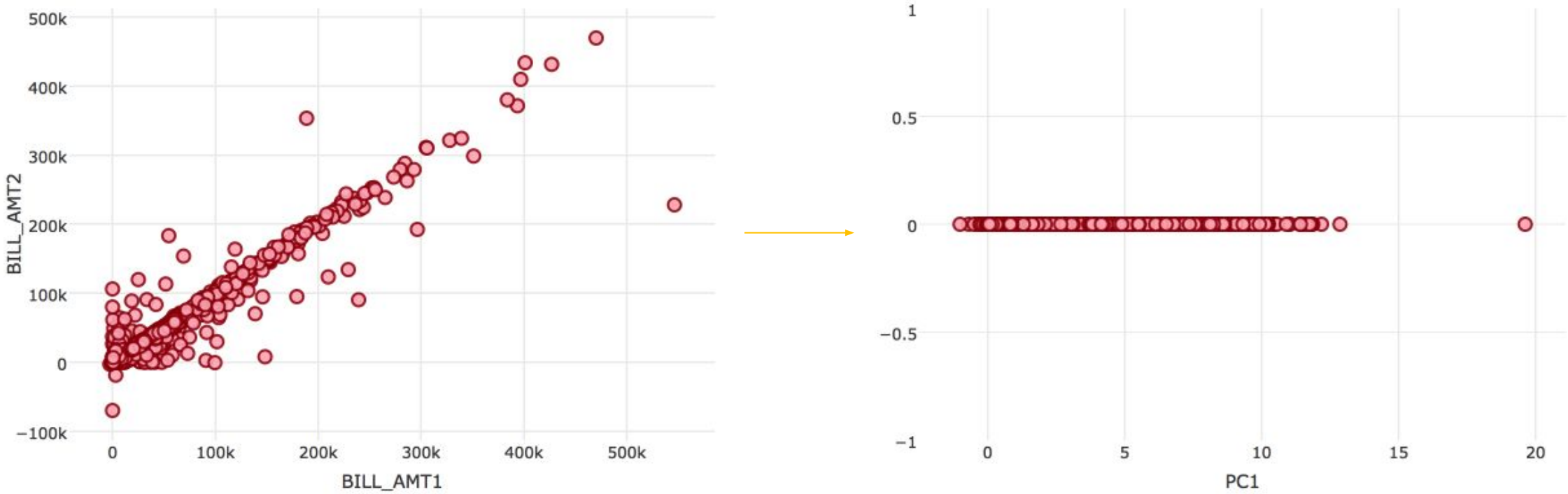
Clustering

Cluster Transformations

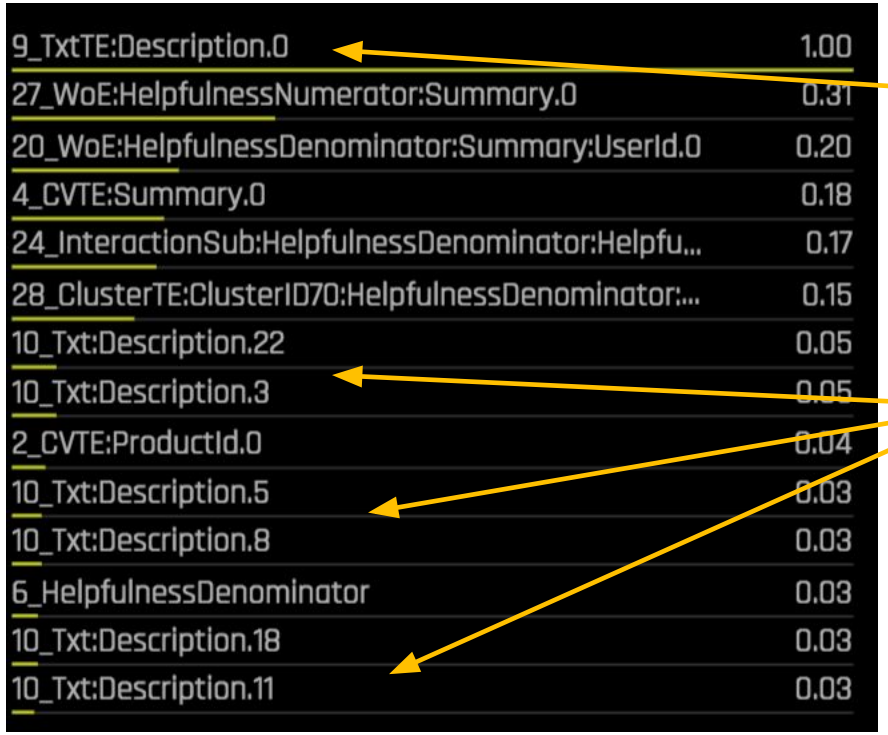
- Distance to a specific cluster
- Cross Validation Target Encoding by Cluster ID



Truncated SVD



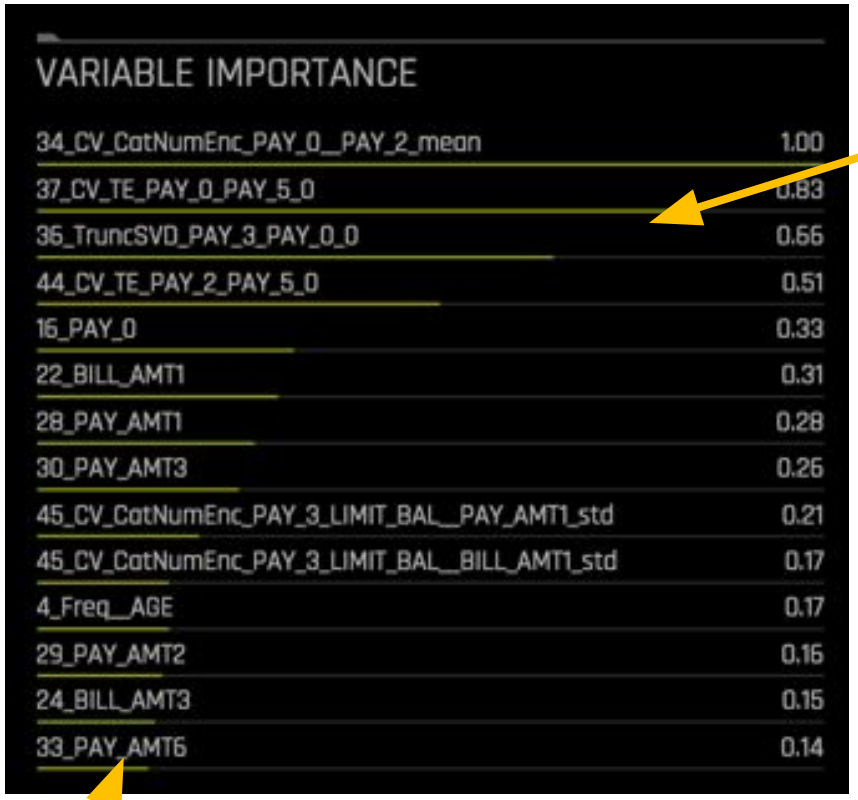
Text Features



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



Generated Features

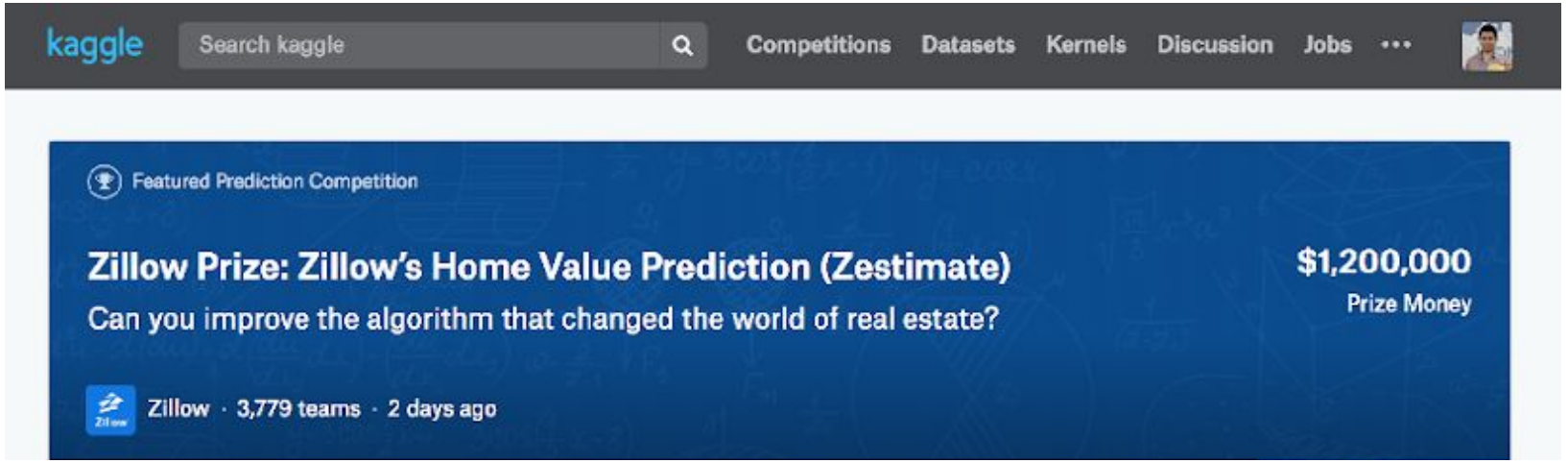
Original Features

Feature Transformations

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

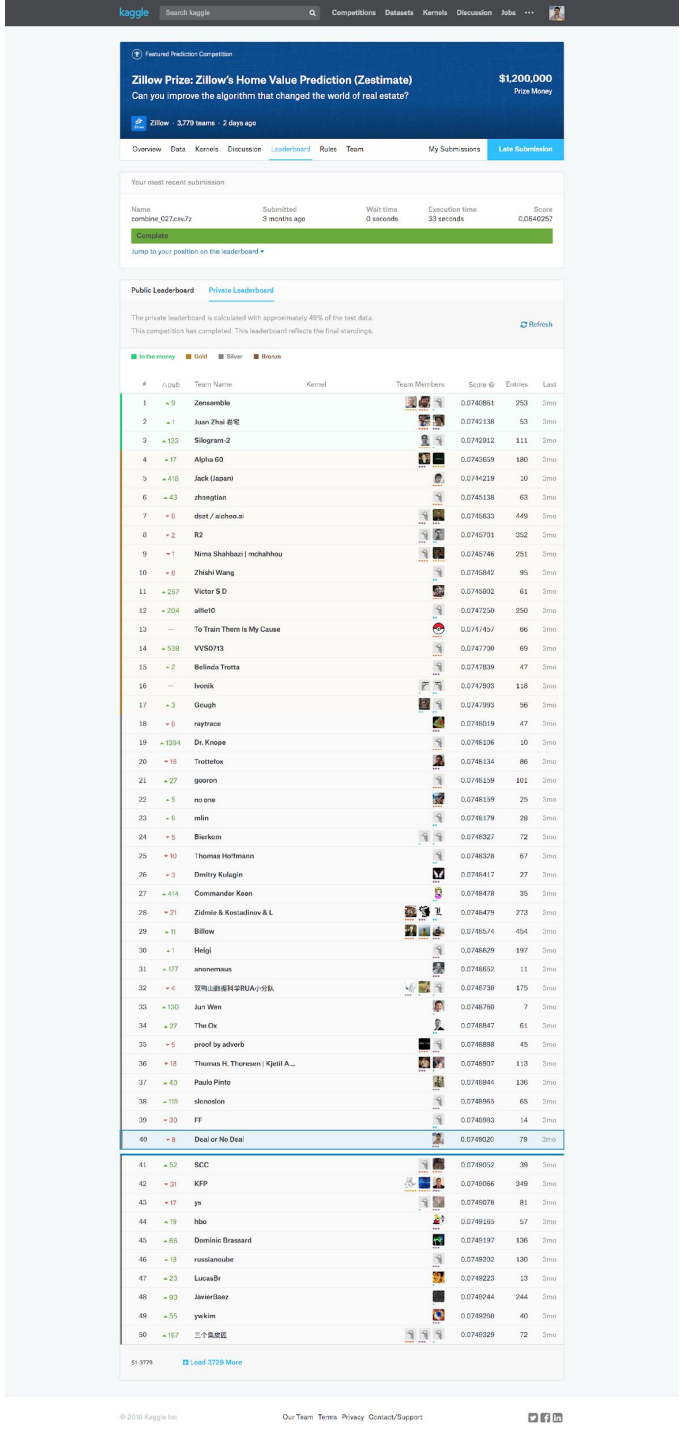
How about Auto Feature Engineering + Marios' StackNet?

Driverless AI + StackNet (9 months ago)



Competition Round One (Top 100 to Next Round)

Finished above my H2O Kaggle Grandmasters colleagues



Live Demo

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

Evaluation

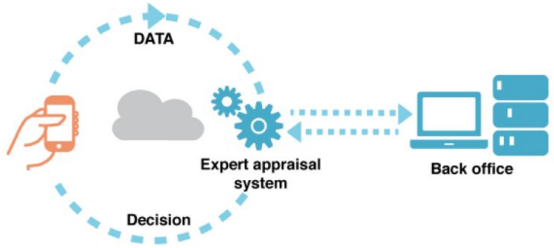
Prizes

Timeline

About Bnp Paribas Cardif

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

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.



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**

- More Info, Code, and Slides
  - bit.ly/ h2o_meetups
- Contact
  - joe@h2o.ai
  - @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

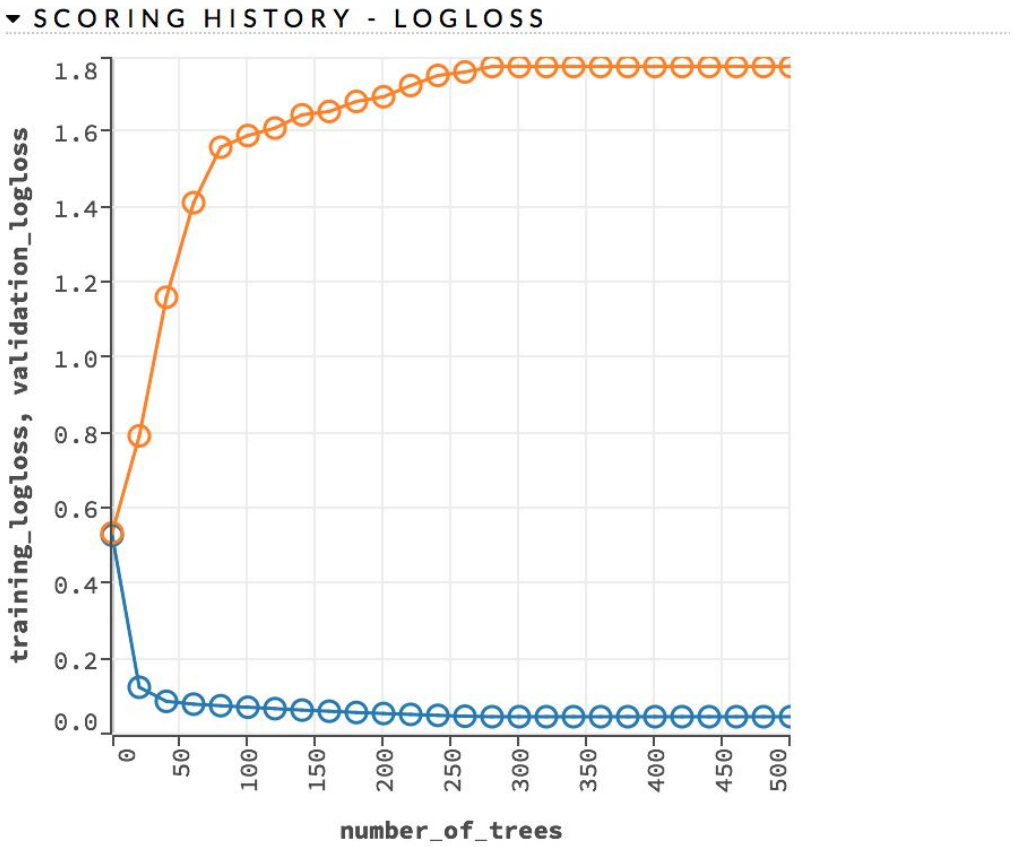
Target Mean Encoding

- 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

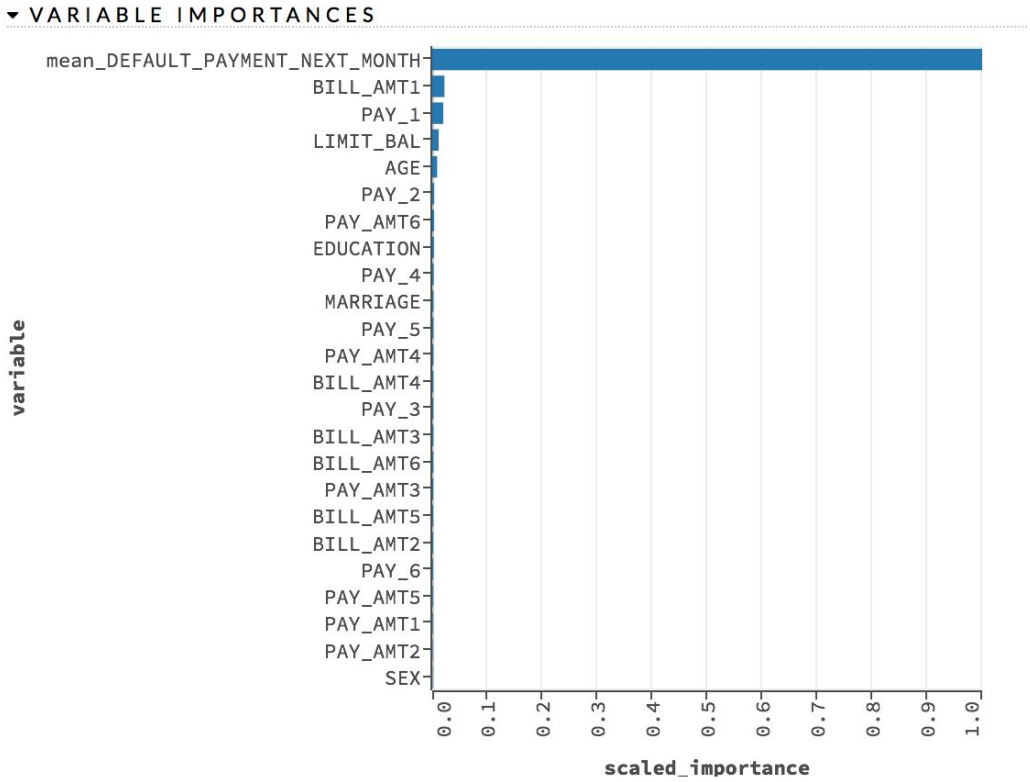
Pay 1	Default Payment	Mean Target Encoding
Missed 5 Mo	1	1

Worst Case Scenario: Response Column = Mean Target Encoding

Effects of Data Leakage



Scoring History: Training vs Testing



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