Explainable AI with H2O Driverless AI's machine learning interpretability module

Michelle Tanco
Data Scientist & Solution Engineer, H2O.ai
Michelle.Tanco@h2o.ai

H2O.ai Seattle Meetup #2 08/20/2019

ABOUT ME

Add profile section ▾ More...

Michelle Gabriele Tanco

Customer Data Scientist & Solution Engineer at H2O.ai

Seattle, Washington · 500+ connections · Contact info

H2O.ai

Ursinus College

About

Michelle likes to solve technical problems and is a Customer Solutions Engineer & Data Scientist for H2O.ai. She focuses on how to apply machine learning to solve cross-industry business problems.

Her background is in pure math and computer science and she is passionate about applying these skills to answer real world questions.

When not coding or thinking of analytics, Michelle can be found hanging out with her dog or playing ukulele.

EXPLAINING AI DECISIONS

H2O.ai

- What is interpretation?
  - The ability to explain or to present in understandable terms toa human
- What is a good explanation?
  - When you can no longer keep asking why
- Why should you care?
  - Understanding of an impactful and quickly expanding set of technologies
  - Addressing accidental or intentional discrimination
  - Preventing malicious hacking and adversarial attacks
  - Enabling regulatory compliance and increased financial margins

Intro

Terminology, scope, and context

H2O.ai Overview

Company	Founded in Silicon Valley in 2012 Funded: $147M Investors: Goldman Sachs, Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
Products	• H2O Open Source Machine Learning (14,000 organizations) • H2O Driverless AI – Automatic Machine Learning
Team	160 AI experts (Expert data scientists, Kaggle Grandmasters, Distributed Computing, Visualization)
Global	Mountain View, London, Prague, India

H2O.ai Product Suite

Open Source

H2O	Spark + H2O SPARKLING WATER	H2O4GPU
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

DRIVERLESSAI

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
- User licenses on a per seat basis (annual subscription)

Terminology, Scope, and Context

H2O.ai

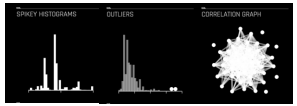
- **Machine Learning Interpretability**
  - *"[Machine learning interpretability] is the ability to explain or present in understandable terms to a human."* –https://arxiv.org/pdf/1702.08608.pdf
- **Structured data**
  - No ~~image~~, ~~video~~ and ~~sound~~ > deep learning typically not used
  - Tabular data and supervised ML
- **Auto ML**
  - H2O Driverless AI (DAI) product (not OSS)
- **MLI module**
  - Solution based on MLI module of H2O Driverless AI

H2O Driverless AI – How it Works

H2O.ai

1 Drag and Drop Data
Bring data in from cloud, big data and desktop systems

2 Automatic Visualization
Understand the data shape, outliers, missing values, etc.



3 Automatic Model Optimization
Use best practice model recipes and the power of high performance computing to iterate across **thousands of possible models** including advanced feature engineering and parameter tuning

4 Automatic Scoring Pipelines
Deploy ultra-low latency Python or Java Automatic Scoring Pipelines that include feature transformations and models

- HDFS
- SQL
- Local
- Amazon S3
- Snowflake
- Google BigQuery
- Azure Blog Storage

Modelling

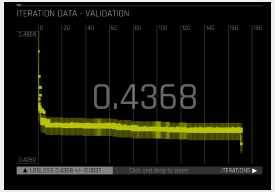
X | Y

Dataset

Automatic Model Optimization

Model Recipes
- i.i.d. data
- Time-series
- More on the way

Survival of the Fittest



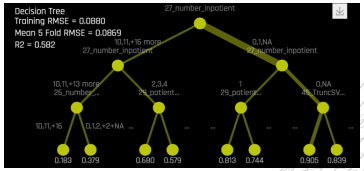
Advanced Feature Engineering + **Algorithm** + **Model Tuning**

Model Documentation

Automatic Scoring Pipeline

Deploy Low-latency Scoring to Production

Machine learning Interpretability



H2O.ai

Why explainability matters

Problem statement

Potential Performance and Interpretability **Trade-off**

(Trade-off)

White box model

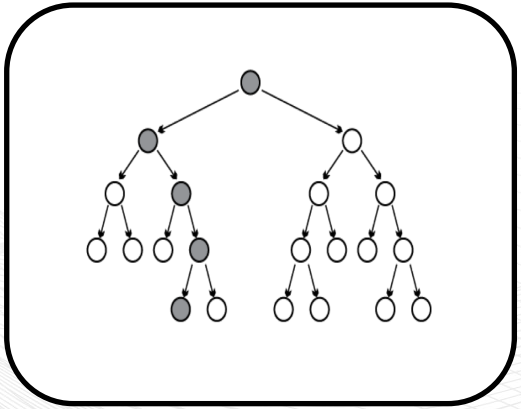
Black box model

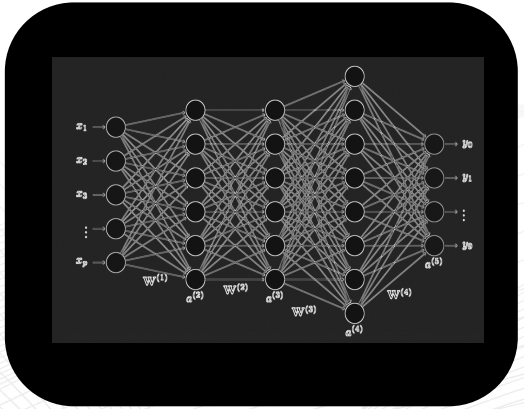
Feature engineering + Algorithm(s)

Balance

Potential Performance and Interpretability **Trade-off**

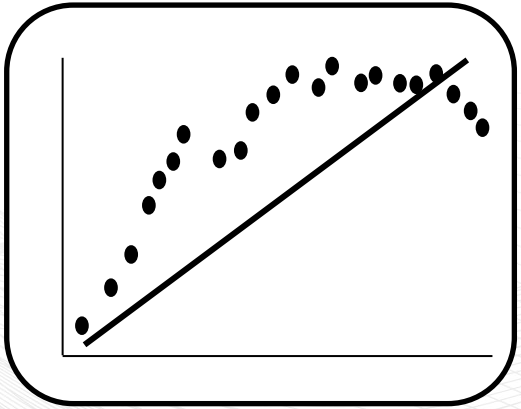
(Trade-off)

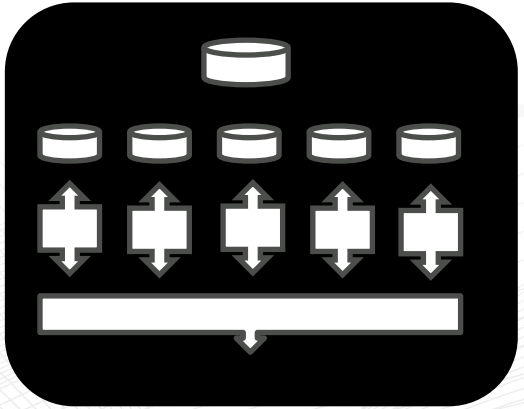




Potential Performance and Interpretability **Trade-off**

(Trade-off)





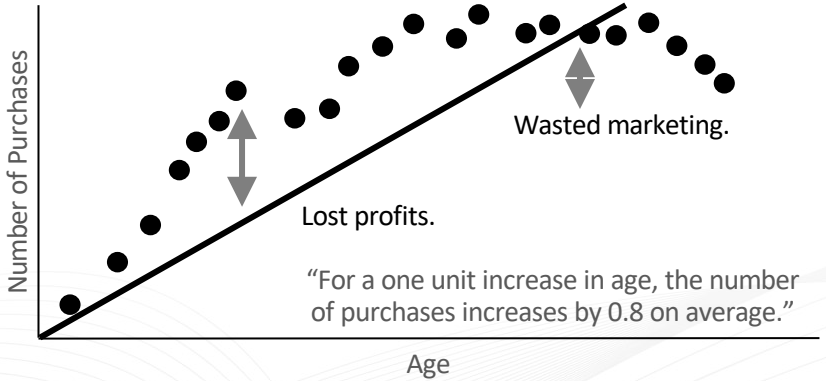
(Trade-off)

Potential Performance and Interpretability **Trade-off**

H2O.ai

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

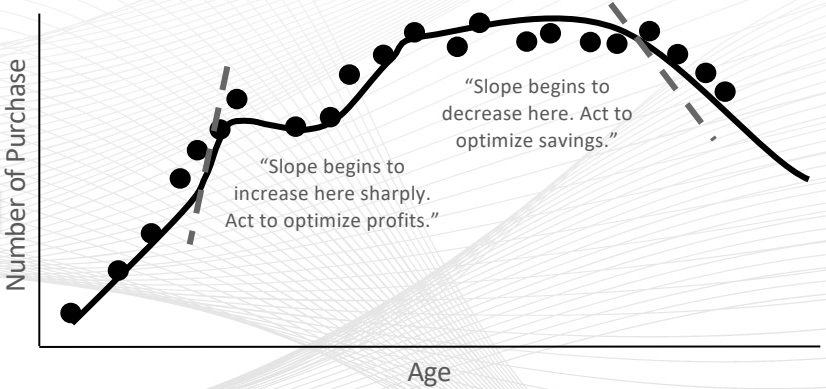
Linear models



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

Sometimes…

Machine learning models

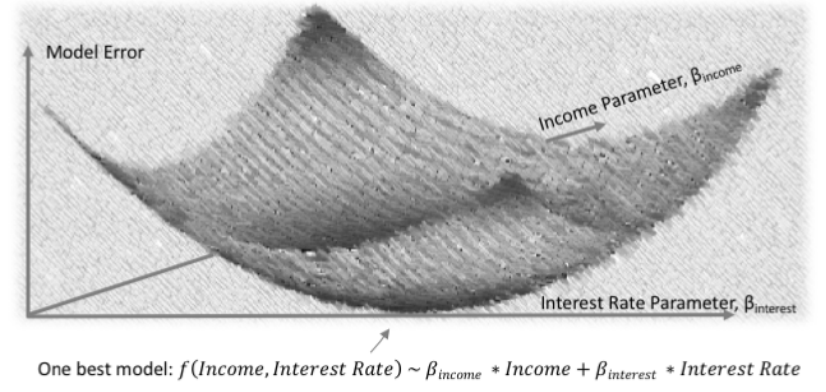


Trade-off

Multiplicity

Multiplicity of Good Models

- For a given well-understood dataset there is usually **one** best linear model, but…



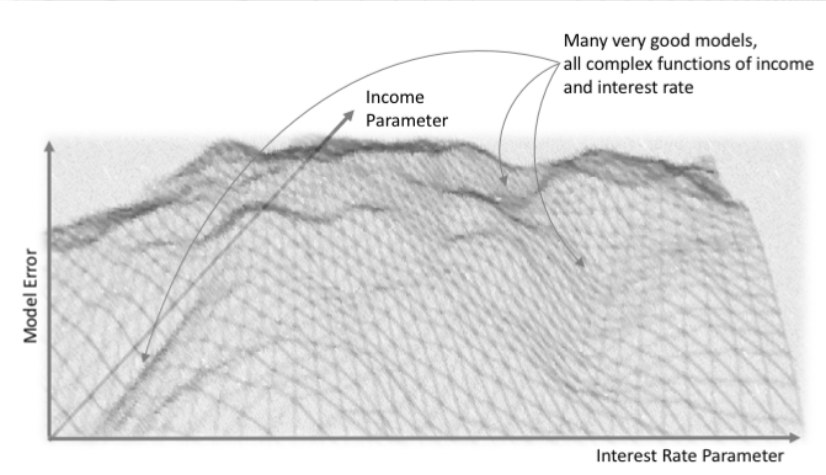
Trade-off

Multiplicity

Multiplicity of Good Models

H$_2$O.ai

- … for a given well-understood dataset there are usually **many good** ML models. Which one to **choose**?
- Same **objective metrics** values, **performance**, …
- This is often referred to as “the **multiplicity** of good models.” -- Leo Breiman



Trade-off

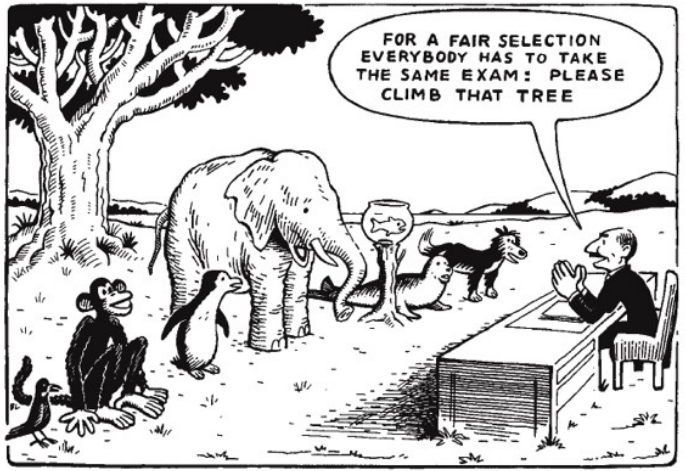
Multiplicity

Fairness

Fairness and Social Aspects

H_2O.ai

- Gender
- Age
- Ethnicity
- Health
- Sexual behavior



- Avoid **discriminatory models** and remediate disparate impact.

Trade-off

Multiplicity

Fairness

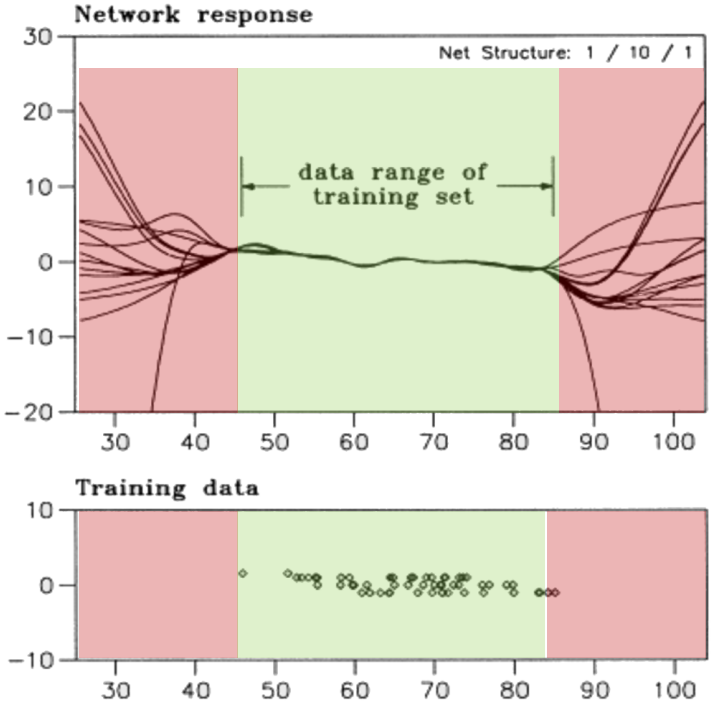
Trust

Trust of model producers & consumers

H2O.ai

- Dataset
  vs.
  real world
- ML **adoption**

- Introspection
- Sensitivity
- Out-of-range
- Diagnostics
- Debugging



Source: http://www.vias.org/tmdatanaleng/

Trade-off

Multiplicity

Fairness

Trust

Security

Security and **Hacking**

H_2O.ai

- Goal: **compromise** model integrity
- Attack types:
  - **Exploratory**
    - Surrogate model trained to identify vulnerabilities ~ MLI
    - Trial and error to learn how models work
  - **Causative**
    - Models trained with poisoned data
    - **Integrity** (compromise system integrity)
      - False negative instance: fraud passes check
    - **Availability** (compromise system availability)
      - False positive instance: blocks access to legitimate instances

Trade-off

Multiplicity

Fairness

Trust

Security

Regulation

Regulated & Controlled Environments

H2O.ai

- Legal requirements
  - Banking, insurance, healthcare, …
- Predictions explanation
  - Decisions justification (reason codes, …)
- Fairness
- Security
- Accuracy first vs. **interpretability** first
  - Competitions vs. real world

Trade-off

Multiplicity

Fairness

Trust

Security

Regulation

Explainability Matters

H2O.ai

- **Balance** performance and interpretability
- **Multiplicity** of good models
- **Fairness** and **social** aspects
- **Trust** of model producers and consumers
- **Security** and **hacking**
- **Regulated/controlled** environments

TAKEAWAYS

- ML interpretability **matters**
- **Multiplicity** of good models
- H2O Driverless AI has **interpretability**
- **Control** model interpretability **end to end**
- Prefer **interpretable models**
- **Test** both your model and explanatory SW
- Use synergy of **local** & **global** techniques
- **Shapley** values

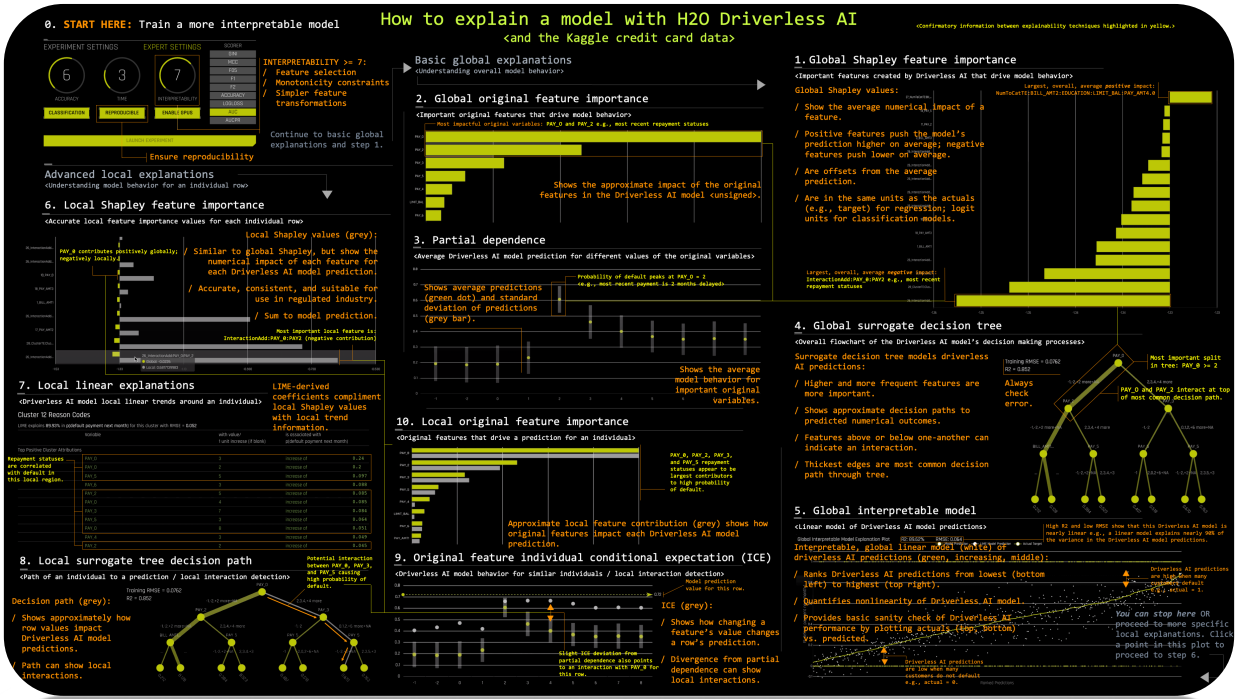
Resources

Books, articles, links and Git repos

Resources

- https://www.h2oai.com/explainable-ai/
- Booklets:
  - Machine Learning Interpretability with DAI
  - Ideas on Interpreting Machine Learning
- Driverless AI's MLI module cheatsheet
- MLI presentations:
  - MLI walkthrough by Patrick Hall
  - Human Friendly Machine Learning by Patrick Hall
- GitHub repositories:
  - MLI Resources
  - H2O Meetups

MLI Cheatsheet



https://github.com/h2oai/mli-resources/blob/master/cheatsheet.png

5 Steps to get started with H2O Driverless AI

- Join the community slack for Questions/Answers, tips and news
- Get a 21-day Driverless AI trial license:
  - Request a 21-day license
  - Cloud Test Drive – two hour session with no install required
- Install Driverless AI
  - To a local machine or server
  - In any major cloud
- Do the tutorials – guided introduction with sample data
- Learn more from:
  - H2O World SF session replays
  - H2O Driverless AI docs