

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



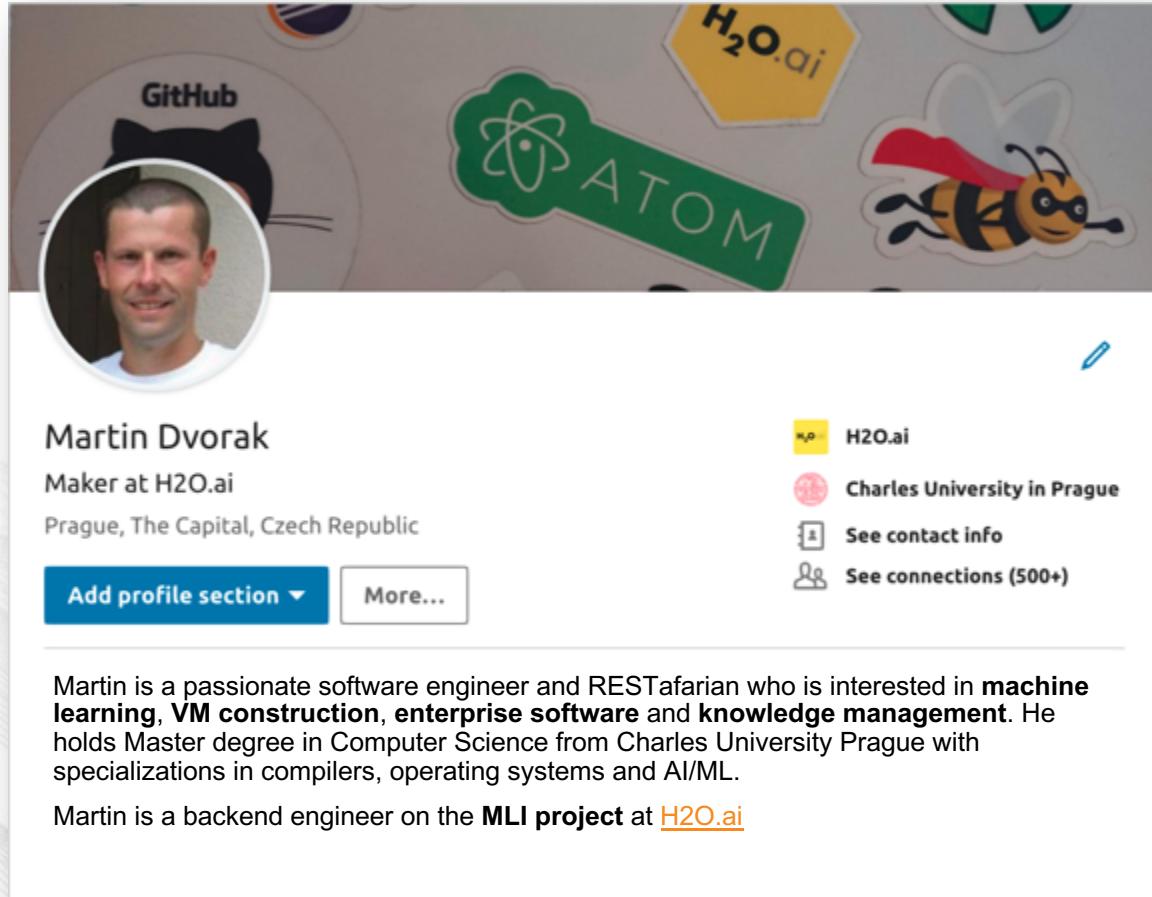
- Explainable AI is in the news, and for good reason. Financial services companies have cited the ability to explain AI-based decisions as one of the critical roadblocks to further adoption of AI for their industry. Transparency, accountability, and trustworthiness of data-driven decision support systems based on AI and machine learning are serious regulatory mandates in banking, insurance, healthcare, and other industries. From pertinent regulations, to increasing customer trust, data scientists and business decision makers must show AI-based decisions can be explained.
- H2O Driverless AI does explainable AI today with its machine learning interpretability (MLI) module. This capability in H2O Driverless AI employs a unique combination of techniques and methodologies to explain the results of both Driverless AI models and external models.

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



Martin Dvorak
Software Engineer, H2O.ai
martin.dvorak@h2o.ai

ABOUT ME



Martin Dvorak

Maker at H2O.ai

Prague, The Capital, Czech Republic

Add profile section ▾ More...

Martin is a passionate software engineer and RESTafarian who is interested in **machine learning**, **VM construction**, **enterprise software** and **knowledge management**. He holds Master degree in Computer Science from Charles University Prague with specializations in compilers, operating systems and AI/ML.

Martin is a backend engineer on the **MLI project** at [H2O.ai](#)

 H2O.ai

 Charles University in Prague

 See contact info

 See connections (500+)

AGENDA

- **Intro**
 - Context and scope.
- **Why**
 - Explainability matters.
- **What**
 - Steps to build human-centered, low-risk models.
- **How**
 - Explaining models using of H2O.ai's solution.

Intro

Terminology, scope and context

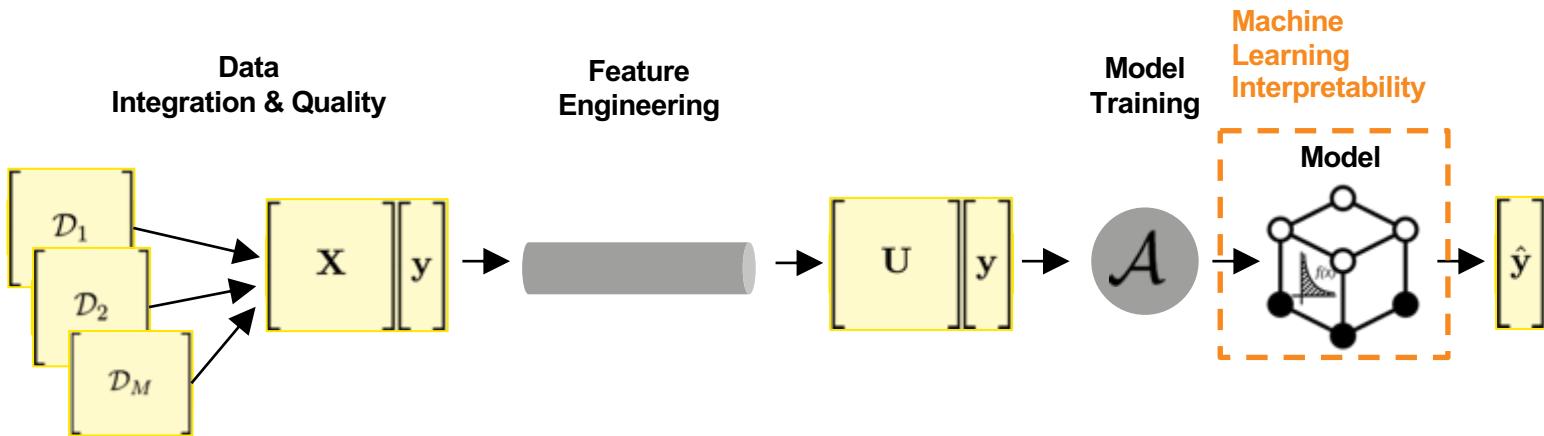
Terminology, Scope and Context

INTRO

- **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.

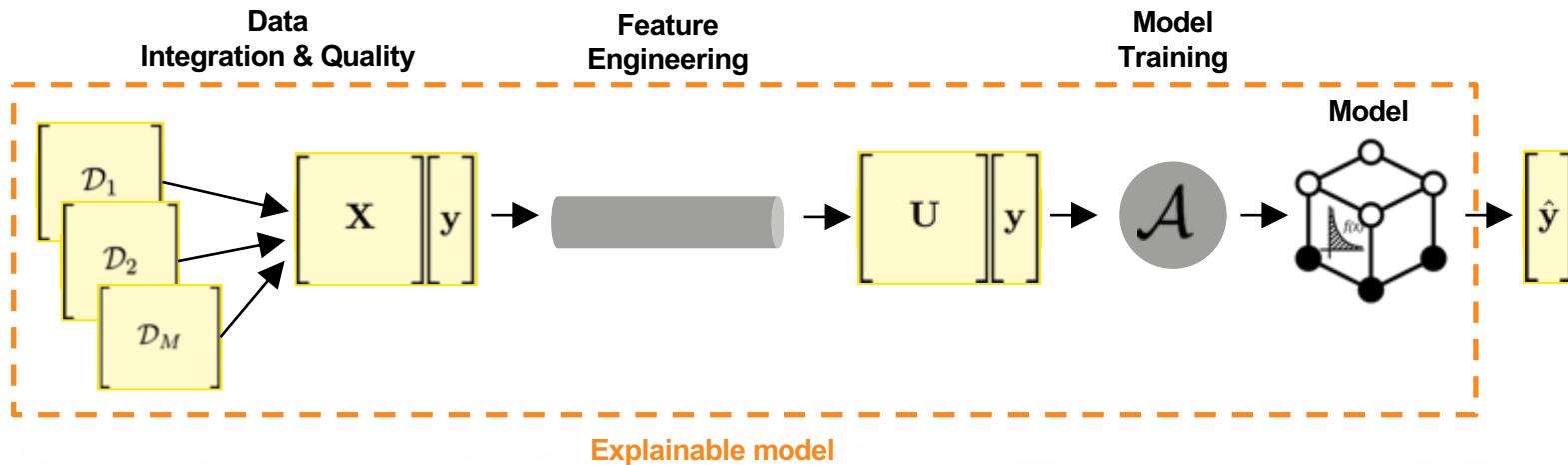
Terminology, Scope and Context

INTRO



Terminology, Scope and Context

INTRO



Why explainability matters

Problem statement

Potential Performance and Interpretability **Trade-off**

(Trade-off)

White box model

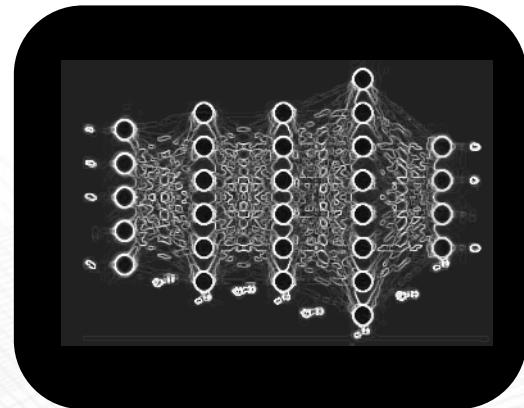
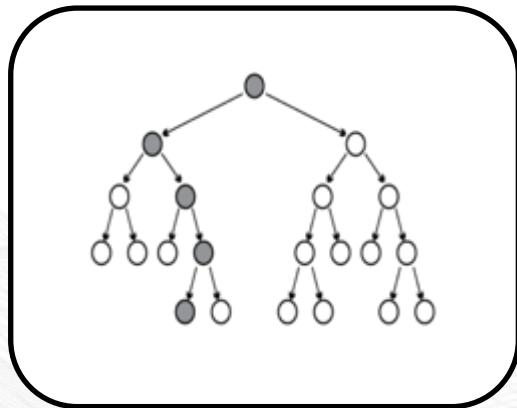
Black box model

Feature engineering + Algorithm(s)

Balance

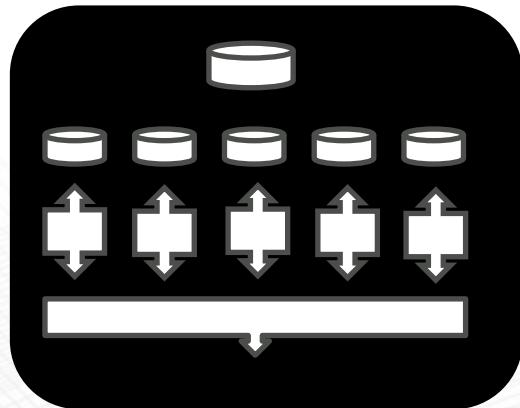
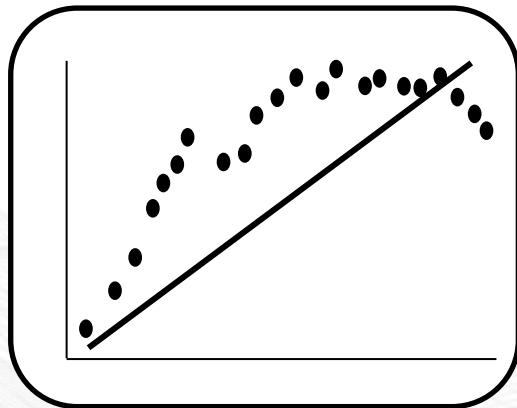
Potential Performance and Interpretability **Trade-off**

(Trade-off)



(Trade-off)

Potential Performance and Interpretability **Trade-off**



Potential Performance and Interpretability Trade-off

(Trade-off)

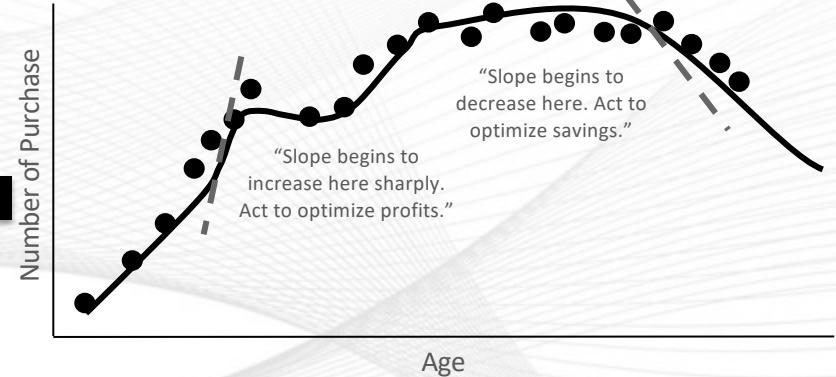
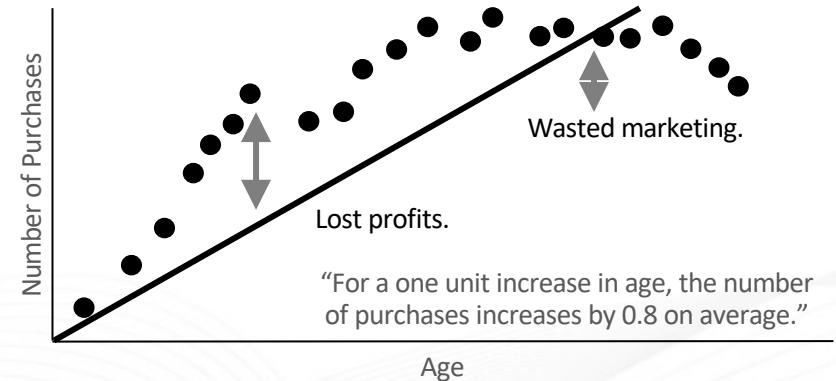
Exact explanations for
approximate models.

Linear models

Approximate explanations for
exact models.

Sometimes...

Machine learning models

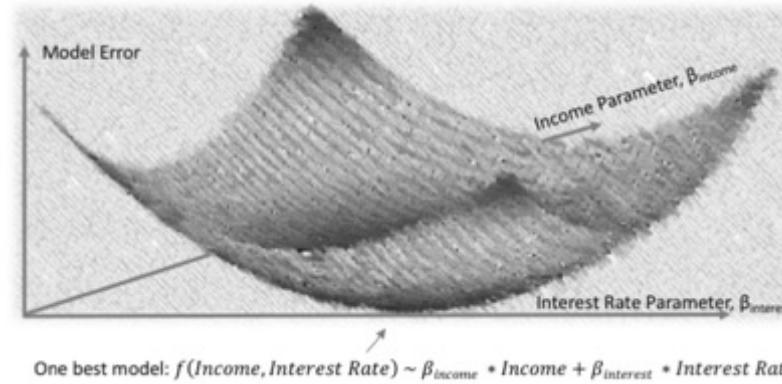


Multiplicity of Good Models

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

Trade-off

Multiplicity

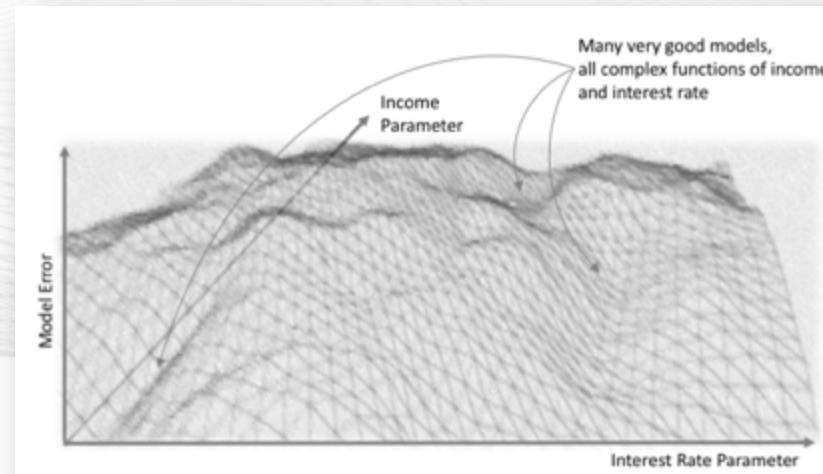


Multiplicity of Good Models

Trade-off

Multiplicity

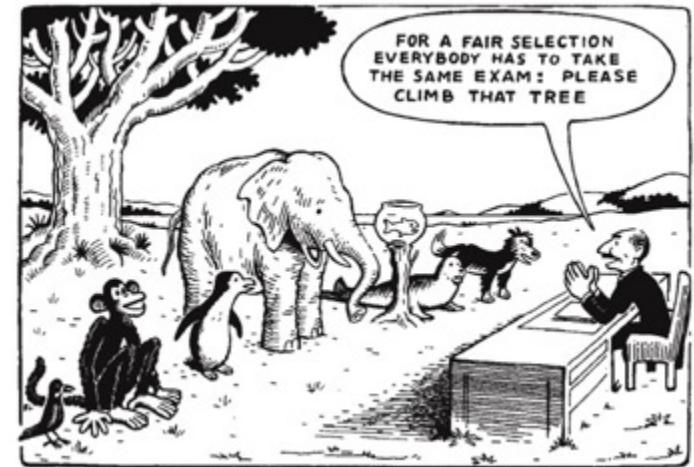
- ... for a given well-understood dataset there are usually **many good** ML models. Which one to **choose**?
- Same **objective metrics** values, **performance**, ...
- This is often referred to as “the **multiplicity** of good models.” -- [Leo Breiman](#)



Trade-off
Multiplicity
Fairness

Fairness and Social Aspects

- Gender
- Age
- Ethnicity
- Health
- Sexual behavior



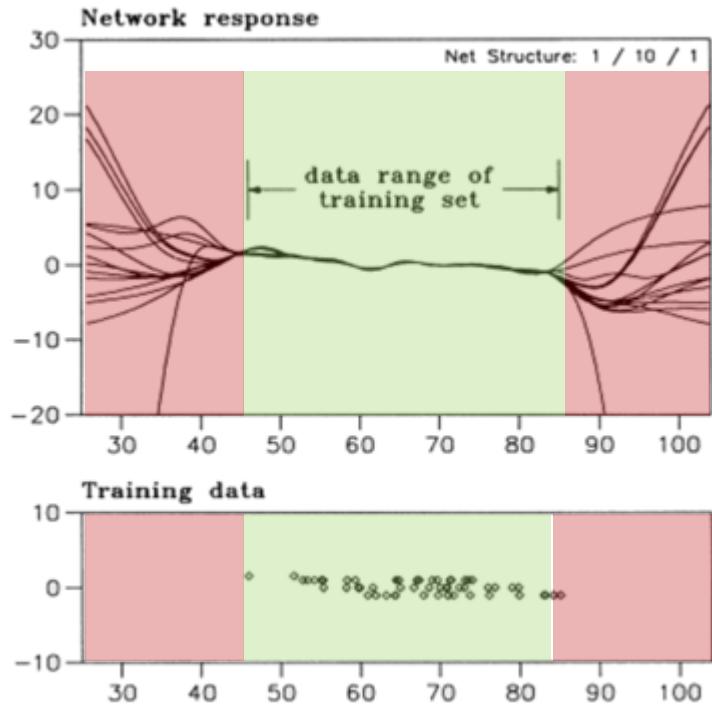
- Avoid **discriminatory models** and remediate disparate impact.

Trade-off
Multiplicity
Fairness
Trust

Trust of model producers & consumers

H₂O.ai

- Dataset vs. **real world**
- ML adoption
- Introspection
- Sensitivity
- OOR
- Diagnostics
- “Debugging”



Trade-off

Multiplicity

Fairness

Trust

Security

Security and Hacking

- Goal: **compromise** model integrity
- Attack types:
 - **Exploratory**
 - Surrogate model trained to identify vulnerabilities ~ MLI.
 - Trial and error (for specific class) x indiscriminate attacks.
 - **Causative**
 - Models trained w/ adversary datasets.
 - Local model > adversarial instance > target model.
 - Standard / continuous learning.
 - **Integrity** (compromise system integrity)
 - False negative instance e.g. fraud passes check.
 - **Availability** (compromise system availability)
 - False positive instance e.g. blocks access to legitimate instances.

Trade-off

Multiplicity

Fairness

Trust

Security

Regulation

Regulated & Controlled Environments

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

Explainability Matters

Trade-off

Multiplicity

Fairness

Trust

Security

Regulation

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

What's needed

Building human-centered, low-risk models

Building Human-Centered, Low-Risk Models

- Big picture.
- Interpretability focused.
- MLI module demo **only**.
- DAI auto ML models.*
- MLI UCs coverage by DAI.
- Techniques and algorithms.
- **Possible** workflow.
- IID and TS.
- Where MLI module fits in E2E.



*) MLI module is not limited to DAI's models

Building Human-Centered, Low-Risk Models

H₂O.ai

Load data

Explanatory Data Analysis and Visualization

Load data

EDA

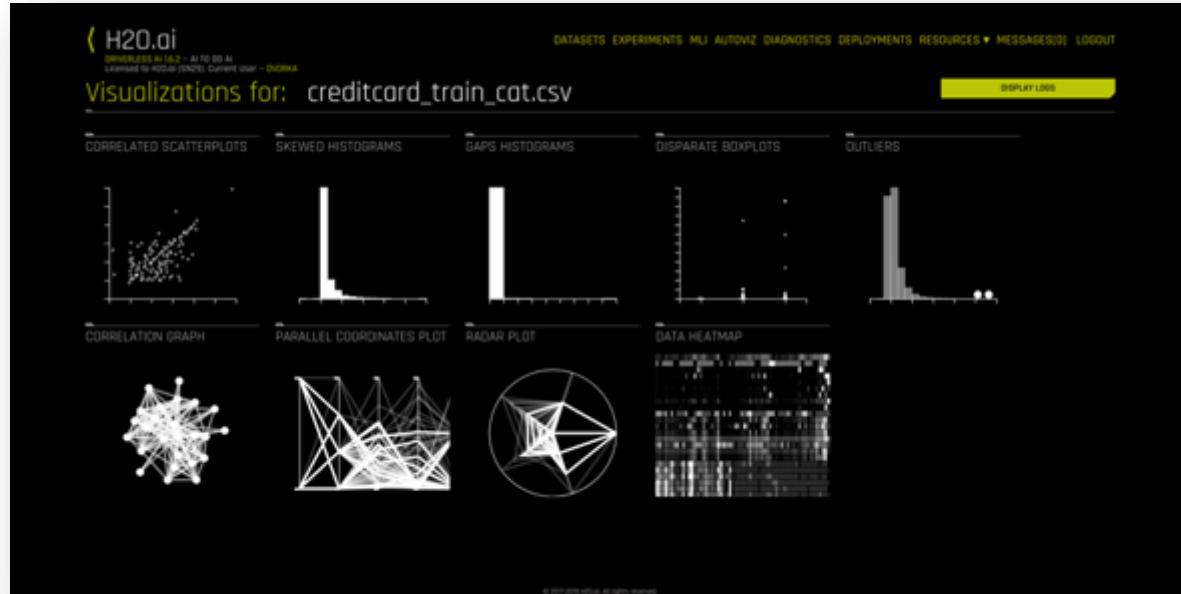


Explanatory Data Analysis and Visualization

H₂O.ai

Load data

EDA

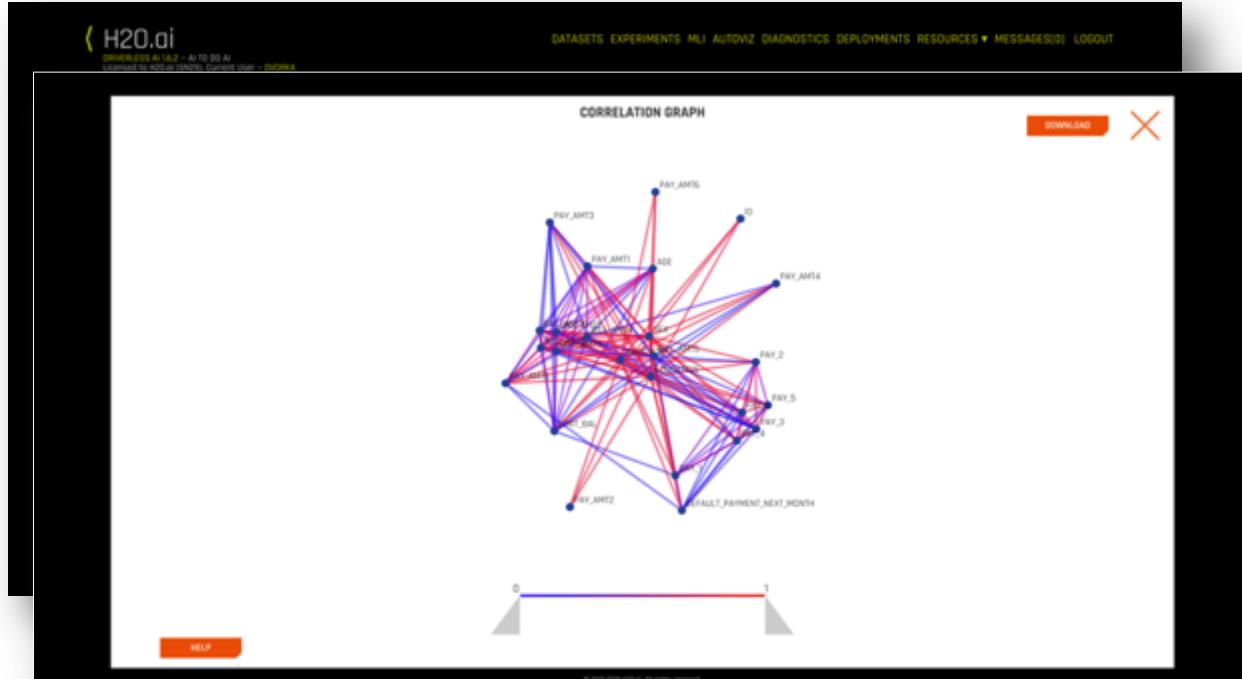


Explanatory Data Analysis and Visualization

H₂O.ai

Load data

EDA

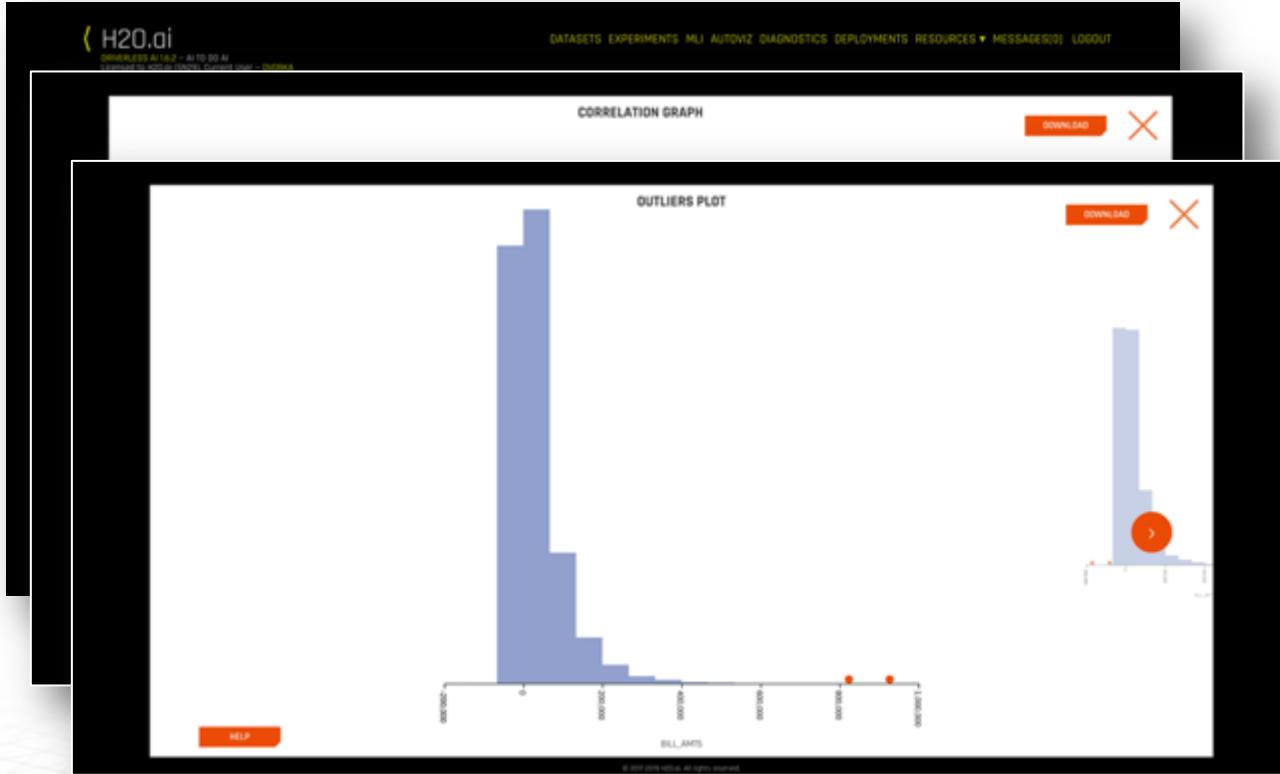


Explanatory Data Analysis and Visualization

H₂O.ai

Load data

EDA



Feature Engineering (Manual & Auto ML)



What do these settings mean?

ACCURACY

- Training data size: 23,999 rows, 25 cols
- Feature evolution: [LightGBM, XGBoost], 1/3 validation split
- Final pipeline: Ensemble (10 models), 5-fold CV

TIME

- Feature evolution: 4 individuals, up to 56 iterations
- Early stopping: After 5 iterations of no improvement

INTERPRETABILITY

- Feature pre-pruning strategy: None
- XGBoost Monotonicity constraints: disabled
- Feature engineering search space (where applicable): [Clustering, Date, FrequencyEncoding, Identity, interactions, isHoliday, NumEncoding, OneHotEncoding, TargetEncoding, Text, TextLin, TruncatedSVD, WeightOfEvidence]

[LightGBM, XGBoost] models to train:

- Model and feature tuning: 32
- Feature evolution: 84
- Final pipeline: 10

Estimated runtime: minutes

DEFAULT_PAYMENT_NE

WEIGHT COLUMN -- TIME COLUMN [OFF]

COUNT 23999 UNIQUE 2 TARGET FREQ 5369

EXPERIMENT SETTINGS

ACCURACY: +6 - (highlighted)

TIME: +4 -

INTERPRETABILITY: +1 - (highlighted)

EXPERT SETTINGS

CLASSIFICATION: REPRODUCIBLE: ENABLE GPU: LAUNCH EXPERIMENT

SCORER

- GNI
- MCC
- F05
- F1
- F2
- ACCURACY
- LOGLOSS
- AUC
- AUCPR

VARIABLE IMPORTANCE	
30_TruncSVD_BILL_AMT5_BILL_AMT5_PAY_AMT1_1	0.92
43_NumToCatTE_AGE_0	0.09
40_CV_TE_AGE_PAY_2_0	0.07
33_PAY_AMT3	0.06
24_CV_TE_AGE_0	0.02
44_CV_CatNumEnc_LIMIT_BAL_PAY_1_PAY_2_std	0.77
44_CV_CatNumEnc_LIMIT_BAL_PAY_1_BILL_AMT2_std	0.75
19_PAY_AMT5	0.74
44_CV_CatNumEnc_LIMIT_BAL_PAY_1_AGE_std	0.71
18_PAY_AMT5	0.71
41_PAY_AMT1	0.71
9_BILL_AMT2	0.69

Black box
model

Feature Engineering



What do these settings mean?

ACCURACY

- Training data size: 23,999 rows, 25 cols
- Feature evolution: [LightGBM, XGBoost], 1/3 validation split
- Final pipeline: [LightGBM, XGBoost]

TIME

- Feature evolution: 4 individuals, up to 56 iterations
- Early stopping: After 5 iterations of no improvement

INTERPRETABILITY

- Feature pre-pruning strategy: FS
- XGBoost Monotonicity constraints: enabled
- Feature engineering search space (where applicable): [Date, FrequencyEncoding, Identity, ISHoliday, NumEncoding, OneHotEncoding, TargetEncoding, Text]

[LightGBM, XGBoost] models to train:

- Model and feature tuning: 32
- Feature evolution: 84
- Final pipeline: 1

Estimated runtime: minutes

DEFAULT_PAYMENT_NE --

WEIGHT COLUMN	TIME COLUMN
--	[OFF]
TYPE	COUNT
bool	23999
UNIQUE	2
TARGET FREQ	5369

EXPERIMENT SETTINGS

ACCURACY	TIME	INTERPRETABILITY
+ 6	+ 4	+ 10
-	-	-

CLASSIFICATION REPRODUCIBLE ENABLE OPUS

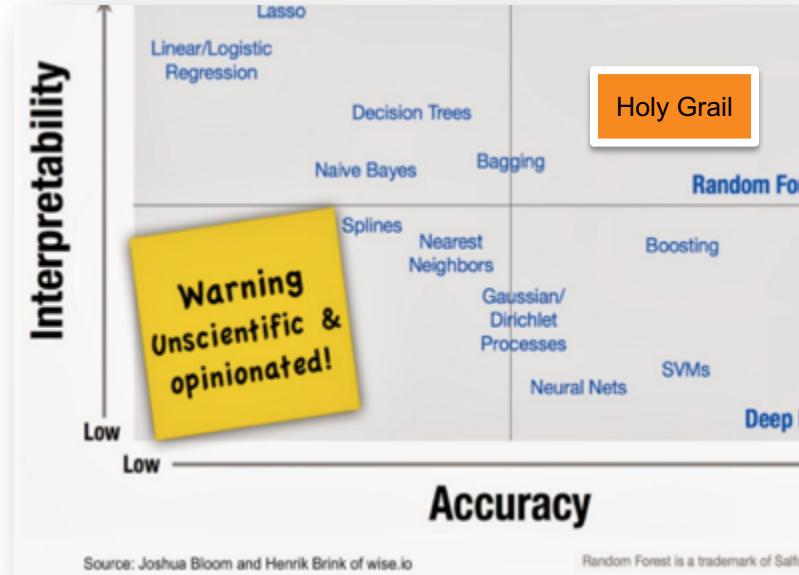
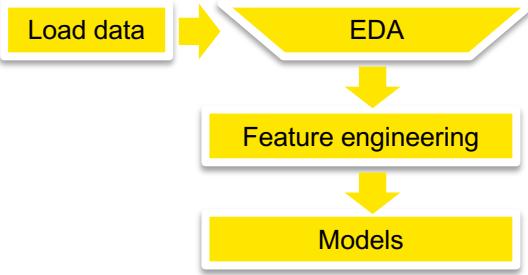
LAUNCH EXPERIMENT

SCORER

- GINI
- MCC
- F05
- F1
- F2
- ACCURACY
- LOGLOSS
- AUC
- AUCPR

White box
model

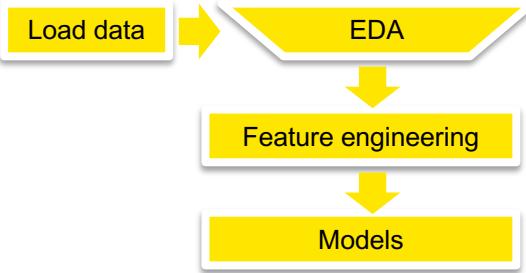
Model Choice: Constrained, Simple, Fair



Black box
model

White box
model

Model Choice: Constrained, Simple, Fair



What do these settings mean?

ACCURACY

- Training data size: 23,999 rows, 25 cols
- Feature evolution: [LightGBM, XGBoost], 1/3 validation split
- Final pipeline: [LightGBM, XGBoost]

TIME

- Feature evolution: 4 individuals, up to 56 iterations
- Early stopping: After 5 iterations of no improvement

INTERPRETABILITY

- Feature pre-pruning strategy: FS
- XGBoost Monotonicity constraints: enabled
- Feature engineering search space (where applicable): {Date, FrequencyEncoding, Identity, Isholiday, NumEncoding, OneHotEncoding, TargetEncoding, Text}

(LightGBM, XGBoost) models to train:

- Model and feature tuning: 32
- Feature evolution: 84
- Final pipeline: 1

Estimated runtime: minutes

DEFAULT_PAYMENT_NE --

WEIGHT COLUMN	TIME COLUMN
--	[OFF]
TYPE	COUNT
bool	23999
UNIQUE	2
TARGET FREQ	5369

EXPERIMENT SETTINGS

ACCURACY	TIME	INTERPRETABILITY
+ 6 -	+ 4 -	+ 10 -

CLASSIFICATION REPRODUCIBLE ENABLE GPU

LAUNCH EXPERIMENT

SCORER

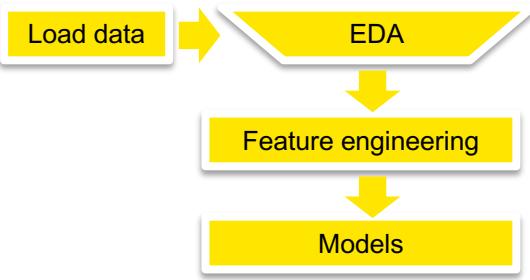
GINI
MCC
F05
F1
F2
ACCURACY
LOGLOSS
AUC
AUCPR

Black box
model

White box
model

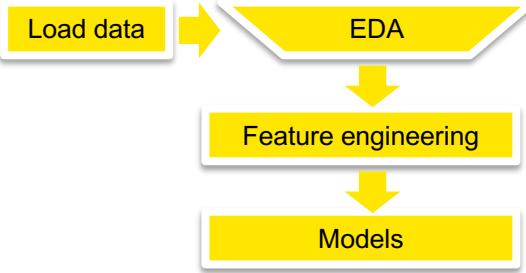
GLM (log regr.), Monotonic GBM (DT), XNN, ...

Model Choice



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

Model Choice



What do these settings mean? DEFAULT_PAYMENT_NE --

Expert Experiment Settings

TIME
INTER
Frequ
Estim

ACCU
- Train
- Feat
- Find

Pipeline Building Recipe
AUTO COMPLIANT

XGBoost GBM models
AUTO ON OFF

LightGBM Random Forest models
AUTO

GLM models
AUTO ON OFF

RuleFit models
AUTO ON OFF

FTRL models
AUTO ON OFF

Ensemble Level (-1 = auto)
-1

Select target transformation of the target for regression problems
AUTO IDENTITY

Data distribution shift detection
ENABLED

Enable Target Encoding
ENABLED

Probability to create non-target lag features
0.5

Number of models during tuning phase (-1 = auto)
-1

Max allowed feature shift (AUC) before dropping feature
0.6

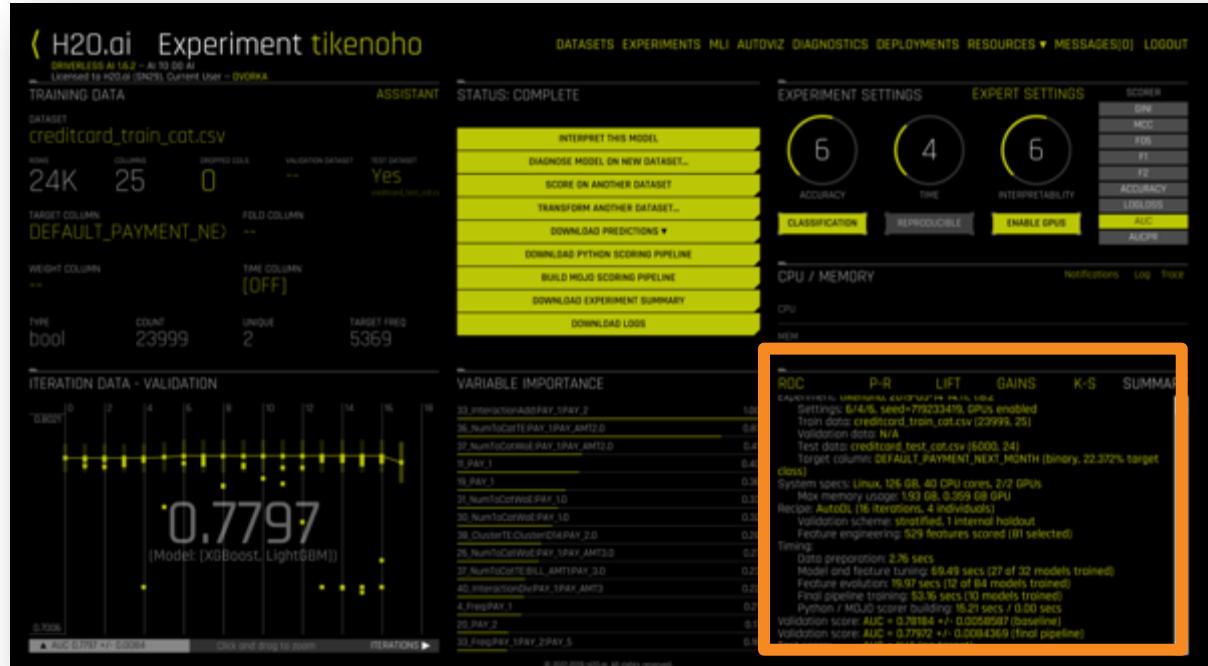
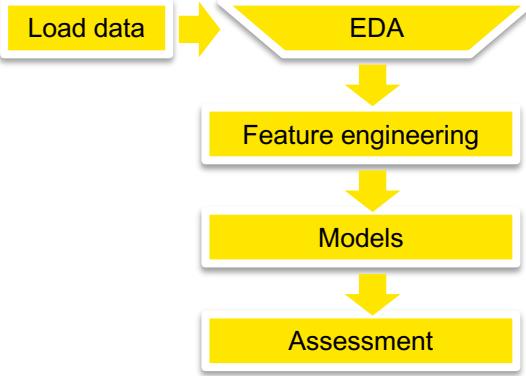
Time-series log-based recipe
ENABLED

Make Python scoring pipeline
ENABLED

SAVE CANCEL

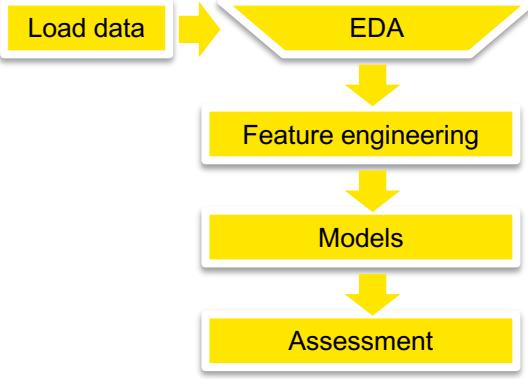
The screenshot shows the "Expert Experiment Settings" dialog box. It contains various configuration options for different machine learning models. Several sections are highlighted with orange boxes: "LightGBM models", "GLM models", and "TensorFlow models". The "LightGBM models" section is currently selected. Other sections include "XGBoost GBM models", "FTRL models", "RuleFit models", "Ensemble Level", "Select target transformation of the target for regression problems", "Data distribution shift detection", "Enable Target Encoding", "Probability to create non-target lag features", "Number of models during tuning phase", "Max allowed feature shift", "Time-series log-based recipe", and "Make Python scoring pipeline". At the bottom are "SAVE" and "CANCEL" buttons.

Traditional Model Assessment and Diagnostics

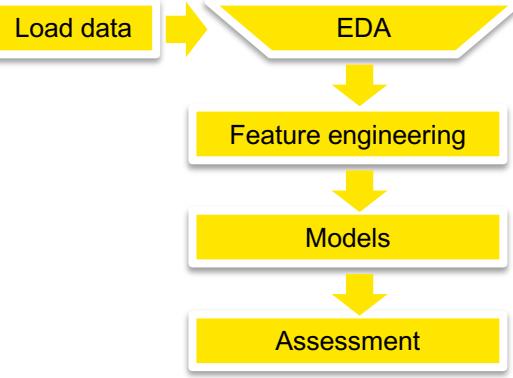


Traditional Model Assessment and Diagnostics

H2O.ai



Traditional Model Assessment and Diagnostics



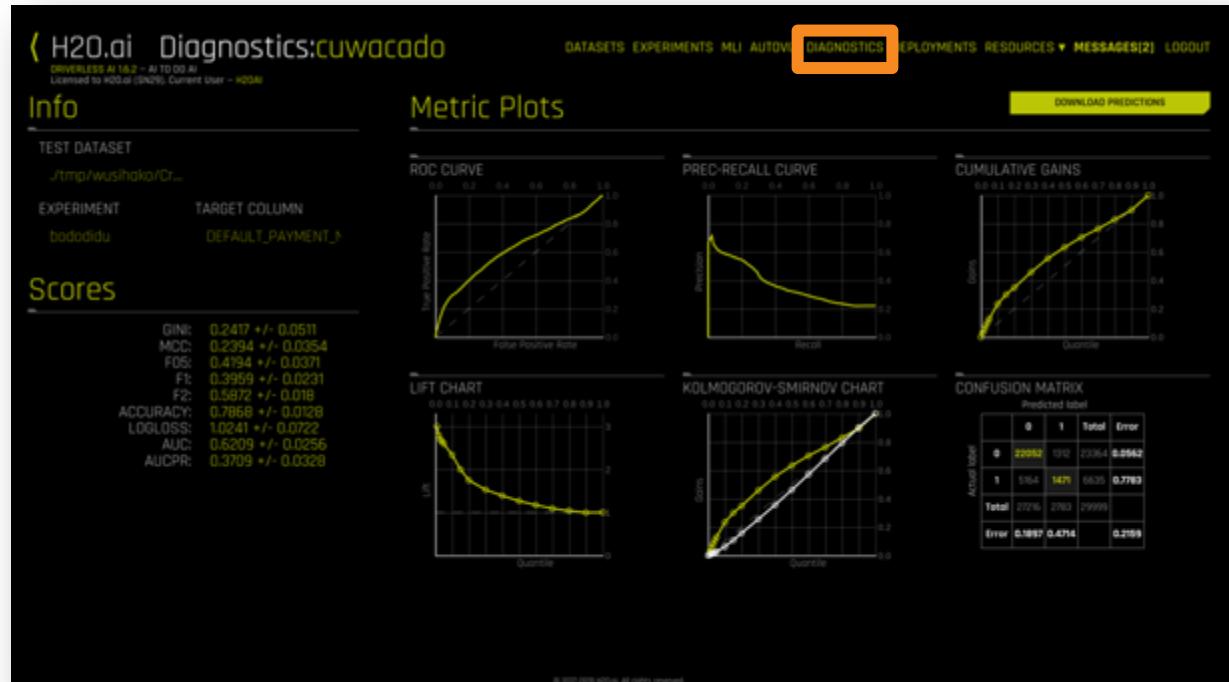
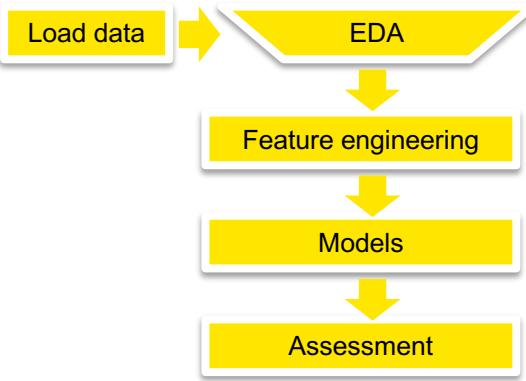
Experiment summary (document + YAML) + AutoDoc

The screenshot shows the H2O Experiment summary interface for a project named 'tikenoho'. The main dashboard includes sections for TRAINING DATA (creditcard_train_cat.csv, 24K rows, 25 columns, validation dataset Yes), ASSISTANT (STATUS: COMPLETE), EXPERIMENT SETTINGS (Accuracy: 6, Time: 4), EXPERT SETTINGS (Score: GME, FOS, F1, Accuracy, LogLoss, AUC, AUCPR), and a central panel for INTERPRET THIS MODEL, DIAGNOSE MODEL ON NEW DATASET, SCORE ON ANOTHER DATASET, TRANSFORM ANOTHER DATASET, DOWNLOAD PREDICTIONS, and DOWNLOAD PYTHON SCORING PIPELINE. A prominent orange box highlights the 'DOWNLOAD EXPERIMENT SUMMARY' button. To the right, a detailed document titled 'Driverless AI Experiment: tikenoho' is displayed, containing sections like Experiment Overview, Performance, and Driverless Settings. The document lists various metrics and configurations used in the experiment.

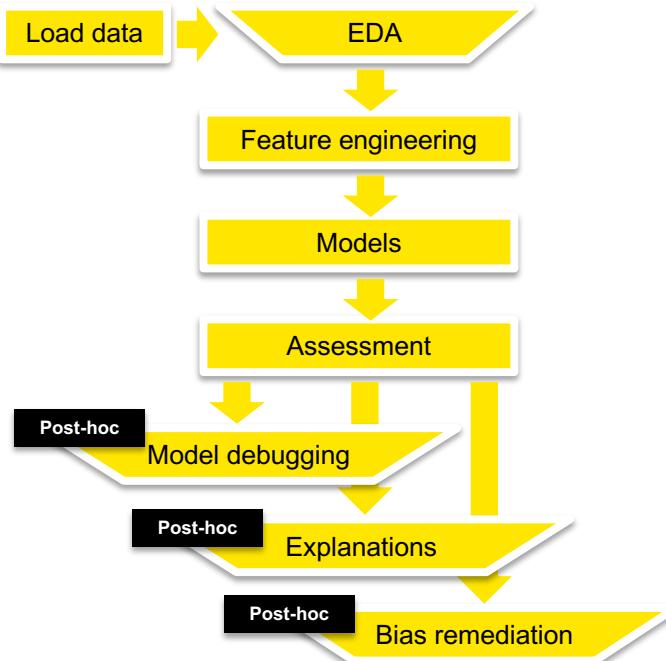
Section	Content		
Experiment Overview	Generated by: dorka Generated on: Tue May 14 14:13:37 2019		
Data Overview	1		
Methodology	3		
Validation Strategy	4		
Model Training	9		
Feature Evolution	11		
Feature Transformations	12		
Final Model	13		
Deployment	14		
Appendix	15		
Performance	16		
Dataset	AUC		
Internal Validation	0.78		
Test Beta	Test Data did not have Target Column		
Driverless Settings			
Dial Settings	Description	Setting Value	Range of Possible Values
Accuracy	Controls accuracy goals of the model	6	1-99

Traditional Model Assessment and Diagnostics

H2O.ai

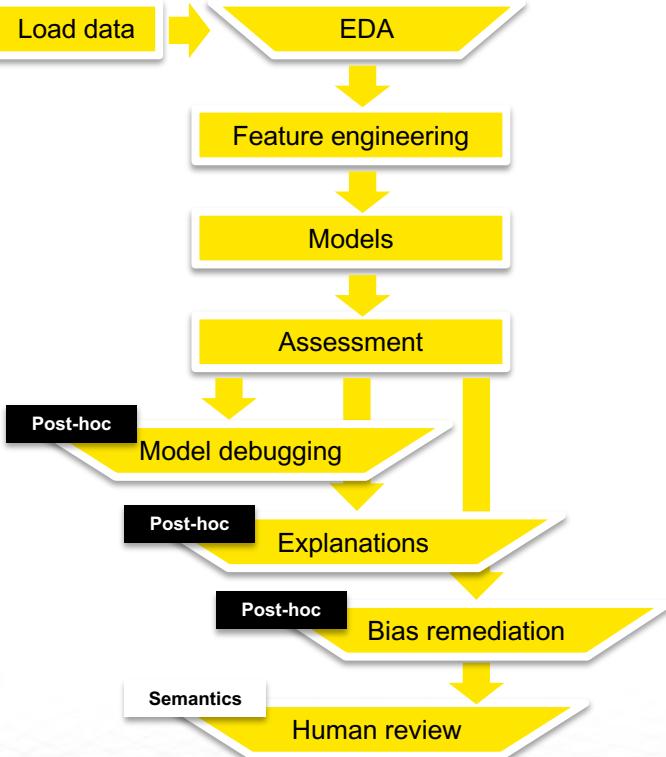


Post-hoc Model Explanations and Debugging

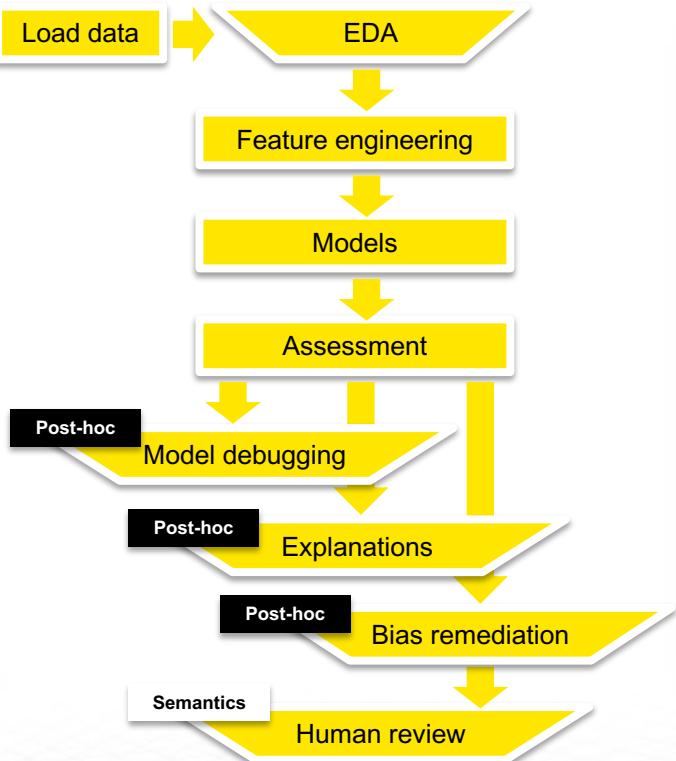


- **Post-hoc model debugging**
 - What-if, sensitivity analysis (accuracy).
- **Post-hoc explanations**
 - Reason codes.
- **Post-hoc bias assessment and remediation**
 - Disparate impact analysis.

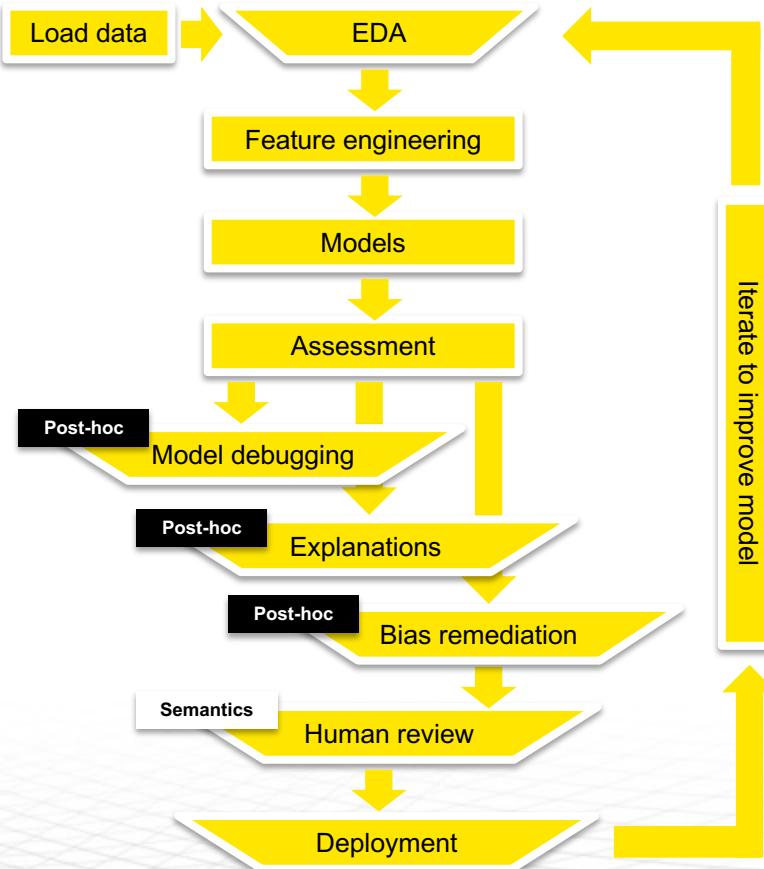
Human Review



Human Review



Iterative Improvement



How

Explaining models - MLI module deep dive

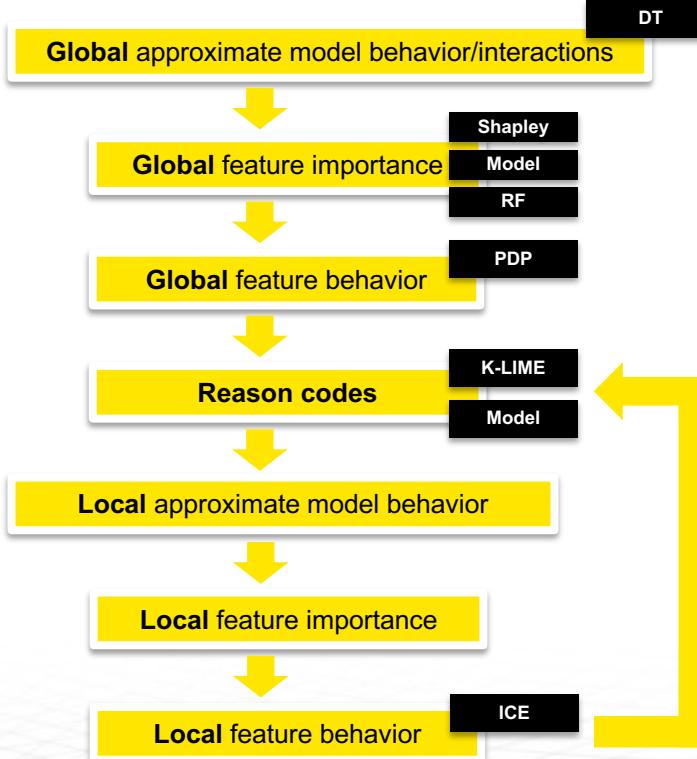
H2O Driverless AI's MLI module



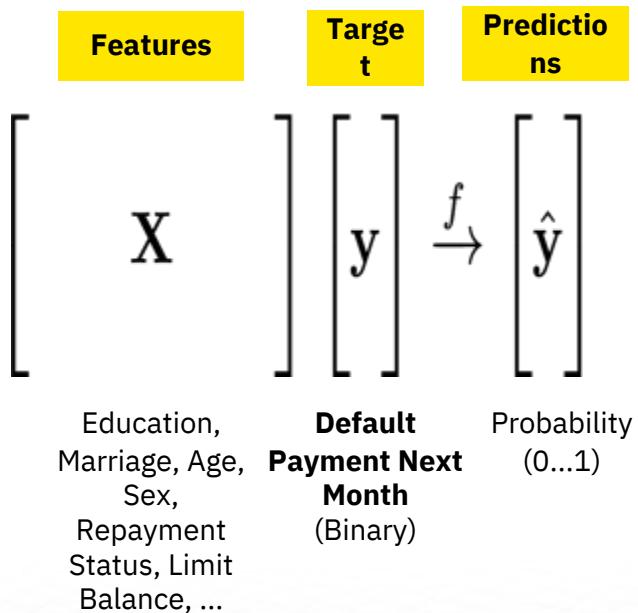
IID and Time Series



H2O Driverless AI's MLI module



Demo Dataset: Credit Card (IID)

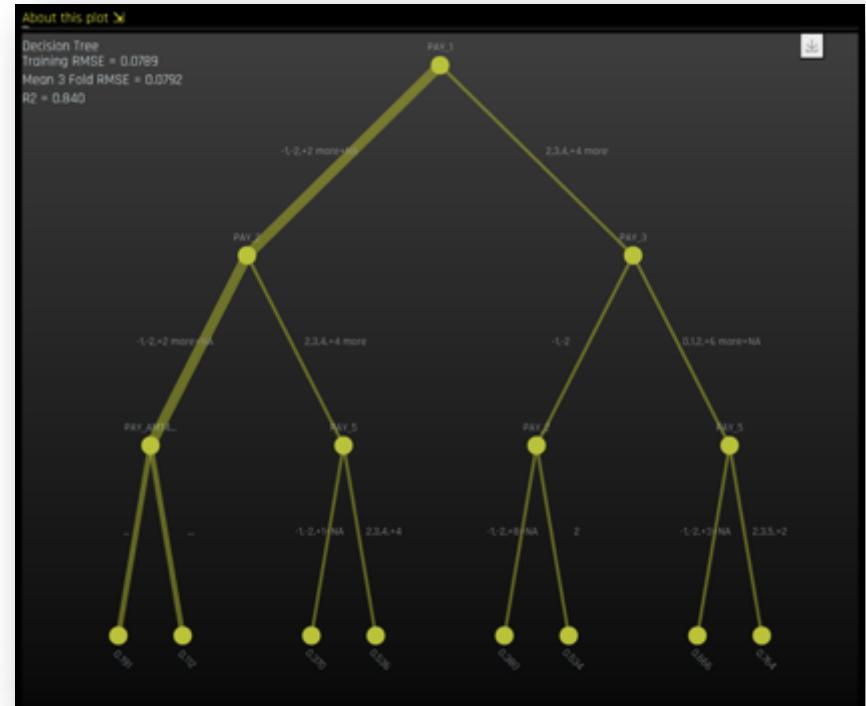


Column Name	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_x {1, ...,6}	Repayment status in August, 2005 – April, 2005 (-1=paid duly,1=payment delay for 1 month, ...,8=payment delay for 8 months)
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)

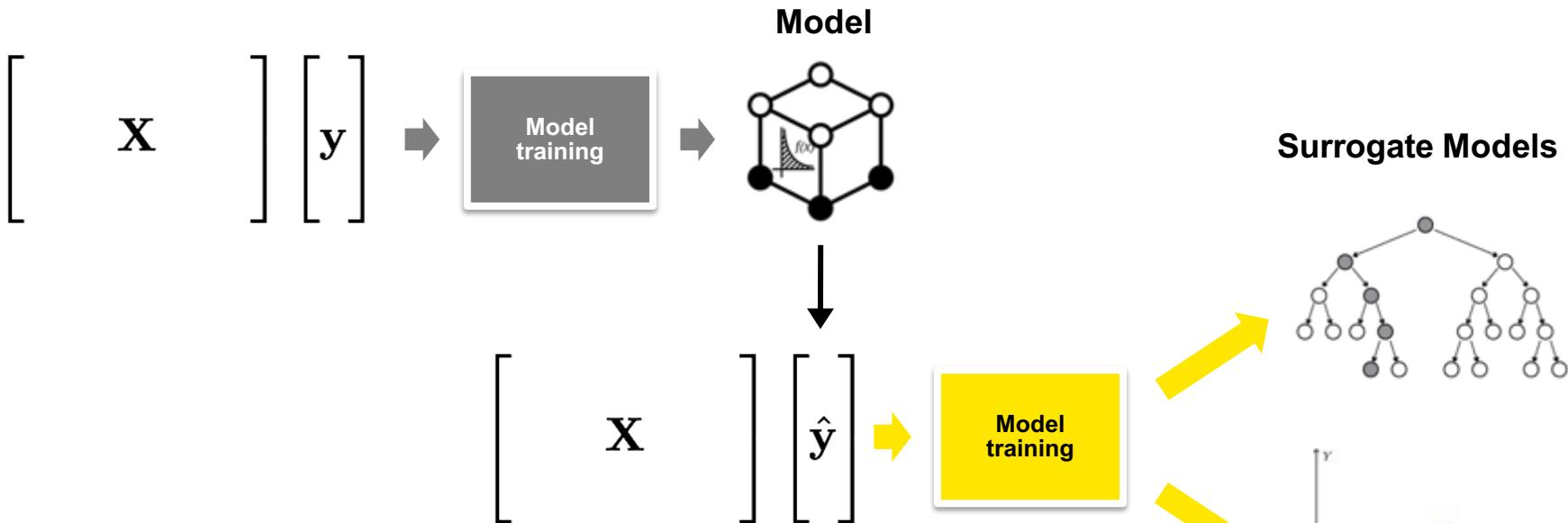


Global Approximate Model Behavior/Interaction

- **Challenge:**
 - Black-box models
 - Original vs. transformed features.
- **Solution:** Surrogate models
 - **Pros**
 - Increases any black-box model's interpretability
 - Time complexity
 - **Cons**
 - Accuracy

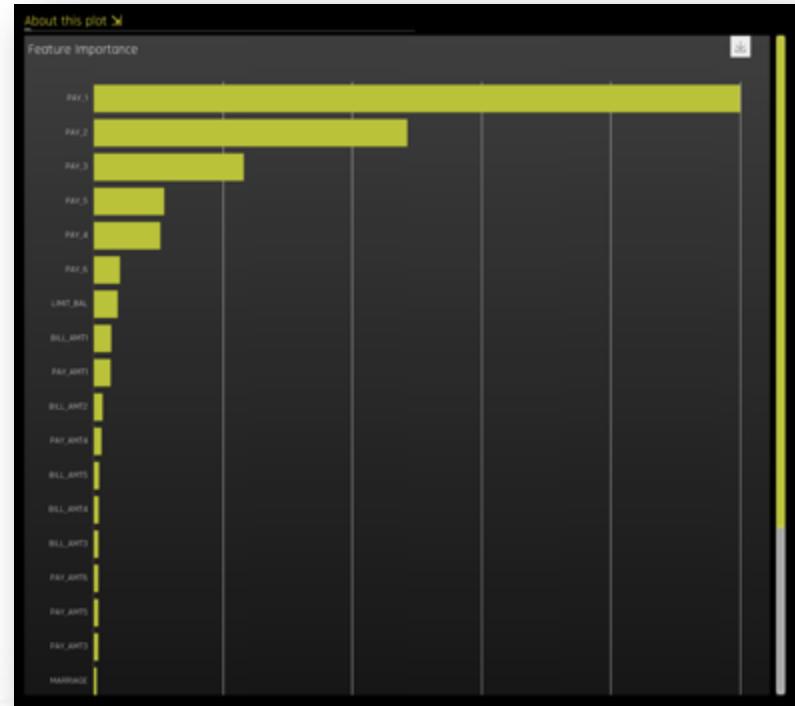


Surrogate Models



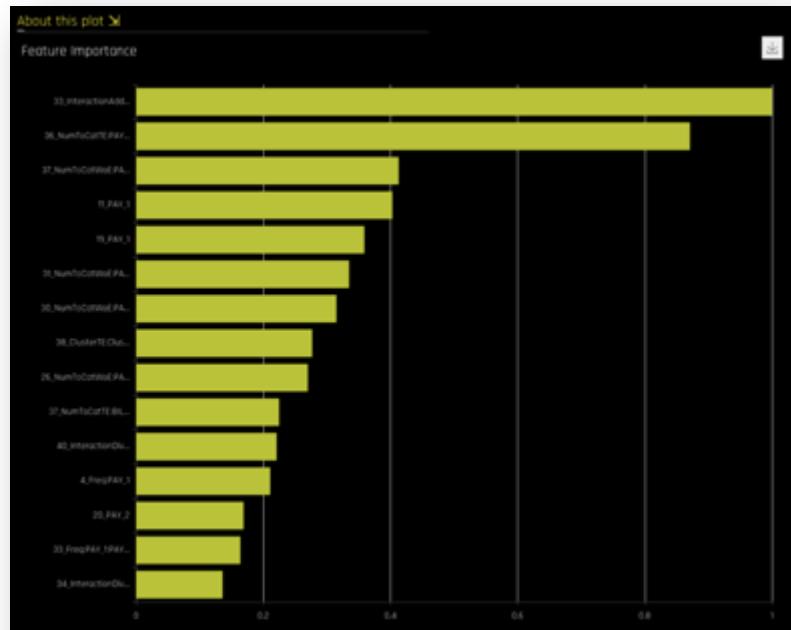
Global Feature Importance: Random Forest

- **Challenges:**
 - Black-box models
 - Original vs. transformed features
- **Solutions:**
 - Surrogate model: RF (introspection)
 - **Pros:**
 - Original features
 - Time complexity
 - **Cons:**
 - Accuracy



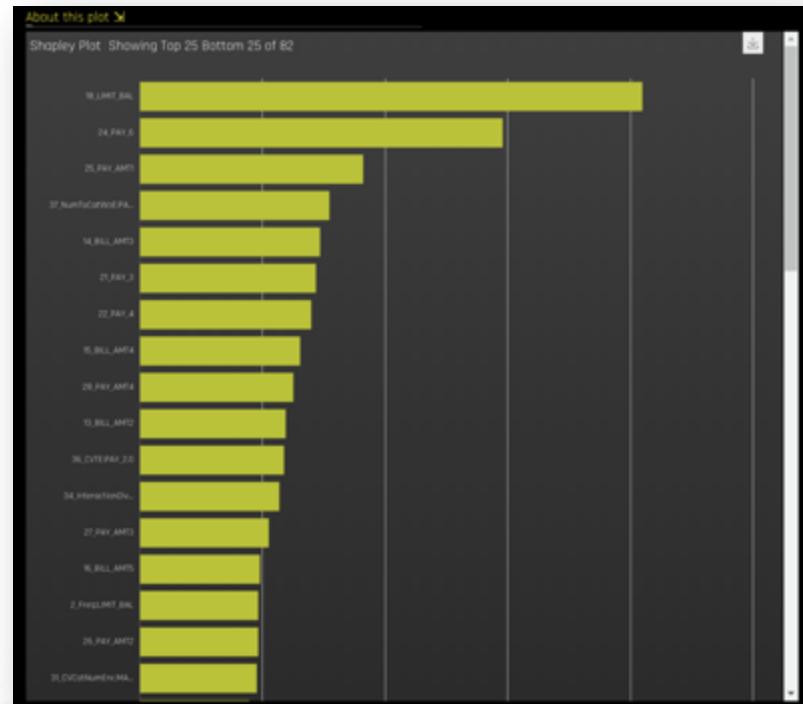
Global Feature Importance: Original Model

- **Challenges:**
 - Black-box models
 - Original vs. transformed features
- **Solutions:**
 - **Original (DAI) Model Introspection**
 - **Pros:**
 - Accuracy
 - **Cons:**
 - Transformed features
 - Global only



Global Feature Importance: Shapley Values

- **Challenge**
 - Black-box models
 - Original vs. transformed features
- **Solutions:**
 - Shapley values
 - Pros:
 - Accuracy
 - Math correctness
 - Cons:
 - Time complexity
 - Transformed features



Shapley Values

- Lloyd Shapley
 - American mathematician who won **Nobel** prize in 2012 (Economics).
 - Shapley values was his Ph.D. thesis written in **50s**.

Shapley values:

- Supported by **solid** mathematical (game) theory.
- Calculation has **exponential** time complexity (number of coalitions).
- Typically **unrealistic to compute** in real world.
- Can be computed in **global** or **local** scope.
- **Guarantee fair distribution** among features in the instance.
- Does **not** work well in **sparse** cases, **all** features must be used.
- Return **single value per feature**, not a model.



ALGORITHM: Shapley value = contribution of feature f in example

Method:

- i) have dataset and chose sample \underline{c} and feature f
- ii) compute marginal contribution of feature f in \underline{c} for every feature coalition
- iii) $f \in \mathcal{F}$ & coalition \underline{c}
 - a) eliminate all features while are not in current coalition \underline{c} & saving value from other randomly selected sample \underline{c}^R
 - b) predict... with feature f in coalition $\rightarrow p^w$
without feature f in coalition $\rightarrow p^w_{no}$
(random select other sample, use c^R / and save value of f from there)
 - c) marginal f contribution in \underline{c} : $p^w - p^w_{no} = \Delta c$
and coalition \underline{c}
- iv) marginal feature contribution is $SHAPLEY(f) = \text{AVG}(\Delta c)_{\underline{c} \in \mathcal{C}}^{2^{\mathcal{F}}} \quad n = \text{number of coalitions} \quad n = \text{exponent } O(2^{\mathcal{F}})$

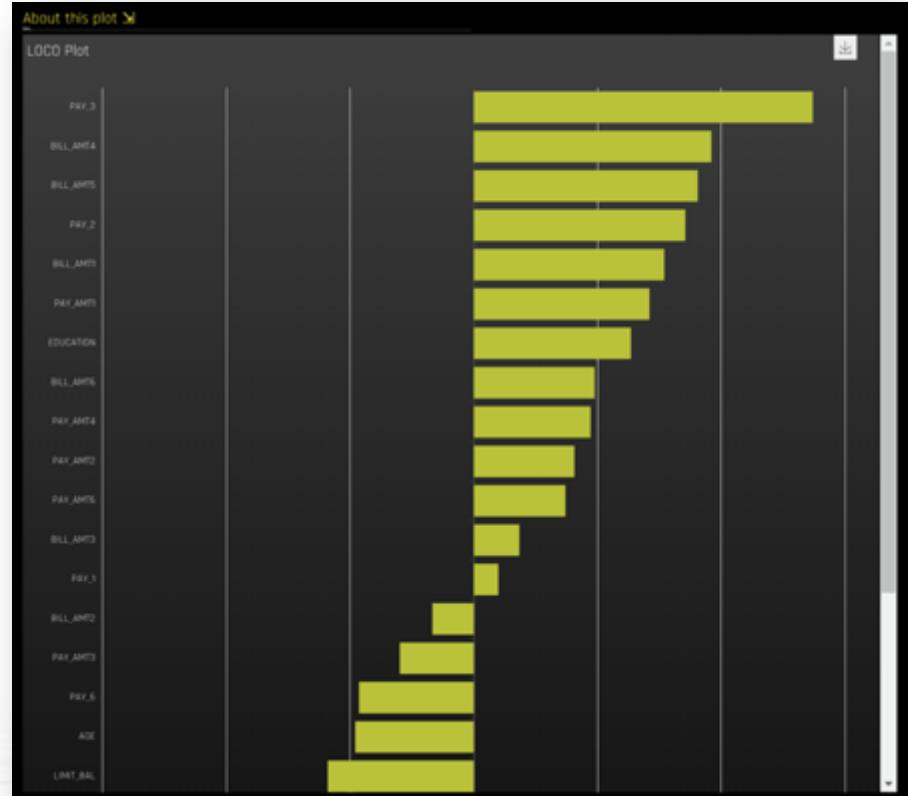
SHAPLEY VALUES

CASE \rightarrow single instance i
GAIN \rightarrow coalition - avg prediction
PLAYERS \rightarrow feature is player;
players cooperate in coalition
to receive gains

Global
With all local
Shapley value sample
all samples \underline{c}_i

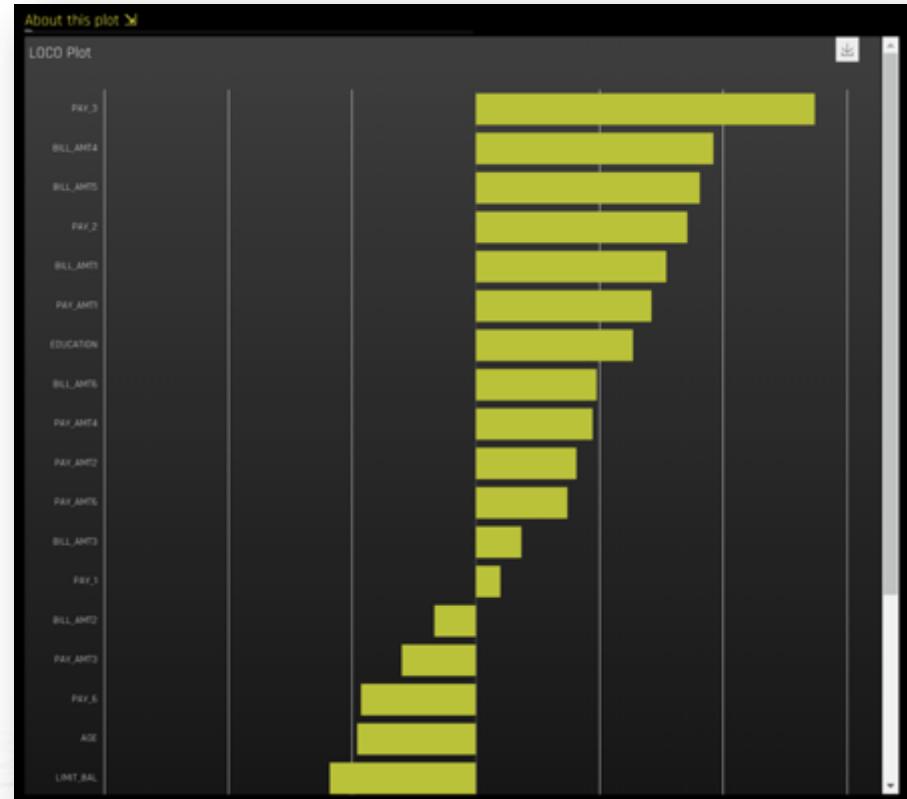
Feature importance: Leave One Covariate Out

- **UC:**
 - Complete other feature importance charts with bias tendency
- **Challenge:**
 - Black-box models
- **Solution:**
 - LOCO



Feature importance: Leave One Covariate Out

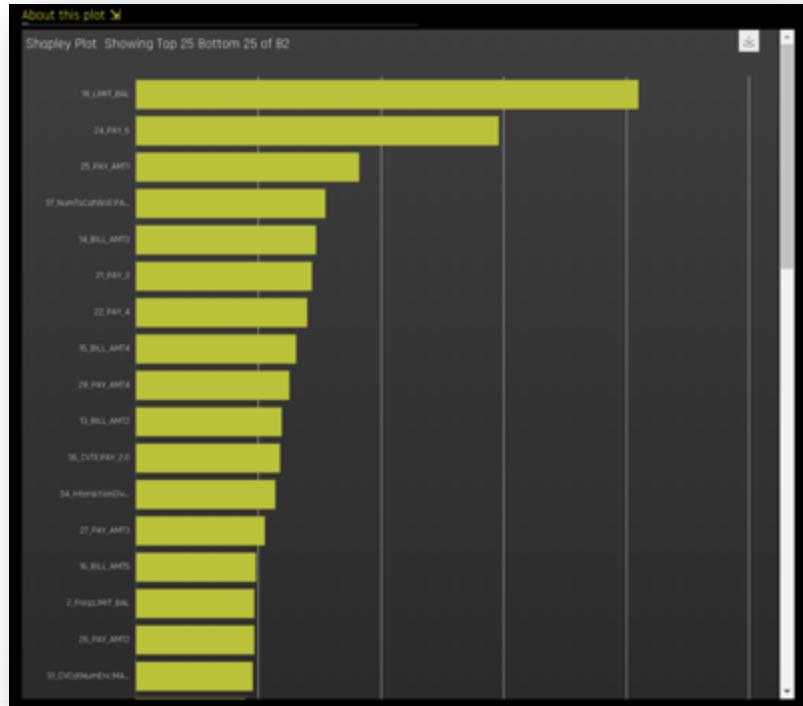
Sex	Age	...	Fare	\hat{y}	$\hat{y}_{(-\text{Sex})}$	$\hat{y}_{(-\text{Age})}$...	$\hat{y}_{(-\text{Fare})}$
M	11	...	8.45	0.2	0.01	0.1	...	0.21
F	34	...	51.86	0.8	0.6	0.65	...	0.78
M	26	...	21.08	0.5	0.2	0.3	...	0.53
...



Global Feature Importance

- **Methods**

- Surrogate models:
 - RF (introspection)
 - Leave One Covariate Out (LOCO)
- Original model (introspection)
- **Shapley values**



Global Feature Behavior: Partial Dependence Plot

- **Solution:** Surrogate model PDP
 - Pros
 - Time complexity
 - Original features
 - White/black model interpretability
 - Cons
 - Accuracy



Global Feature Behavior: Partial Dependence Plot

H₂O.ai

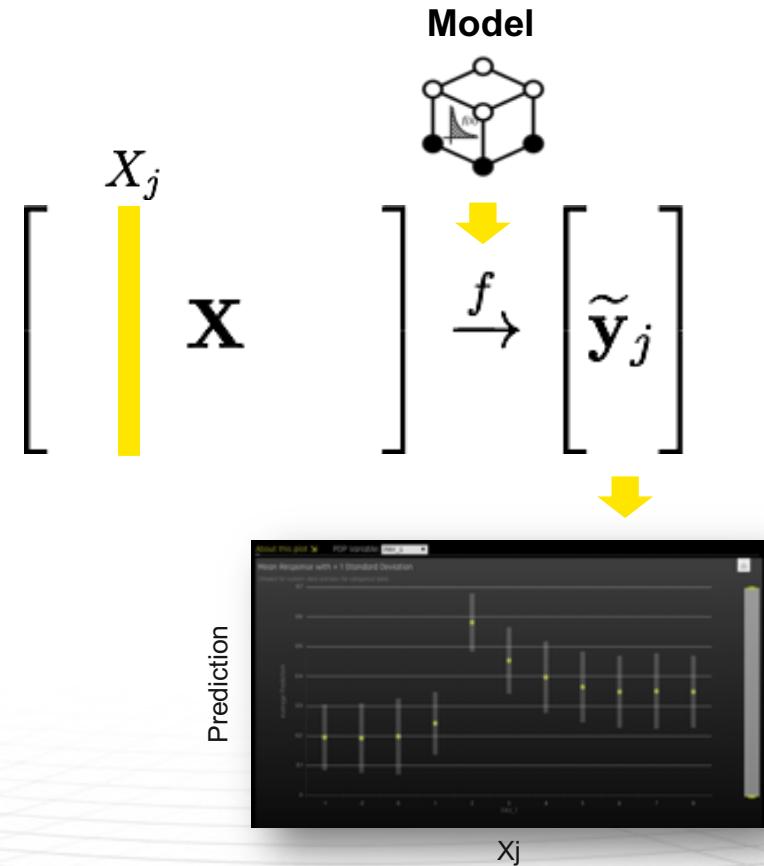
- **Solution:** Surrogate model PDP

- **Pros**

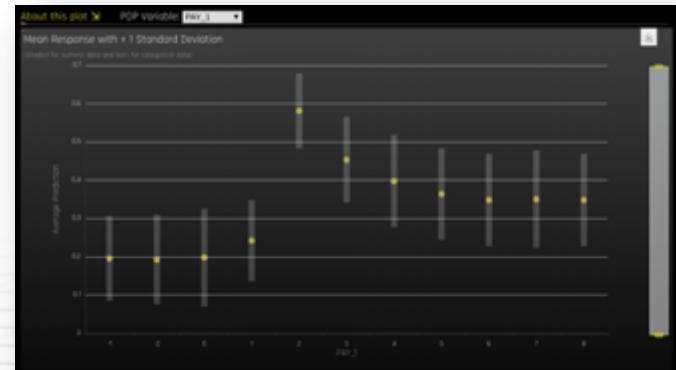
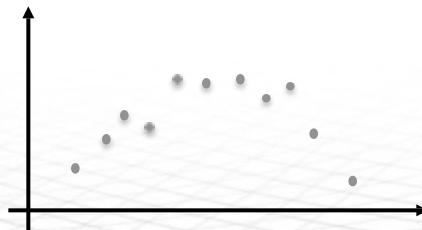
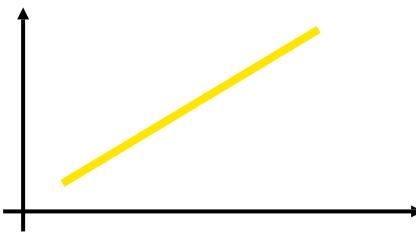
- Time complexity
 - Original features
 - White/black model interpretability

- **Cons**

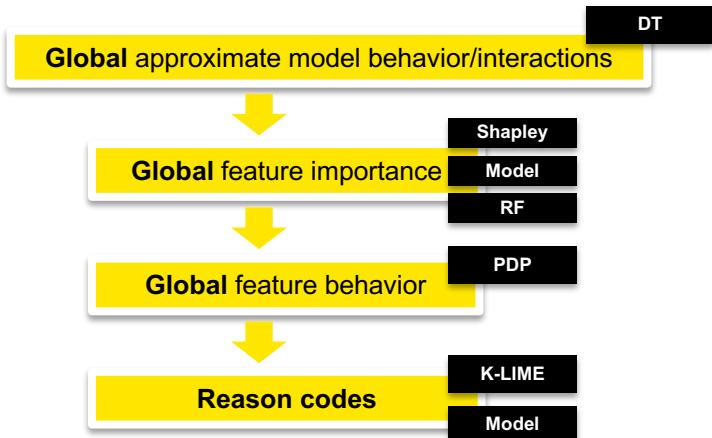
- Accuracy



PDP: Character of the Feature Behavior



Reason codes: Local Feature Importance



Reason codes: Local Feature Importance

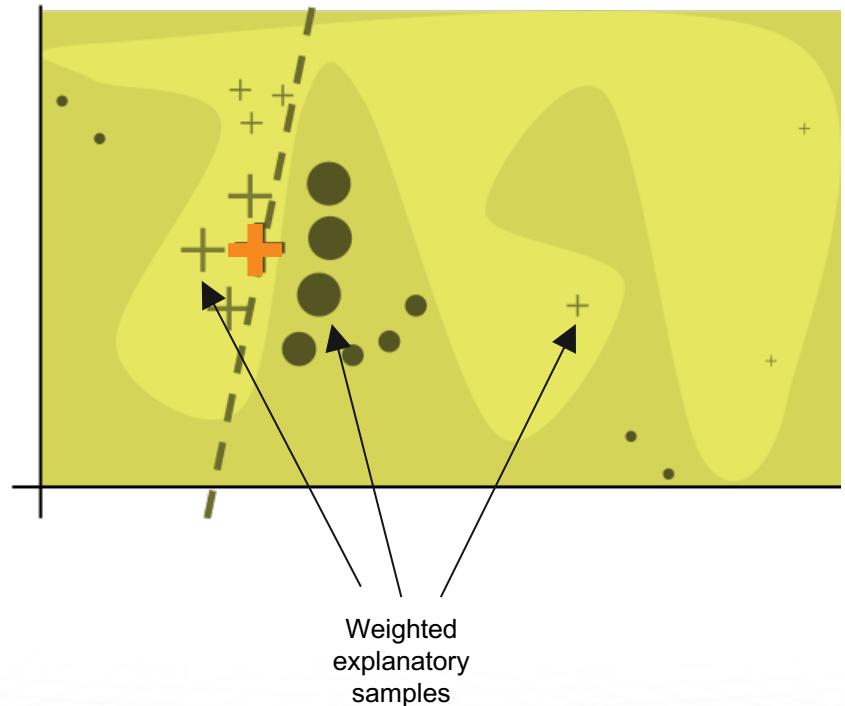
- UCs:
 - Predictions explanations
 - Legal
 - Debugging
 - Drill-down,
 - ...
- From **global** to **local** scope
- Surrogate methods:
 - **K-LIME** (K-means)
 - LIME-SUP (trees)



LIME: Local Interpretable Model-agnostic Explanations

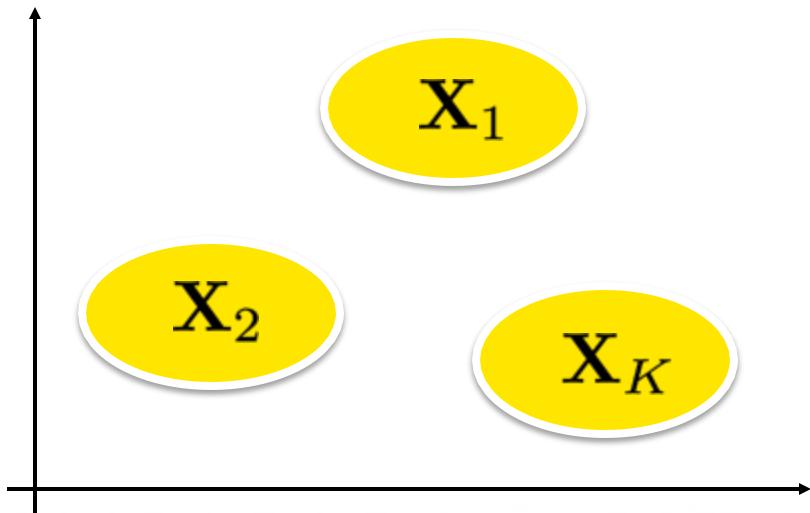
H₂O.ai

- Weighted **linear** surrogate model used to explain **non-linear** decision boundary in **local region**.
- **Single prediction.**
- **+** example:
 - Set of explainable records are scored using the **original model**.
 - To interpret a decision about another record, the explanatory records are **weighted by their closeness** to that record.
 - L1 regularized linear model is trained on this weighted explanatory set.
 - The **parameters of the linear model** then help **explain** the prediction for the selected record.

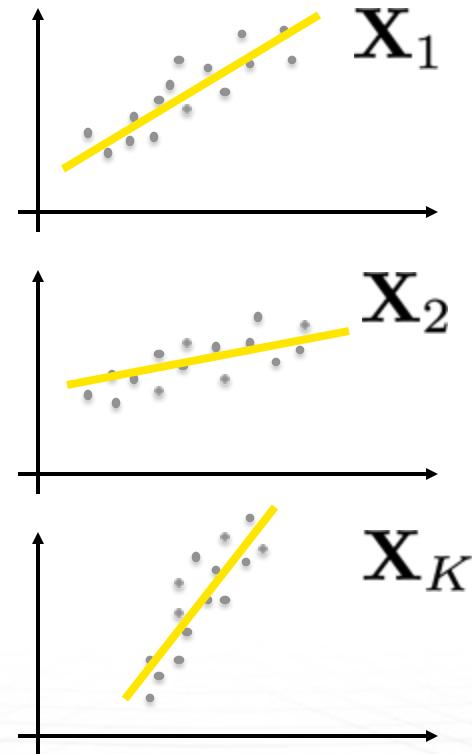


Source: <https://github.com/marcotcr/lime>

K-LIME: Clustered LIME



$$\{\mathbf{X}_1 \cup \mathbf{X}_2 \cup \dots \cup \mathbf{X}_K\} = \mathbf{X}$$



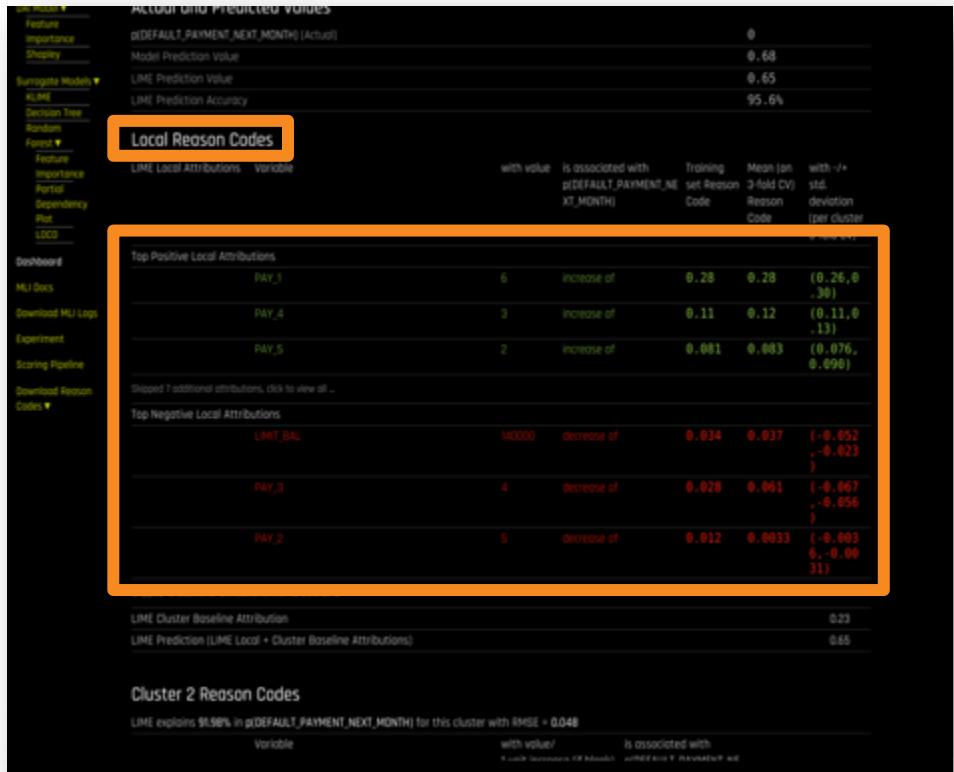
Reason codes: Local Feature Importance

- UCs:
 - Predictions explanations
 - Legal
 - Debugging
 - Drill-down,
 - ...
- From **global** to **local** scope
- From global explanatory model to **cluster-scoped** explanatory model.



Reason codes: Local Feature Importance

- Challenges:**
 - Black-box models
 - Original vs. transformed features
- Solutions:**
 - Surrogate model: **K-LIME**
 - Pros:
 - Original features
 - Time complexity
 - Cons:
 - Accuracy



$$g(\mathbf{x}^{(i)}) \approx h_{\text{GLM},k}(\mathbf{x}^{(i)}) = \beta_0^{[k]} + \sum_{j=1}^P \beta_j^{[k]} x_j^{(i)}$$

reason code

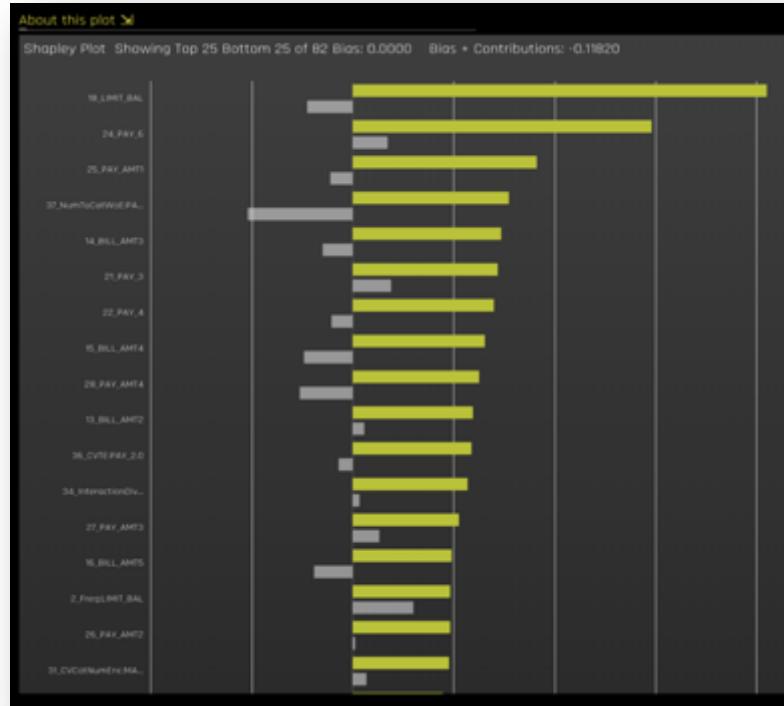
Local Approximate Model Behavior/Interaction

- **UC:**
 - Particular instance explanation
 - Note path segments thickness.
- **Challenge:**
 - Black-box models
- **Solution:** Surrogate models
 - **Pros**
 - Black-box models interpretability
 - Time complexity
 - **Cons**
 - Accuracy



Local Feature Importance

- Mean absolute value vs. **local contributions**
- **Challenge**
 - Black-box models
 - Original vs. transformed features
- **Solutions:**
 - Surrogate models:
 - RF (introspection)
 - Leave One Covariate Out (LOCO)
 - **Shapley values**



Local Feature Behavior: ICE

- **Solution:** Surrogate model ICE

- Pros

- Time complexity
 - Original features
 - White/black model interpretability

- Cons

- Accuracy
(dotted line vs. gray dot discrepancy)



ICE: Individual Conditional Expectations

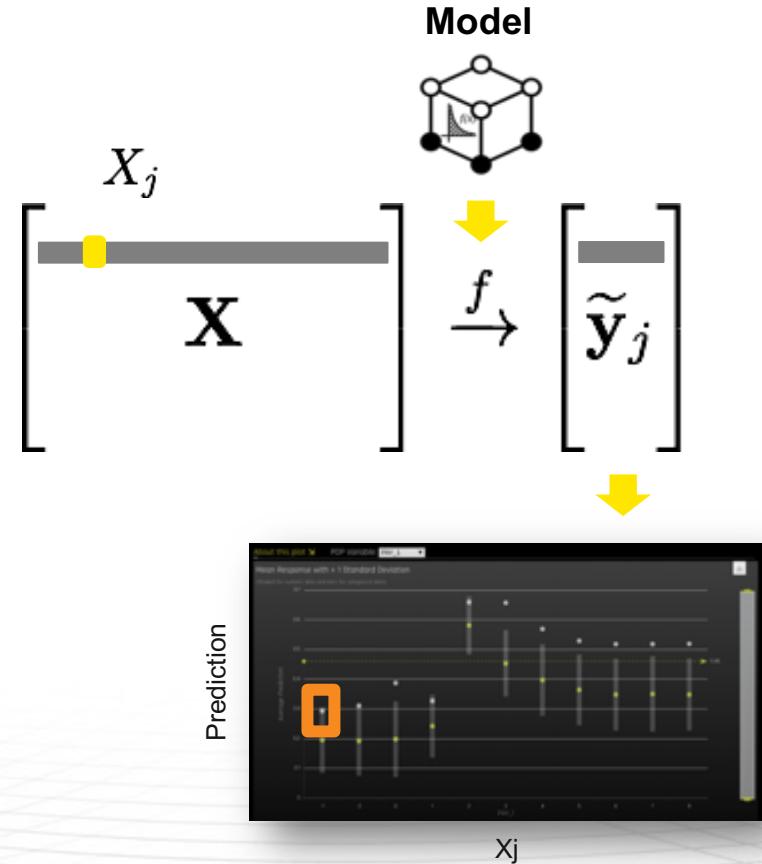
- **Solution:** Surrogate model ICE

- **Pros**

- Time complexity
 - Original features
 - White/black model interpretability

- **Cons**

- Accuracy



Driverless AI Experiment: cuwecoga

Generated by: username
Generated on: Wed May 15 12:04:26 2019

Experiment Overview.....	1
Data Overview.....	3
Methodology.....	5
Data Sampling.....	10
Validation Strategy.....	10
Model Tuning.....	12
Feature Evolution.....	13
Feature Transformations.....	13
Final Model.....	14
Alternative Models.....	17
Deployment.....	18
Partial Dependence Plots.....	19
Appendix.....	19

Experiment Overview

Driverless AI built 1 LightGBMModel to predict DEFAULT_PAYMENT_NEXT_MONTH given 23 original features from the input dataset CreditCard_Cat-train.csv. This classification experiment completed in 1 minutes and 19 seconds (0:01:19), using 20 of the 23 original features, and 4 of the 90 engineered features.

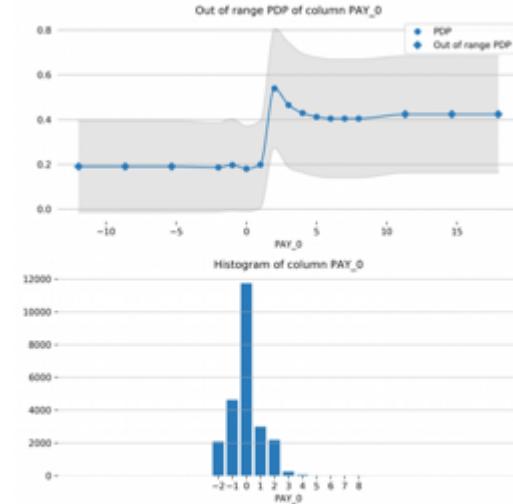
Performance

Dataset	AUC
Provided Validation Data	0.771
Test Data	Test Data not Provided

Driverless Settings

Dial Settings	Description	Setting Value	Range of Possible Values
Accuracy	Controls accuracy	3	1-10

Feature PAY_0



MLI for Time Series

- Time series experiments
 - Test dataset
- Explainability:
 - Original model
 - Global and per-group
 - Forecast horizon
 - Feature importance
 - Per-group
 - Local Shapley values



MLI Cheatsheet

8. START HERE: Train a more interpretable model

INTERPRETABILITY vs. Accuracy / Feature selection / Monotonicity constraints / Nonlinear feature transformations

Ensure reproducibility

Advanced local explanations (containing overall behavior for an individual row)

6. Local Shapley feature importance (accurate local feature importance values for each individual)

Local Shapley values (grey):
 - Similar to global Shapley, but show the individual impact of each feature for each Driverless AI model prediction.
 - Accurate, consistent, and suitable for use in regulated industry.
 - Sum to model prediction.
 - Interactions are built-in.

7. Local linear explanations (Driverless AI model local linear trends around an individual)

Local coefficients (grey):
 - LIME-style coefficients compliant with local trend information.
 - Local Shapley values with local trend information.

8. Local surrogate tree decision path (Effects of an individual to a prediction / Local interaction detection)

Decision path (grey):
 - Shows approximately how new values impact Driverless AI model predictions.
 - Path can show local interactions.

How to explain a model with H2O Driverless AI (and the Kaggle credit card data)

Basic global explanations (containing 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 (consolidated).

3. Partial dependence (change Driverless AI model prediction for different values of the original variables)

Show average predictions (ignores std. and standard deviation of predictions).
 - Sum to model prediction.
 - Interactions are built-in.

4. 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.

5. Global surrogate decision tree (overall structure of the Driverless AI model's decision-making process)

Surrogate decision tree models drivelines 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 errors.

6. Global interpretable model (lower model of Driverless AI model predictions)

Interpretable global linear model (GLM) for 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 actual vs. predicted.

You can stop here OR proceed to more specific Driverless AI model types (e.g., logistic regression, random forest, etc.) to proceed to step 5.

7. Global linear model (lower model of Driverless AI model predictions)

Global linear model (GLM) for Driverless AI predictions (blue, decreasing, middle):
 - Ranks Driverless AI predictions from lowest (bottom left) to highest (top right).
 - Quantifies nonlinearity of Driverless AI model.

8. Global quadratic model (lower model of Driverless AI model predictions)

Global quadratic model (GQM) for Driverless AI predictions (orange, U-shaped, middle):
 - Ranks Driverless AI predictions from lowest (bottom left) to highest (top right).
 - Quantifies nonlinearity of Driverless AI model.

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.

10. Local original feature importance (original features that drive a prediction for an individual)

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

Continuous interaction terms are highlighted in yellow.

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

MLI Functional Architecture

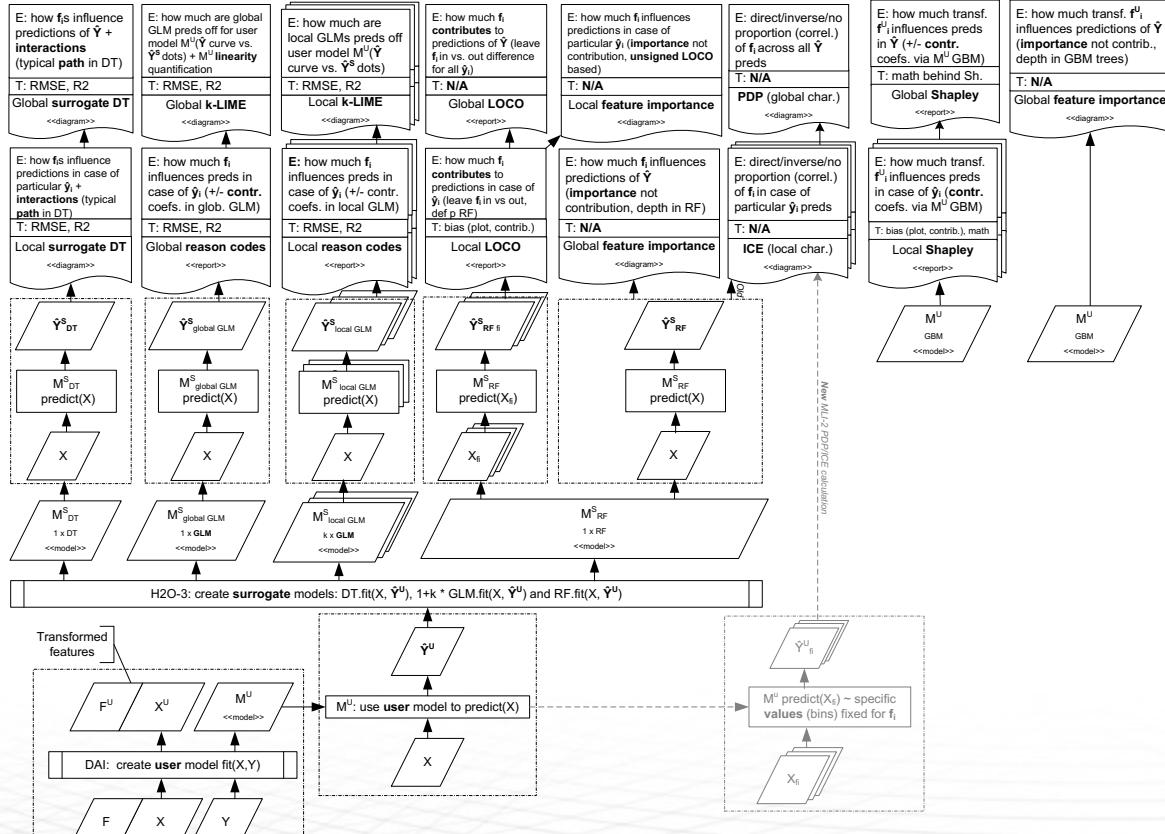


Figure: MLI-2 functional architecture flow diagram

Conclusion

Takeaways

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.

MLI TEAM



Patrick



Navdeep



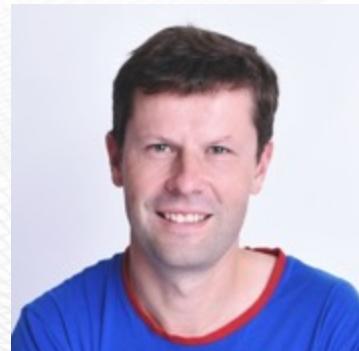
Mateusz



Zac



Laco



Martin

H₂O.ai

Thank you!

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](#)

