MOODY'S ANALYTICS

SUMMIT 2019

Machine Learning Interpretability Techniques in Credit Risk Modeling

November 2019

Craig Peters
Sr Dir-Research Model Validation

SUMMIT 2019

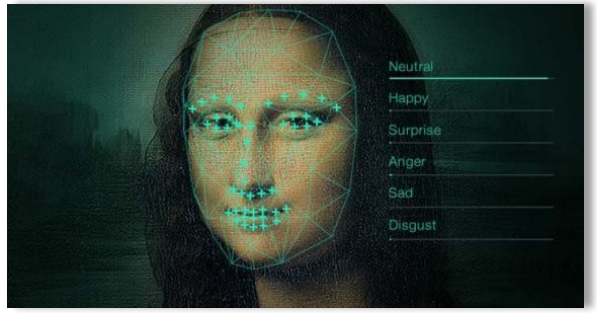
Craig Peters

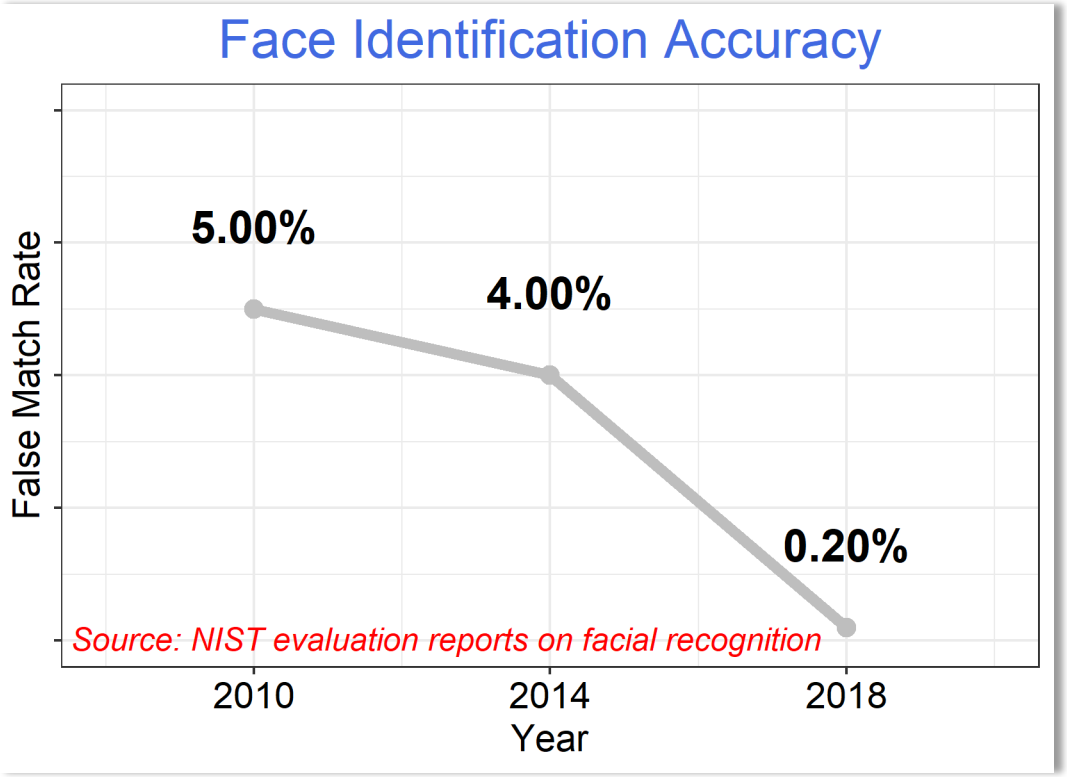
Senior Director - Research Model Validation

Moody's Analytics

Research Motivation

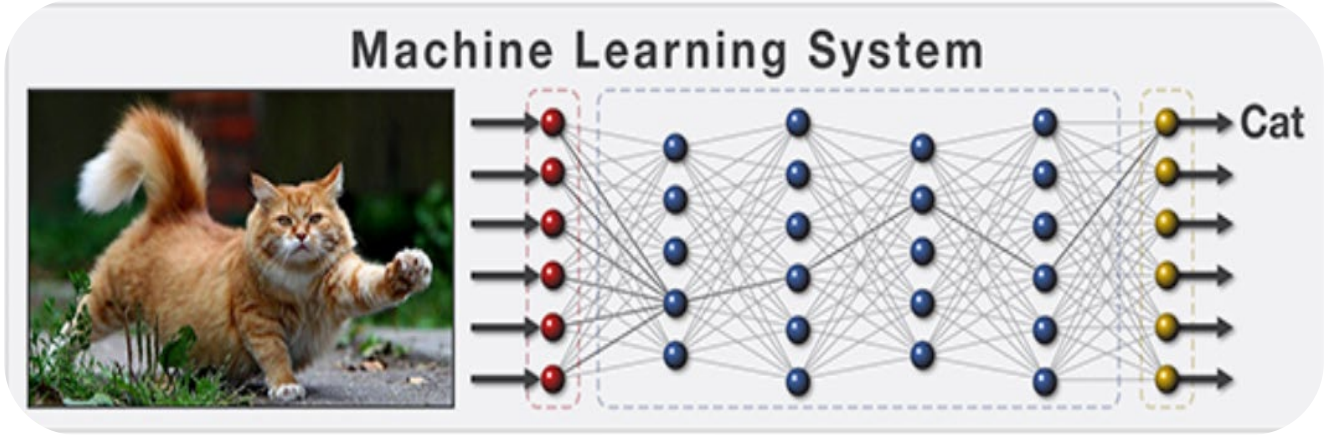
ML algorithms improve prediction accuracy over traditional statistical models





Research Motivation

ML algorithms are often criticized as black-box models



Agenda

1. Problem Setting
   - Dataset
   - Generalized Additive Model (GAM) vs XGBoost (XGB)
2. Global Interpretability
   - Feature Importance
   - Feature Effect
   - Feature Interaction
   - Alternate GAM Model
3. Local Interpretability
   - Local Interpretable Model-agnostic Explanation (LIME)
   - Shapley value
4. Take-aways & Questions

1

Problem Setting

A Probability of Default Model

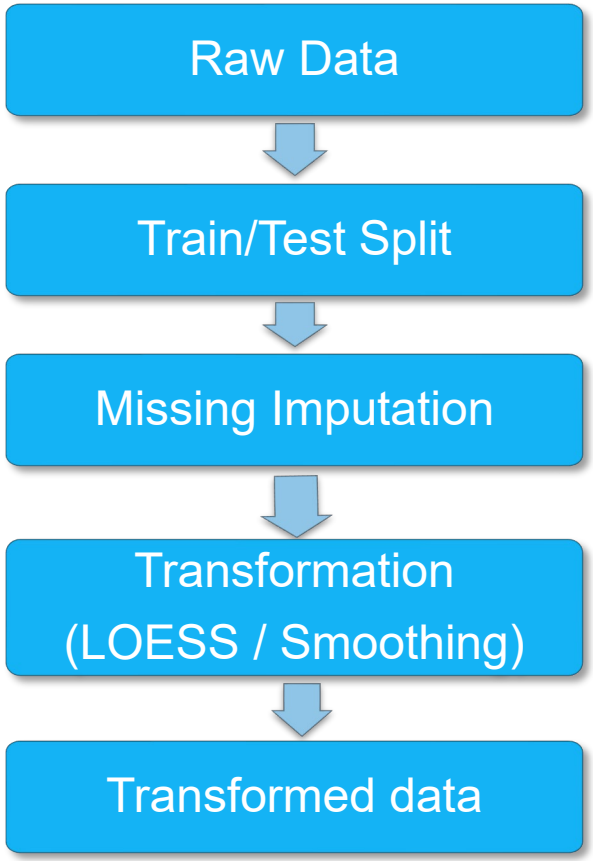
Problem Setting
Dataset

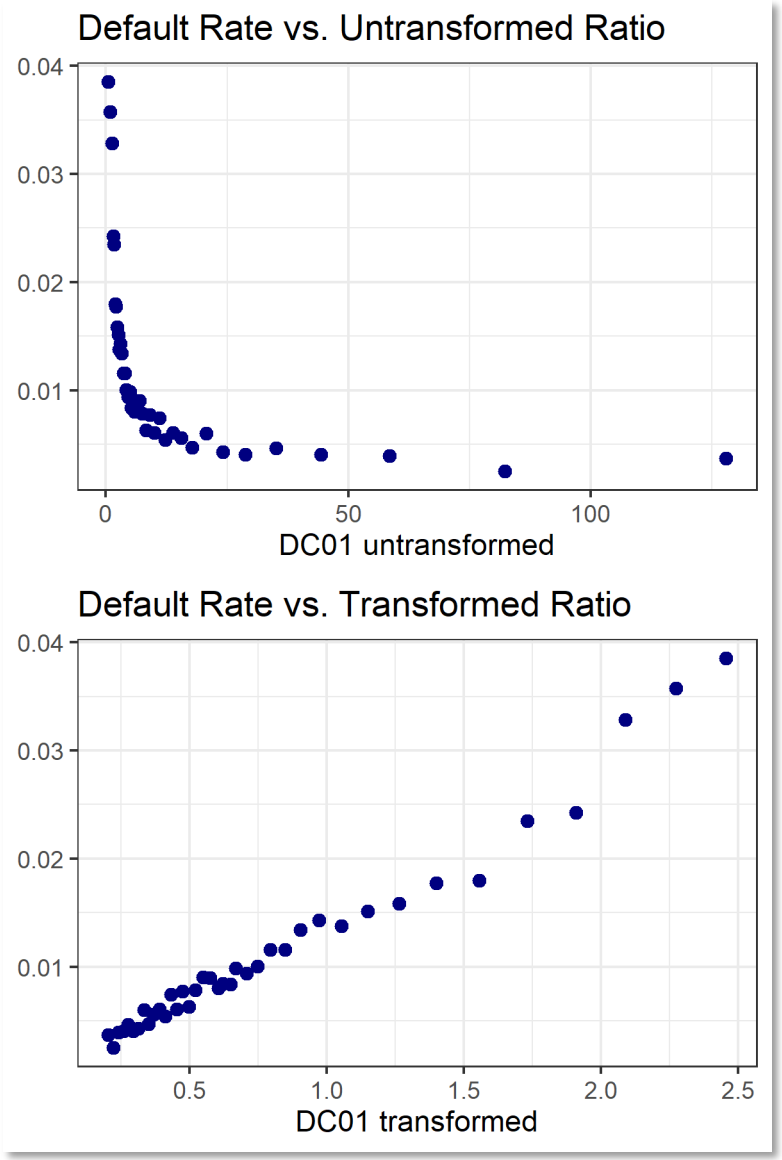
Category	Ratio Name	Ratio Description
Activity	A03	Inventories to Sales
	A08	Current Liabilities to Sales
	A18	Change in Working Capital over Sales
Debt Coverage	DC01**	EBITDA over Interest Expense
Growth	GROW01**	Sales Growth: Sales(t)/Sales(t-1) – 1
	GROW04	Change in ROA
Leverage	LEV12**	Retained Earnings to Current Liabilities
	LEV13**	LT Debt to (LT Debt plus Net worth)
Liquidity	LIQ05**	Cash and Marketable Securities to Total Assets
Profitability	PFT01**	ROA/ Net Income to Total Assets
Size	SIZE01**	Total Assets
Sector	SECTOR	14 Sectors
DUMDEF	PD	Default flag (1=default)

** feature of interest (to be covered later)
** important features (to be covered later)

Problem Setting

Data Processing





Problem Setting

Methodology and Results

SUMMIT 2019

GAM

A generalized linear model (GLM) on transformed predictors $(T_i(x_i))$

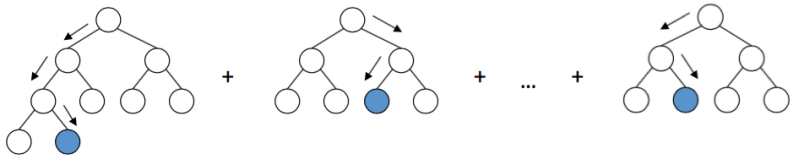
$T_i(x_i)$: Loess transformation

$$PD = \phi(\beta_0 + \beta_1 T_1(x_1) + \cdots + \beta_N T_N(x_N))$$

AR: 0.579 (train)
0.575 (test)

XGB

Ensemble tree methodology involving both bagging and boosting



AR: 0.700 (train)
0.605 (test)

3 pts.

2 Global Interpretability

Global Interpretability

Feature Importance

Permutation Test

Feature Effects

Partial Dependence Plots (PDP)

Accumulated Local Effects (ALE)

Feature Interaction

Friedman's H-statistic

Alternate GAM model

Splines

Interactions

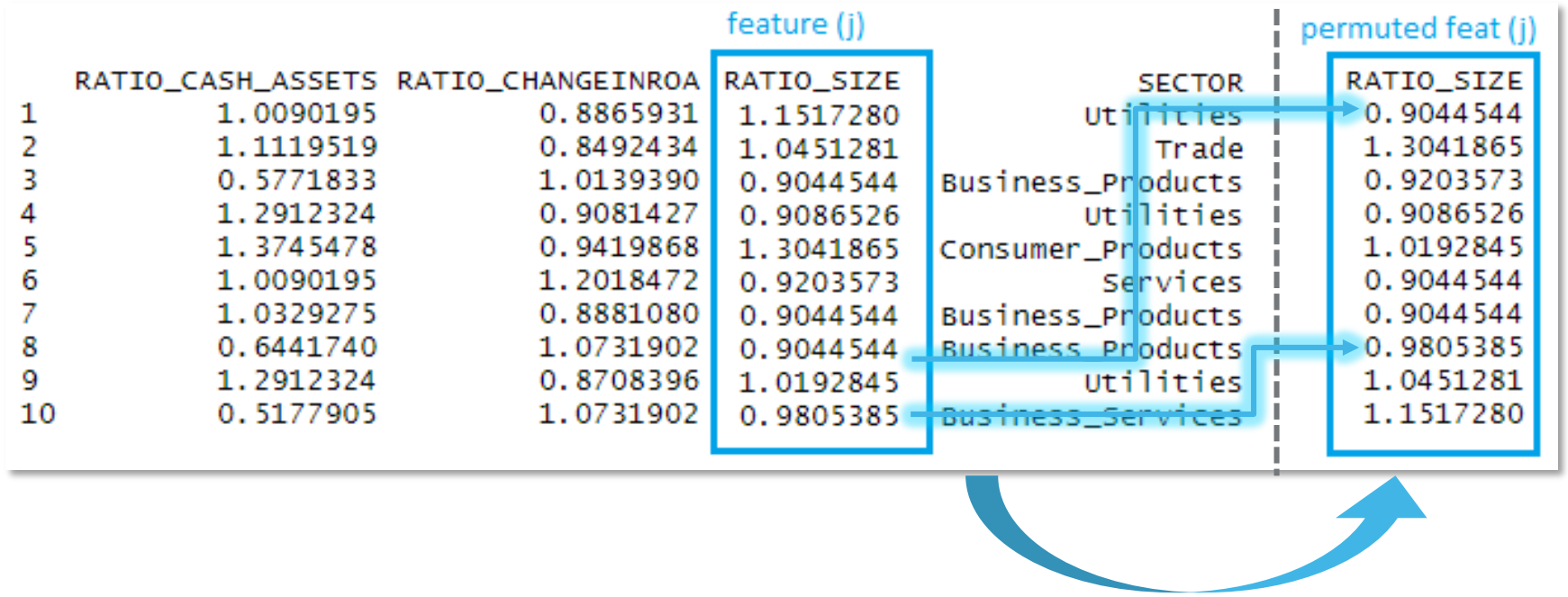
Global Interpretability

Feature Importance

Permutation Test

Permute feature(s) → **Change in AR** → **Rank features**

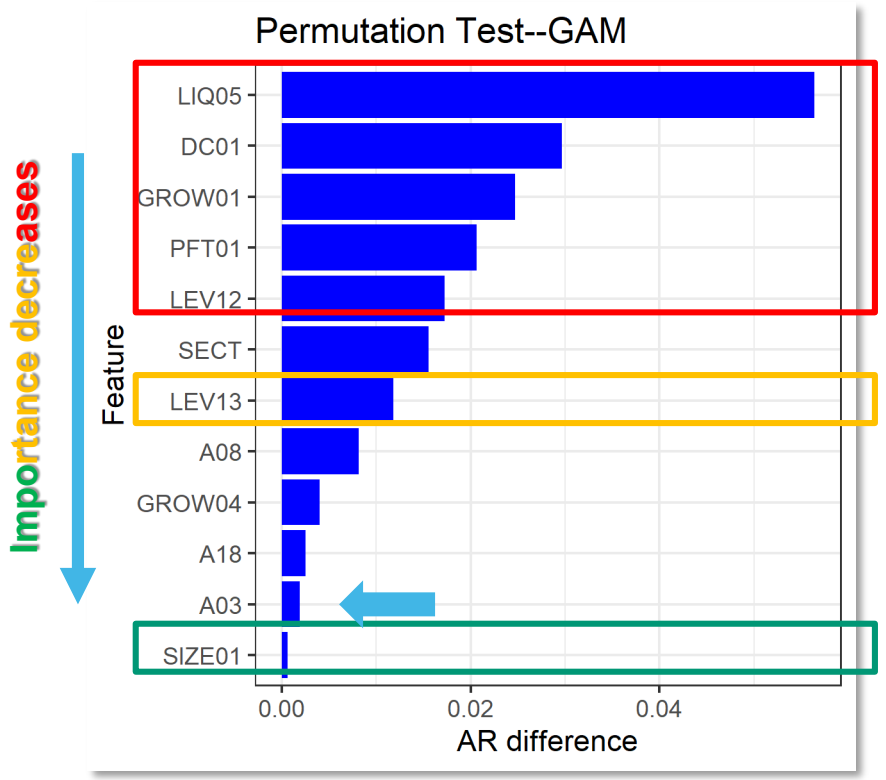
Most important features produce the **largest difference** in AR

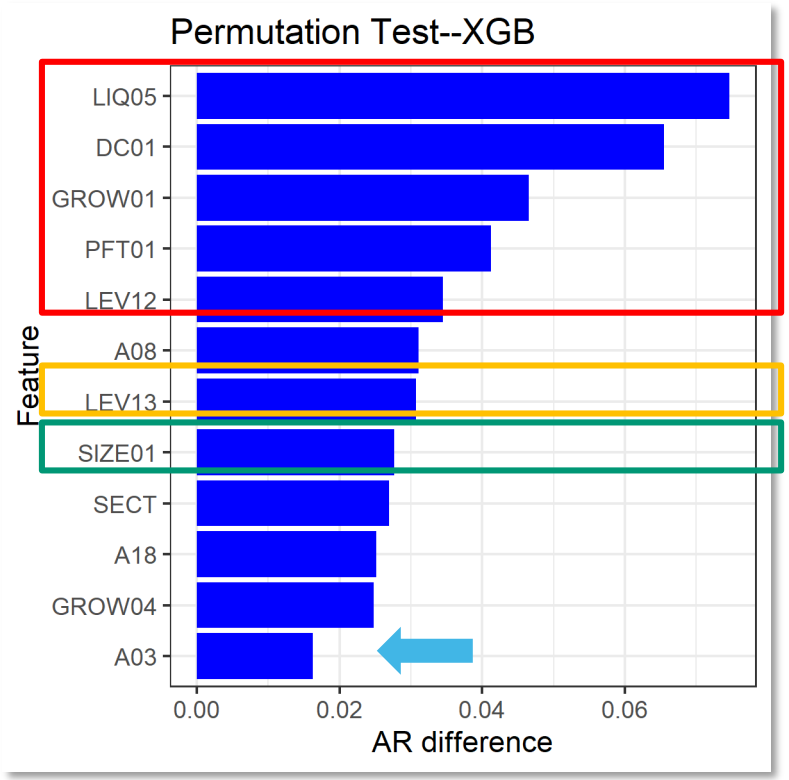


	RATIO_CASH_ASSETS	RATIO_CHANGEINROA	RATIO_SIZE	SECTOR	RATIO_SIZE
1	1.0090195	0.8865931	1.1517280	Utilities	0.9044544
2	1.1119519	0.8492434	1.0451281	Trade	1.3041865
3	0.5771833	1.0139390	0.9044544	Business_Products	0.9203573
4	1.2912324	0.9081427	0.9086526	Utilities	0.9086526
5	1.3745478	0.9419868	1.3041865	Consumer_Products	1.0192845
6	1.0090195	1.2018472	0.9203573	Services	0.9044544
7	1.0329275	0.8881080	0.9044544	Business_Products	0.9044544
8	0.6441740	1.0731902	0.9044544	Business_Products	0.9805385
9	1.2912324	0.8708396	1.0192845	Utilities	1.0451281
10	0.5177905	1.0731902	0.9805385	Business_Services	1.1517280

Global Interpretability

Feature Importance: Permutation Test



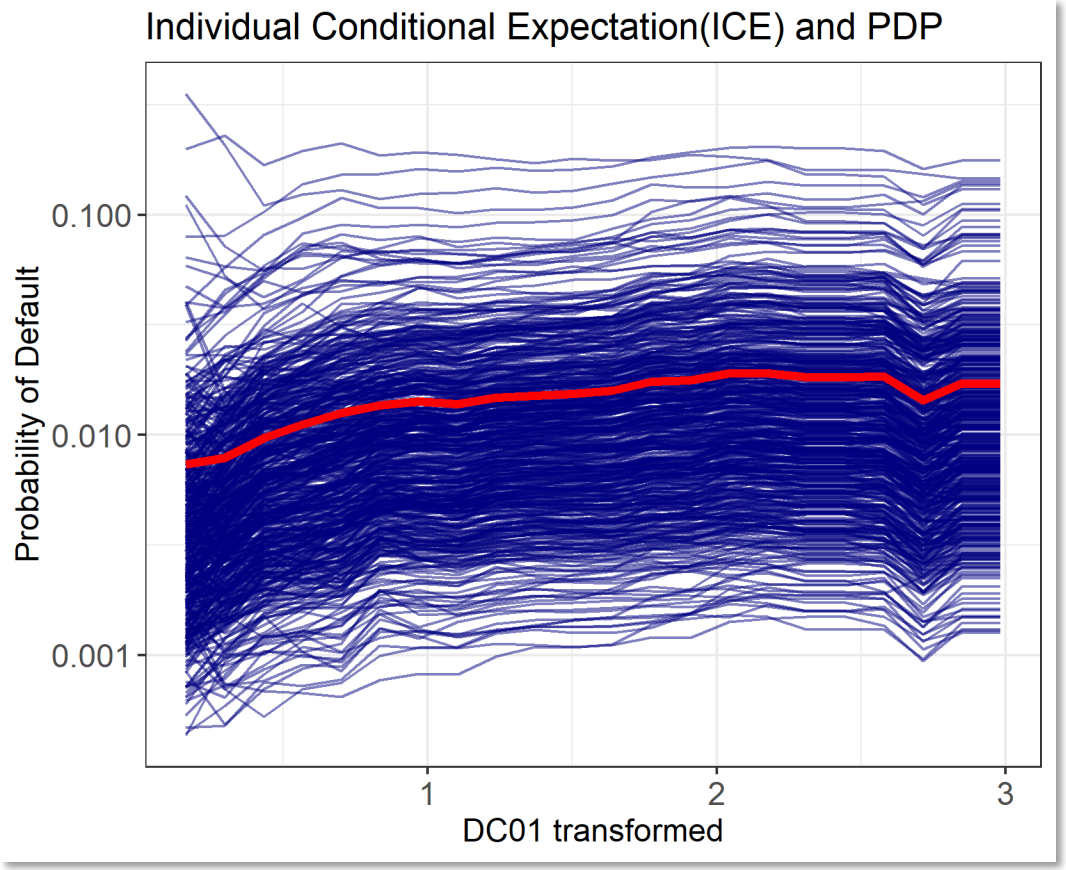


- The **top 5 important features** (**LIQ05, DC01, GROW01**, . .) are the **same**
- **SIZE01** becomes **more important** in **XGB vs GAM**
- **More area** is covered by bar chart in **XGB vs GAM**

Global Interpretability

Feature Effects: Partial Dependence Plot (PDP)

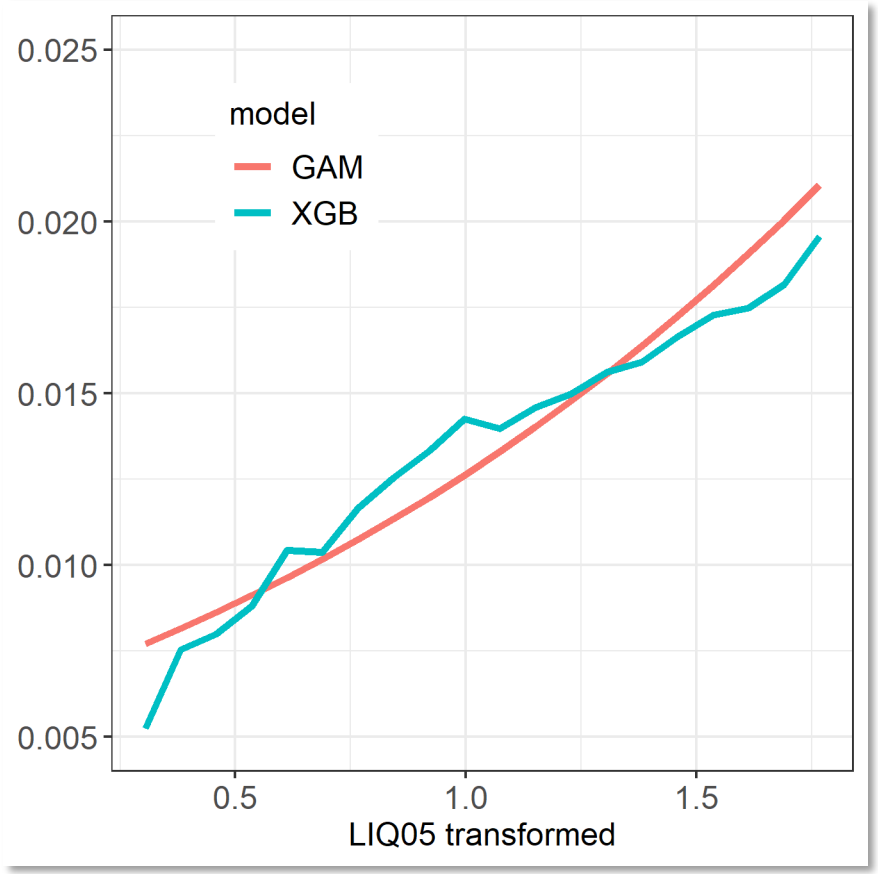
PDP shows the marginal/**partial effect** of feature(s) on the predicted outcome.

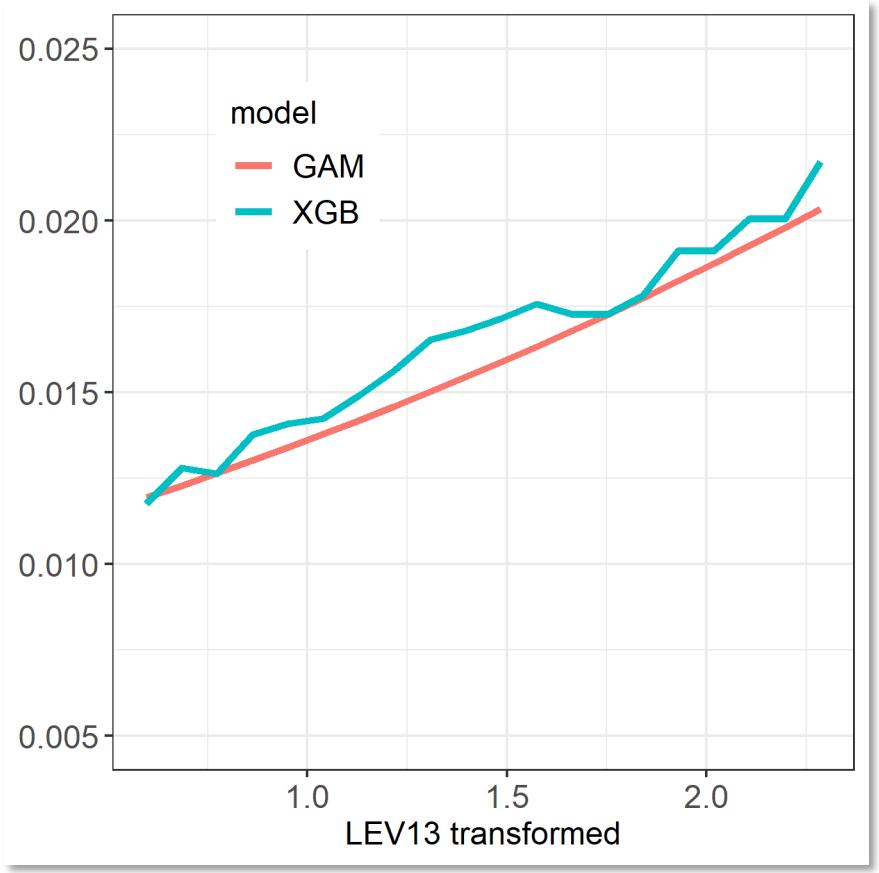


Global Interpretability

Feature Effects: Partial Dependence Plot (PDP)

Analogous behavior (GAM and XGB)



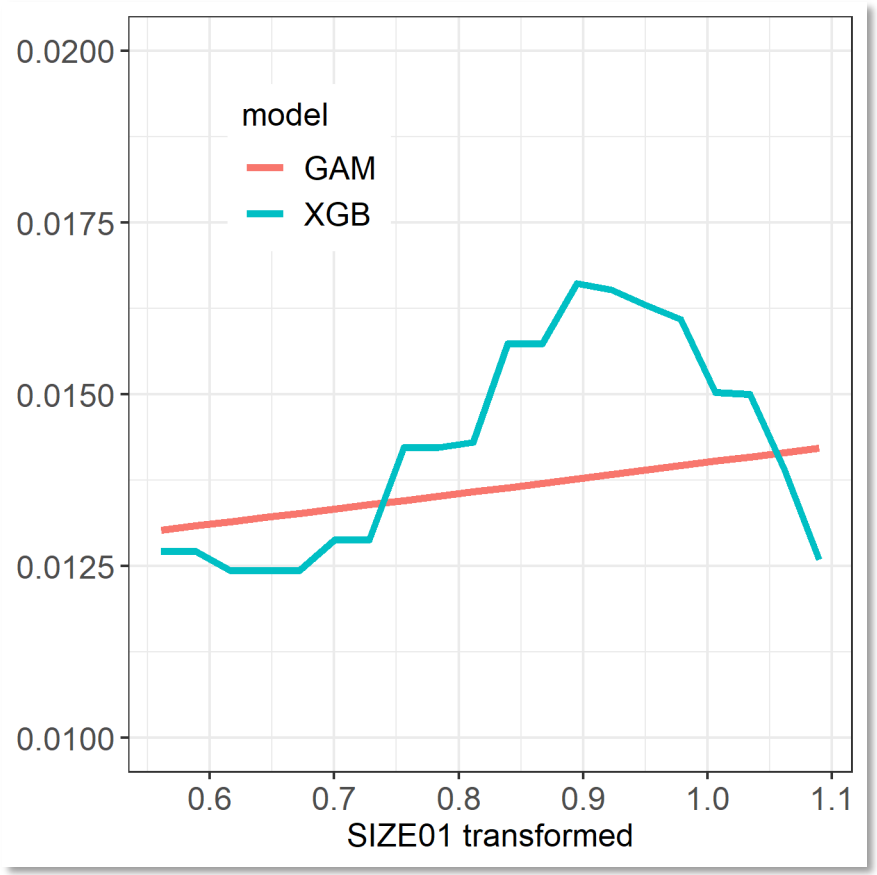


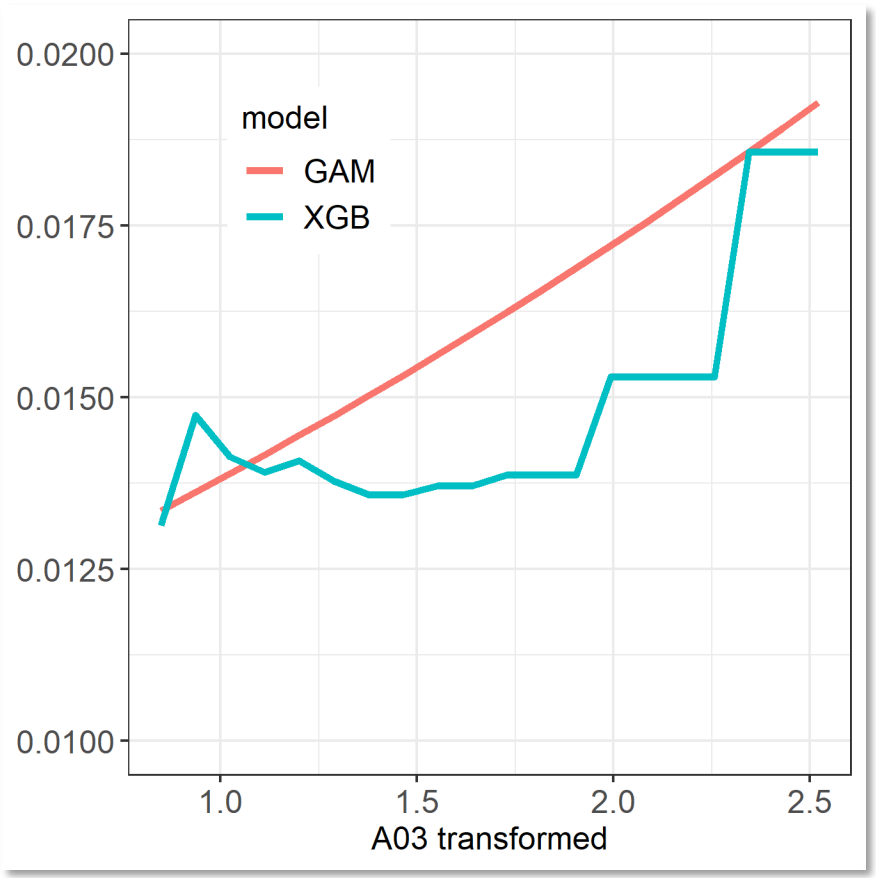
LIQ05, LEV13 were among the **top important** (also common) features for both GAM and XGB

Global Interpretability

Feature Effects: Partial Dependence Plot (PDP)

Non-analogous behavior (GAM vs XGB)





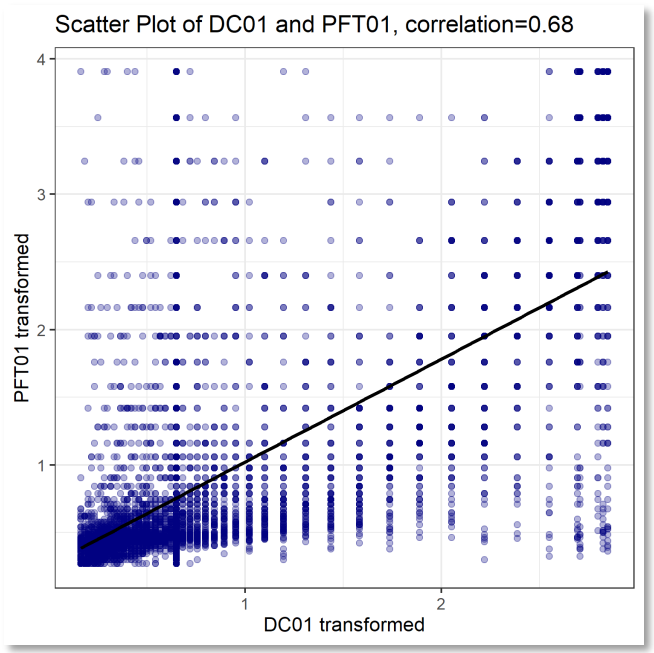
Size01 becomes more important in XGB, and **A03** has higher AR drop from permutation test in XGB

Global Interpretability

Feature Effects: Accumulated Local Effect (ALE)

PDP ignores correlations among features

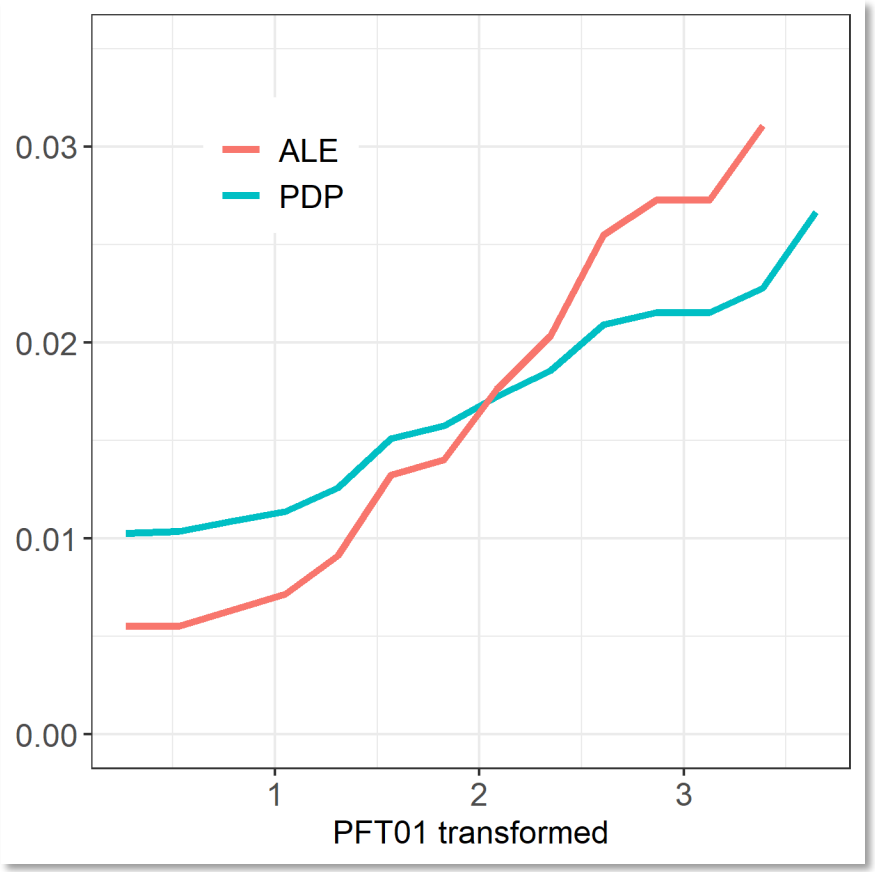
ALE solves this problem



ALE	PDP
1. Conditional distribution	1. Marginal distribution
2. Considers correlation of features	2. Less informative if features are correlated
3. Global view of sum of local effects	3. Global view of global effects

Global Interpretability

Feature Effects: PDP vs. ALE--XGB



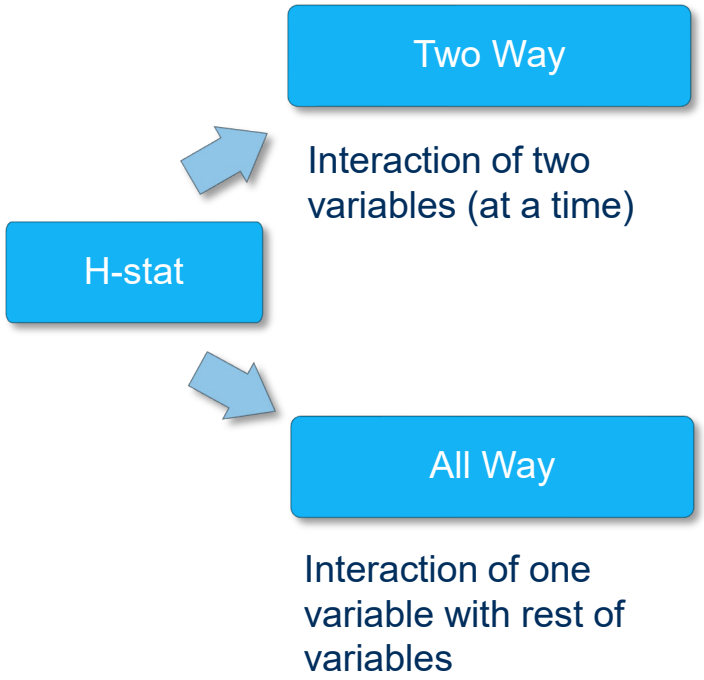
PDP and ALE show different effects of average PD changes in response to changes in **PFT01**

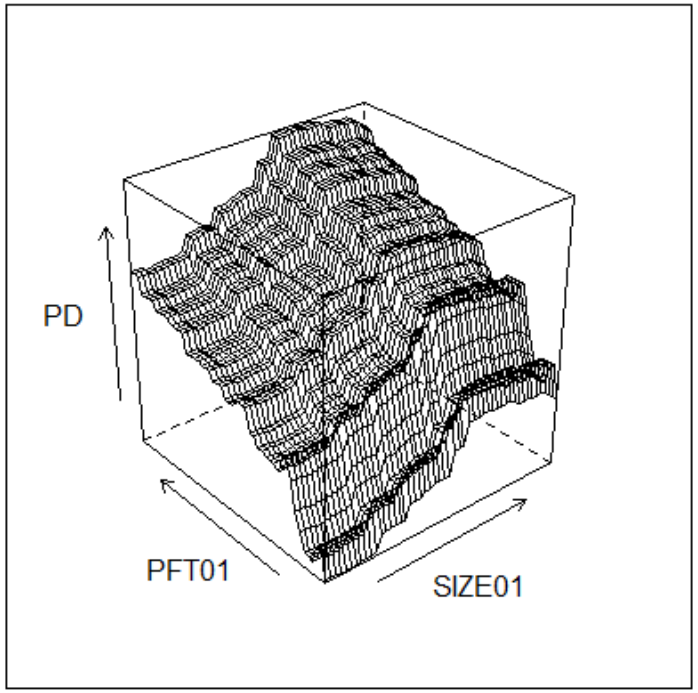
Global Interpretability

Feature Interaction

SUMMIT 2019

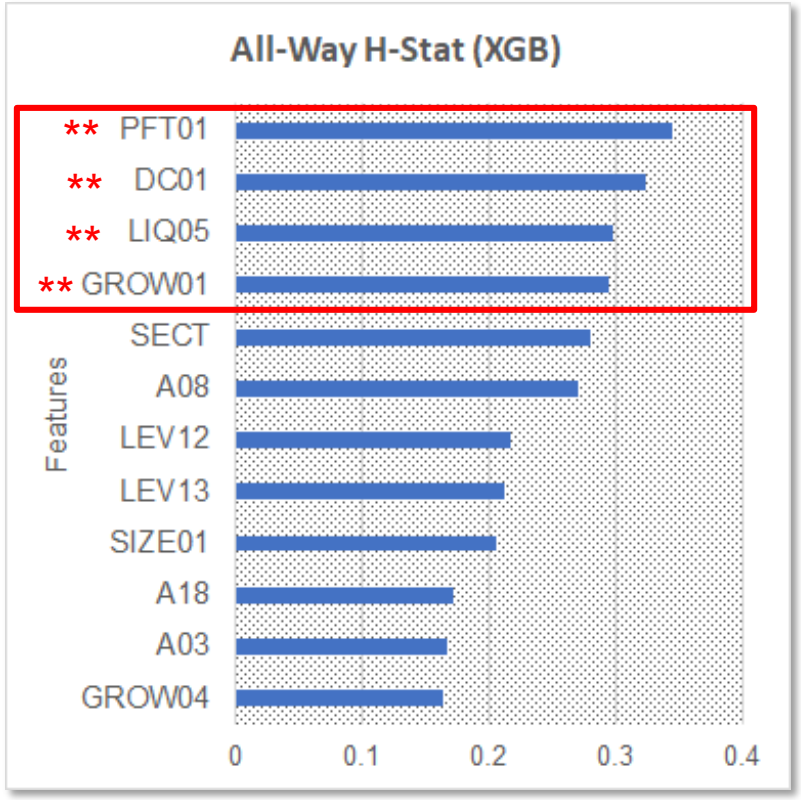
[Friedman's H-statistic]

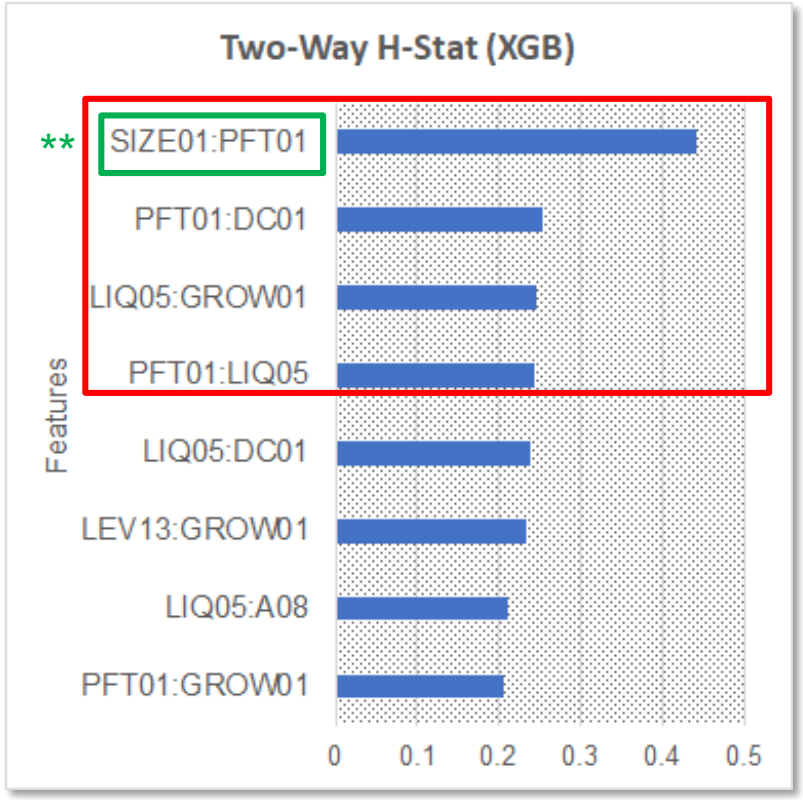




Global Interpretability

Feature Interaction: Friedman's H-statistic

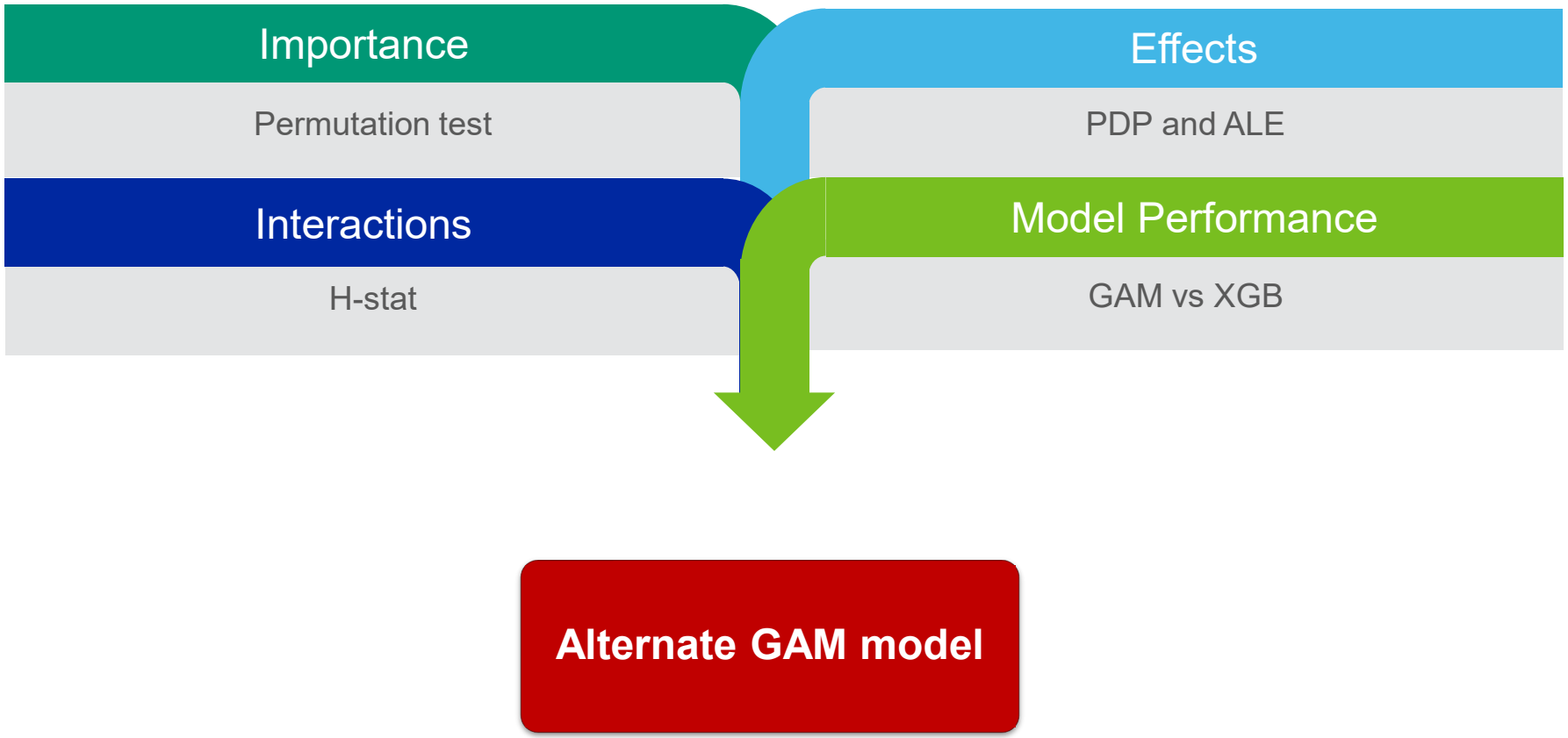




- All Way: Strong interaction **of PFT01, DC01, LIQ05, GROW01** with rest of variables
- Two way: Pairwise **SIZE01:PFT01, PFT01:DC01, LIQ05:GROW01. . .** strong interaction observed

Global Interpretability

Alternate GAM Model

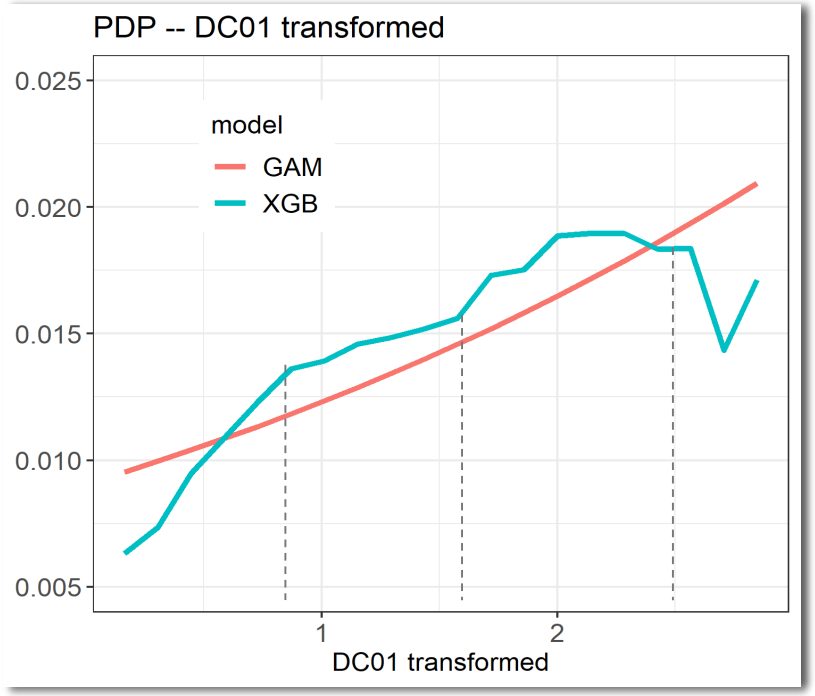


Global Interpretability

Alternate GAM Model: Non-linearities

Original GAM

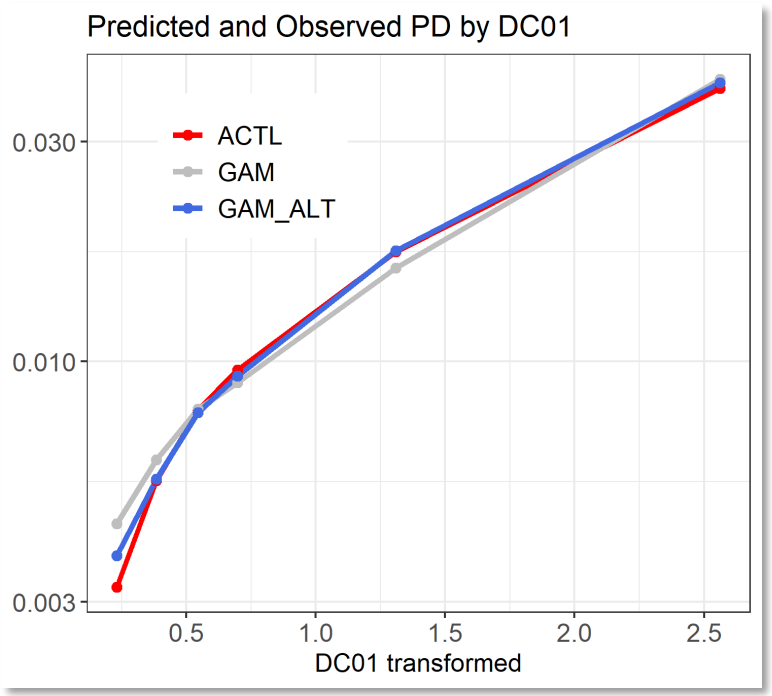
AR: 0.579 (train)
0.575 (test)



Original GAM + Splines

AR: 0.589 (train)
0.584 (test)

1 pts.



Global Interpretability

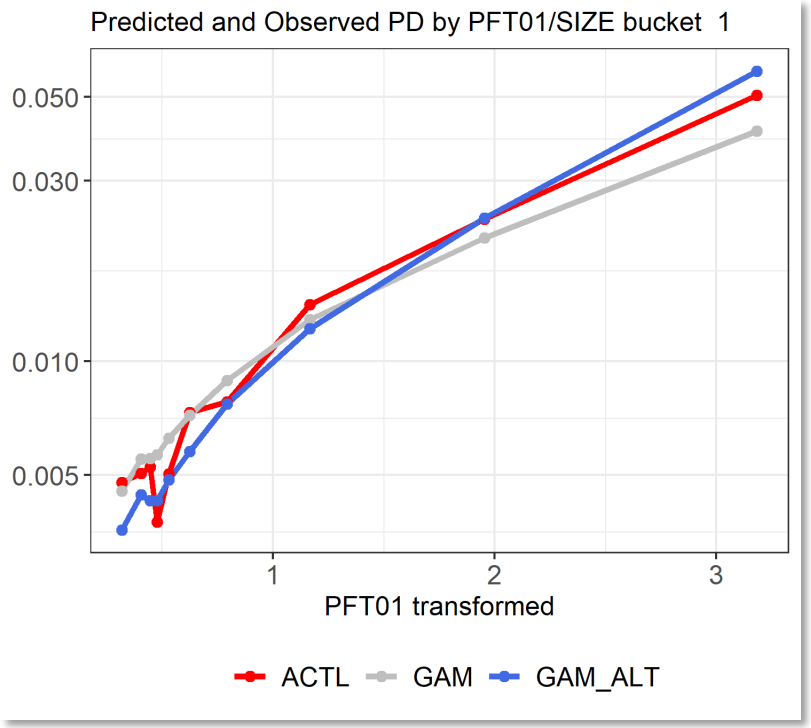
Alternate GAM Model: Interactions

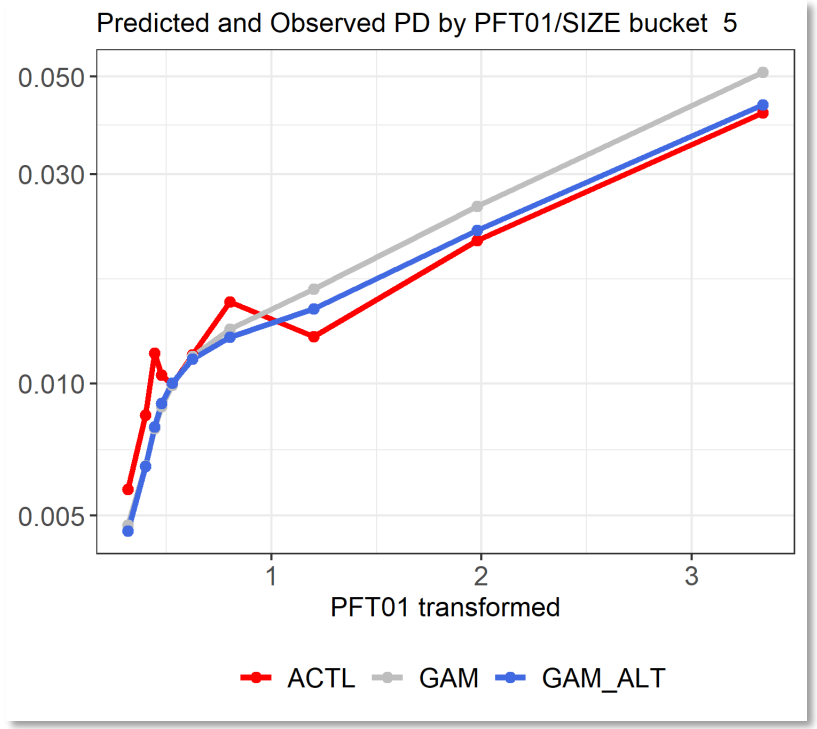
Original GAM → Original GAM + Splines → Original GAM + Splines + Interactions

AR: 0.579 (train)
0.575 (test)

AR: 0.589 (train)
0.584 (test) ⬆ **1 pts.**

AR: 0.594 (train)
0.589 (test) ⬆ **1.5 pts.**





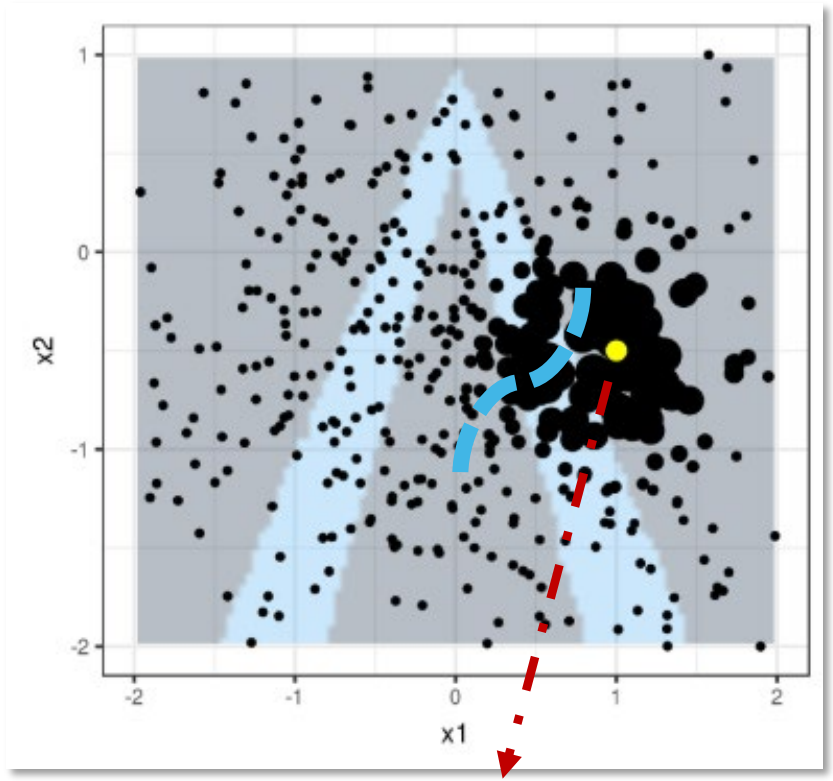
3

Local Interpretability

Local Interpretability

Local Interpretable Model-agnostic Explanations (LIME)

- **Simulate** points near specific observation
- **Generate** model predictions at these points
- **Use** model predictions as Y variable
- **Weight** new observations by proximity
- **Build** weighted linear regression (or other interpretable model)
- **Interpret** the local surrogate model



Molnar, C. (2018)

Advantage:

- Conceptually Intuitive
- Easy to interpret

Disadvantage:

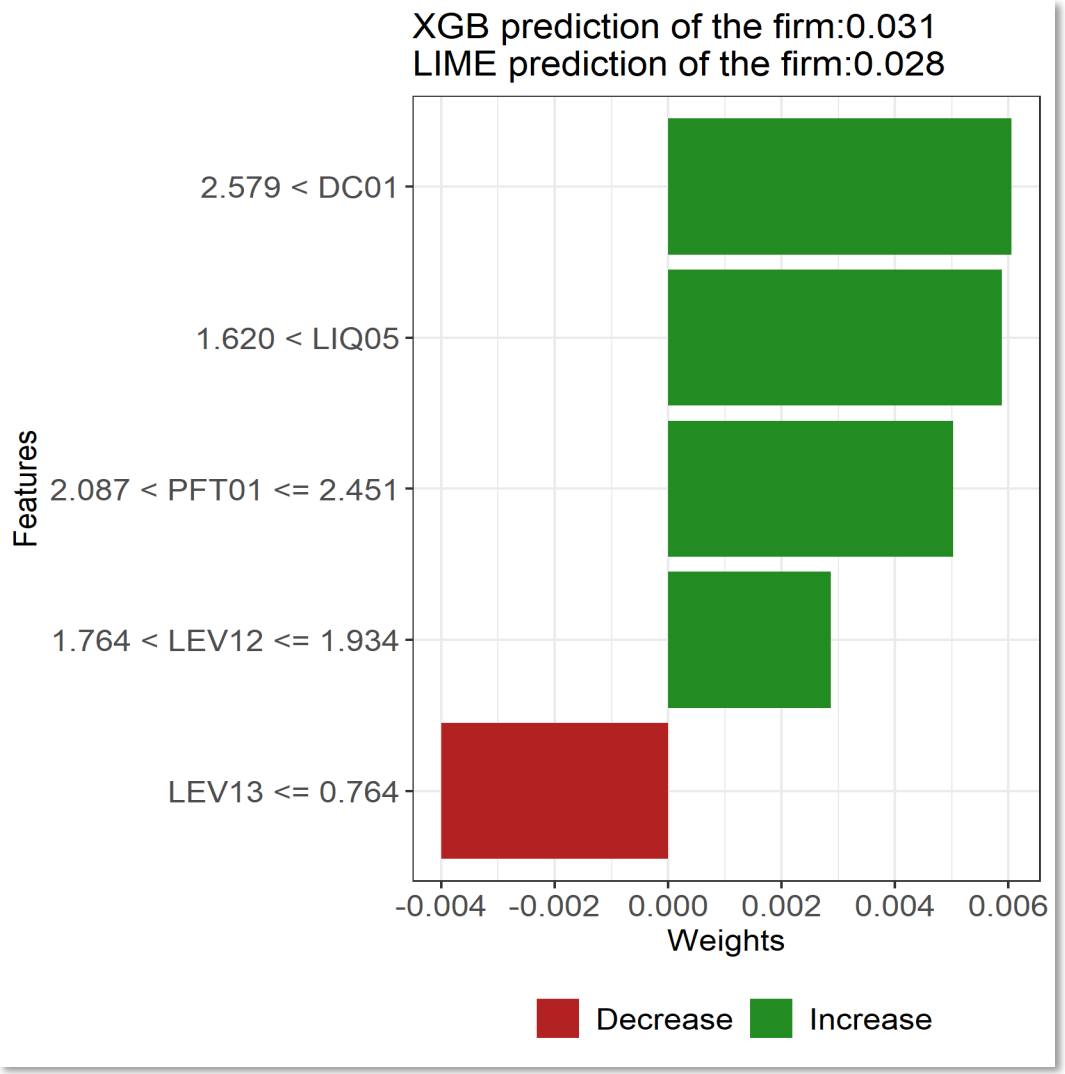
- Simulating “good” nearby points
- Unstable results

Local Interpretability

LIME Example

Firm Profile

Ratio	Value
A03	0.85
A08	1.04
A18	1.29
DC01	**2.69**
GROW01	0.95
GROW04	1.12
LEV12	**1.86**
LEV13	0.63
LIQ05	**1.77**
PFT01	**2.40**
SIZE01	1.05
SECTOR	Business Services



Local Interpretability

Shapley Value

Originally from game theory to attribute **the value of a team effort to individual members**

Unlike LIME, uses the **same original model** in a local space.

Explains:

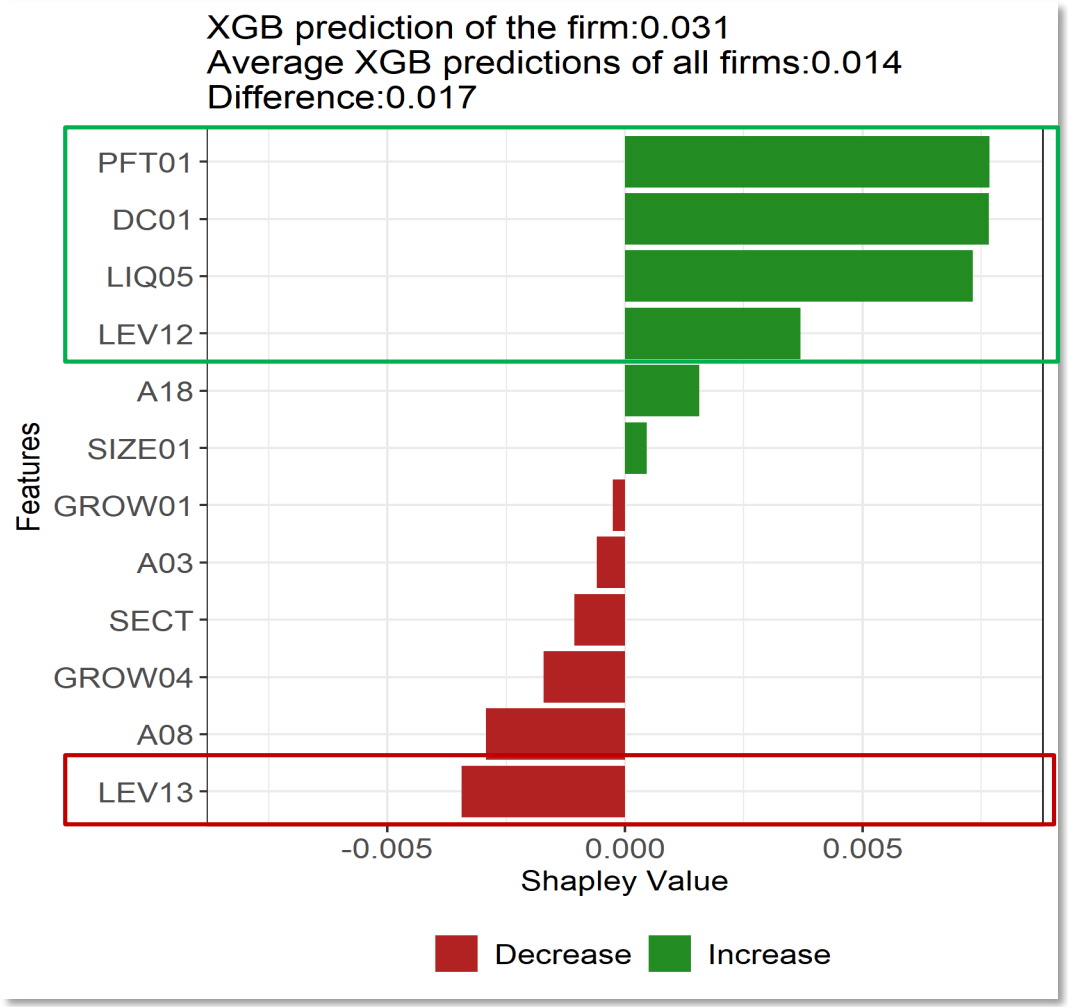
- **Individual vs. Average PD**
- **Feature contribution** towards the difference

Local Interpretability

Shapley Value Example

Firm Profile

Ratio	Value
A03	0.85
A08	1.04
A18	1.29
DC01	**2.69**
GROW01	0.95
GROW04	1.12
LEV12	**1.86**
LEV13	0.63
LIQ05	**1.77**
PFT01	**2.40**
SIZE01	1.05
SECTOR	Business Services



4 Take-aways

Take-aways

1. Interpretability techniques can help explain and predict black box model output
2. Model-Agnostic methods can be applied to any model enabling a broader range of methodologies
3. Interpretability techniques can help make today's black boxes tomorrow's interpretable models
4. The tradeoff between interpretability and accuracy is real and can only be mitigated

References

Permutation feature importance
Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Partial Dependency Plot (PDP)

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Accumulated Local Effect (ALE)

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.

Friedman's H-statistics

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916-954.

Local interpretable model-agnostic explanations (LIME)

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM.

Shapley Value

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*(3), 647-665.

Lundberg, S., & Lee, S. I. (2016). An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*.

Machine Learning Interpretability with H2O Driverless AI (K-LIME)

Patrick Hall, Navdeep Gill, Megan Kurka, & Wen Phan, Edited by: Angela Bartz

Discussions of Machine Learning Interpretation framework and taxonomy

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80-89). IEEE.

Molnar, C. (2018). Interpretable machine learning: A guide for making black box models explainable. *E-book at< https://christophm. github. io/interpretable-ml-book/>, version dated, 10.*

Appendix

Appendix

Partial Dependence Plots

- The partial dependence function is defined as:

integrate over all x_C

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

x_S
feature(s) of interest

x_C
other features used in the model

Accumulated Local Effects

$$\hat{f}_{x_S,ALE}(x_S) = \int_{z_{0,1}}^{x_S} E_{X_C|X_S}\left[\hat{f}^S(X_s, X_c)|X_S = z_S\right] dz_S = \int_{z_{0,1}}^{x_S}\int_{x_C} \hat{f}^S(z_s, x_c)\mathbb{P}(x_C|z_S) dx_C dz_S$$

Conditional distribution

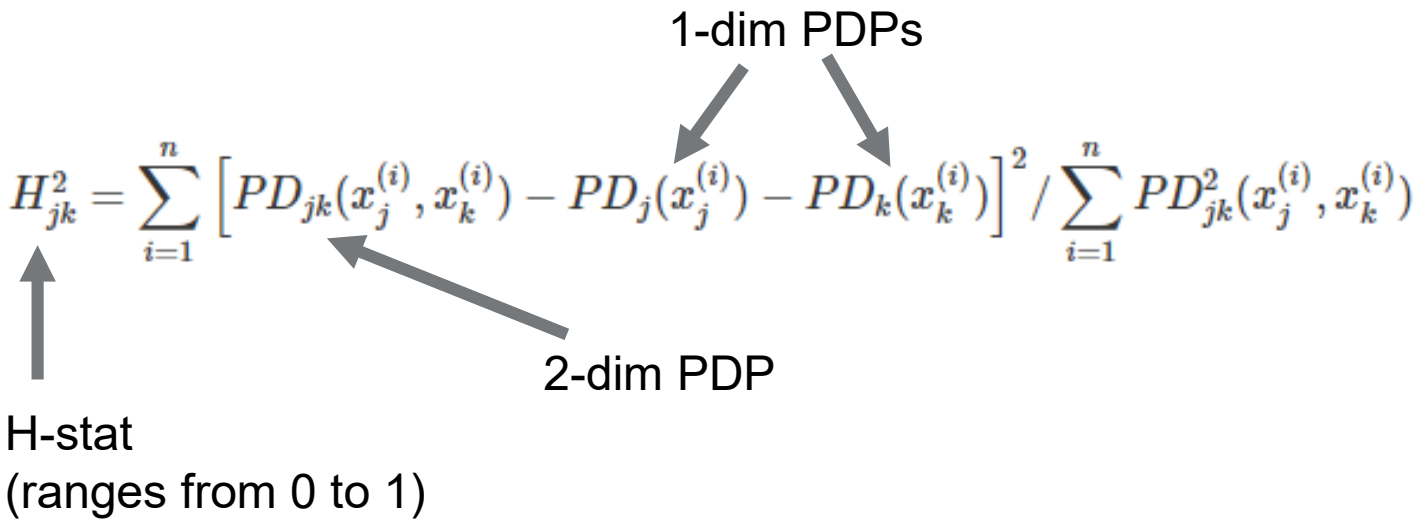
where: $\hat{f}^S(x_s, x_c) = \frac{\delta \hat{f}(x_S, x_C)}{\delta x_S}$

Differential/change in PD

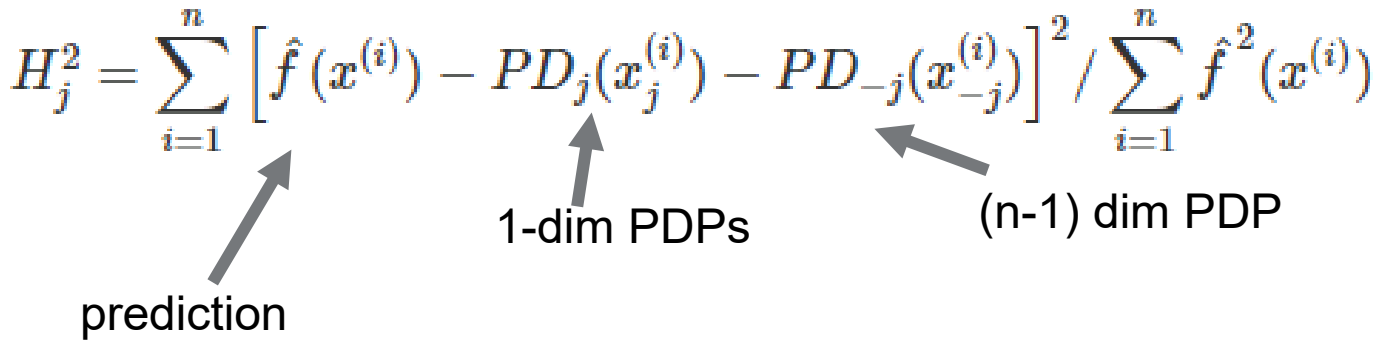
Appendix

H-statistic

- Two Way:



- All Way:



MOODY'S ANALYTICS

Craig Peters
Sr Dir-Resrch Model Validation
Craig.Peters@moodys.com