OPTIMAL RULE-FIT ALGORITHM (ORFA)

Machine Learning Under an Optimization Lens

Ryan Lucas & Paul Roeseler

MIT SLOAN SCHOOL OF MANAGEMENT

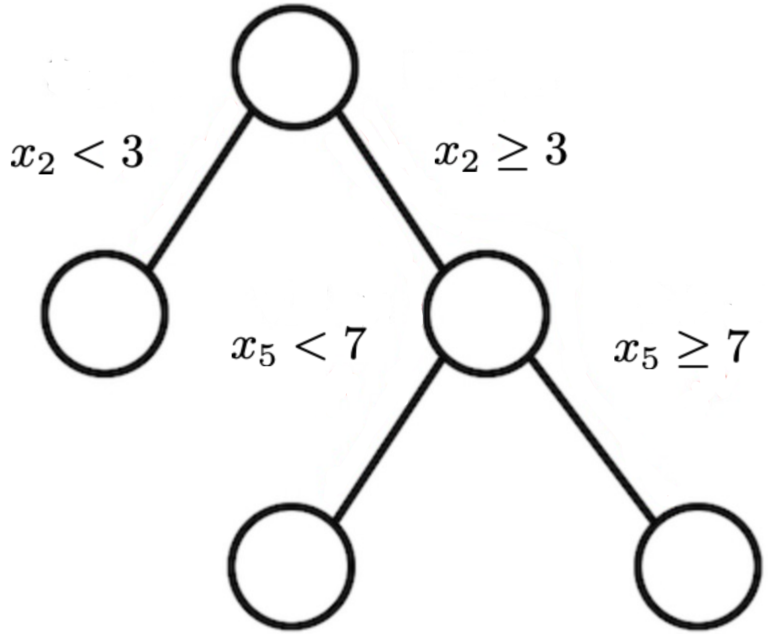
MIT MANAGEMENT SLOAN SCHOOL

MOTIVATION

Decision trees and linear models uncover different types of effects

Decision trees

- Uncover interaction effects



Linear models

- Uncover linear relationships

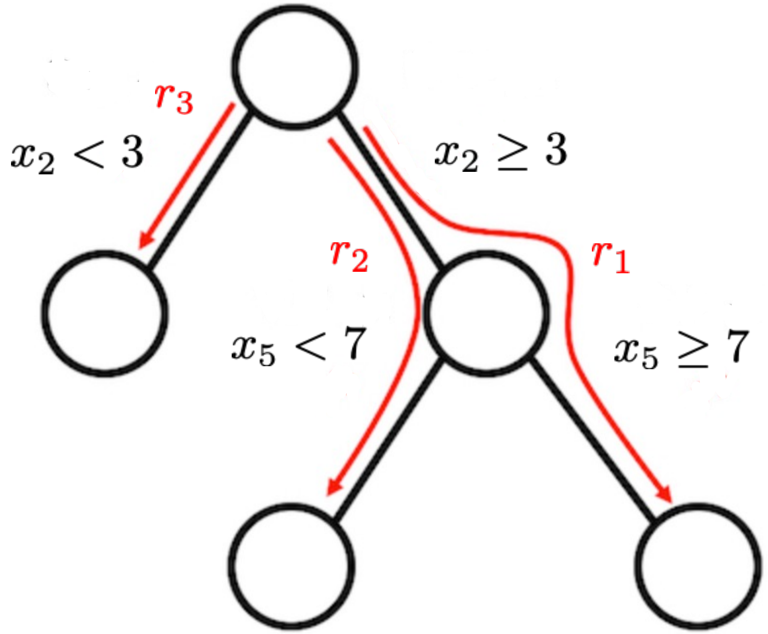
$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

But what if both types of effects are present?

MOTIVATION

RuleFit algorithm (Friedman and Popescu, 2008)

Interpretable machine learning method using rules from a decision tree as features for a linear regression model



$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\delta}_1(\mathbb{1}\{x_2 \geq 3\} \cdot \{x_5 \geq 7\})$$

RuleFit adds rules as interaction features…

MAJOR DRAWBACK

Greedy tree building methods (e.g., CART) require many splits to achieve strong performance – leads to great number of rules and overly sparse features

Few/Short Rules ↓ Interpretability

Trade-off

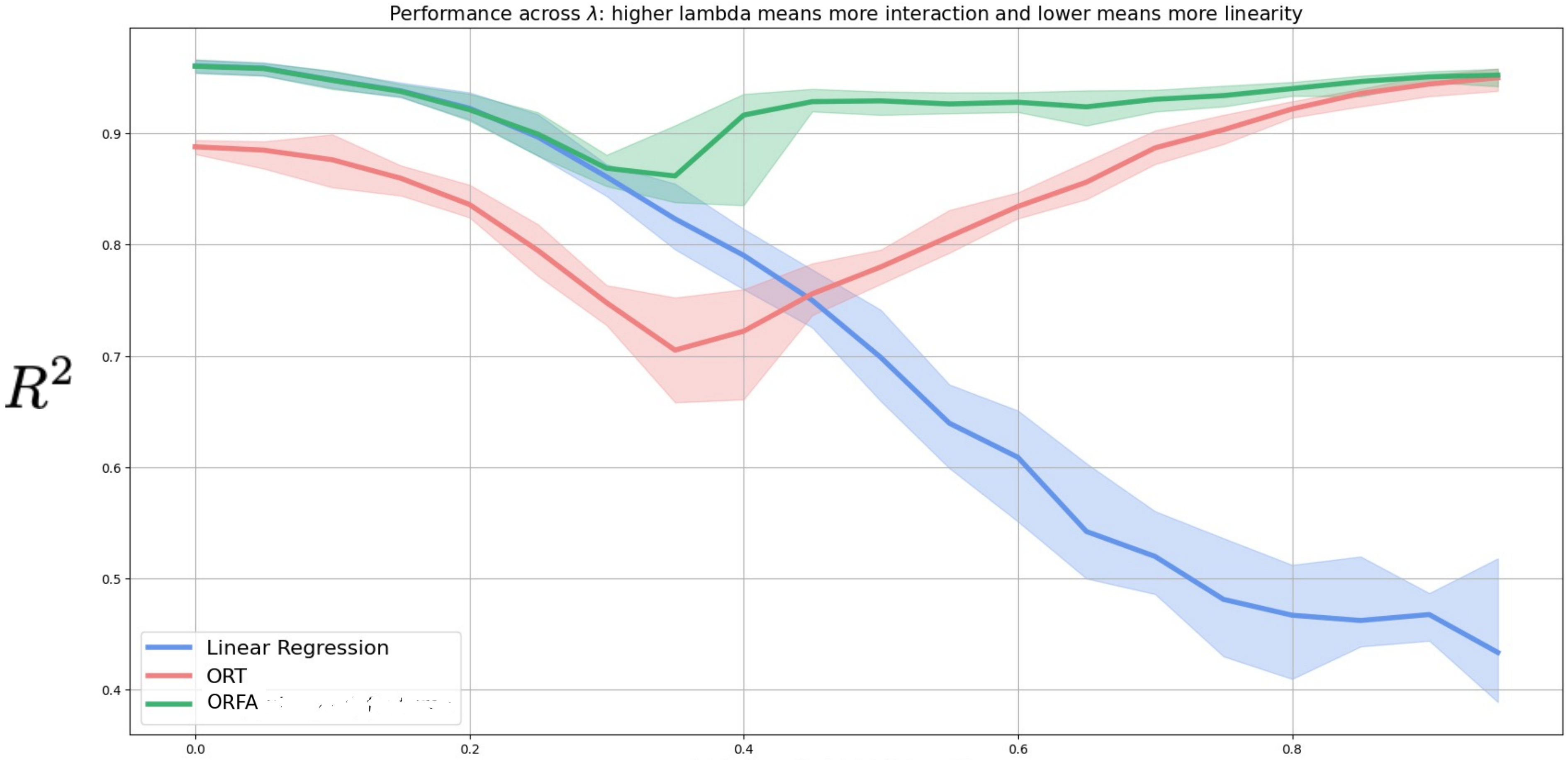
Many/Long Rules ↓ High Performance

Optimal Regression Trees (ORTs) uncover true interaction effects in efficient number of splits and require only a single tree, resulting in fewer, more interpretable rules.

We propose An Optimal RuleFit Algorithm (ORFA), combining ORTs and Linear Regression in a similar fashion to RuleFit

SIMULATIONS

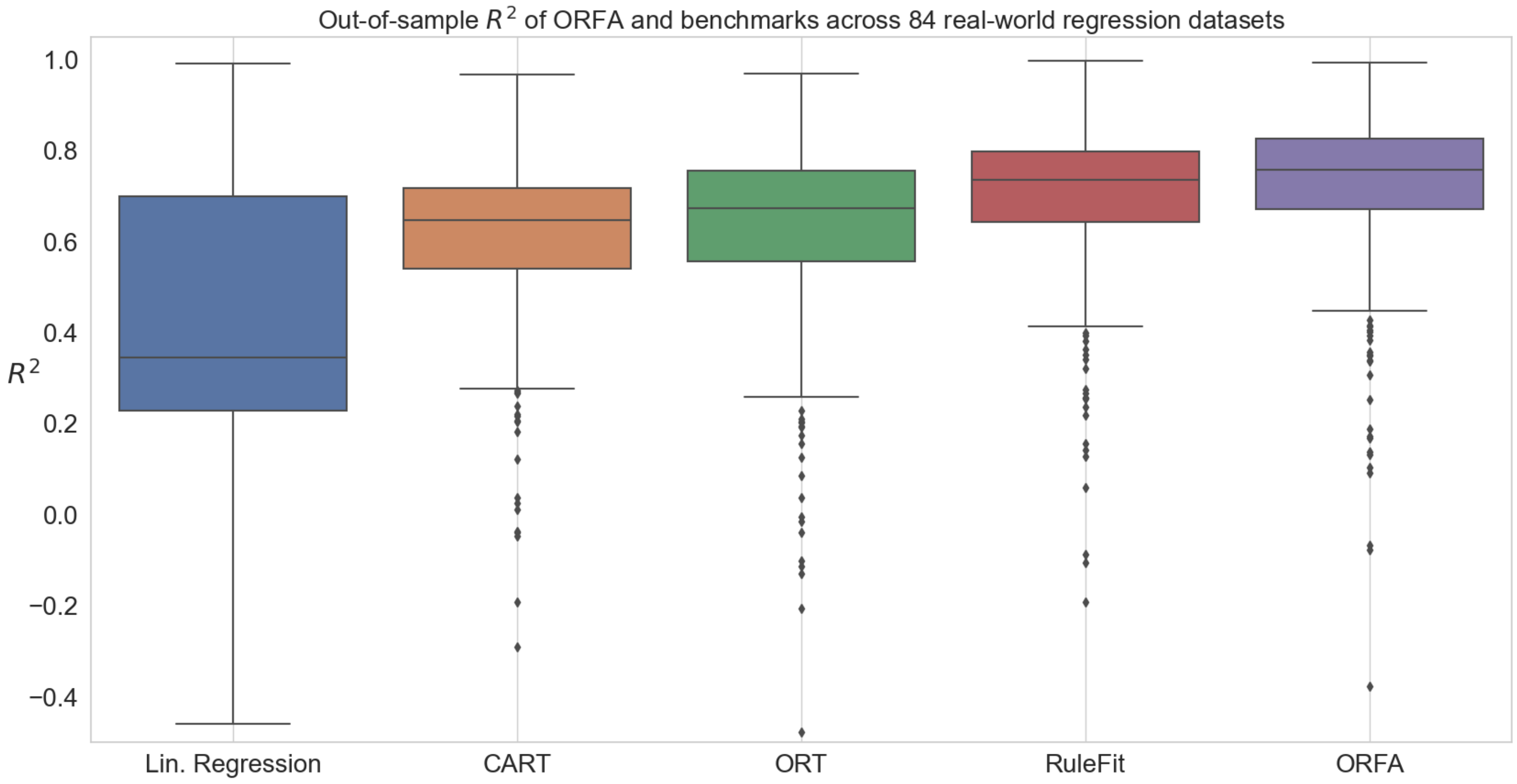
$$\boldsymbol{y}_3 = \lambda \times \underbrace{\mathbb{1}\{\boldsymbol{x}_3 \le 0.3\} \times \mathbb{1}\{\boldsymbol{x}_4 \ge -0.5\}}_{\text{interaction terms}} + (1-\lambda) \times \underbrace{0.5\boldsymbol{x}_1 + 0.1\boldsymbol{x}_2}_{\text{linear terms}} + \boldsymbol{\varepsilon}$$



$\leftarrow$ high linearity λ high interactions $\rightarrow$

BENCHMARK

Across 84 real-world regression datasets, provided by [PLMB], ORFA consistently ranks among the best methods



BENCHMARK

Across 84 real-world regression datasets, provided by PLMB, ORFA consistently ranks among the best methods

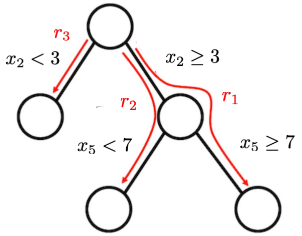
		Lin. Regression	CART	ORT	RuleFit	ORFA
Dataset	CPU	0.721	0.951	0.956	**0.976**	**0.978**
	Automobile	0.759	0.847	**0.879**	0.806	**0.874**
	Rabe	0.984	0.882	0.882	**0.986**	**0.987**
	Puma	0.375	0.567	**0.601**	0.571	**0.608**
	PW	0.710	0.780	0.762	**0.820**	**0.822**
	Wind	**0.754**	0.663	0.667	**0.754**	0.753
	Sleep Apnea	0.193	**0.845**	0.852	0.836	**0.844**
	Bodyfat	**0.974**	0.944	0.946	**0.974**	0.973
	CPU Small	0.707	0.936	0.947	**0.963**	**0.969**
	FRI	0.265	0.580	**0.684**	0.614	**0.749**
	Chatfield	0.851	0.704	0.679	0.781	**0.750**
	Geyser	**0.800**	**0.775**	0.755	0.779	0.762
⋮	⋮	⋮	⋮	⋮	⋮	
	Average	0.424	0.625	0.633	**0.707**	**0.724**

Table 1: Out-of-sample R^2 across 84 real-world regression datasets provided by PMLB. The best performer on each dataset is hilighted in **blue**, while **purple** denotes the second best.

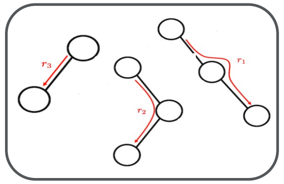
MANAGEMENT
SLOAN SCHOOL

BEYOND A HEURISTIC

The ORFA training is **disaggregated**. Rules are fed to regression and a new problem is solved.



Problem 1

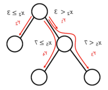


Rules $\boldsymbol{r}$

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta}} \mathcal{L}((\boldsymbol{X},\boldsymbol{y}),\boldsymbol{r})$$

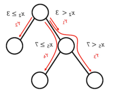
Problem 2

But this is *Machine Learning under an* ***Optimization Lens.*** What if we solve just one problem?

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{r}} \mathcal{L}(\boldsymbol{X},\boldsymbol{y},$$  $$)$$

Single Problem

INTEGRATED ORFA (IORFA)

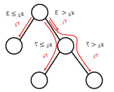
Introducing IORFA, an integrated approach to solving $\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{r}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{y},$  $)$

IORFA is a modification of the MIO problem of ORT, introducing a **rule** term to the objective:

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{z}} \sum_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \sum_{t \in T_L} \delta_t z_{i,t} \right)^2 \qquad z_{i,t} = \mathbb{1}\{x_i \in \text{leaf } t\}$$

subject to the usual constraints on $\boldsymbol{z}$...

INTEGRATED ORFA (IORFA)

Introducing IORFA, an integrated approach to solving $\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{r}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{y},$  $)$

IORFA is a modification of the MIO problem of ORT, introducing a **rule** term to the objective:

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{z}} \sum_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \sum_{t \in T_L} \delta_t z_{i,t} \right)^2 \qquad z_{i,t} = \mathbb{1}\{x_i \in \text{leaf } t\}$$

subject to the usual constraints on $\boldsymbol{z}$...

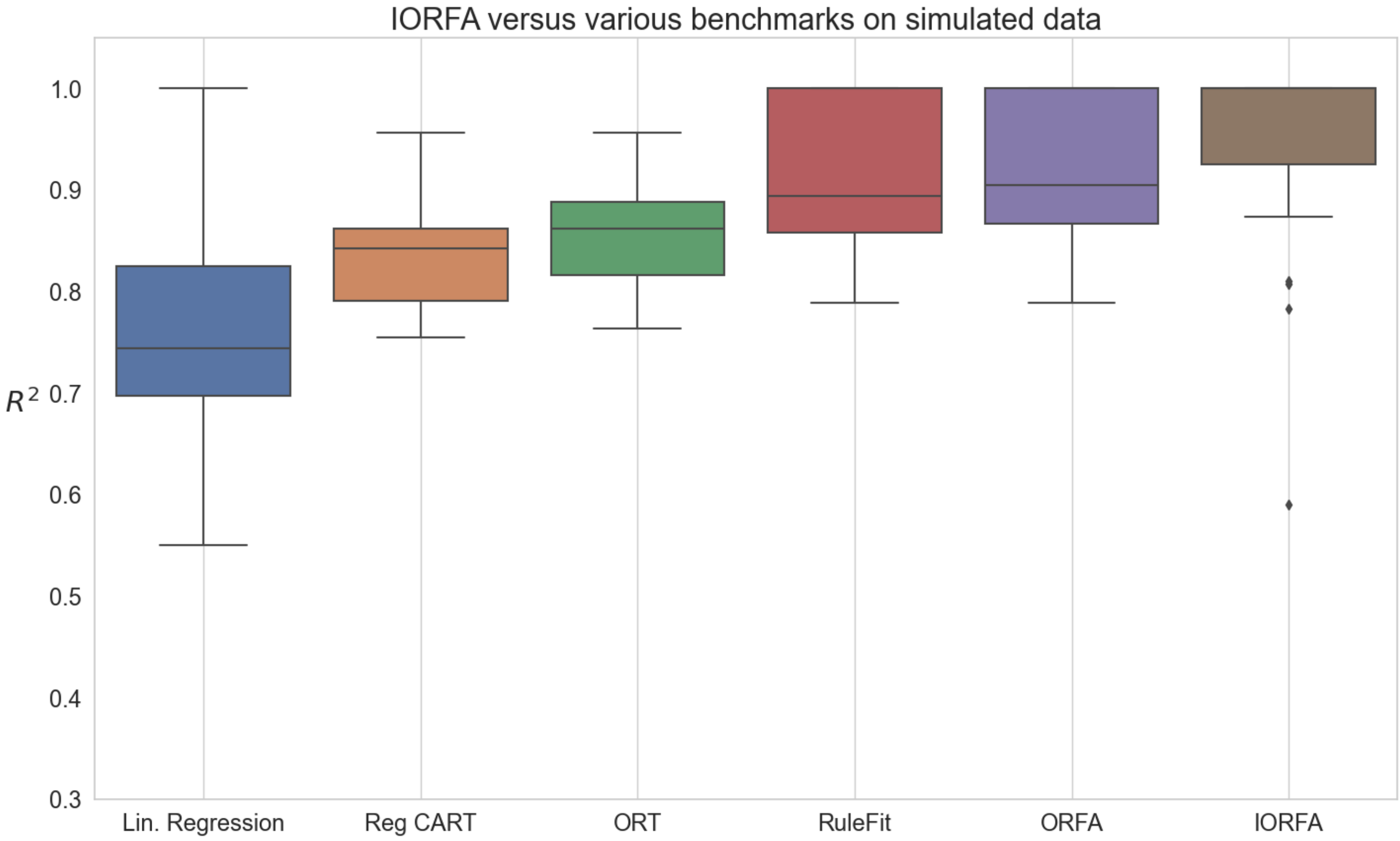
Think of δ_t as fitting a coefficient to every group belonging to leaf nodes $t \in T_L$

- Equivalent to fitting a parameter to every rule, as we would like to do in RuleFit!
- See Appendix B for the complete MIO formulation

INTEGRATED ORFA (IORFA): RESULTS

Our resulting algorithm (IORFA) outperforms ORFA and RuleFit on a small simulated dataset

- We plan to extend these trials to real-world datasets in the coming weeks



MIT MANAGEMENT SLOAN SCHOOL

APPENDIX A: INTERPRETATION

Predicting bodyfat with one rule from OCT:

$$\hat{\text{Bodyfat}}_i = 24.2 + 2.29 \cdot \text{Age}_i +, \ldots, + 8.8 \cdot (\mathbb{1}\{\text{Weight}_i < 183.77\} \cdot \mathbb{1}\{\text{Age}_i < 37\} \cdot \mathbb{1}\{\text{Height}_i > 182.65\})$$

If weight is less than 183.77 lbs, age is less than 38 and height is greater than 182.65 cm, then predicted bodyfat decreases by 8.8%, when all other feature values remain fixed.

This rule identifies a subgroup of tall, athletic young people with high weight but low bodyfat.

MIT MANAGEMENT SLOAN SCHOOL

APPENDIX B: IORFA MIO FORMULATION

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{z}} \sum_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \sum_{t \in T_L} \delta_t z_{i,t} \right)^2$$

subject .to.

$$N_t = \sum_{i=1}^{n} z_{it}, \quad \forall t \in T_L$$

$$\boldsymbol{a}_m^\top \boldsymbol{x}_i \geq b_t - (1 - z_{it}), \quad i = 1, \ldots, n, \quad \forall t \in T_B, \quad \forall m \in A_R(t),$$

$$\boldsymbol{a}_m^\top (\boldsymbol{x}_i + \boldsymbol{\epsilon}) \leq b_t + (1 + \epsilon_{\max})(1 - z_{it}), \; i = 1, \ldots, n, \forall t \in T_B, \forall m \in A_L(t)$$

$$\sum_{t \in T_L} z_{it} = 1, \quad i = 1, \ldots, n,$$

$$z_{it} \leq l_t, \quad \forall t \in T_L,$$

$$\sum_{i=1}^{n} z_{it} \geq N_{\min} l_t, \quad \forall t \in T_L,$$

$$\sum_{j=1} a_{jt} = d_t, \quad \forall t \in T_B,$$

$$0 \leq b_t \leq d_t, \quad \forall t \in T_B,$$

$$d_t \leq d_{p(t)}, \quad \forall t \in T_B \backslash \{1\}$$

$$z_{it}, l_t \in \{0, 1\}, \quad i = 1, \ldots, n, \quad \forall t \in T_L,$$

$$a_{jt}, d_t \in \{0, 1\}, \quad j = 1, \ldots, p, \quad \forall t \in T_B$$