Targeted Learning for Variable Importance

Xiaohan Wang (Cornell University)
Yunzhe Zhou (UC Berkeley)
Giles Hooker (University of Pennsylvania)

Variable Importance and Uncertainty

Two dichotomies in uncertainty and explainable AI:

- Do we explain global patterns or individual decisions?
- Do we describe the model, or the world that generated it?

All combinations appropriate for different purposes.

This talk: global patterns about the world

- Variable Importance for model an estimate for population
- Need to define target of estimation
- Quantify uncertainty associated with data generation and model fitting.

Conditional Simulation Importance

- Data $\{Y_i, X_i, Z_i\}_{i=1}^n \sim P_{Y,X,Z}$
- Model $\hat{f}(x, z)$ to predict y with loss $L(y, \hat{y})$

For "importance" of X, generate $X_i^C \sim X|Z_i$ indep of Y_i and

$$\widehat{VI}_X^C = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \hat{f}(X_i^C, Z_i)) - L(Y_i, \hat{f}(X_i, Z_i))$$

increase in loss when using uninformative X.

Alternatives:

- Permutation importance: $X_i^\pi \perp (Z_i, Y_i)$
- LOCO: compare to (Y_i, Z_i) model

Note similar mechanism to SHAP

The Estimand of VI

Assuming

$$\hat{f}(x,z) \approx \hat{y}(x,z) = E(Y|X=x,Z=z)$$

then $\widehat{VI}_X^C$ targets

$$\begin{aligned}\Psi^C(P) &= \mathbb{E}L(Y,\hat{y}(X^C,Z)) - \mathbb{E}L(Y,\hat{y}(X,Z)) \\ &= \Psi_0^C(P) - \Psi(P)\end{aligned}$$

treated as a functional of the data distribution.

Aim: de-bias $\widehat{VI}_X^C$ and provide confidence intervals.

Framework: targeted learning

A Targeted Learning Primer

Start with:

- Data distribution P, and estimand $\Psi(P)$, estimated $\hat{P}$
- Influence function $\psi(x; P)$ given by

$$\psi(x; P) = \frac{d}{d\epsilon}\Psi((1-\epsilon)P + \epsilon\delta_x)$$

- Independent data with distribution P_n

Initialize $\tilde{P} = \hat{P}$, iterate

- 1d maximization over η

$$\tilde{\eta} = \text{argmax} \sum \log \tilde{P}(X_i)(1 + \eta\psi(X_i; \tilde{P}))$$

- Update $\tilde{P} = \tilde{P}(X_i)(1 + \tilde{\eta}\psi(X_i; \tilde{P}))$

Return $\Psi(\tilde{P}) \pm 2\text{sd}_{P_n}\psi(X; \tilde{P})/\sqrt{n}$

Influence Functions

Key component in targeted learning satisfies

$$\frac{d}{d\epsilon}\Psi(P + \epsilon(Q - P)) = \int \psi(x; P) d(Q - P)$$

for all Q.

Most Ψ can derived from

$$\psi(x; P) = \frac{d}{d\epsilon}\Psi((1 - \epsilon)P + \epsilon\delta_x)$$

for each x.

Functional equivalent of gradient; expresses direction of greatest sensitivity in Ψ.

Justification

von Mises expansion:

$$\Psi(\hat{P}) - \Psi(P) = \frac{1}{n}\sum_{p_n} \psi(X_i; P) - \frac{1}{n}\sum_{p_n} \psi(X_i; \hat{P}) + \int (\psi(X; P) - \psi(x; \hat{P}))d(P_n - P) + R$$

- $\sqrt{n}\frac{1}{n}\sum \psi(X_i; P) \xrightarrow{d} N(0, \sigma^2)$ = uncertainty quantification
- $\frac{1}{n}\sum \psi(X_i; \tilde{P}) = 0$ after TL iteration
- Cross-product = $o(1/\sqrt{n})$ if $\hat{P}$, P_n independent
- $R = o(1/\sqrt{n})$ if $|\hat{P} - P| = o(n^{-1/4})$

When naively subtracting bias $\frac{1}{n}\sum \psi(X_i; \hat{P})$; CI's do not account for bias correction.

Influence Functions for Variable Importance

Break into:

$$\psi_0(X, Y, Z) = (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z)) P(y|X, Z) dy + L(Y, \hat{y}(X, Z)) - \Psi_0(P).$$

and

$$\begin{aligned}\psi_0^C(X, Y, Z) = &\int L'(y, \hat{y}(X, Z))(Y - \hat{y}(X, Z)) p(y|Z) dy \\ &+ \int L(y, \hat{y}(X, Z)) p(y|Z) dy \\ &- \int L(y, \hat{y}(x, Z)) p(y|Z) p(x|Z) dx dy \\ &+ \int L(Y, \hat{y}(x, Z)) p(x|Z) dx - \Psi_0^C(P).\end{aligned}$$

Notes

- In squared error case some terms drop out or cancel.
  - ψ_0 evaluates to mean of squared error
  - removed by comparisons between features
- ϕ_0^C requires estimates of $p(y|Z)$, $p(x|Z)$, obtained through empirical distribution weighted by random forest kernel:
  - Initialize $P(Y = Y_i|Z) = w_i(z)$ from in-leaf proximity weight.
  - Monte-Carlo approximation to integrals in ψ_0^C by weighted bootstrap.
  - TL update = update $w_i(z)$.
- Approximate update:

$$Y_i = \epsilon_0 f(X_i, Z_i) + \epsilon_1 \psi_0^C(X_i, Y_i, Z_i) + \eta$$

- Shrinkage: use $\epsilon_1/10$ for numerical stability.
- Cross validation: split into training/TL update over 10 folds.

Algorithm

Require: $\{Y_i, X_i, Z_i\}$ for $i = 1, \ldots, n$, I_1, I_2, I_3 such that $I_1 \cup I_2 \cup I_3 = \{1, \ldots, n\}$ and $I_1 \cap I_2 \cap I_3 = \emptyset$, initial estimates $\hat{f}_{I_1}$, $\hat{P}(x|z), \hat{P}(y|z)$.

1: **for** each iteration t **do**

2: Sample $\{X_j^*\}_{j=1,\ldots,m}$ from $\hat{P}(x|z)$, $\{Y^*\}_{k=1,\ldots,m}$ from $\hat{P}(y|z)$

3: Calculate $\hat{\Psi}^C_{I_2,0} = \frac{1}{|I_2|}\sum_{i\in I_2}(Y_i - \hat{f}(X_i^C, Z_i))^2$. and

$$\hat{\psi}^C_{I_2,0}(X_i, Y_i, Z_i; \hat{P}) = \frac{1}{m}\sum_{j=1}^{n} L(Y_i, \hat{f}(X_j^*, Z_i))$$

$$- \frac{1}{m^2}\sum_{j=1}^{n}\sum_{k=1}^{n} L(Y_k^*, \hat{f}(X^*, Z_i)) + \frac{1}{m}\sum_{k=1}^{n} L(Y_k^*, \hat{f}(X_i, Z_i))$$

4: Find $\hat{\epsilon}$ by regressing Y_i on $f(X_i, Z_i)$ and $\psi_0^C(X_i, Y_i, Z_i; \hat{P})$ using I_2

5: Update $\hat{P} = c(\hat{\epsilon})(1 + \hat{\epsilon}\hat{\psi}^C_{I_2,0})\hat{P}$

6: Repeat the above iteration until convergence.

7: **Return:** $\hat{\Psi}(\hat{f}_{I_1}, P_{\varepsilon^{k_n}})$ and variance $\sqrt{\frac{1}{n}\sum_{i\in I_3}\hat{\psi}^C_{I_2,0}}$ based on I_3.

Theory

1. Bounded number k_n of updates.
2. von Mises expansion for Ψ
3. Consistency of $\hat{f}$
4. Sample Splitting or
5. Donsker classes for $\hat{f}$

Theorem 1

Assume that Assumptions 1-3 hold, and Assumption 4 or 5 hold. Our final estimator $\hat{\Psi}(\hat{f}_{l_1}, P_{\varepsilon_n^{k_n}})$ *is asymptotically linear and satisfies:*

$$\hat{\Psi}(\hat{f}_{l_1}, P_{\varepsilon_n^{k_n}}) - \Psi(P^*) = P_n \psi_{P^*} + o_P(1/\sqrt{n}),$$

where $\psi(P^*)$ *is the efficient influence function.*

Why *Conditional* Permutation Importance?

Permutation importance IF:

$$\psi_0^{\pi}(X, Y, Z) = (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z)) \frac{P(X)P(y, Z)}{P(X, Z)} dy$$
$$+ \int L(Y, \hat{y}(x', Z)) P(x') dx'$$
$$+ \int L(y, \hat{y}(X, z)) P(y, z) dy dz - 2\Psi_0^{\pi}(P),$$

Ratio in first term can be large if (X, Z) associated.

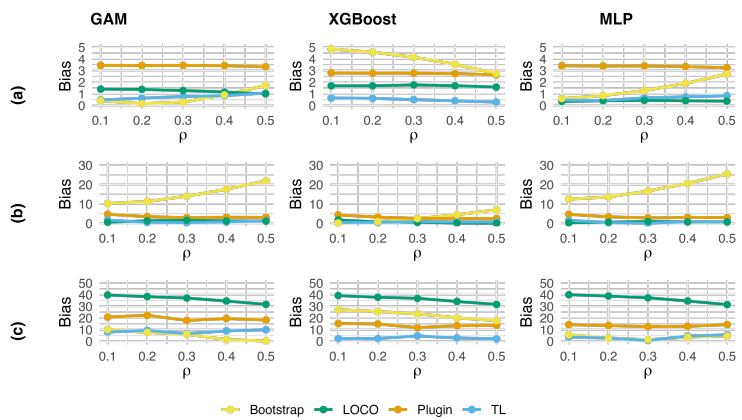
LOCO importance uses 2 overlapping models, does not fit neatly into TL.

Experiments and Results

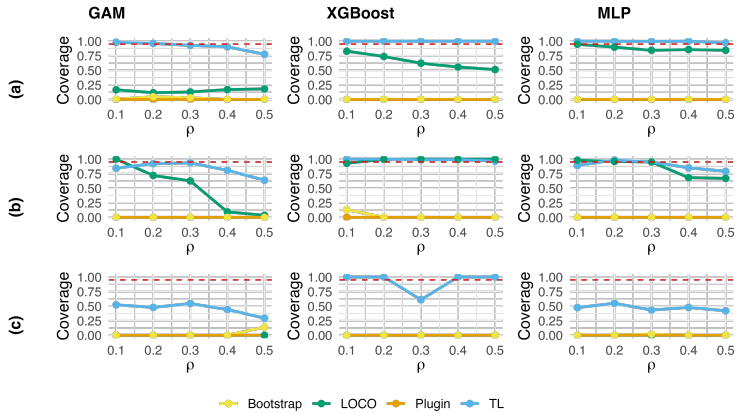
Repeated 240 times

- 1000 obs, 10-d X
- vary x_1,x_2 correlation
- Y_i: just x_1 (a), linear (b), nonlinear (c).
- Bootstrap, LOCO, Plugin, TL

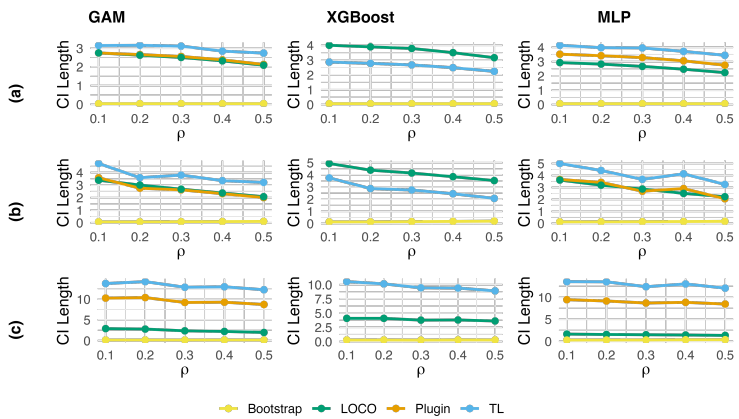
Bias



Coverage

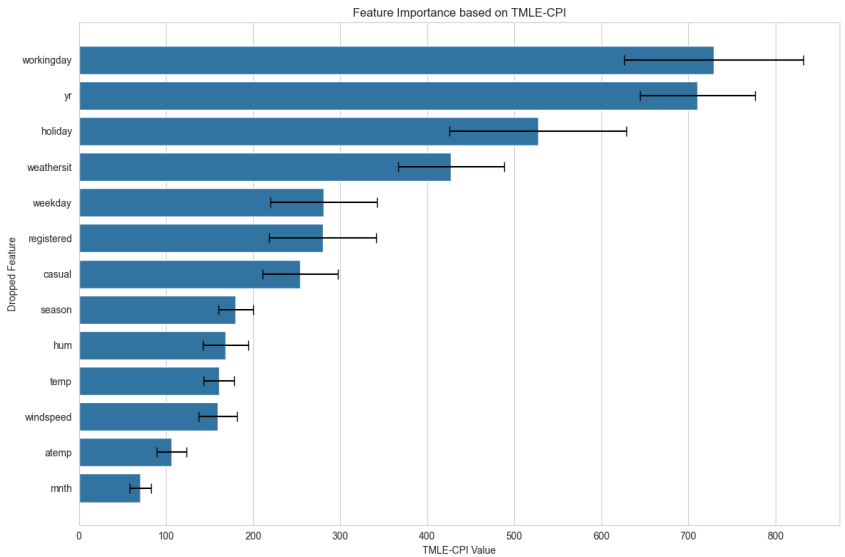


CI Length

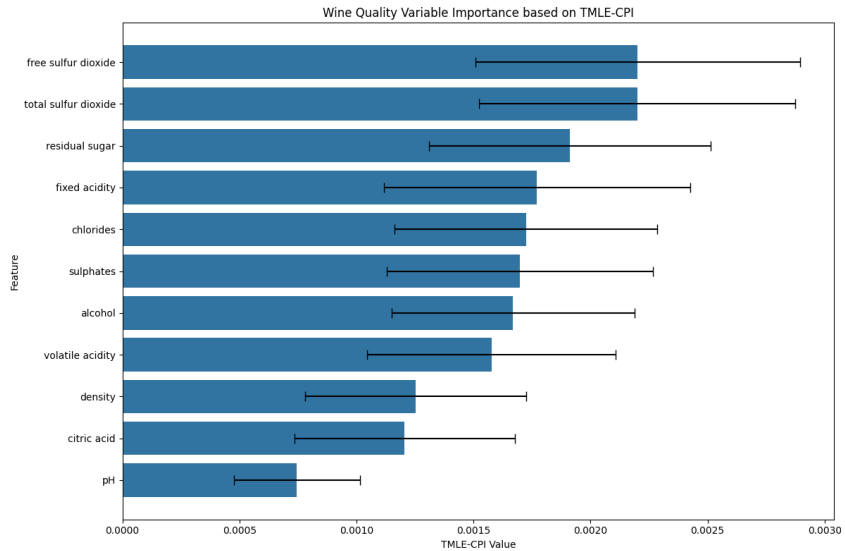


Real World Demonstrations

Bike Share



Wine Quality



Extensions

Targeted learning approaches readily extendable to

- Certifying feature orderings
- Testing strength of interactions
- UQ for partial dependence plots
- Model distillation/approximations
- Assessing fairness
- Large-scale economic/social consequences

But: regularity requirements can restrict range of application.

Discussion

Uncertainty in AI explanations needs care:

- Relevant uncertainty depends on purpose.
- Data uncertainty can be challenging.

Targeted learning provides a route to data uncertainty!

General framework but needs

- estimand with sufficient regularity
- often auxiliary quantities to evaluate influence function
- some additional analysis
- sufficient convergence of the ML method.

But many potential applications.