"Understanding" the Output of Machine Learning Methods and Warnings

Giles Hooker

Conference on Applied Statistics in Ireland
Killarney, May 15, 2023

Diagnostic Methods in Machine Learning

- Machine learning $\equiv$ high dimensional non-parametric prediction.
- Enormously successful over past 30 years.
- But, deliberately avoids assumptions:
  - Results in algebraically complex "black box" prediction functions.
  - Provides little guidance as to what features are important or how they affect predictions.
- Historically, ML philosophy opposed to interpretability as a consideration.
  - But heuristics (often from statisticians) often improved popularity
  - e.g. Gradient Boosting (Friedman 2001) and Random Forests (Brieman 2001).
- Recent (last 5 years) more general rise in interest.

The Doctor Just Won't Accept That

- Rise in publicly-explicit use of ML, increased demand for explanations of black box models.
- Partly driven by professional fears.
- But explanations/diagnostics $\Rightarrow$ software popularity long before.

The Doctor Just Won't Accept That!

Zachary C. Lipton
Carnegie Mellon University,
Amazon AI
zlipton@cmu.edu

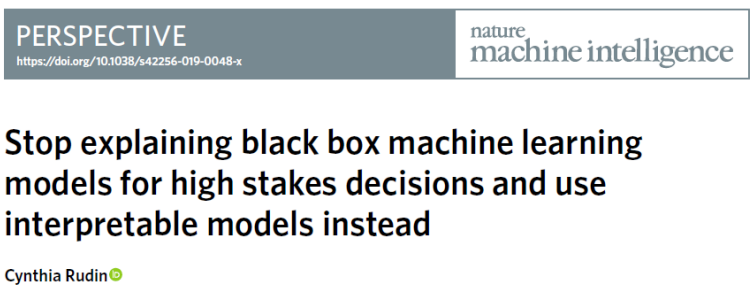
"I work with medical data. We work with doctors and they're interested in predicting risk of mortality, recognizing cancer in radiologic scans, and spotting diagnoses based on electronic health record data. We can train a model, and it can even give us the right answer. But we can't just tell the doctor "my neural network says this patient has cancer!" The doctor just won't accept that! They want to know why the neural network says what it says. They want an explanation. They need interpretable models."

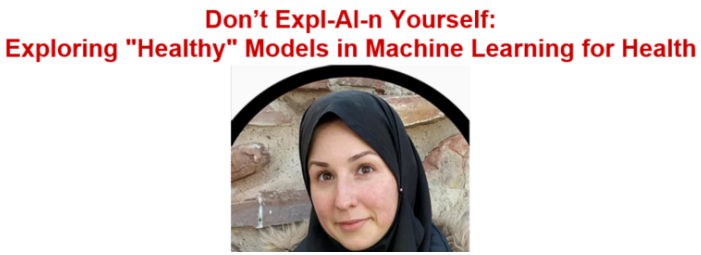
Why Care About Explanation?

Reasons to value insight into the black box:

- Confidence-building exercise (marketing)
- Basis for evaluating disagreement between experts
- Detection of anomalous/non-causal predictive behavior
- Explanation/description of causal relationships
- Subject access/transparency, legal obligations

But





Agenda

- Introduction
- Part I: Global Interpretation
  - Variable Importance
  - Feature Effects
- Part II: Distillation
  - Interpretable models
  - Approximation by interpretable models
  - Stability and when to care about it
- Part III: Local Explanations
  - Local importances: LIME, SHAP, saliency
  - Counterfactual Explanations
  - From local to global
- Discussion

Most methods available in the `iml` package in `R`. See also Molnar, 2022, *Interpretable Machine Learning*

Types of Explanation Strategy

Distinctions:

global/local : global patterns across whole populations vs "What drove this particular prediction?"

model/summary/example : Mechanics of making a prediction (human computability) vs indicator of important effects vs how can I change prediction?

- Global diagnostics usually about understanding a system
  - Hypothesis generation/pattern discovery/inference.
  - Sanity checks.
- Local explanations driven by practitioner needs
  - Justification, sanity check, recommendation
- Can provide very different information.
- Common approaches to both have often been poorly thought out.

Some Notation and Nomenclature

- Assume that we have a data set of *n observations*:

$$\{(X_i, Y_i)\}_{i=1}^{n}$$

  (also examples, realizations, ...)

- Each X_i is a vector of *p covariates*

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$$

  (also features, predictors, variables,...)

- Data set is used to estimate (learn, train,...) a *prediction function* $F(x_1, \ldots, x_p)$
- Use: X_i = row of data set, $X_{\cdot j}$ = column (values for covariate j), x_j = variable value, x_{-j} = vector without element j.

Desired: some way to "understand" $F(x_1, \ldots, x_p)$

Beijing Housing Data

Used for illustrations, predict log(totalPrice) from

- Lat, Lng
- Days On Market
- online followers
- square m
- Number of
  - livingRoom
  - drawingRoom
  - kitchen
  - bathRoom
- Building Type
- Construction Date
- Renovation
- Building structure
- Ladder Ratio (resident to elevator capacity)
- Elevator
- Ownership $>$ 5 years
- Subway access
- District

`randomForest/R` used for demonstrations, but observations apply to all ML.

Part I: Variable Importance

Which Features Matter and How Much?

Can be thought of in two ways

- How much difference does changing the value of this feature make?
- How much information does including this feature add?

In linear models

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

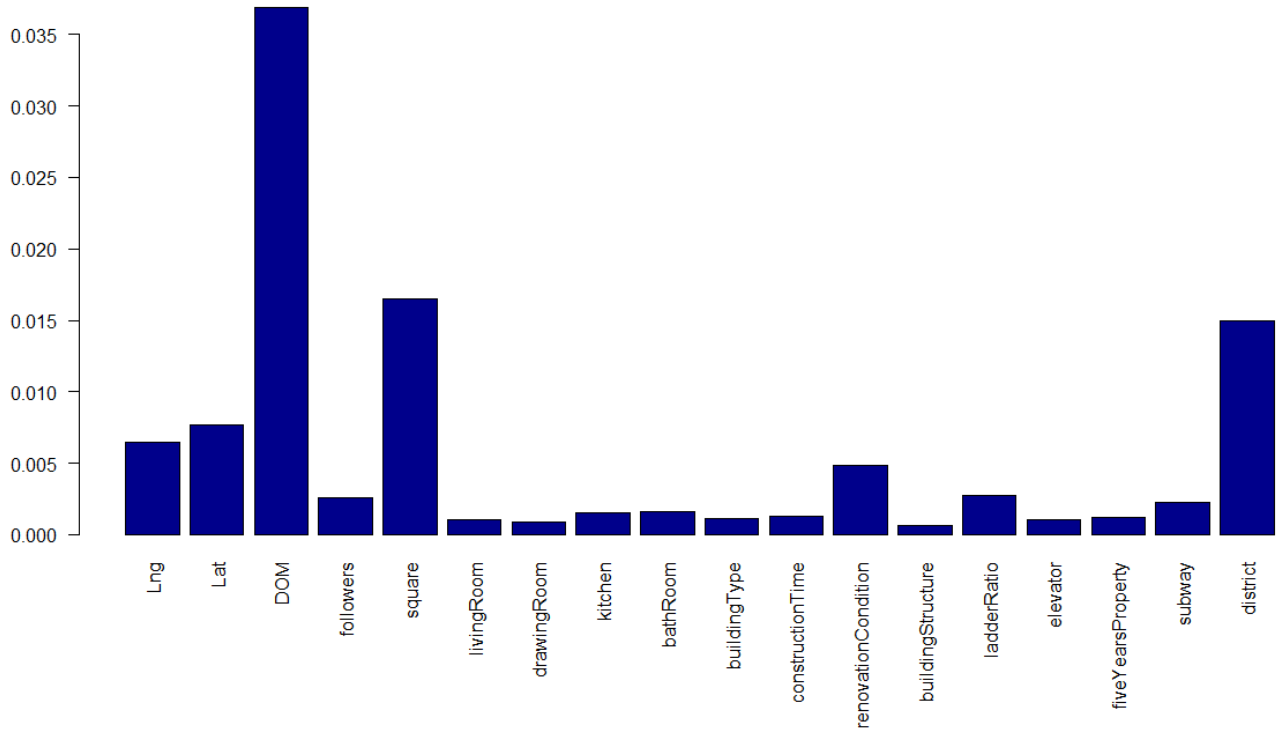
interpretation of β_j = how does changing X_j affect prediction?

Tests of $H_0 : \beta_j = 0$ are relative to the other features included.

Variable Importance as Added Information

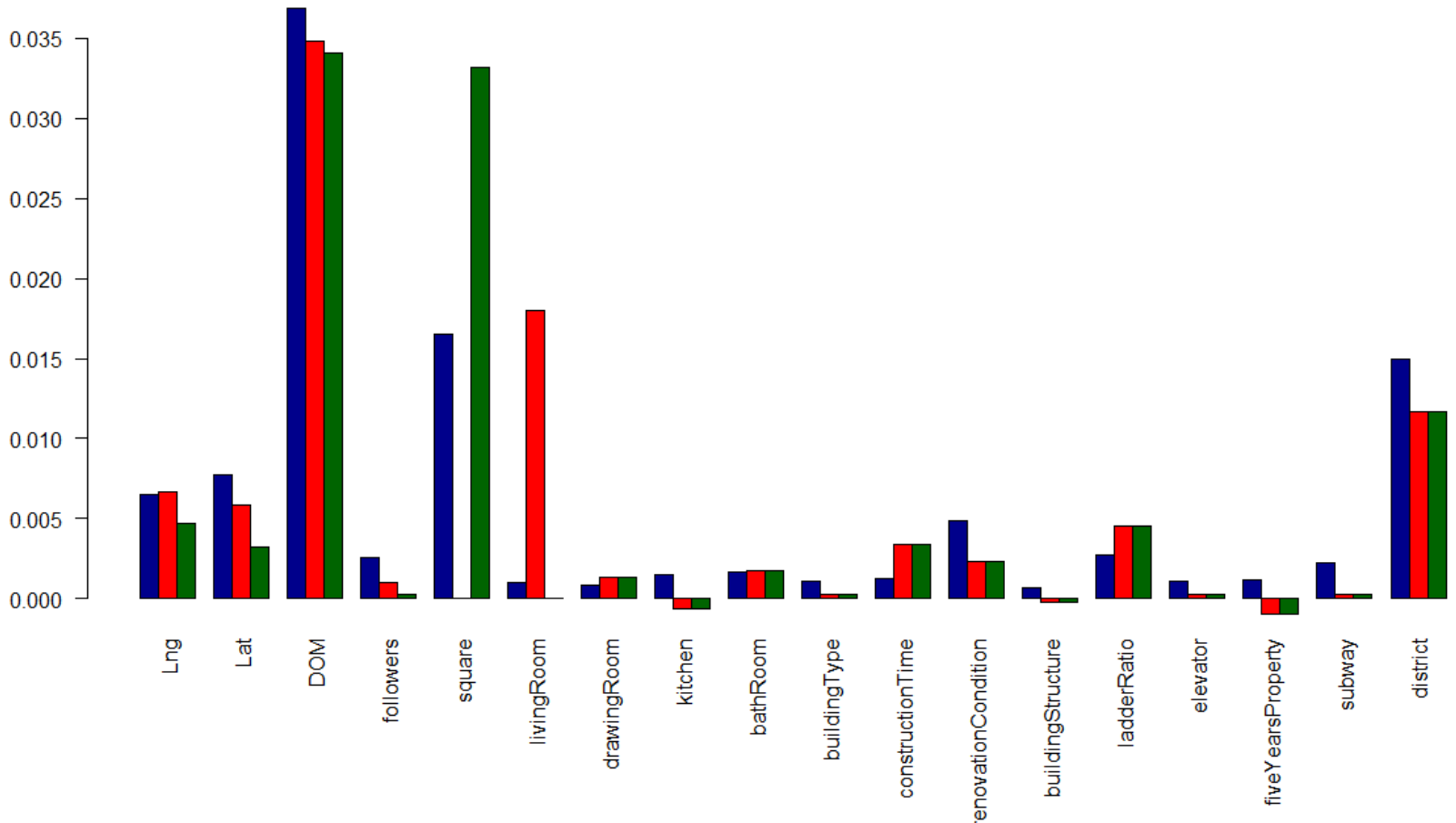
"Does x_1 contribute to predictive accuracy"?

Measure difference in test-set performance when training with versus without X_1.



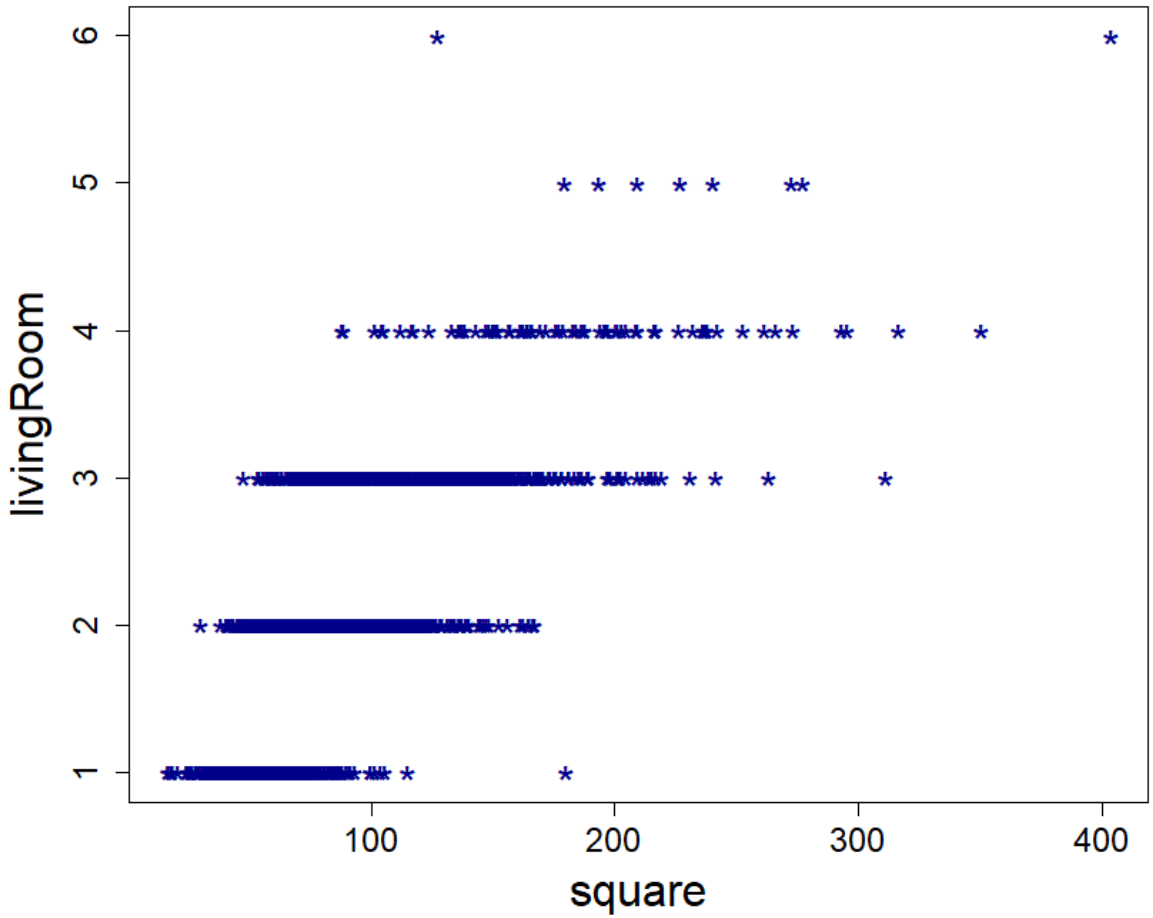
Dependence on Feature Set

Repeat with `livingRooms` removed, or with `Square` removed



Look Out for Feature Distributions

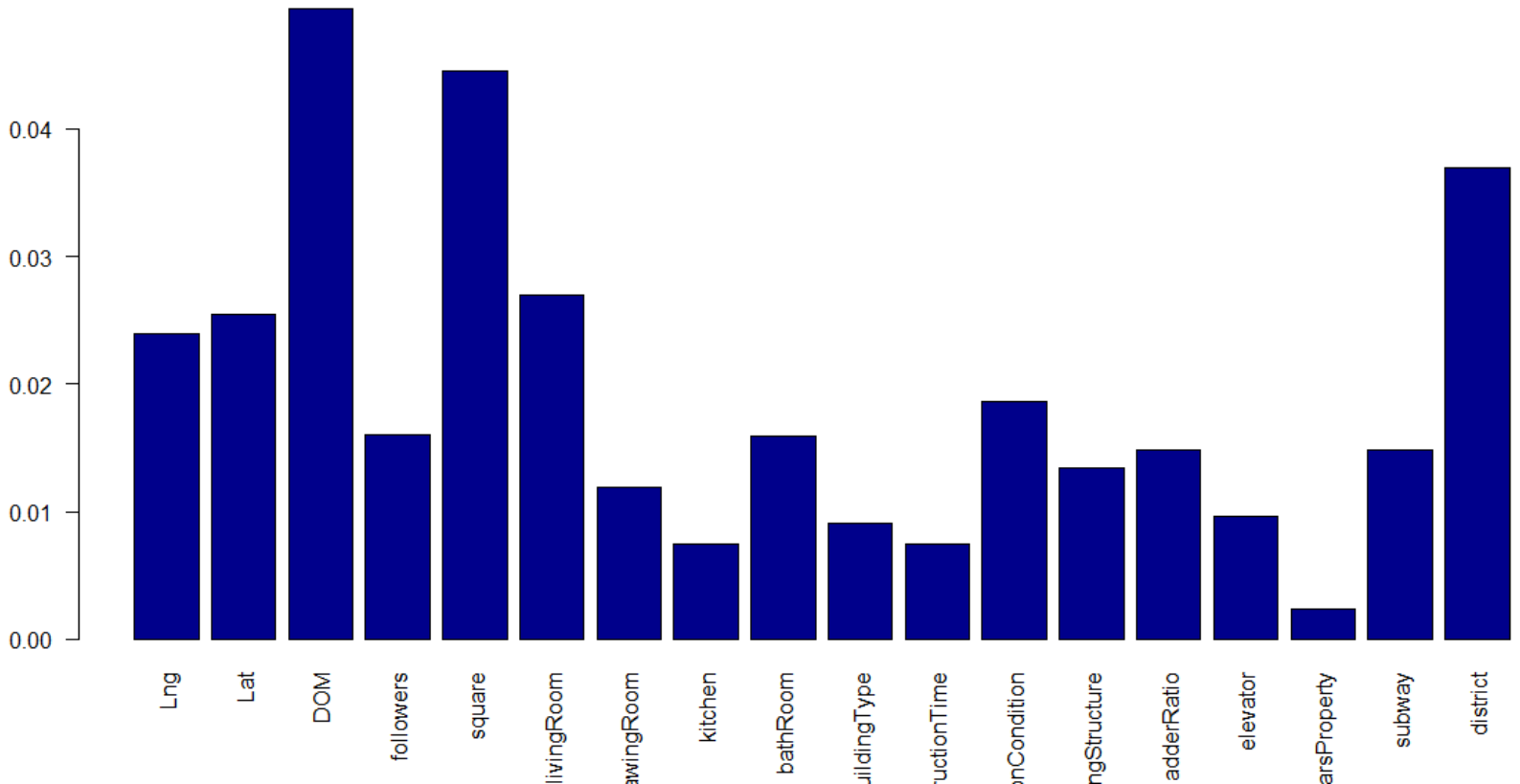
`livingRoom` associated with `Square` $\Rightarrow$ removing one transfers "signal" to the other.



Shapley Values

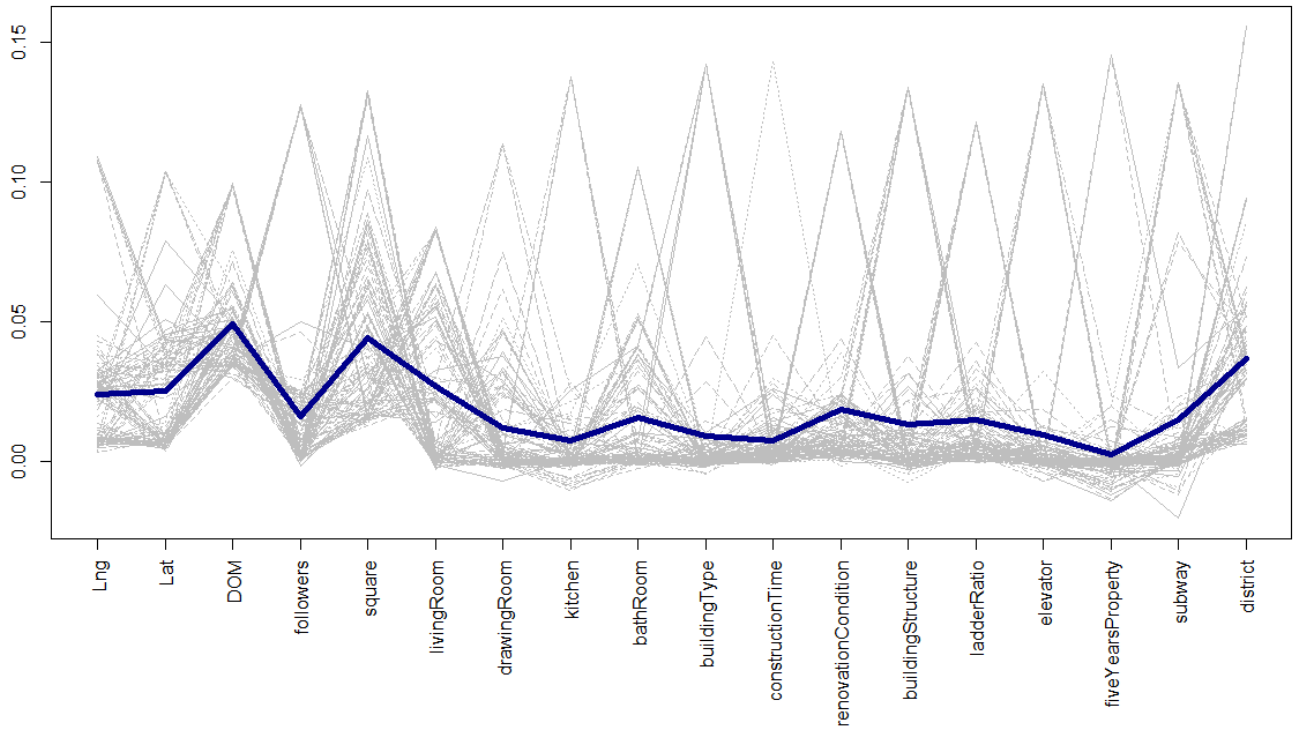
What feature set to measure against? All of them!

- Any ordering of features
- Importance of X_j = change in test-accuracy when including it versus those before it.
- Average over orderings.



Shapley Values

Shapley values average contributions to prediction, but can be helpful to show spread.



SAGE: Shapley Additive Global importancE, to distinguish from local SHAP values.

Kernel Shapley Calculations

Shapley values motivated from co-operative game theory:

The most equitable way of sharing revenues among a set of actors.

- Original Shapley calculation = Monte Carlo average
- But for any subset S, $\sum_{j \in S} \phi_j$ = improvement in accuracy from 0 to using X_S.
- Motivates least-squares criterion:

$$\phi_1, \ldots, \phi_p = \text{argmin} \sum \frac{p-1}{\binom{p}{|S|}|S|(p-|S|)} \left(v(S) - \phi_0 - \sum_{j \in S} \phi_j \right)^2$$

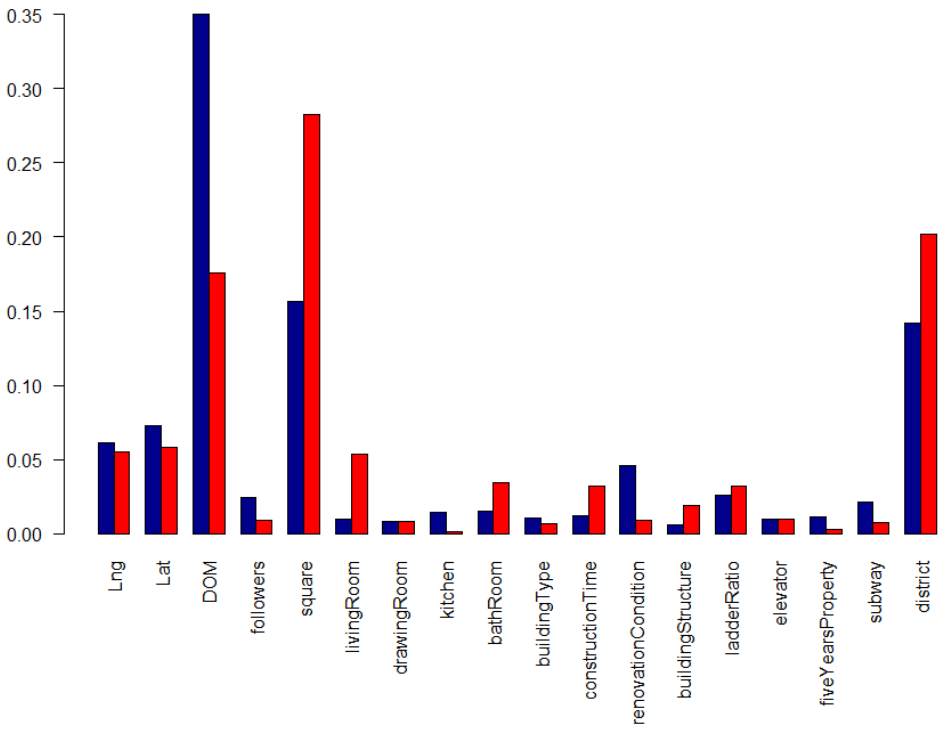
  for $v(S)$ = the "value" of S
- Motivates sampling S at random, performing linear regression on indicators $I(j \in S)$.

Permutation Importance

Rather than re-train, can we remove information from X_1?

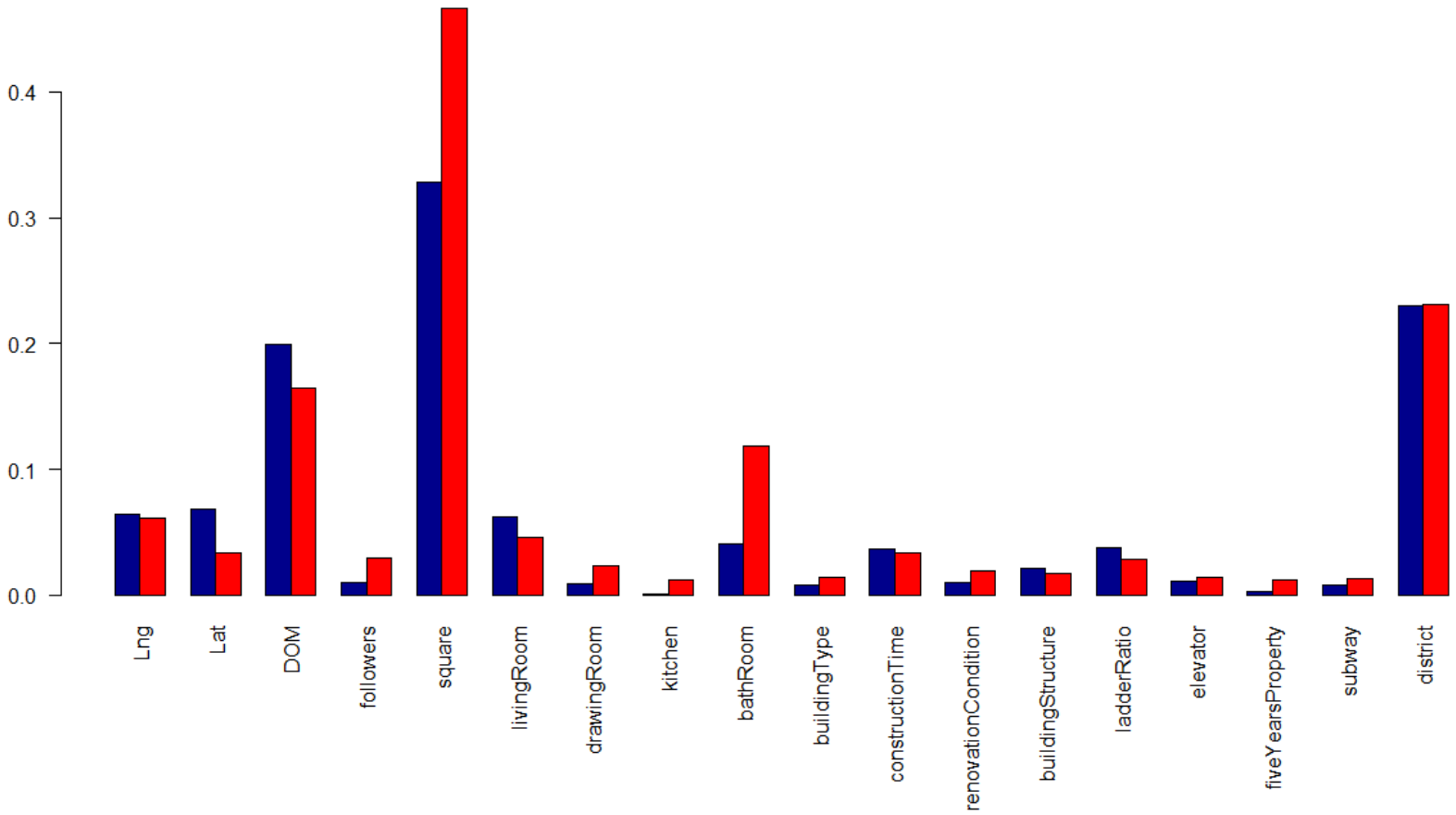
1. Permute the values of $X_{\cdot 1}$ in the data set (relative to other covariates) to get $X_{\cdot 1}^{\pi}$.
2. Measure change in test-set accuracy on permuted data

$$\text{VI}_1 = \frac{1}{n}\sum_{i=1}^{n} L\left(Y_i, F(X_{i1}^{\pi}, \ldots, X_{ip})\right) - L\left(Y_i, F(X_{i1}, \ldots, X_{ip})\right)$$



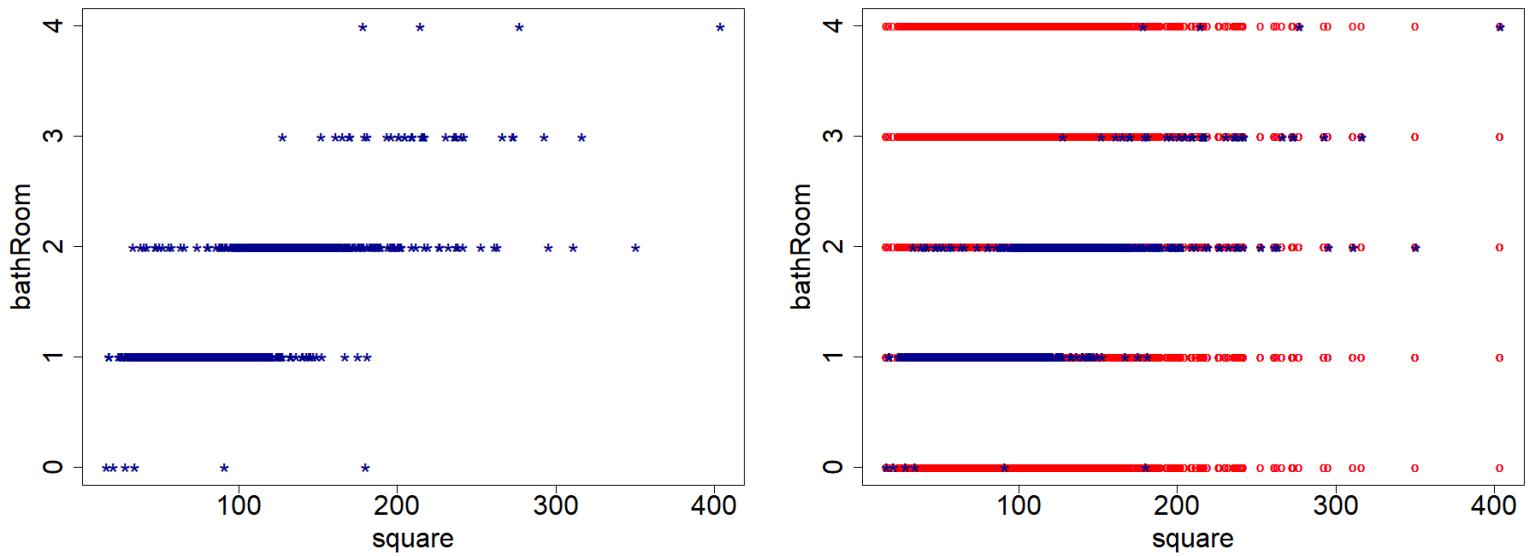
The Problem with Permutation

- In a linear model $y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$, $VI_j(F) = 2\beta_j^2 \text{var}(x_j)$.
- What if we simulate from a linear model, and train a random forest to learn the linear model?



Feature distributions *again*!

- `bathRoom` and `square` correlated.
- Permuting one creates combinations of 0 bathrooms in a huge house or 4 bathrooms in something tiny.
- F has no data near these combinations to tell it what to do.



Permutation importance tends to over-emphasize correlated features (but different reasons for different learners)

Tests and Variations

Alternatives to permutations

- Conditional permutation:
  - $X_i^{cj} \sim X_i | X_{i,-j}$ - simulate from conditional distribution
  - Measure $\sum \tilde{L}(\tilde{Y}_i, F(\tilde{X}_i)) - L(\tilde{Y}_i, F(\tilde{X}_i^{cj}))$
- Re-learn $F^{\pi j}(\mathrm{x})$ from permuted data, or $F^{cj}(\mathrm{x})$ with conditional simulation. Measure Loss.

Under squared error

- Target for F is $E(Y|X)$
- Target for F^d, F^π, F^c is $E(Y|X_{\cdot,-j})$

but statistical properties vary.

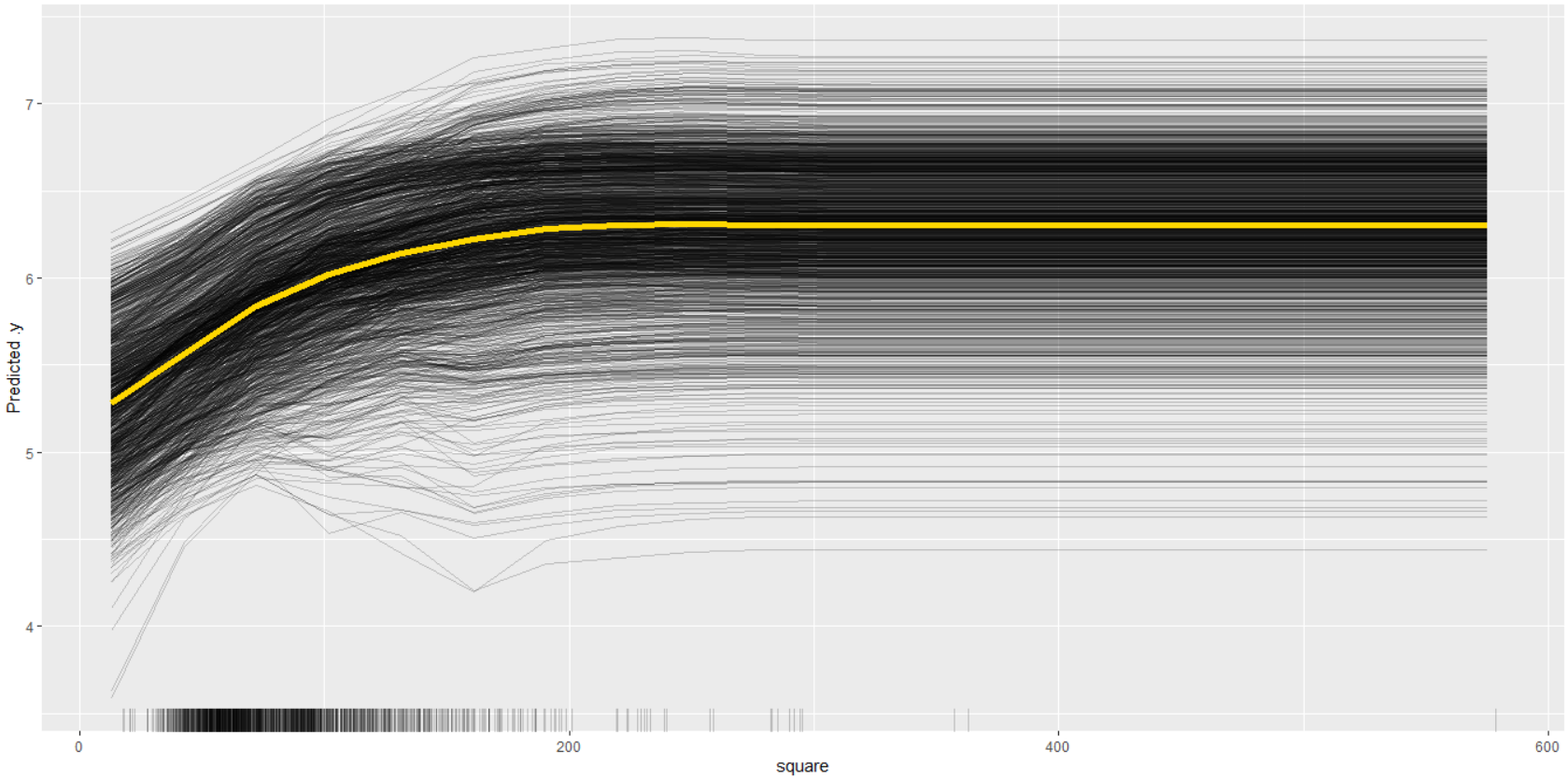
See Uncertainty Quantification for tests.

But What Does the Feature Do?

Define *Individual Conditional Expectation* of x_j for obs i by

$$\text{ICE}_{ij}(x_j) = F(x_j, X_{i,-j})$$

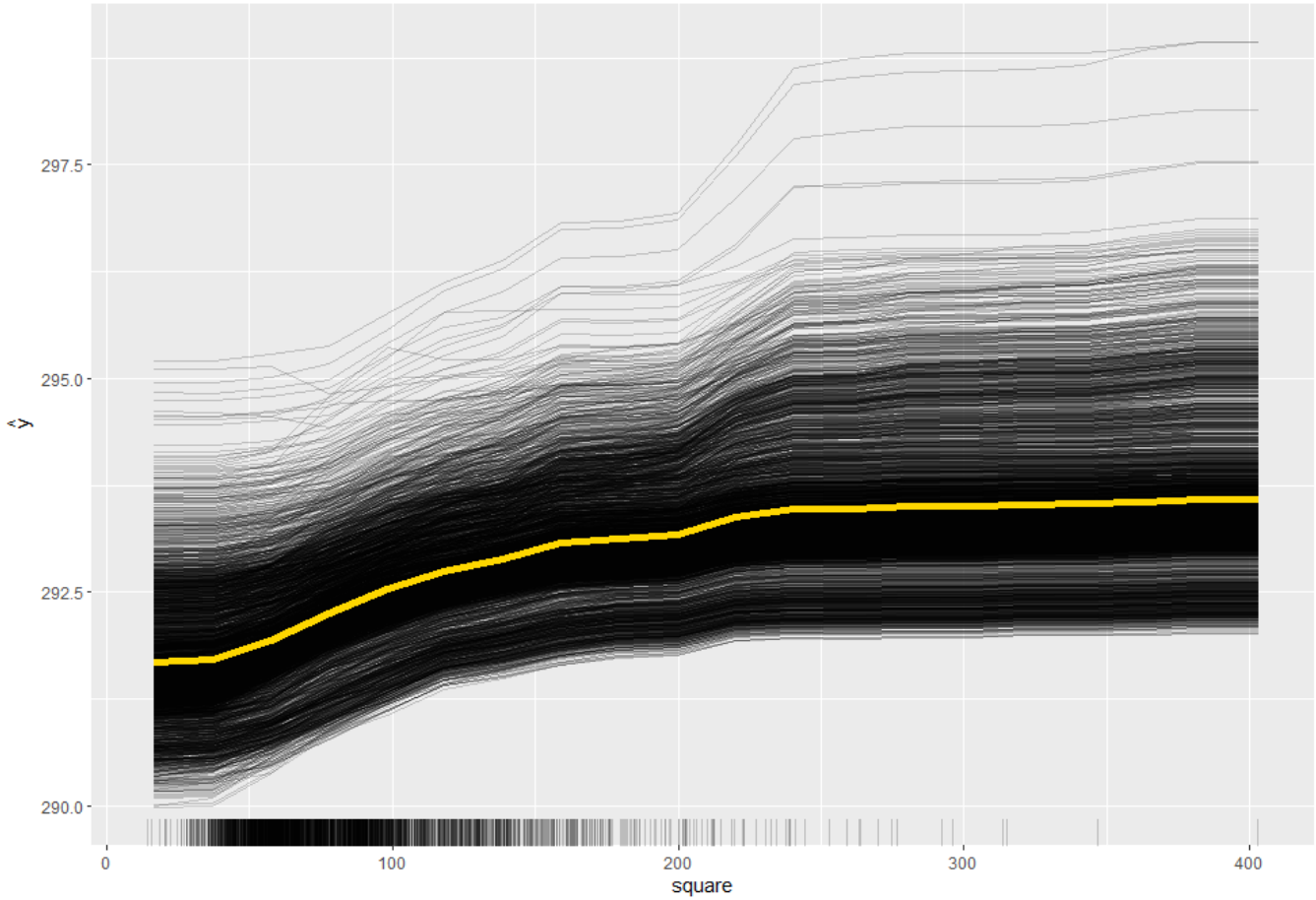
and the *Partial Dependence Function* as the average $\text{PD}_j(x_j) = \overline{\text{ICE}_{\cdot j}(x_j)}$.



Feature Distributions Again

In linear models, PD and ICE plots should also be linear.

When derived from RF trained on linear model:



Gradient Based Alternatives

Conditional distribution-based summaries designed to

- focus on places we will want to make predictions
- avoid extrapolation

but require a model for $X_j|X_{-j}$.

How about using gradients instead?

$$\text{VI}^{\partial} = \sum \left(\frac{\partial F}{\partial x_j}(X_i) \right)^2$$

Accumulated Local Effects re-integrate to get gradients

$$\text{ALE}_j(x_j) = \int_{-\infty}^{x_j} \int \frac{\partial F}{\partial x_j}(\text{x}) dP(\text{x}|x_j) dx_j$$

In practice, done by discretising range of x_j and often using finite differences.

Specifics

For a generic, non-differentiable F

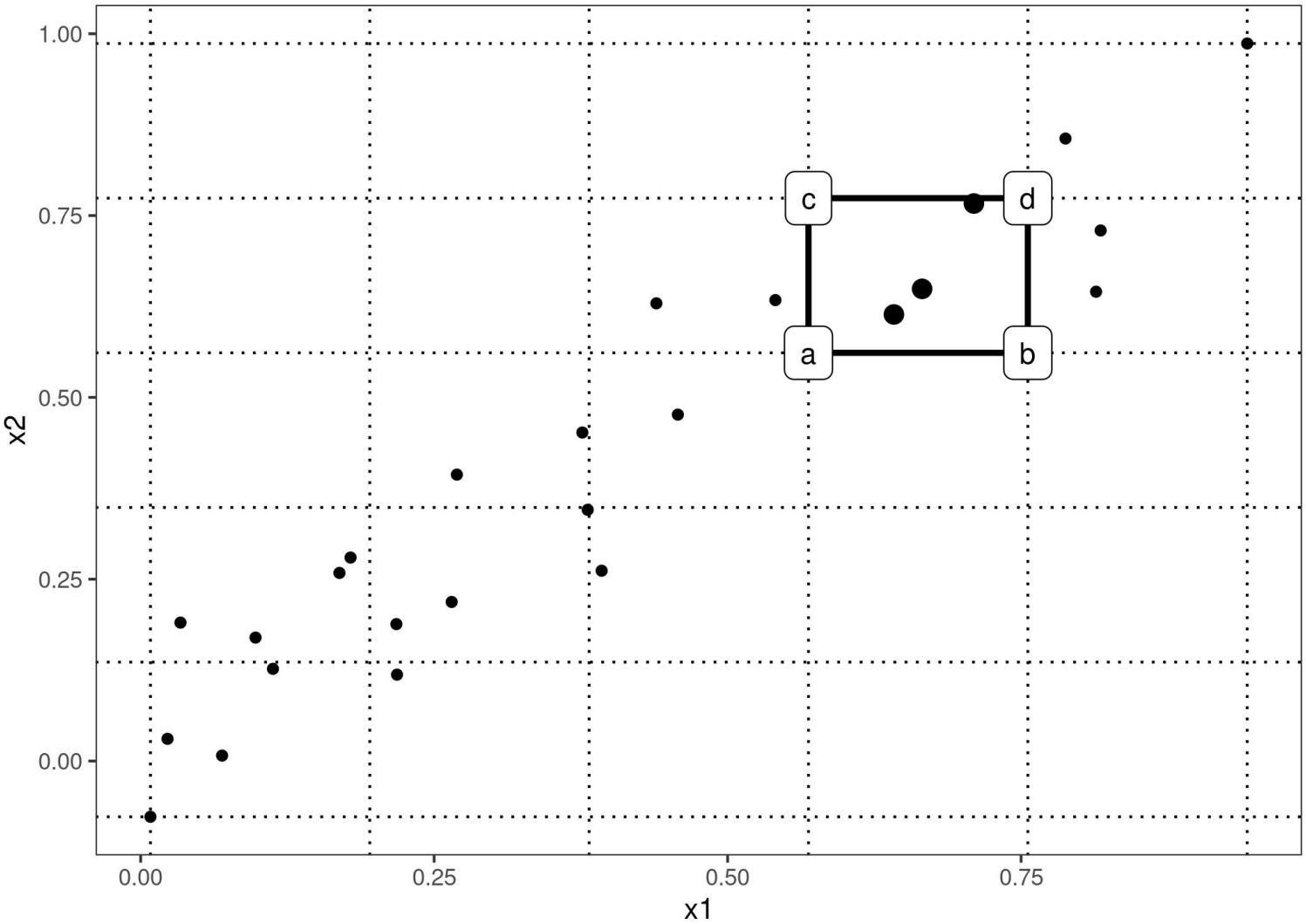
1. Divide range of x_j into k_j bins of with h with end points z_{lj}, $l = 1, \ldots, k_j$.
2. Calculate the average finite difference between bin-end points over observations with x_j in bin l

$$\delta F_{lj} = \frac{1}{N_{lj}} \sum_{z_{lj} < X_{ij} < z_{(l+1)j}} \frac{F(z_{(l+1)j}, X_{i,-j}) - F(z_{lj}, X_{i,-j})}{h}$$

   for $N_{lj} = \sum(z_{lj} < X_{ij} < z_{(l+1)j})$.
3. Now record integral and center

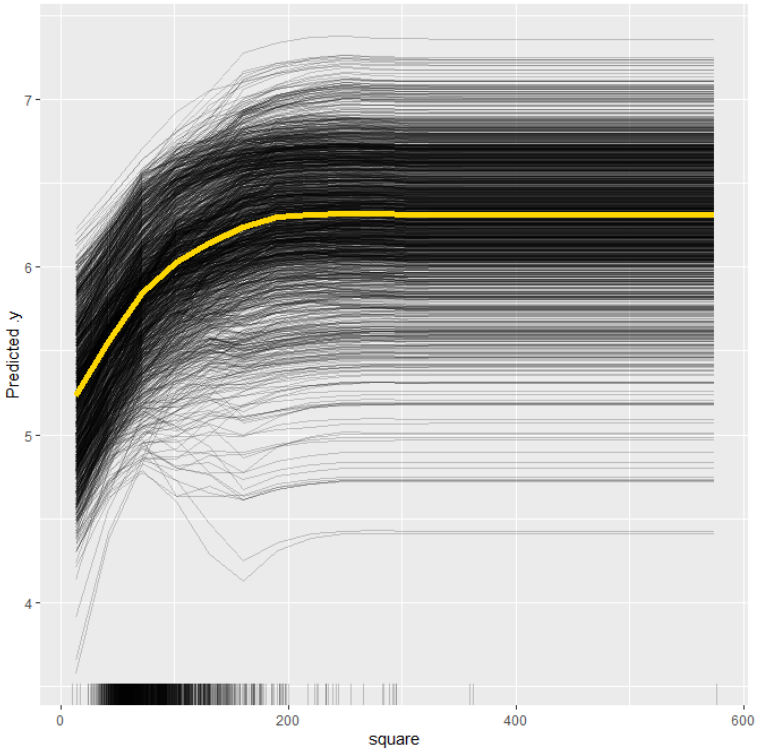
$$\text{ALE}_j(x_j) = \sum_{z_{lj} < x_j} h \delta F_{lj}$$

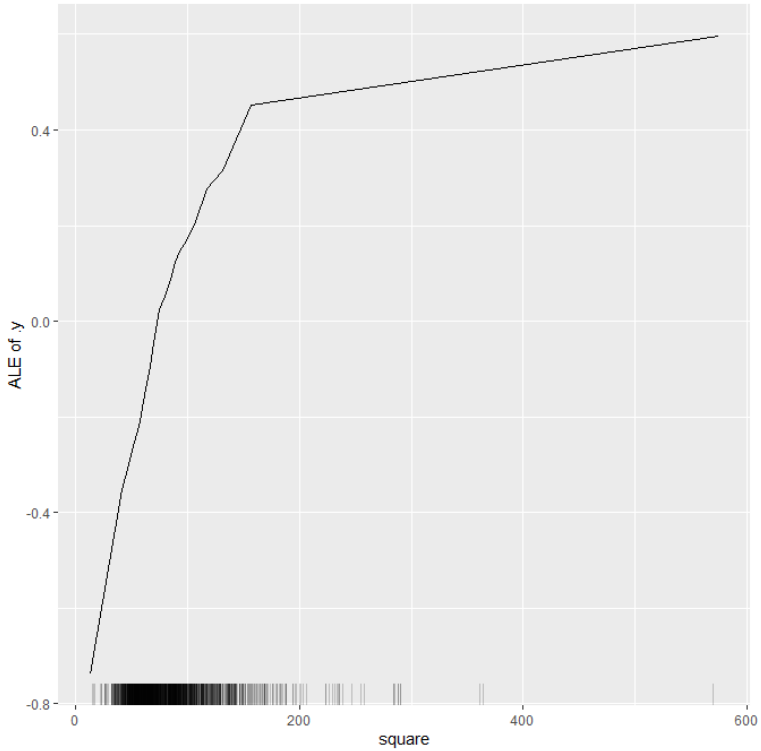
Illustrated



Avoids extrapolation, need not recover additive effects.

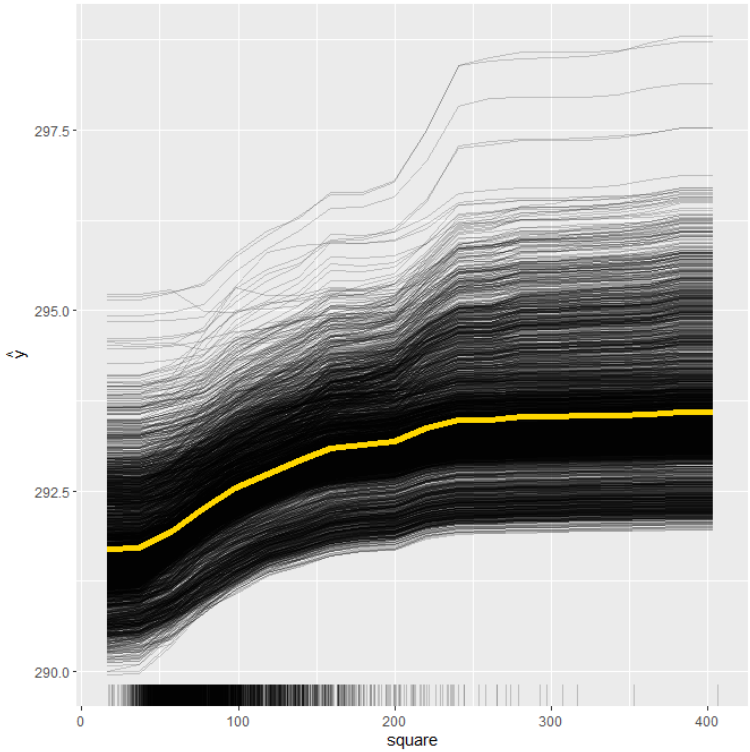
Comparison

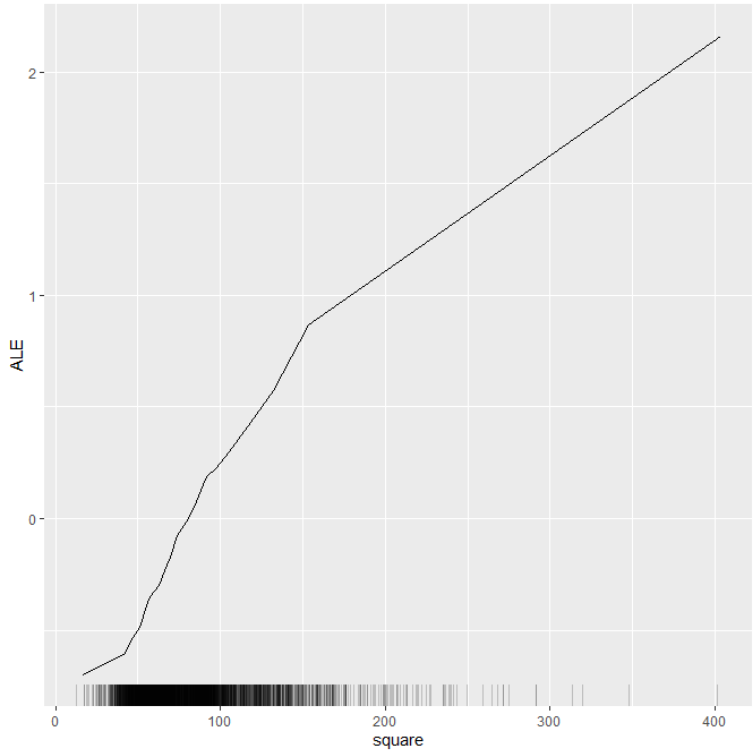

Predicted .y
7
6
5
4
0
200
400
600
square


ALE of .y
0.4
0.0
-0.4
-0.8
0
200
400
600
square

Does it Help with Extrapolation?

When we simulate from a linear model:





Part II: Model Distillation

Instead of *summarizing* a model, *approximate* it with something you can understand.

- Obtain or generate feature examples (pseudo-data)
- Black box "teacher" provides responses to be mimicked by "student"

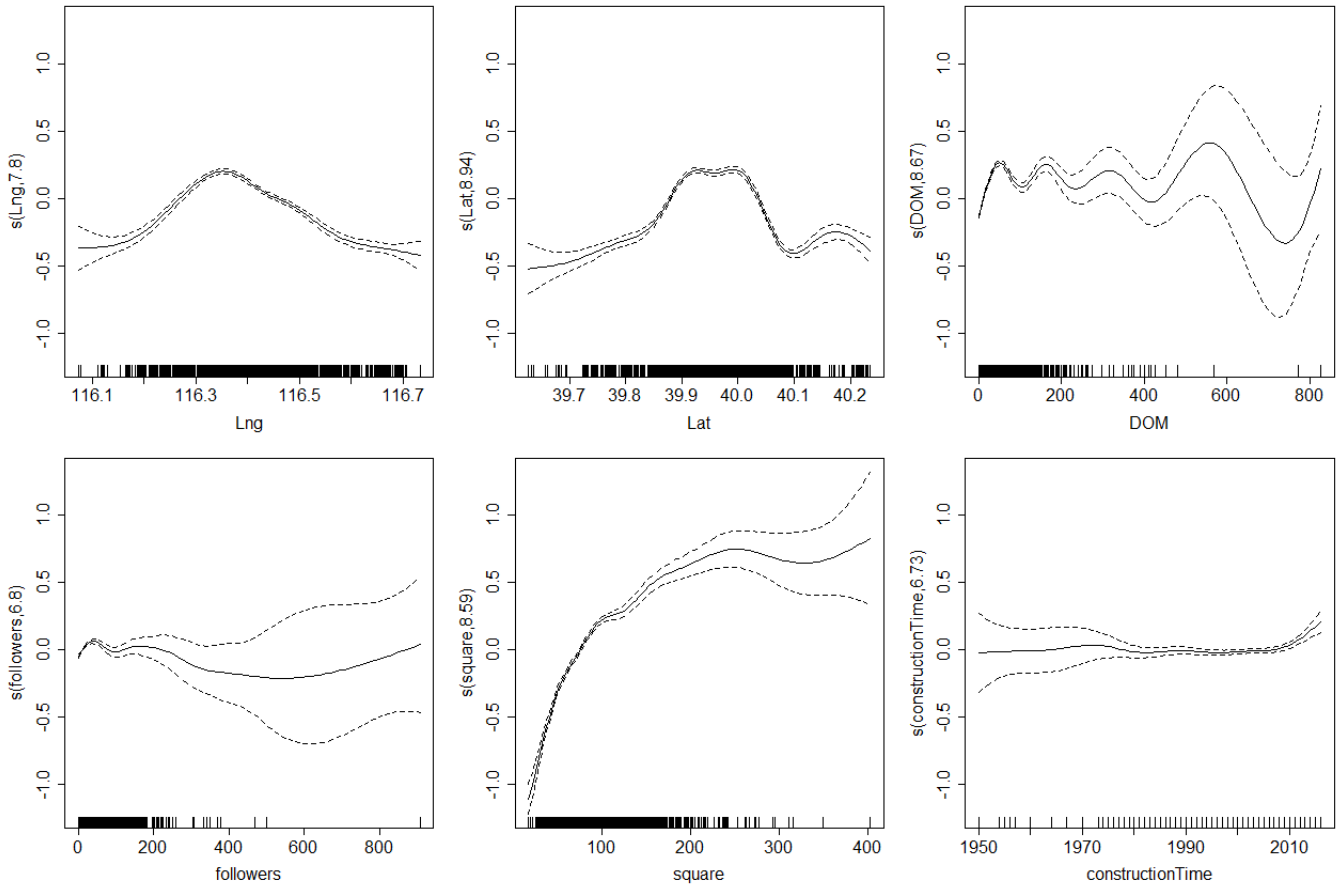
Why not just train the student using the original data?

- Student may serve as approximation only
  - to aid understanding of large patterns
  - as an indicator of spurious behavior
- Student may not perform well at data sizes available, especially if it searches over structures.
- We may want to different distributions over features, eg to localize.

Generalized Additive Models

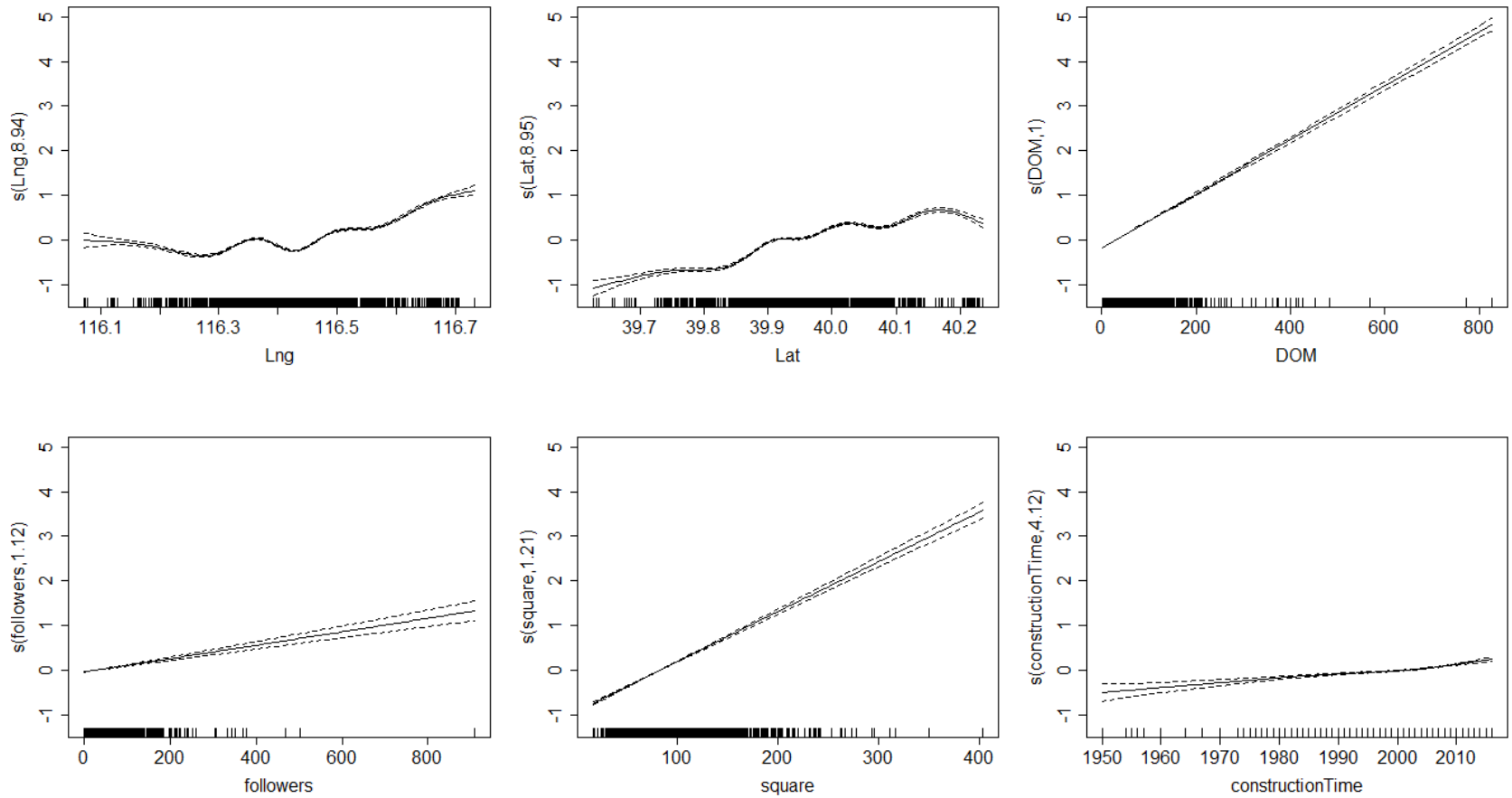
$$f(\boldsymbol{x}) = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

flexible + visualisable univariate functions, but leaves out interactions



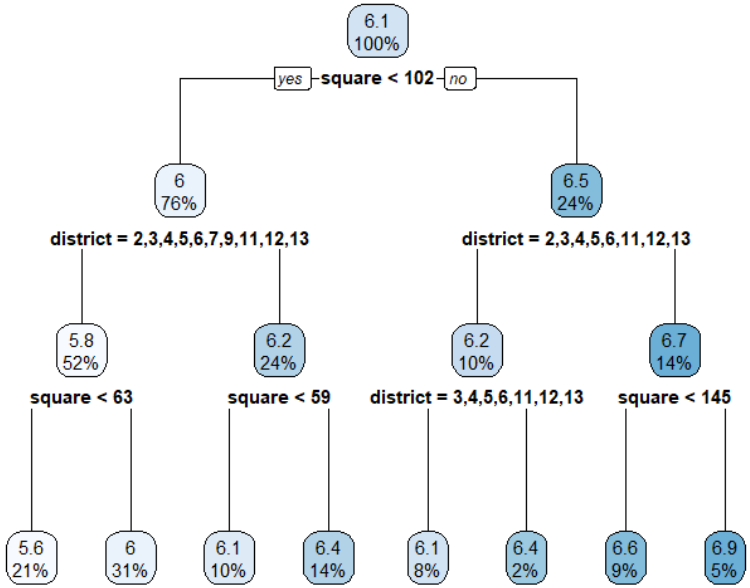
Generalized Additive Models

Distilling into GAMS avoids PD extrapolation issues



Distilling Decision Trees

- CART/C4.5 performs badly because highly variable/divides data
- Distillation $\Rightarrow$ generate as much data as needed for good performance
- Handy explanation for decision: last node before leaf.



Case Study II

Gibbons *et. al.*, 2013, Zhou[2] and Hooker, 2018

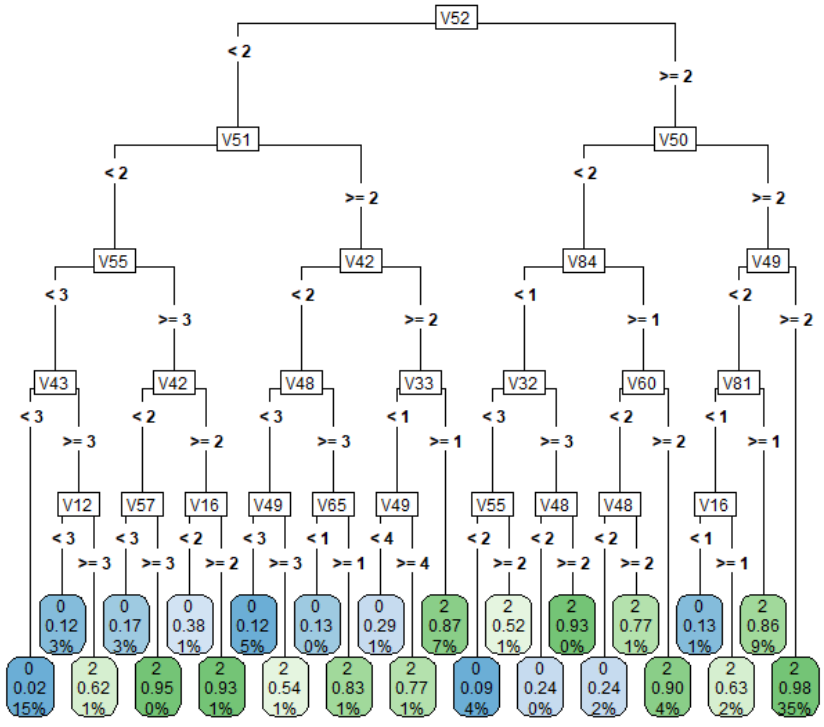
- Psych questionnaires pose significant response burden (depression Q runs to 88 items)
- Can we shorten for screening purposes?
- Decision trees = sequence of questions.
- *Adaptive*: not everyone sees the same items

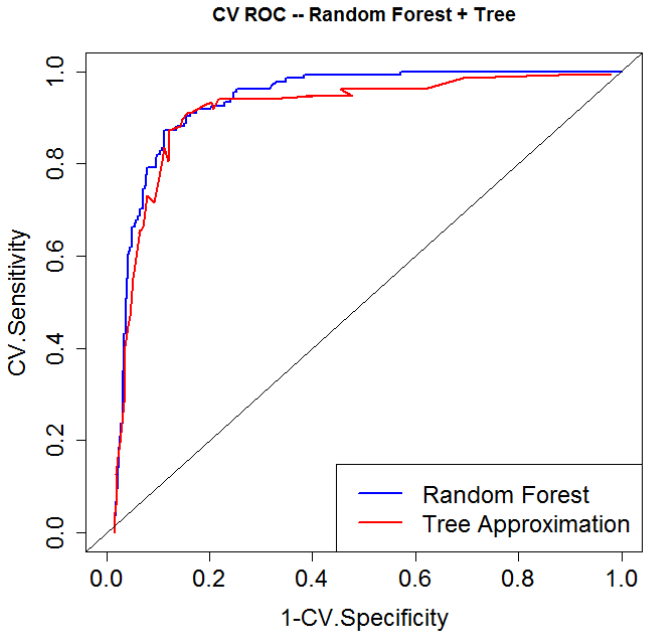
But trees are pretty bad predictors!

- Build random forest to predict depression based on 800 observations
- Generate 12,000 new data points, build tree to predict random forest.

Depth 5 trees = RF accuracy, sensitivity/specificity > 0.8

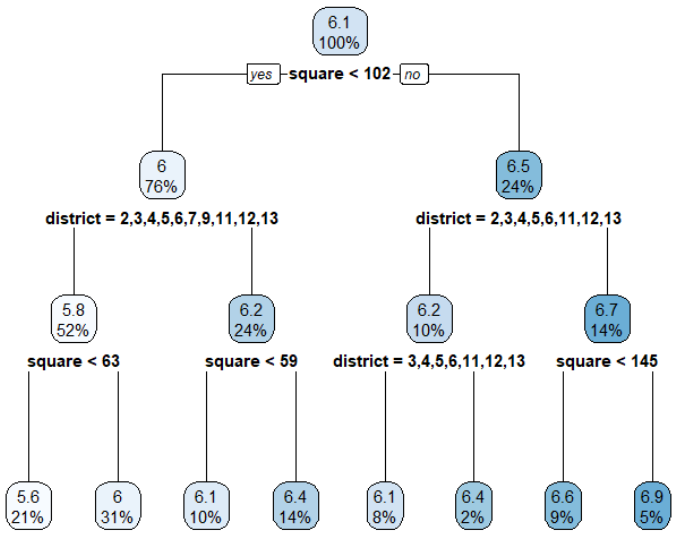
CAD-MDD Tree

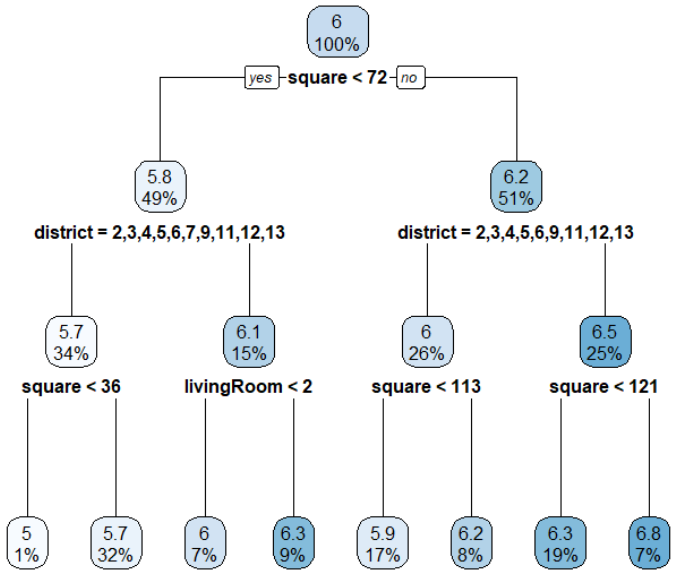

V52
< 2
>= 2
V51
V50
V55
V42
V84
V49
V43
V42
V48
V33
V32
V60
V81
V12
V57
V16
V49
V65
V49
V55
V48
V48
V16
0 0.02 15%
2 0.62 1%
2 0.95 0%
2 0.93 1%
2 0.54 1%
2 0.83 1%
2 0.77 1%
0 0.09 4%
0 0.24 0%
0 0.24 2%
2 0.90 4%
2 0.63 2%
2 0.98 35%
0 0.12 3%
0 0.17 3%
0 0.38 1%
0 0.12 5%
0 0.13 0%
0 0.29 1%
2 0.87 7%
2 0.52 1%
2 0.93 0%
2 0.77 1%
0 0.13 1%
2 0.86 9%


CV ROC -- Random Forest + Tree
CV.Sensitivity
1-CV.Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest
Tree Approximation

Distillation Reproducibility

RF trained with 3,000 points, trees distilled using 20,000 but still get different answers.





When Does Reproducibility Matter?

Unstable distilled model may be ok if:

- Student replaces teacher for prediction; e.g. for compression.
- "Here is our formula" suffices as an explanation.
- Student is not re-distilled (or not frequently).

But may be problematic when:

- Student model only used as approximation
- Explanations are intended as *justifications* (usually based on causal reasoning).
- Explanations are intended to motivate actions.
- Student is re-distilled frequently.
- Uncertainty due to distillation not easily represented (eg searches over structures).

Particularly problematic for local distillation.

Part III: Local Explanations

Much of recent attention around individual predictions

Why was my loan not approved?

Designed to

- Satisfy a "right to an explanation"
- Provide recourse for adverse decisions
- Provide a basis for disagreements (eg in treatment recommendations)
- Used as a surrogate for/alternative to global understanding.

Local/Global diagnostics can provide very different information:

- Global variable importance: *What large-scale changes make most difference across the data set?*
- Summaries of local explanations: *What small-changes make most difference to individual predictions?*

But many of the same considerations apply.

Local Interpretable Model-agnostic Explanations (LIME)

What features are most important $\approx$ derivative of prediction w.r.t. x_j, but

- not all models are differentiable
- derivatives can be unstable
- large feature set = need to select a few to present

LIME builds a local LASSO model:

$$\beta = \sum_{q=1}^{Q} w(X_i, Z_q) \left(f(Z_q) - \beta_0 - \sum \beta_j Z_{qj} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

for Z_q generated (or weighted) locally around X_i.

λ chosen to return 5 to 8 (tunable) features.

Distillation Stability Again

LIME is a distillation method, but surely linear regression is pretty stable? Let's use 1000 pseudo-examples.

Here we have applied LIME to the first point in the test set:

square	livingRoom	district6	district8	district10
0.00467	0.03950	-0.11350	0.07707	0.10783

We'll repeat the exercise but re-draw the 1000 pseudo-examples:

square	livingRoom	drawingRoom	district6	district10
0.002760	0.06624	0.001653	-0.02000	0.001565

Distilling with enough data will stabilize, but sample sizes needed are big.

Local Explanations: SHAP

Rather than how predictions change with the value of a feature

How does knowing the value of a feature change the prediction?

Define

$$f_S(x_S) = \int f(x_S, x_{-S})\mu(x_{-S}|x_S)dx_{-S}$$

by integrating out the subset of features x_{-S}.

- Over the marginal distribution of x_{-S}, independent of x_S (see $\text{PD}_j(x)$).
- Over the conditional distribution, estimated with a kernel density
- By re-learning to predict y from X_S.

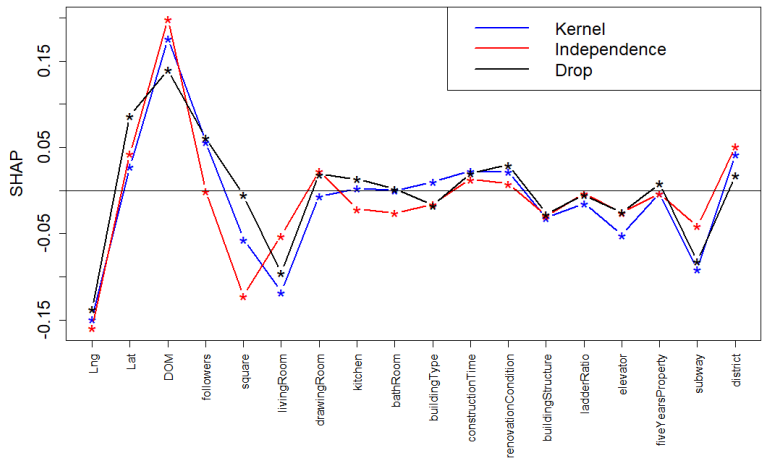
Examine change in prediction when adding x_j to x_S:

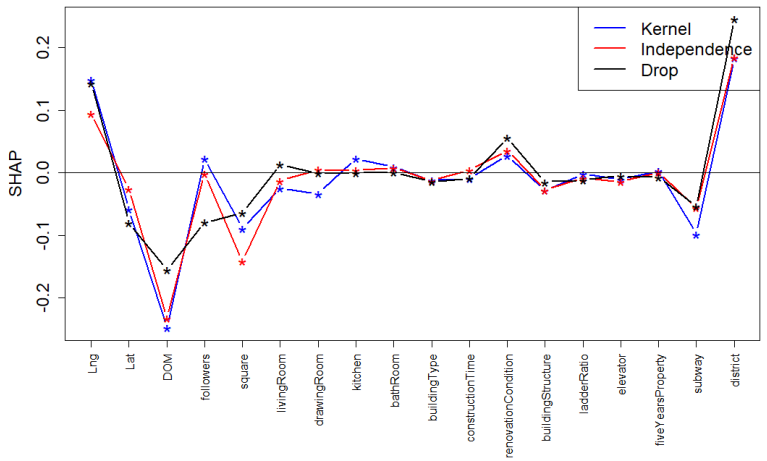
$$\Delta_S^j(x) = f_{S\cup j}(x_{s\cup j}) - f_S(x_s)$$

and follow Shapley by averaging over sequence of features to add.

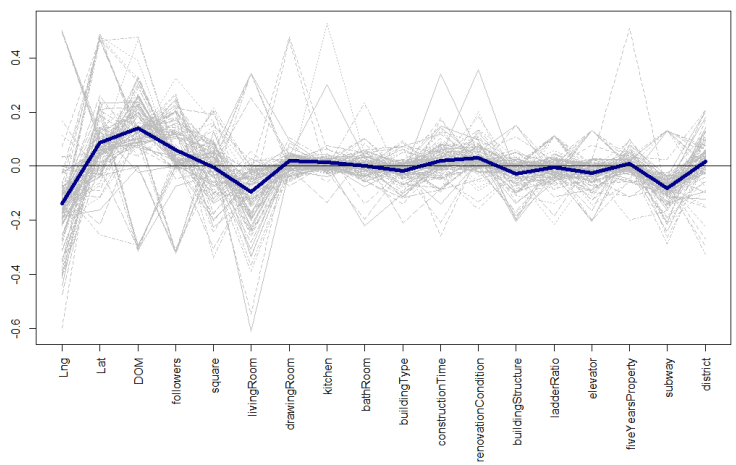
SHAP to Explain Test Points

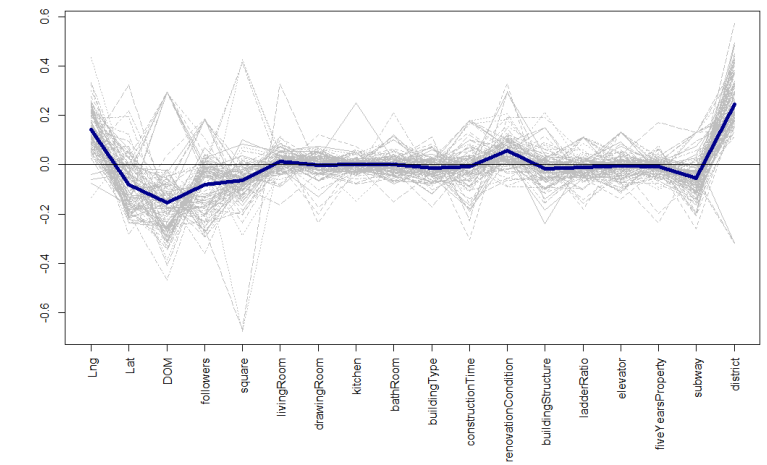
Different integration operators:





$\Delta_S^j(x)$ over different S:





Local Explanations for Deep Learning: Saliency

- Explanations require features to be individually meaningful.
- Eg image data:
  - no pixel values are individually interpretable
  - but *patterns* of what influences prediction most might be.
- Instead, consider local gradients $\nabla f(x)$
- In deep learning, fits neatly into back-propagation.

Greyhound (vanilla)

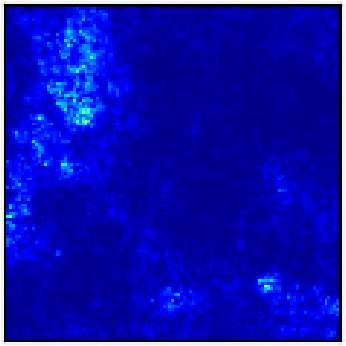
Soup Bowl (vanilla)

Eel (vanilla)

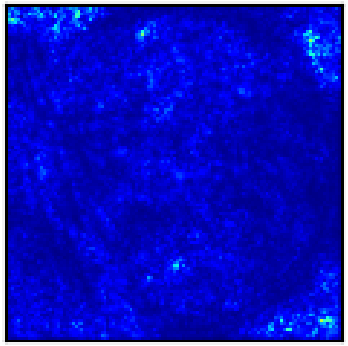
(from https://christophm.github.io/interpretable-ml-book)

With Attributions

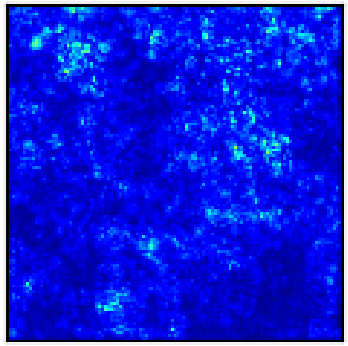
Greyhound (vanilla)



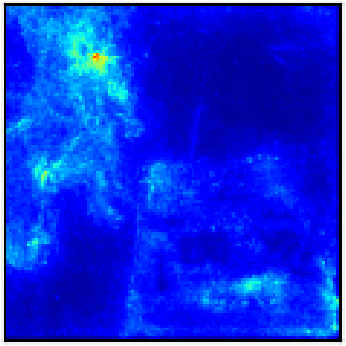
Soup Bowl (vanilla)



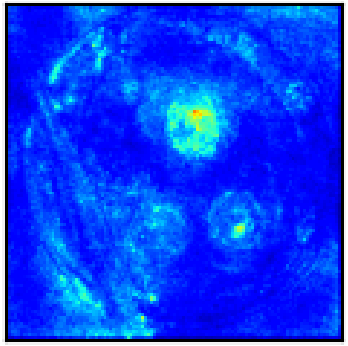
Eel (vanilla)



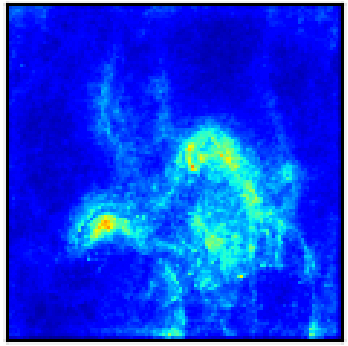
Greyhound (Smoothgrad)



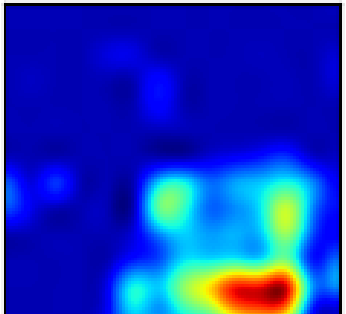
Soup Bowl (Smoothgrad)



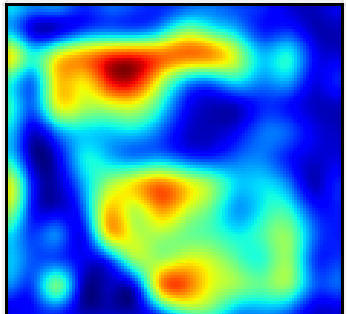
Eel (Smoothgrad)



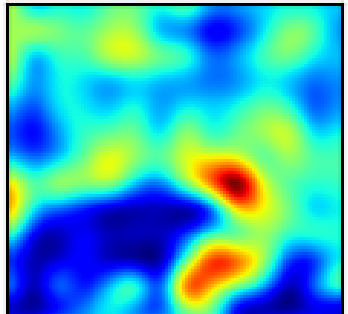
Greyhound (Grad-Cam)



Soup Bowl (Grad-Cam)



Eel (Grad-Cam)



Variations on Gradients

Empirically:

- saliency maps very unstable to perturbing x
- can find imperceptible perturbations that significantly change explanation without changing classification

SmoothGrad solution

- Add noise to each pixel and calculate gradients
- average over many realizations
- Target = convolution $\int \nabla f(\mathrm{x}+\mathrm{z})\phi_\sigma(\mathrm{z})dz$
- Expensive + need to pick noise variance

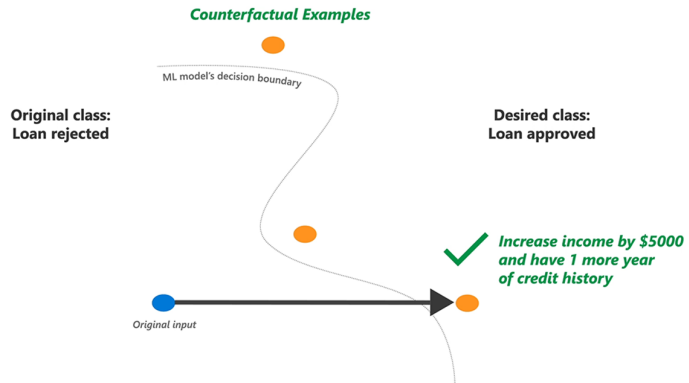
Grad-CAM: focusses on convolutional layers, and thresholds by direction towards a class of interest.

Counterfactual Explanations

How could I change this decision?

If at x, find nearest x^* so that

- $f(x^*) =$ desired outcome
- x and x^* are close
- $x - x^*$ is sparse
- (x^* is realistic?)



But:

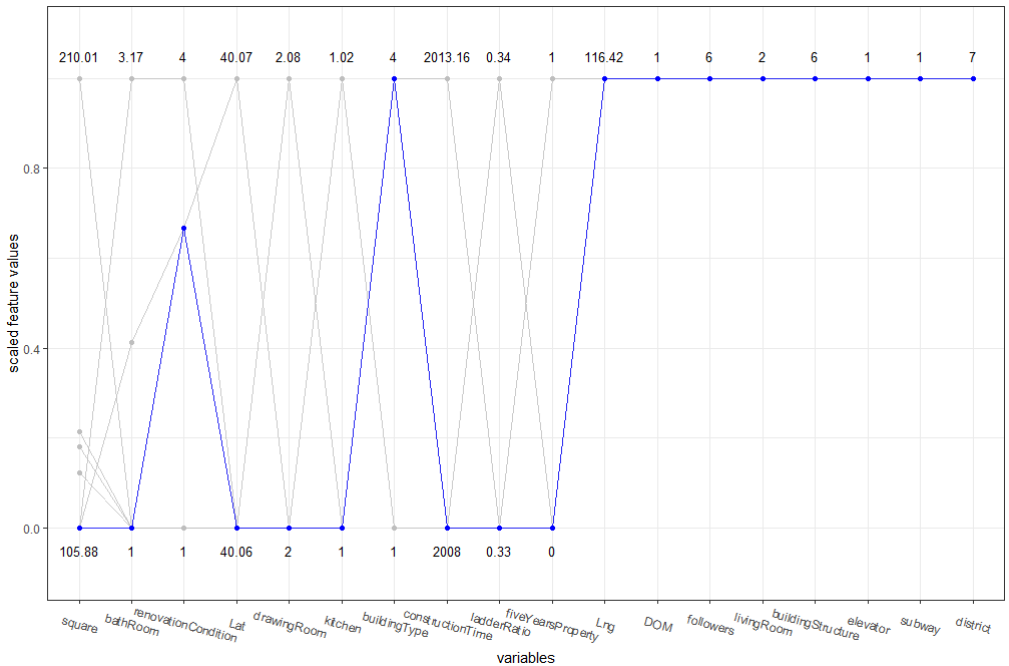
- Challenging (model-specific) optimization
- "close" = ??

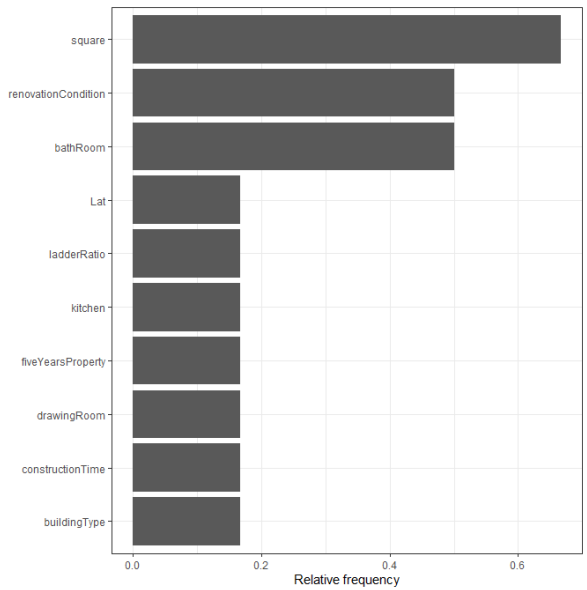
But proposed as legal way to satisfy provision of recourse.

Counterfactual Explanations

How do I get more than 300 for my apartment?

	square	drawingRoom	kitchen	bathRoom	Type	Time	Cond	ladderRatio
1:	105.8800	2.08373	1.000000	1.8945558	4	2008	3	0.333000
2:	128.0949	2.00000	1.000000	1.0000000	4	2008	1	0.333000
3:	124.7427	2.00000	1.000000	1.0000000	4	2008	1	0.333000
4:	210.0132	2.00000	1.020913	0.9969316	1	2008	3	0.343342
5:	105.8800	2.00000	1.000000	3.1720178	4	2013	4	0.333000
6:	118.6187	2.00000	1.000000	1.0000000	4	2008	3	0.333000





From Local to Global

- Most local explanations = some form of feature attribution
- Some explicit: saliency/gradCAM, LIME, SHAP
- Some less so:
  - Counterfactuals – feature difference with nearest positive class.
  - Anchors (local rules) – use a subset of features (could also provide weights).
- Framework: for each input x, $f(x)$ also comes with attribution $A(x)$
- Summarize collection of $A(X_i)$ for global understanding.

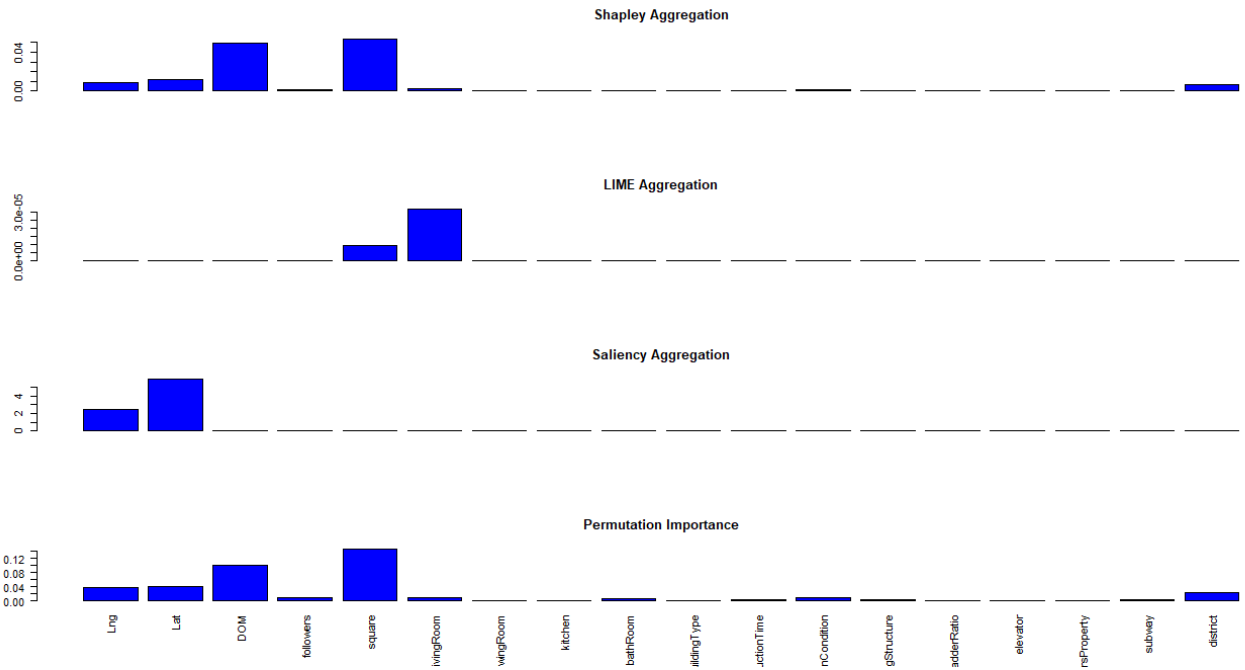
Will use Shapley, LIME, Saliency (from finite differences) for RF trained on Beijing Housing Data for convenience.

Variable Importance

Some summary of distribution:

$$V_j = E_X S(A_j(X))$$

estimated from training/test/uniform data.

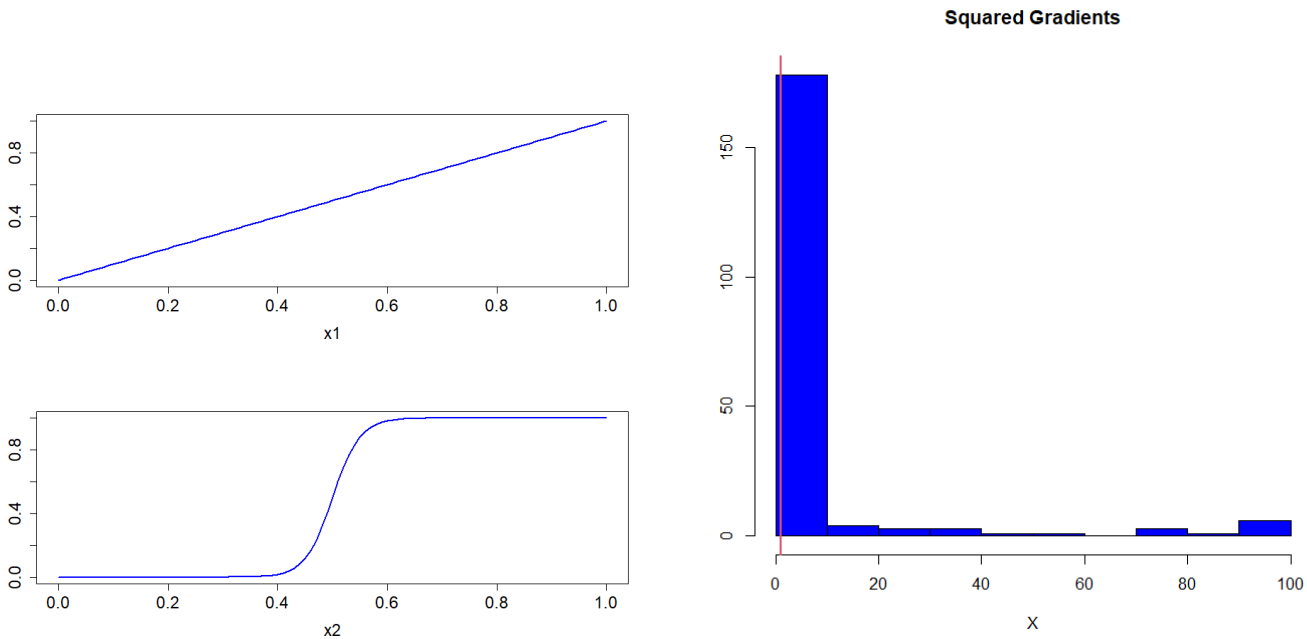


Global *versus* Local

Accumulating local models tells you what is important for each local effect; can be different from global importance:

$$f(x_1, x_2) = x_1 + \text{logit}(10x_2)$$

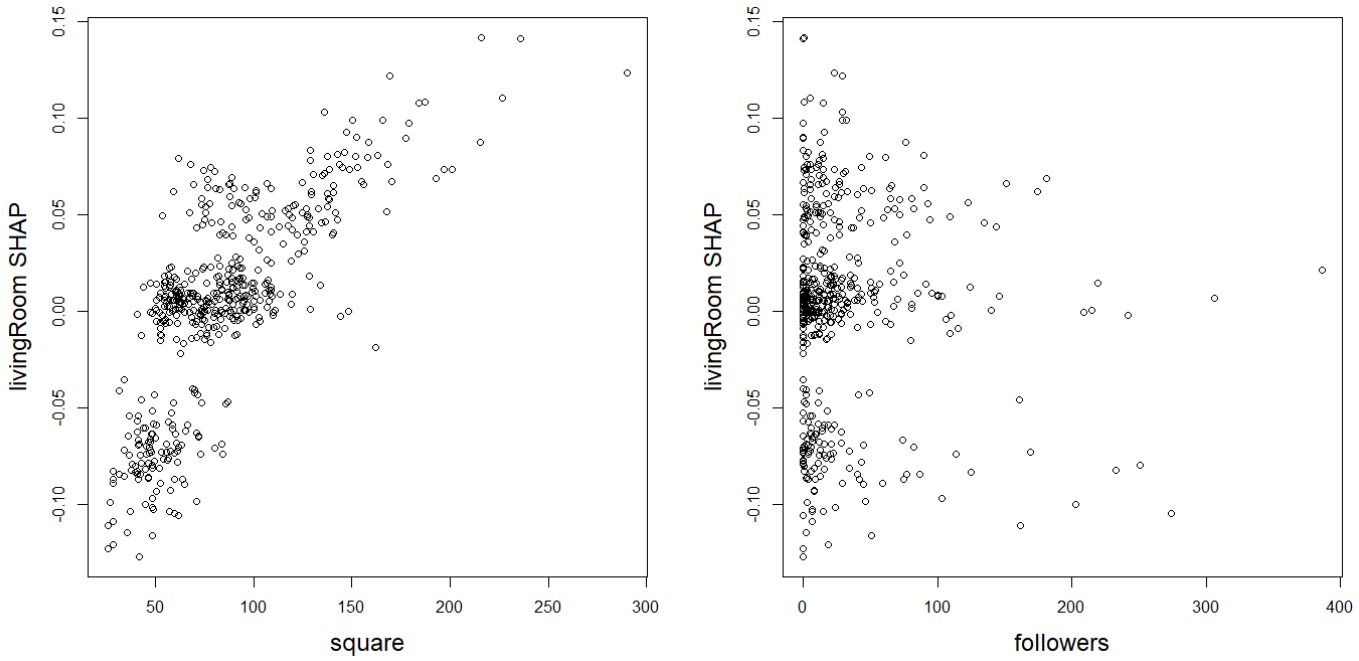
With, $x_1, x_2 \sim U[0, 1]$ x_2 gradients mostly much smaller (3e-6 vs 1) although mean squared gradient is still large (6 vs 1).



For most points, x_2 makes little difference, but global variance is large.

Relationships with Features

How do feature attributions change across feature space?



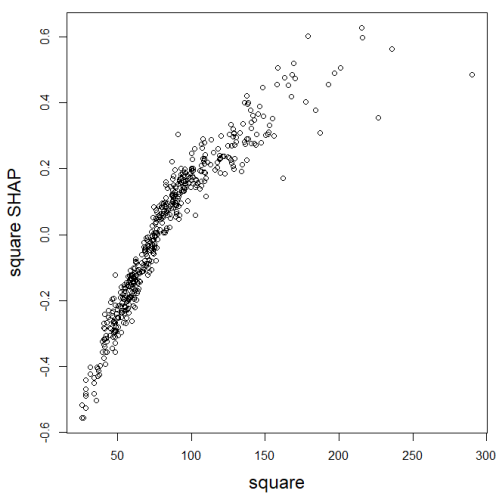
Build (interpretable?) model to predict importance of x_j from other features.

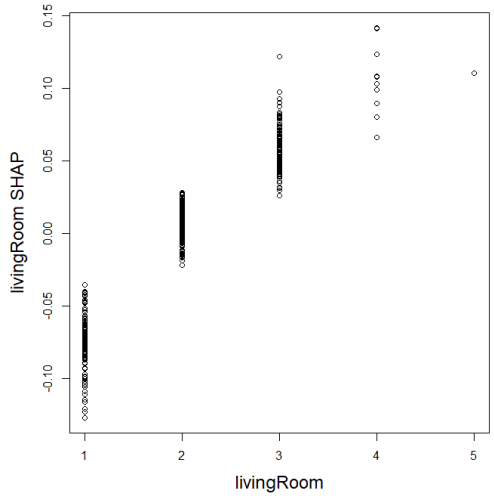
- Partial dependence plots *within* regions of a tree.
- Cluster explanations to select prototypes.

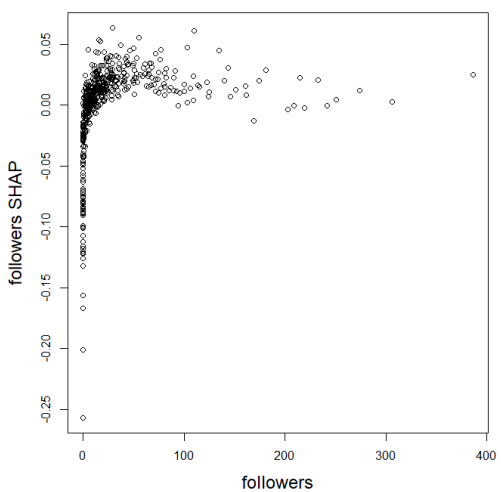
Connections to Additivity

$$f(x) = \sum g_j(x_j)$$

- Most $A_j(x)$ preserve some aspect of $g_j(x)$, and ignore $g_k(x)$.
- (Especially if $X_j \perp X_k$)
- $\Rightarrow (X_{ij}, A_j(X_i))$ should be 1:1
- Non-additivity measured by spread.







Summary and Messages

- Field fast moving, many proposals, not all thought-out
- Warnings:
  - Feature dependencies make a difference
  - Beware of creating unreasonable feature combinations
  - Searching for structure produces instability
- Simple checks:
  - Does this method give me what I ought to find if I start from a known model? (Apply to both model and ML alg that has tried to learn it).
  - Do I get the same answer if I re-run with a different seed? (Not always sufficient).
- Questions of strategy
  - What do I want to know about this model? Does this approach answer that?
  - Who is the audience for this explanation? Will they understand what they are seeing? Do I want them to?

Happy Playing! But be careful.

Highly Curated Relevant Papers

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Leo Breiman, 2001, "Random Forests", *Machine Learning*

Jerome Friedman, 2001, "Greedy function approximation: a gradient boosting machine", *Annals of Statistics*

Carolin Strobl, Anne-Laurie Boulestreix, Achim Zeileis and Torsten Hothorn, "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution", *BMC Bioinformatics*

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

Slack, Dylan, et al. "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841.

Self-Serving Bibliography

Hooker, G., L. Mentch and S. Zhou, 2021, "Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance", Statistics and Computing, in press.

Zhou, Z., G. Hooker and F. Wang, 2021, "S-LIME: Stabilized-LIME for Model Explanation", KDD21

Zhou, Y., Z. Zhou and G. Hooker, 2018, "Approximation Trees: Statistical Stability in Model Distillation", arxiv.org/abs/1808.07573

Tan, S., R. Caruana, G. Hooker and Y. Lou, 2018, "Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation"", AAAI/ACM Artificial Intelligence, Ethics, and Society 2018.

RD Gibbons, G. Hooker, MD Finkelman, DJ Weiss, PA Pilkonis, E. Frank, T. Moore and DJ Kupfer, 2013, "Computerized Adaptive Diagnosis of Depression Using the CAD-MDD", Journal of Clinical Psychiatry, 74(7):669-674.

G. Hooker, 2007. "Generalized Functional ANOVA Diagnostics for High Dimensional Functions of Dependent Variables". Journal of Computational and Graphical Statistics, 16:709-732.