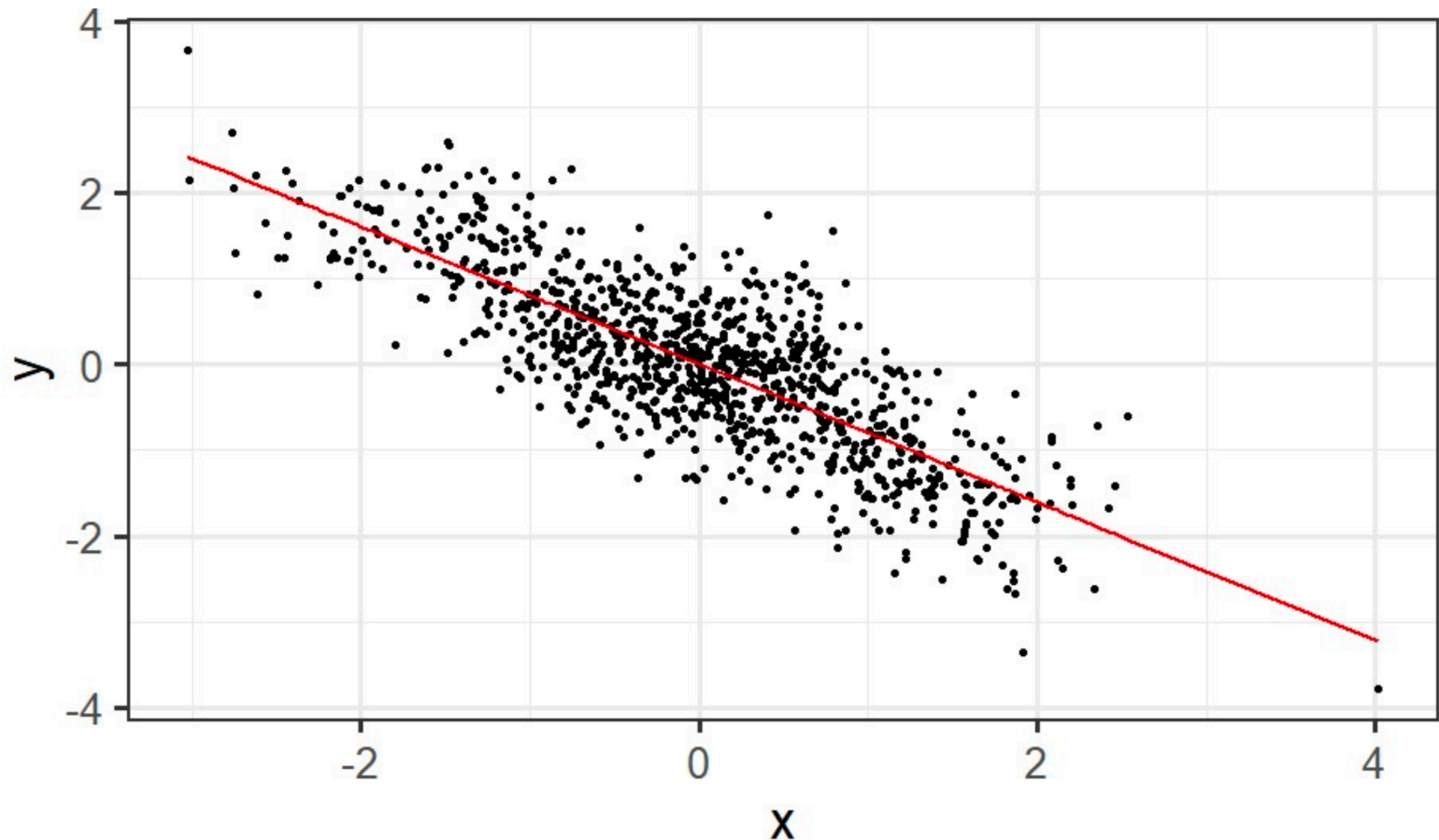


Model Agnostic Interpretability

Making sense of complex models

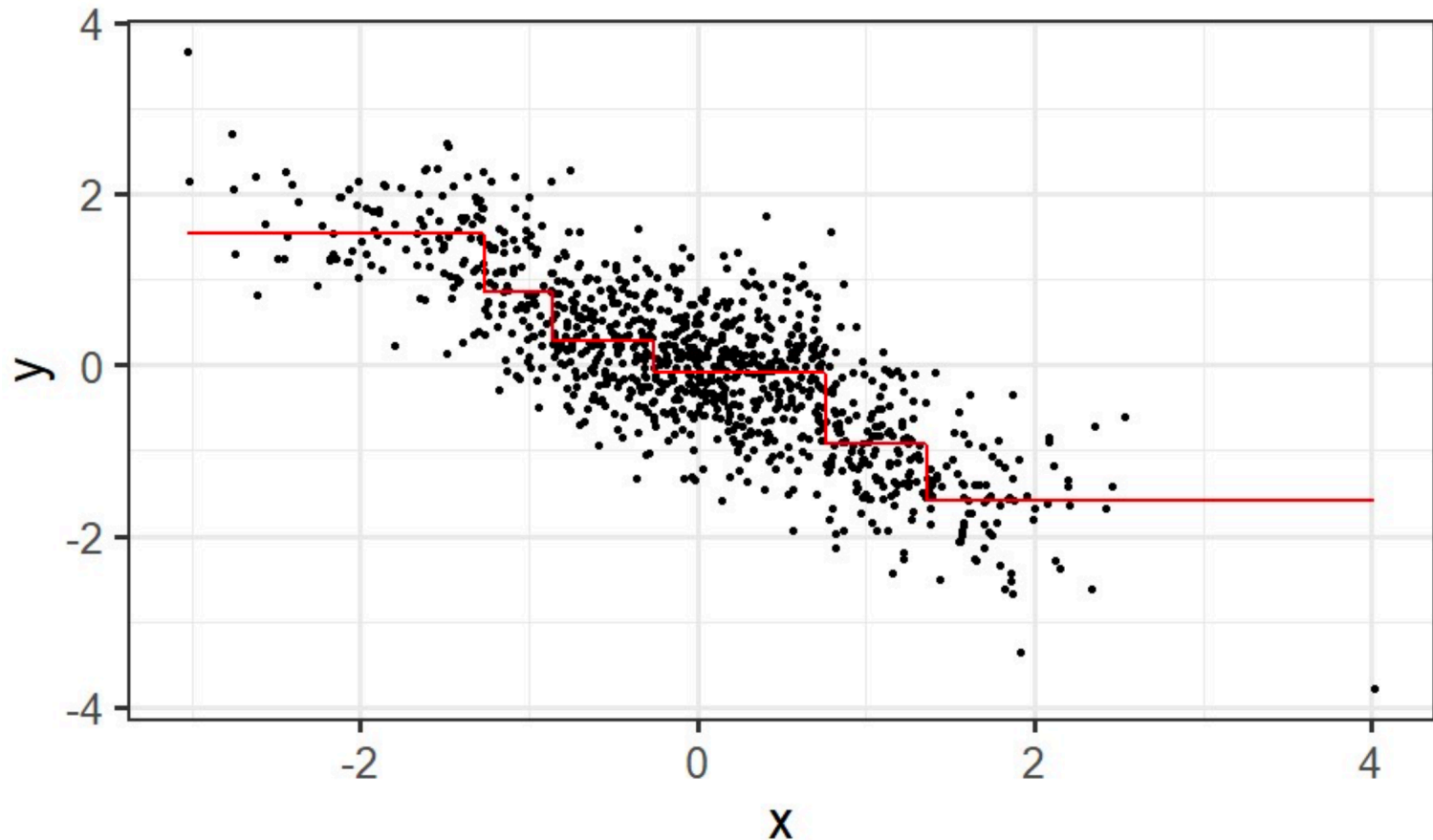
Interpretable Models

Generalized Linear Models: If x increases by 1 unit, y increases by β units



Interpretable Models

Decision Trees: If conditions A, B, and C are satisfied, then then y is If conditions !A and D are satisfied, then y is....



Other Models

Random Forests?

No.

Gradient Boosting Machines?

Sorry, no.

Neural Networks?

Try again.

Support Vector Machines?

Nah.

Naive Bayes?

 *Nope!*

Motivation

- Humans want to interpret and understand model behavior
- We have questions!
 - *Why* was this person's loan application rejected?
 - *Why* is the symptom occurring in this patient?
 - *Why* is the stock price expected to go down?
- Interpretations can be model and context dependent
 - Model dependent: variable importance in regression has different implications than variable importance in trees.
 - Context dependent: the effect of, say age, on a response may depend on an individual's age and other factors.
(i.e. nonlinearity)

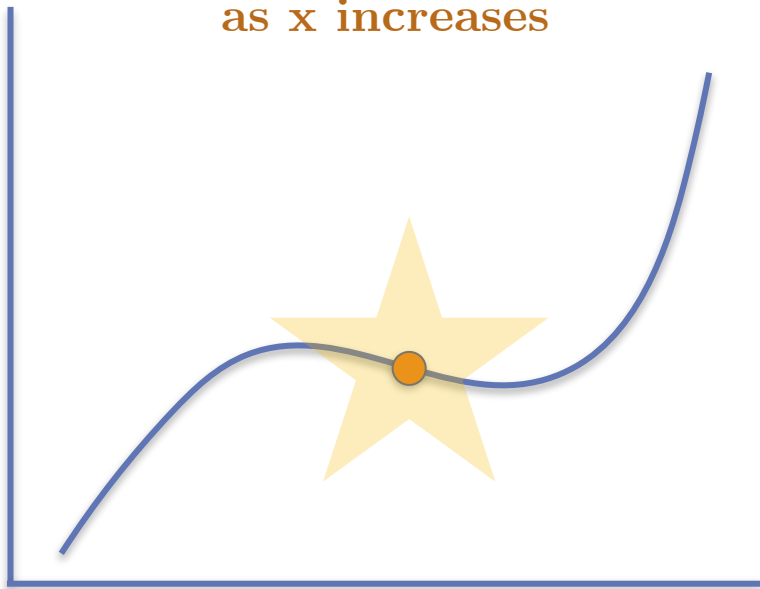
Why so important?

- Fairness/Transparency
 - Understanding modeled decisions improves consumer trust
 - Interpretations reveal model behavior on different groups of people (including marginalized groups)
- Model Robustness and Integrity
 - Interpretability methods can reveal odd model behavior or issues with overfitting.
- Adverse Action notice requirements
 - Equal Credit Opportunity Act (ECOA)
 - Fair Credit Reporting Act (FCRA)
 - More likely to come with the tide of Ethical Machine Learning.

Types of Model Interpretability

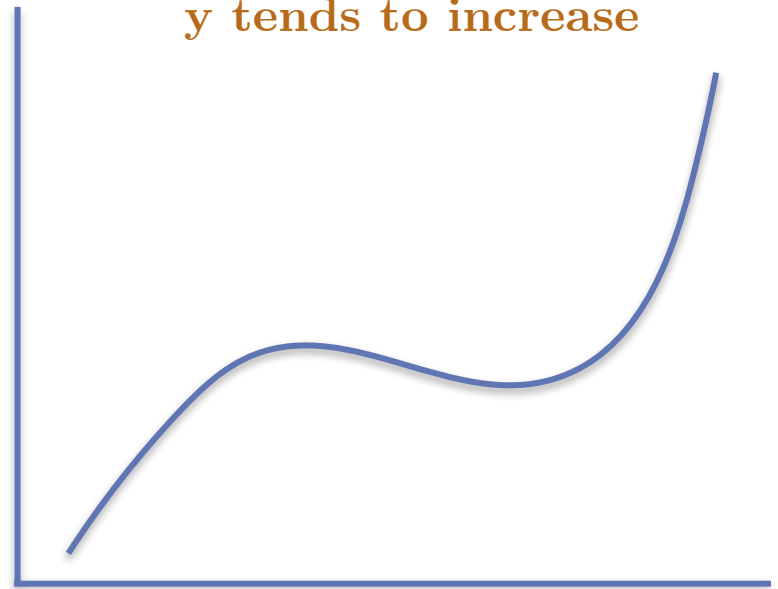
Local

When $x=10$, y decreases
as x increases



Global

As x increases,
 y tends to increase



Types of Model Interpretability

	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE LIME Shapley Values	Permutation Importance Partial Dependence ALE

Types of Model Interpretability

	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE LIME Shapley Values	Permutation Importance Partial Dependence ALE

Boston Dataset

```
#' crim -- per capita crime rate by town.  
# ' zn -- proportion of residential land zoned for lots over 25,000 sq.ft.  
# ' indus -- proportion of non-retail business acres per town.  
# ' chas -- Charles River dummy variable (= 1 if tract bounds river; 0  
otherwise). nox nitrogen oxides concentration (parts per 10 million).  
# ' rm -- average number of rooms per dwelling.  
# ' age -- proportion of owner-occupied units built prior to 1940.  
# ' dis -- weighted mean of distances to five Boston employment centres.  
#rad -index of accessibility to radial highways.  
# ' tax -- full-value property-tax rate per \$10,000.  
# ' ptratio -- pupil-teacher ratio by town.  
# ' black --  $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of black residents  
by town.  
# ' lstat lower status of the population (percent).  
# ' medv -- median value of owner-occupied homes in \$1000s.
```

Predicting target nox – *nitrogen oxides concentration (parts per 10 million)*.

Types of Model Interpretability

	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE LIME Shapley Values	Permutation Importance Partial Dependence ALE

Permutation Importance

...

“Let me show you how much worse the predictions of our model get if we input randomly shuffled data values for each variable”

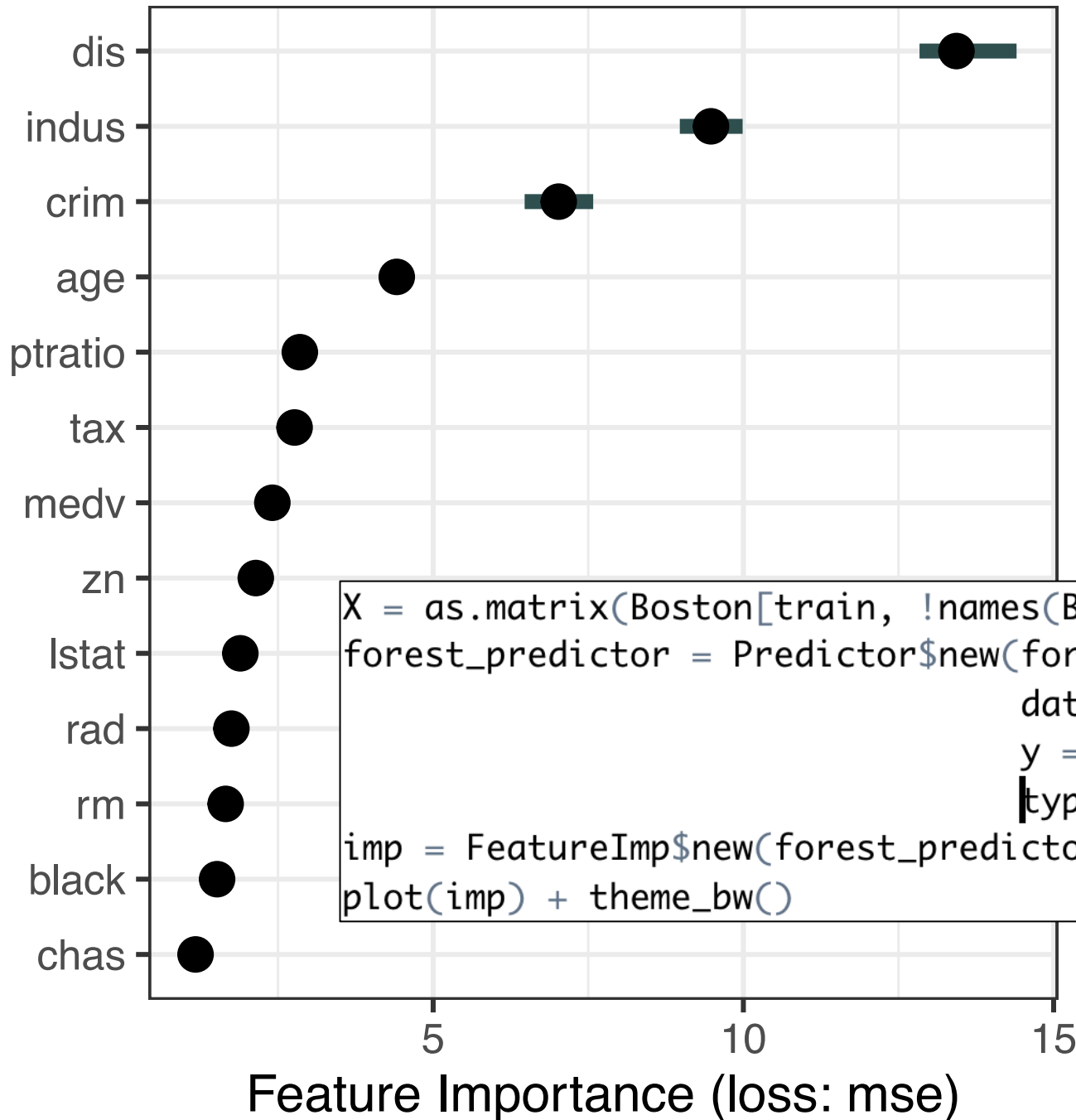
Permutation Importance

- If a variable is important, the model should get worse when that variable is removed.
- To make a direct comparison, rather than *remove* the variable from the model, we'll *destroy its signal*.
- By randomly permuting that values in that column of data, we *break* the true relationship and make it nonsense.
- How much worse does the model get when we do?
(on average, over default n=5 permutations)
=> Permutation Feature Importance

```
train = sample(c(T,F),nrow(Boston),replace=T,p=c(0.75,0.25))  
forest = randomForest(nox~.,data=Boston[train,])
```

```
# Linear Model for Comparison  
f = lm(nox~., data=Boston[train,])
```

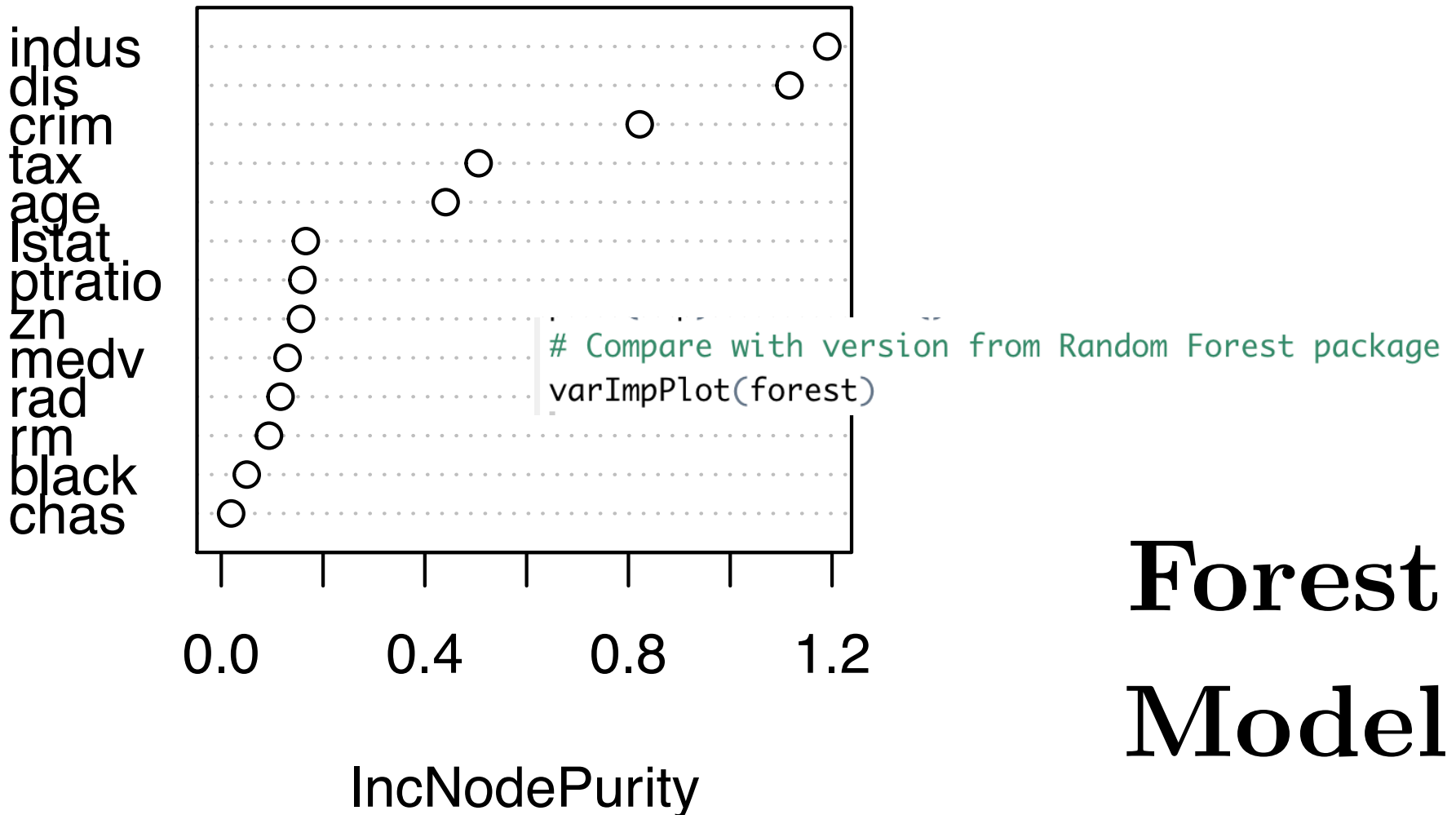
Forest Model



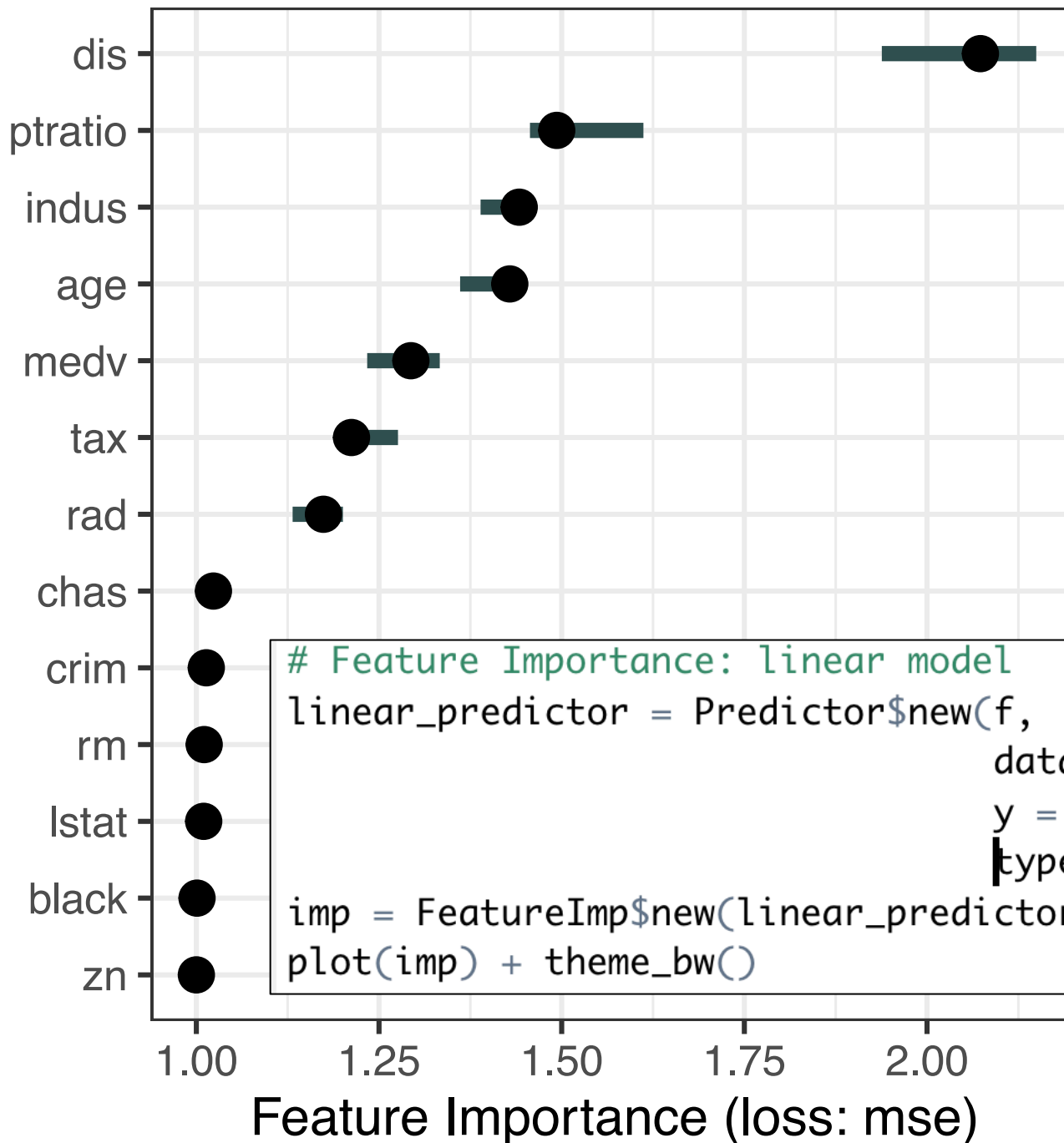
```
X = as.matrix(Boston[train, !names(Boston) %in% c("nox")])
forest_predictor = Predictor$new(forest,
                                data = as.data.frame(X),
                                y = Boston[train, "nox"],
                                type = "response")
imp = FeatureImp$new(forest_predictor, loss = "mse")
plot(imp) + theme_bw()
```

Compare to randomForest

Variable Importance



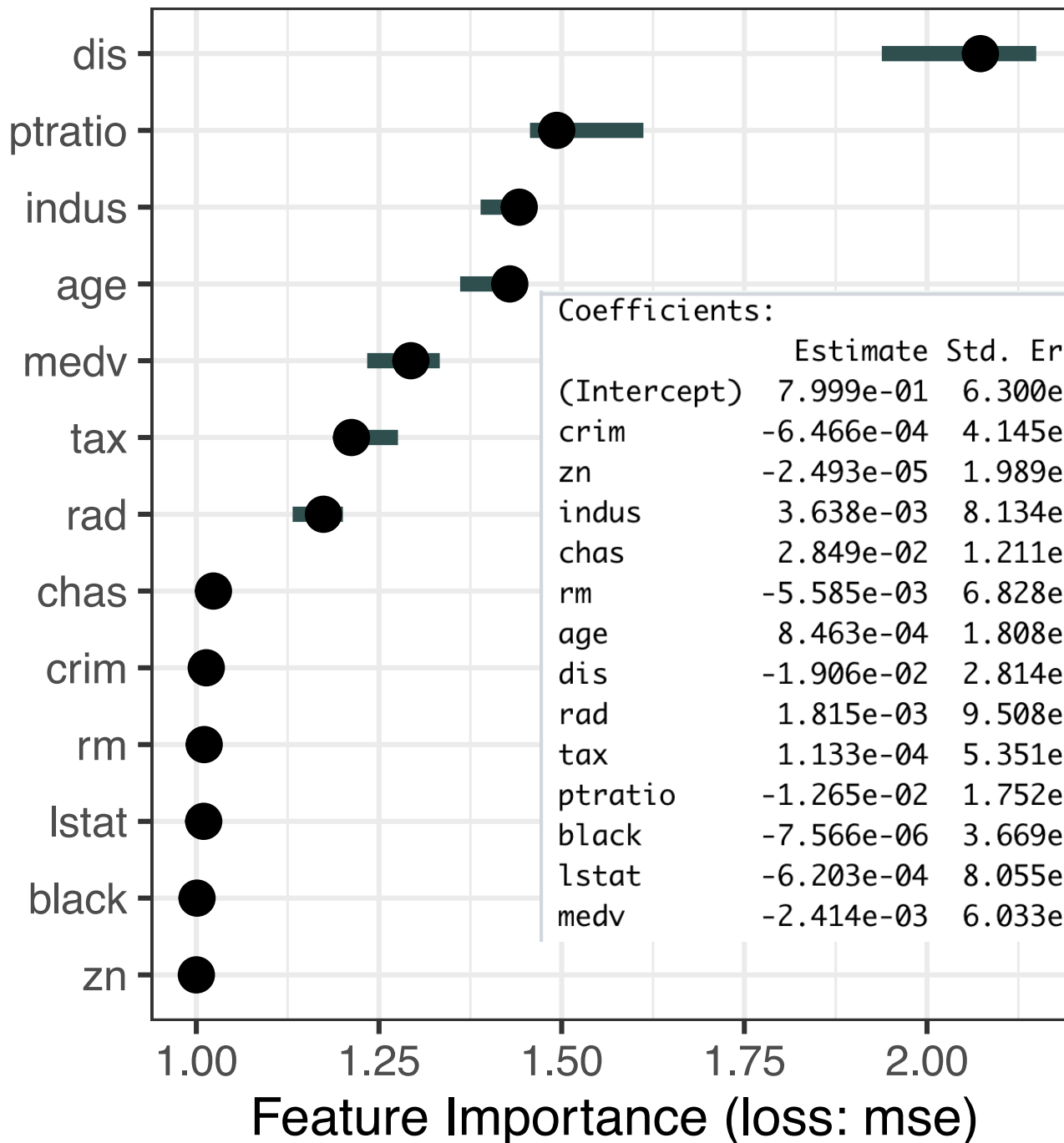
Linear Model



```
# Feature Importance: linear model
linear_predictor = Predictor$new(f,
                                data = as.data.frame(X),
                                y = Boston[train,"nox"],
                                type = "response")

imp = FeatureImp$new(linear_predictor, loss = "mse")
plot(imp) + theme_bw()
```


Linear Model



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	7.999e-01	6.300e-02	12.698	< 2e-16	***
crim	-6.466e-04	4.145e-04	-1.560	0.1196	
zn	-2.493e-05	1.989e-04	-0.125	0.9003	
indus	3.638e-03	8.134e-04	4.472	1.05e-05	***
chas	2.849e-02	1.211e-02	2.352	0.0192	*
rm	-5.585e-03	6.828e-03	-0.818	0.4139	
age	8.463e-04	1.808e-04	4.682	4.08e-06	***
dis	-1.906e-02	2.814e-03	-6.773	5.36e-11	***
rad	1.815e-03	9.508e-04	1.909	0.0571	.
tax	1.133e-04	5.351e-05	2.117	0.0349	*
ptratio	-1.265e-02	1.752e-03	-7.223	3.21e-12	***
black	-7.566e-06	3.669e-05	-0.206	0.8367	
lstat	-6.203e-04	8.055e-04	-0.770	0.4418	
medv	-2.414e-03	6.033e-04	-4.002	7.68e-05	***

Types of Model Interpretability

	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE LIME Shapley Values	Permutation Importance  Partial Dependence ALE

Individual Conditional Expectation (ICE)

...

“Let me show you how the predictions for each observation change if we vary the feature of interest.”

Individual Conditional Expectation (ICE)

- This is a *local* method because it visualizes the dependence of an *individual prediction* on a given input variable.
- Fix all other variables for a single observation while varying the feature of interest.
- Plot the resulting prediction vs the feature of interest.

Individual Conditional Expectation (ICE)

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	234	4.3	0.469	0.454
5	17.2	41	9.1	71	326	2.5	0.538	0.512
6	20.1	31	15.2	88	222	5.1	0.458	0.470
2	15	22	5.2	45	430	6.3	0.556	0.561

Choose a variable of interest and a single observation.

Individual Conditional Expectation (ICE)

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	234	4.3	0.469	0.454
5	17.2	41	9.1	71	326	2.5	0.538	0.512
6	20.1	31	15.2	88	222	5.1	0.458	0.470
2	15	22	5.2	45	430	6.3	0.556	0.561

Choose a **variable of interest** and a **single observation**.

Individual Conditional Expectation (ICE)

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	

Replicate single observation, holding constant data values on other variables.

Individual Conditional Expectation (ICE)

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	200	4.3	0.469	
3	12.1	13	22.6	90	201	4.3	0.469	
3	12.1	13	22.6	90	202	4.3	0.469	
3	12.1	13	22.6	90	203	4.3	0.469	
3	12.1	13	22.6	90	204	4.3	0.469	
3	12.1	13	22.6	90	205	4.3	0.469	
3	12.1	13	22.6	90	206	4.3	0.469	

Fill in values for variable of interest
across the entire range of the variable

Individual Conditional Expectation (ICE)

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	200	4.3	0.469	0.448
3	12.1	13	22.6	90	201	4.3	0.469	0.449
3	12.1	13	22.6	90	202	4.3	0.469	0.450
3	12.1	13	22.6	90	203	4.3	0.469	0.450
3	12.1	13	22.6	90	204	4.3	0.469	0.451
3	12.1	13	22.6	90	205	4.3	0.469	0.452
3	12.1	13	22.6	90	206	4.3	0.469	0.452

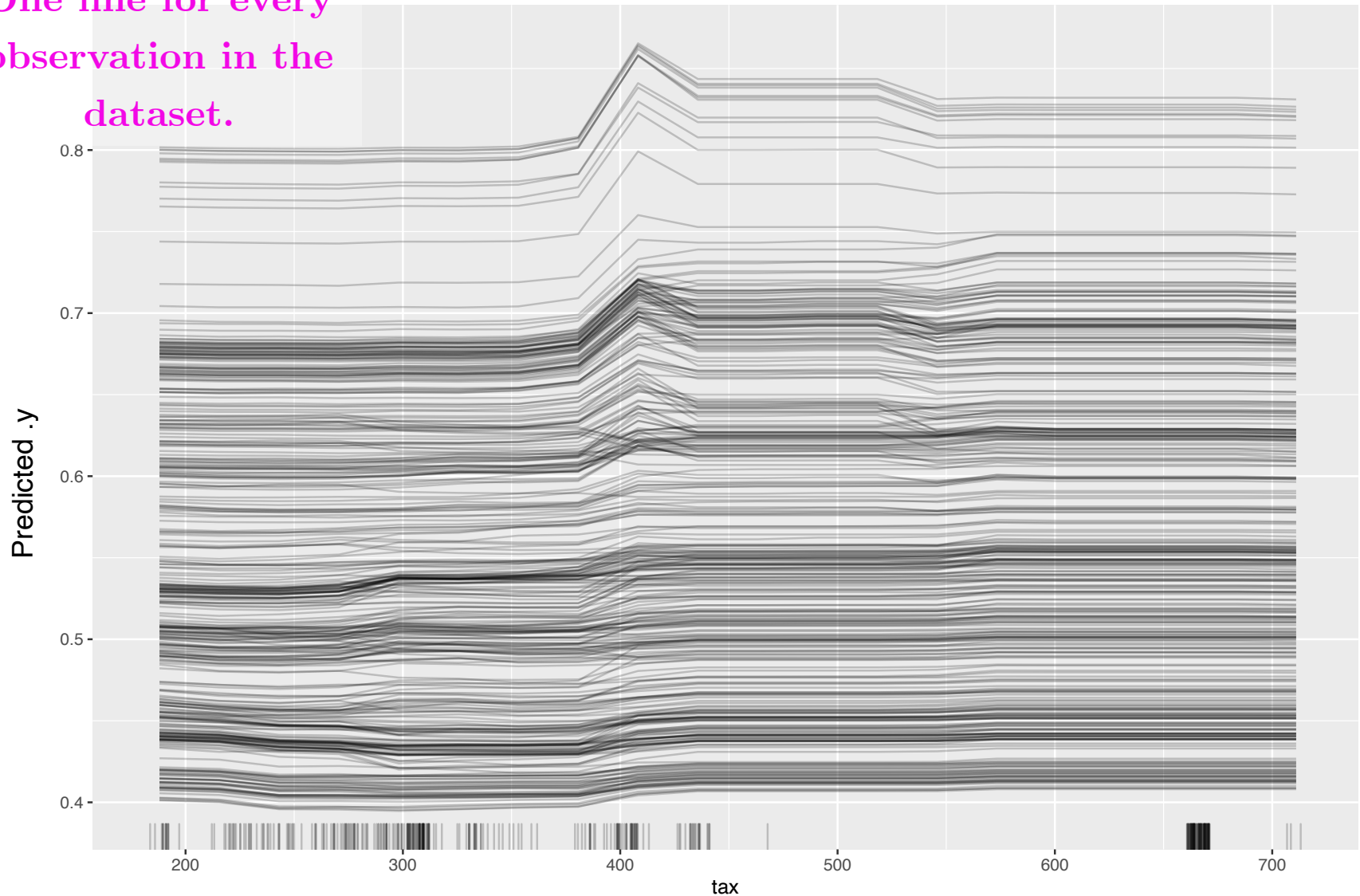
Use the model to generate predictions for this simulated data.
Typically repeat for each observation (or large sample)

Individual Conditional Expectation (ICE) Plot

```
#' Individual Conditional Expectation (ICE) Plots  
#'  
set.seed(13)  
pdps = FeatureEffects$new(forest_predictor, method='ice')  
pdps$plot() #All charts  
pdps$plot(c("tax")) #Subset of Charts
```

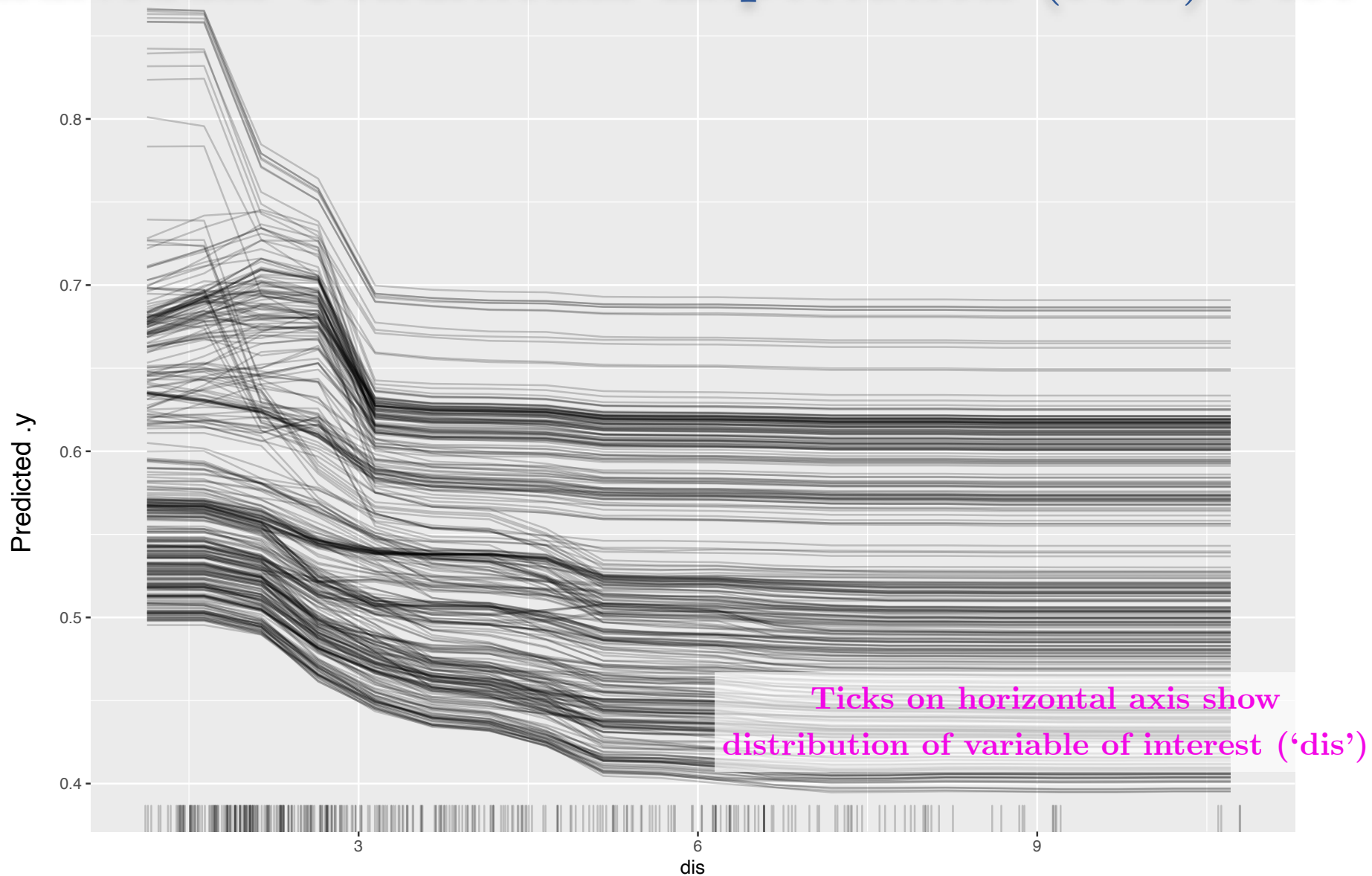
Individual Conditional Expectation (ICE) Plot

One line for every
observation in the
dataset.



An ICE plot displays the relationship between the prediction and a feature for *each* observation separately, resulting in one line per observation

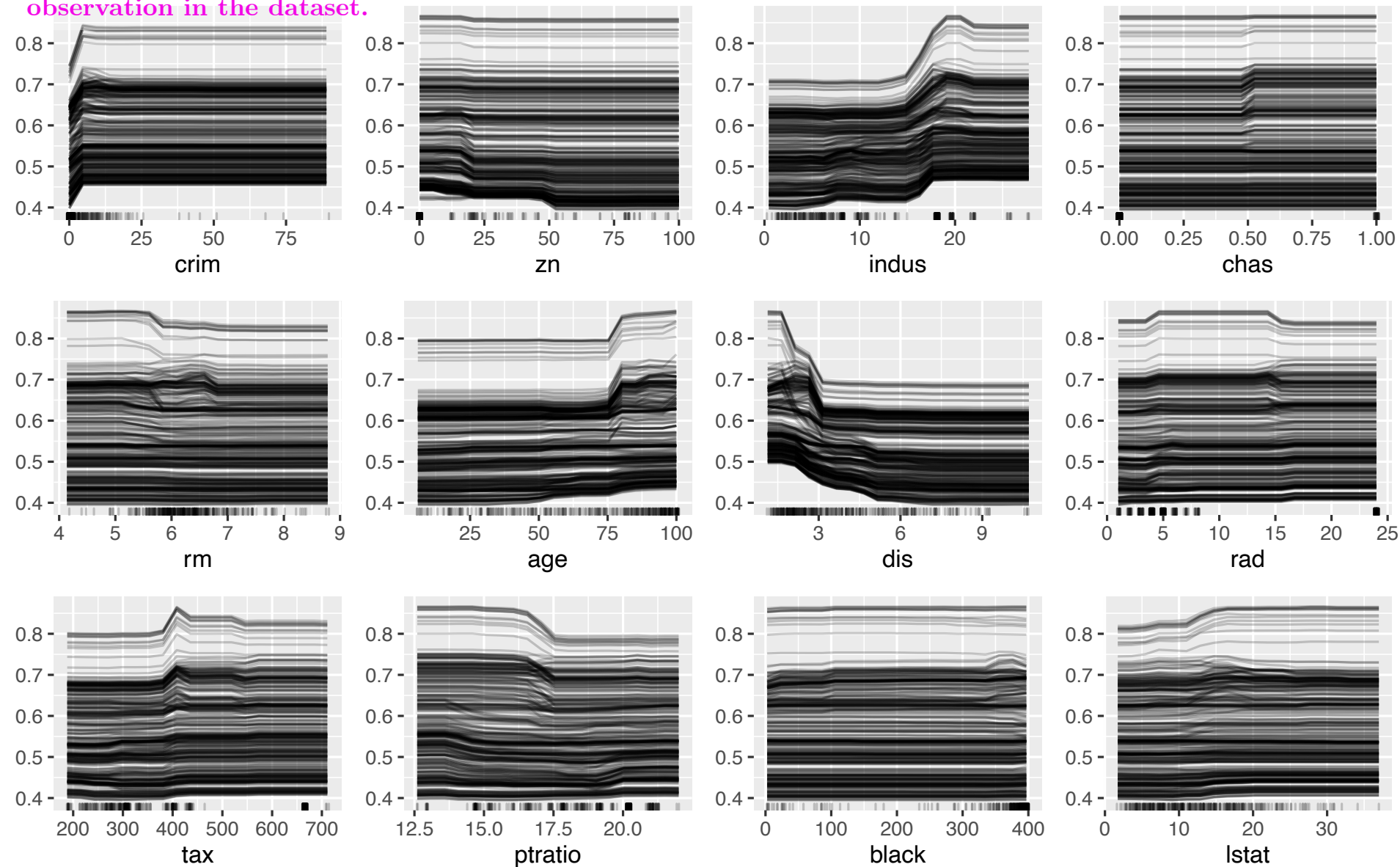
Individual Conditional Expectation (ICE) Plot



An ICE plot displays the relationship between the prediction and a feature for *each* observation separately, resulting in one line per observation

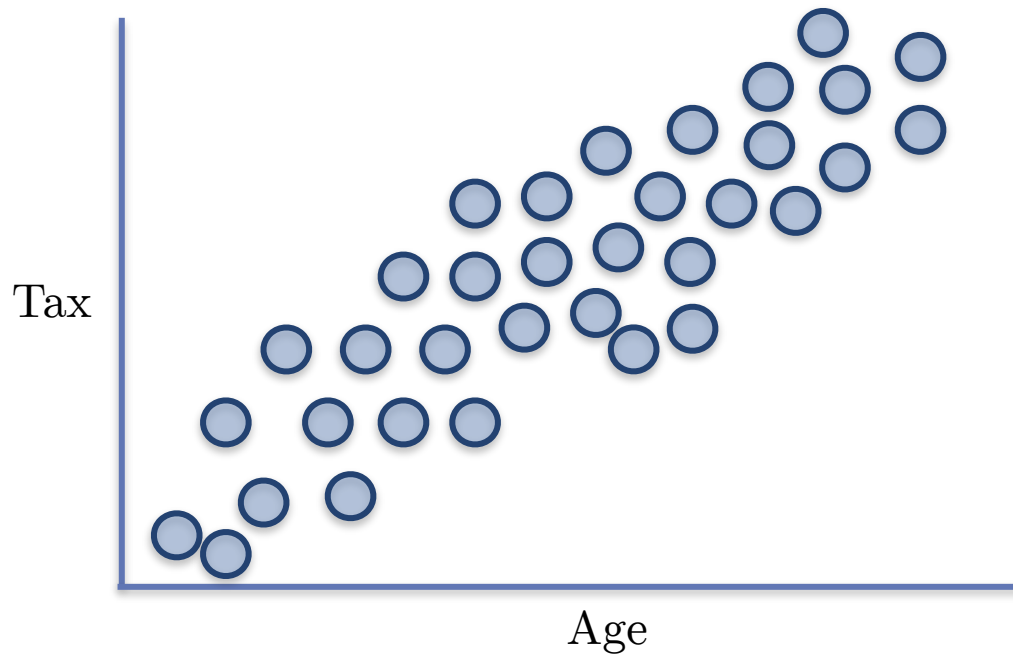
Individual Conditional Expectation (ICE) Plots

One line for every
observation in the dataset.



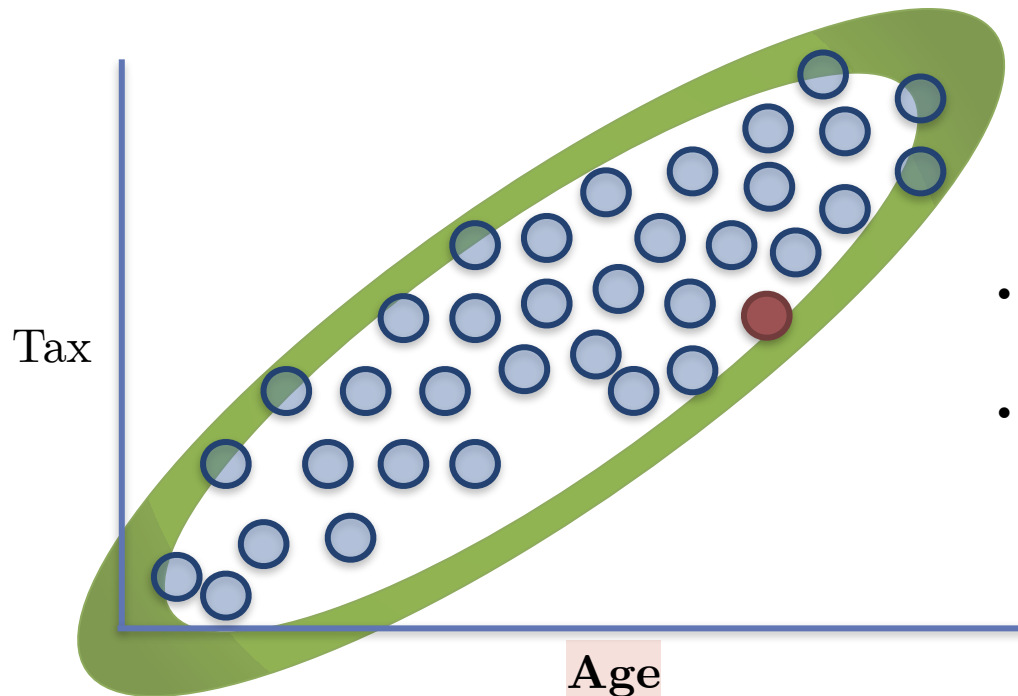
ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



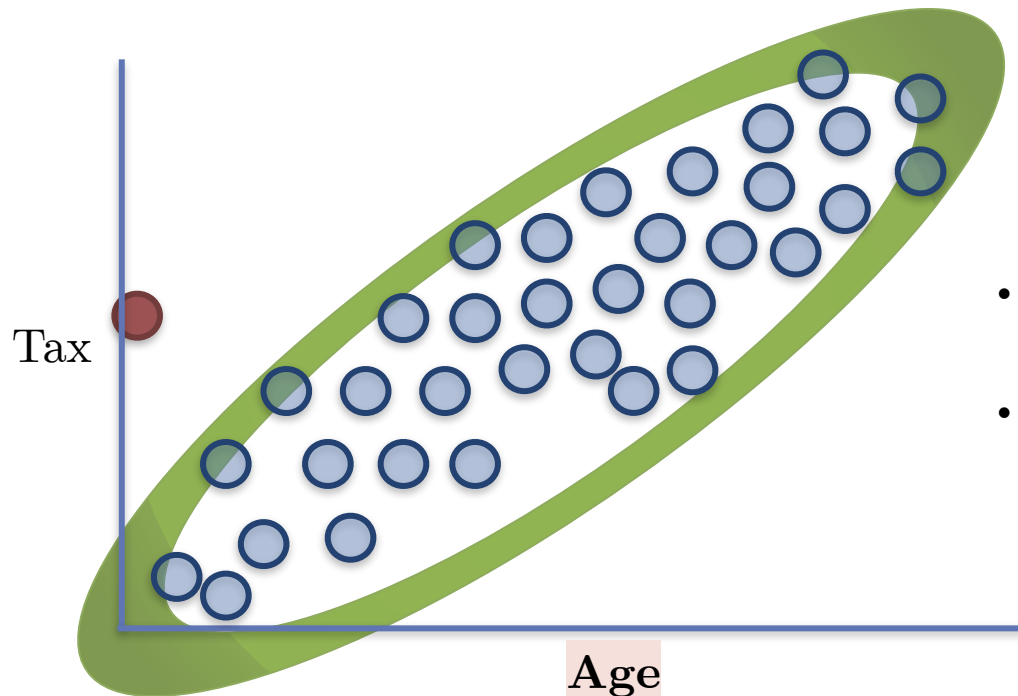
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



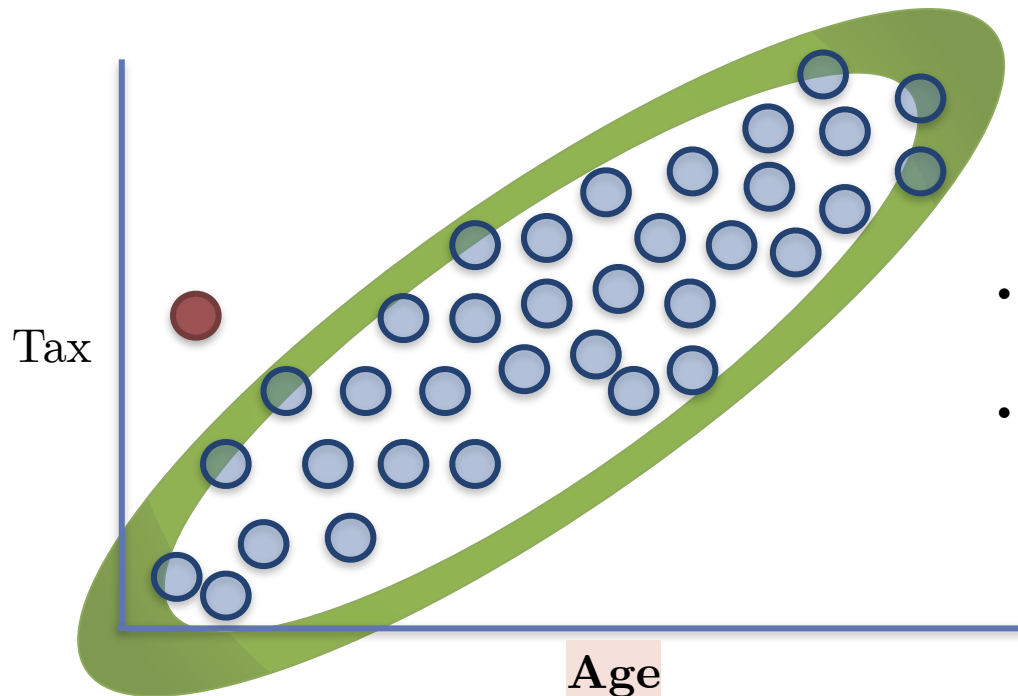
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



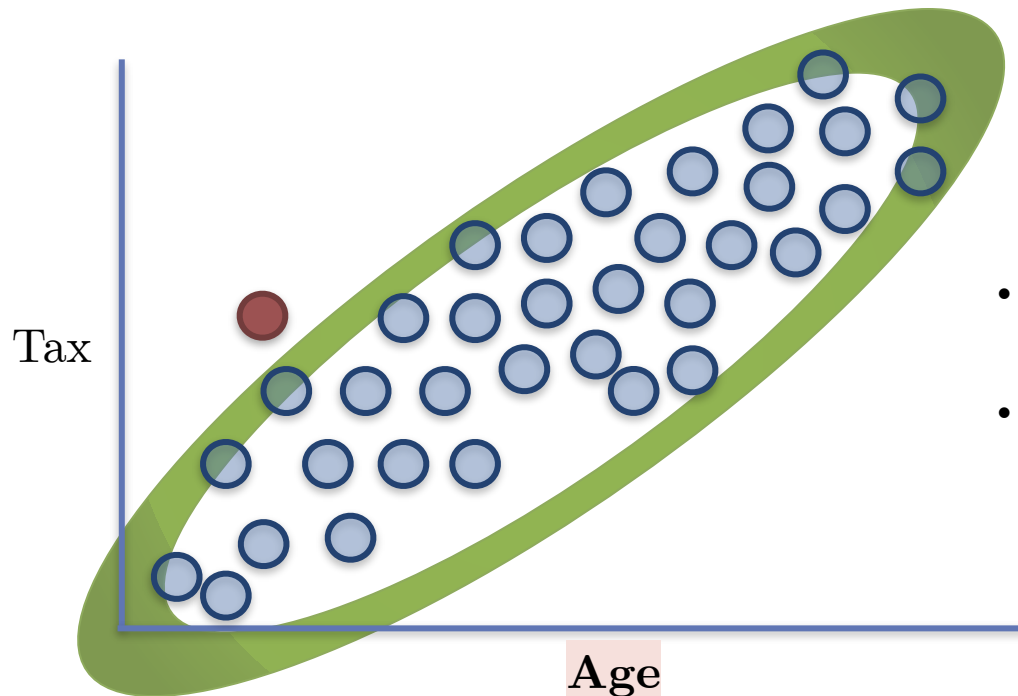
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



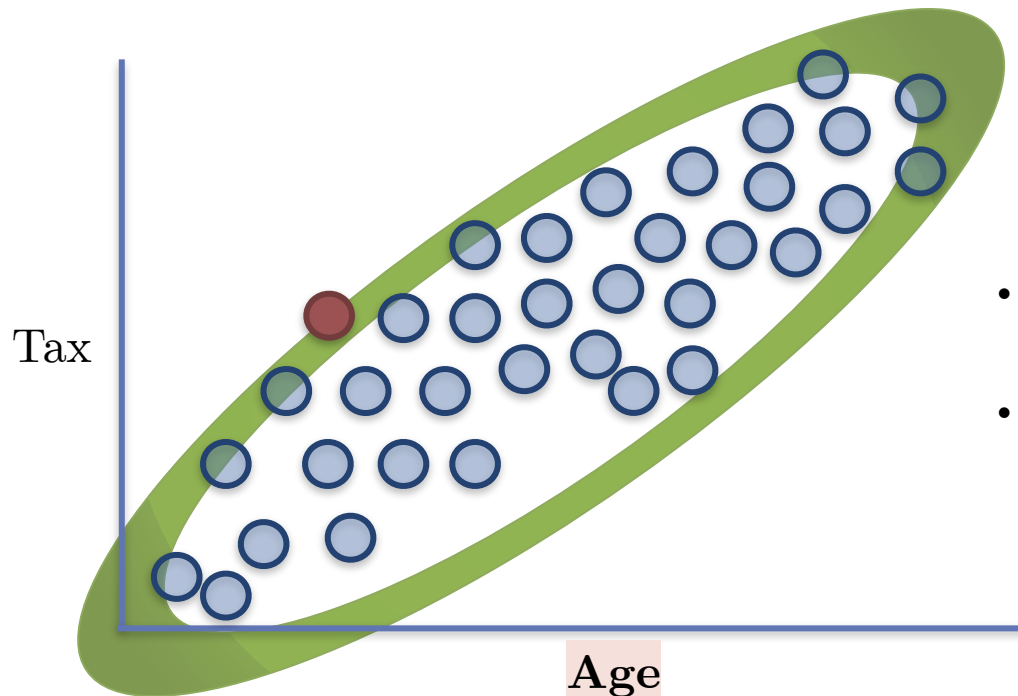
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



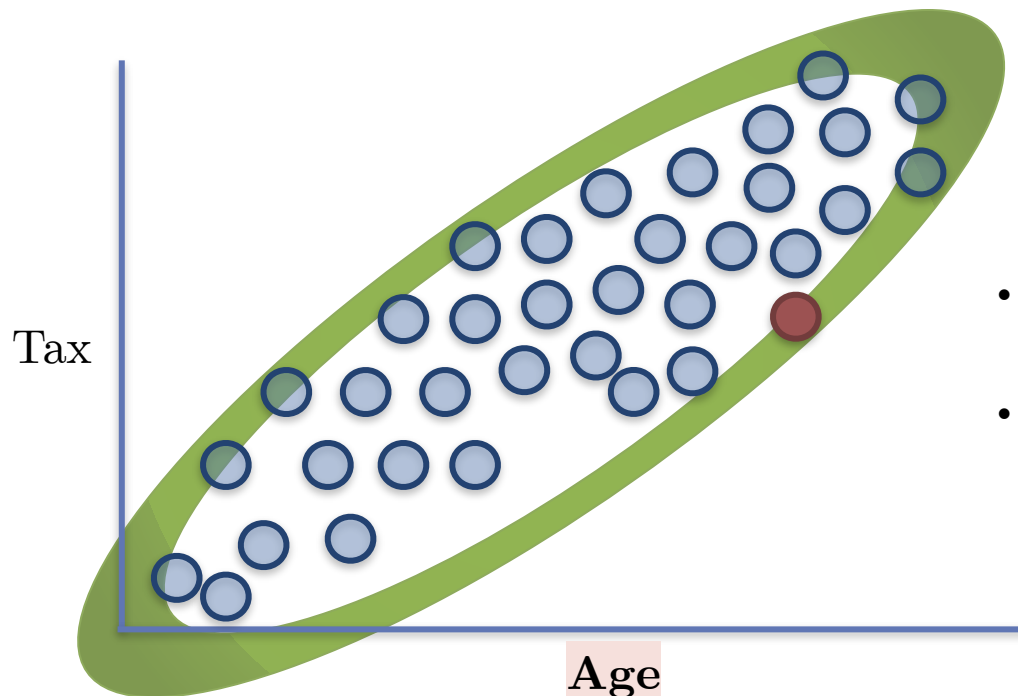
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



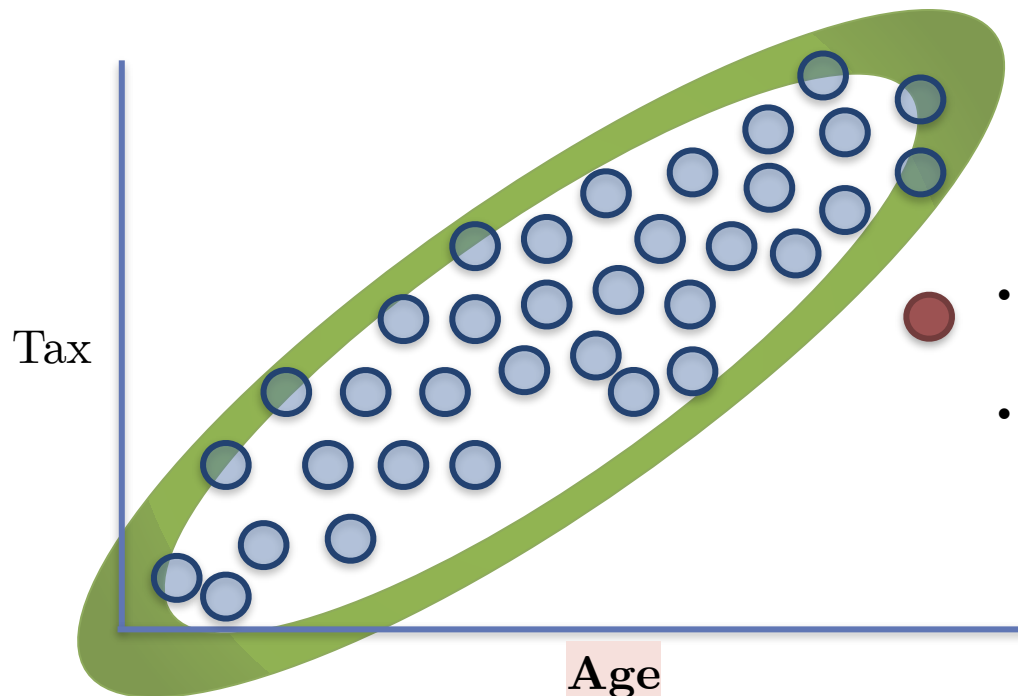
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

If variable of interest is correlated with other inputs, some of the simulated data may be invalid!



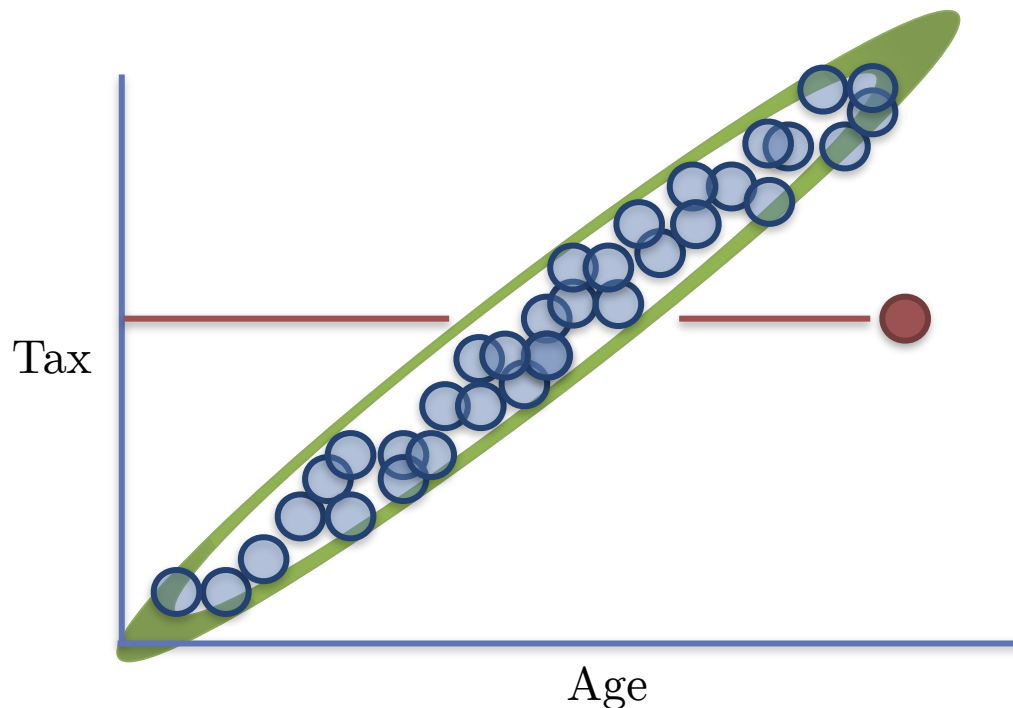
For this red point of interest, and for the variable of interest **Age**:

- We'll keep everything the same, and *vary the age across some replicates*.
- Computes $F(y|\mathbf{x})$ for each replicate \mathbf{x}

Green Ellipse shows what is *likely* to occur based on the *bivariate* distribution of Age and Tax.

ICE: Problems with Multicollinearity

More Severe Correlation \Rightarrow More Severe Problems.



Summary: Individual Conditional Expectation (ICE)

Advantages

- Intuitive to understand. One line represents to predictions for one observation if we vary the feature of interest.
- Capable of uncovering *heterogeneous* relationships (when the feature of interest has different impact for different observations)

Summary: Individual Conditional Expectation (ICE)

Disadvantages

- Difficult to process (computationally and visually) with too many observations, but sampling may lose interesting signal.
- If variable of interest is correlated with other inputs, some of the simulated data may be invalid!
- Can sometimes reveal interesting interactions, but requires a *lot* of exploration.
- Can only meaningfully display relationship of one variable

Types of Model Interpretability

	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE ✓ LIME Shapley Values	Permutation Importance ✓ Partial Dependence ALE

Partial Dependence

...

"Let me show you what the model predicts on average when each observation has the value v for that feature. We'll ignore whether the value v makes sense for all data instances."

Partial Dependence Plots

Attempts to show the marginal effect of inputs on the target.

Marginal: relating to a random variable that is obtained from a function of several random variables by averaging over all possible values of the other variables

In other words, what is the *expectation* (mean/average) of our predictive model $f(\mathbf{x})$ across values of a single variable x_i ?

In OTHER words, what if we *average* all those lines on the ICE plot?

Partial Dependence Plots

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	234	4.3	0.469	0.454
5	17.2	41	9.1	71	326	2.5	0.538	0.512
6	20.1	31	15.2	88	222	5.1	0.458	0.470
2	15	22	5.2	45	430	6.3	0.556	0.561

Choose a variable of interest.

Partial Dependence Plots

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
5	17.2	41	9.1	71		2.5	0.538	
6	20.1	31	15.2	88		5.1	0.458	
2	15	22	5.2	45		6.3	0.556	

Replicate your *dataset*, holding constant all values except the variable of interest.

Partial Dependence Plots

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	200	4.3	0.469	
5	17.2	41	9.1	71	200	2.5	0.538	
6	20.1	31	15.2	88	200	5.1	0.458	
2	15	22	5.2	45	200	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	204	4.3	0.469	
5	17.2	41	9.1	71	204	2.5	0.538	
6	20.1	31	15.2	88	204	5.1	0.458	
2	15	22	5.2	45	204	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	201	4.3	0.469	
5	17.2	41	9.1	71	201	2.5	0.538	
6	20.1	31	15.2	88	201	5.1	0.458	
2	15	22	5.2	45	201	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	205	4.3	0.469	
5	17.2	41	9.1	71	205	2.5	0.538	
6	20.1	31	15.2	88	205	5.1	0.458	
2	15	22	5.2	45	205	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	202	4.3	0.469	
5	17.2	41	9.1	71	202	2.5	0.538	
6	20.1	31	15.2	88	202	5.1	0.458	
2	15	22	5.2	45	202	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	206	4.3	0.469	
5	17.2	41	9.1	71	206	2.5	0.538	
6	20.1	31	15.2	88	206	5.1	0.458	
2	15	22	5.2	45	206	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	203	4.3	0.469	
5	17.2	41	9.1	71	203	2.5	0.538	
6	20.1	31	15.2	88	203	5.1	0.458	
2	15	22	5.2	45	203	6.3	0.556	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	207	4.3	0.469	
5	17.2	41	9.1	71	207	2.5	0.538	
6	20.1	31	15.2	88	207	5.1	0.458	
2	15	22	5.2	45	207	6.3	0.556	

Fill in values for variable of interest for each replicated dataset across the range of the variable.

Partial Dependence Plots

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	200	4.3	0.469	0.426
5	17.2	41	9.1	71	200	2.5	0.538	0.532
6	20.1	31	15.2	88	200	5.1	0.458	0.445
2	15	22	5.2	45	200	6.3	0.556	0.550

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	204	4.3	0.469	0.510
5	17.2	41	9.1	71	204	2.5	0.538	0.532
6	20.1	31	15.2	88	204	5.1	0.458	0.495
2	15	22	5.2	45	204	6.3	0.556	0.532

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	201	4.3	0.469	0.436
5	17.2	41	9.1	71	201	2.5	0.538	0.542
6	20.1	31	15.2	88	201	5.1	0.458	0.445
2	15	22	5.2	45	201	6.3	0.556	0.523

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	205	4.3	0.469	0.426
5	17.2	41	9.1	71	205	2.5	0.538	0.532
6	20.1	31	15.2	88	205	5.1	0.458	0.485
2	15	22	5.2	45	205	6.3	0.556	0.561

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	202	4.3	0.469	0.526
5	17.2	41	9.1	71	202	2.5	0.538	0.532
6	20.1	31	15.2	88	202	5.1	0.458	0.475
2	15	22	5.2	45	202	6.3	0.556	0.561

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	206	4.3	0.469	0.469
5	17.2	41	9.1	71	206	2.5	0.538	0.532
6	20.1	31	15.2	88	206	5.1	0.458	0.495
2	15	22	5.2	45	206	6.3	0.556	0.532

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	203	4.3	0.469	0.426
5	17.2	41	9.1	71	203	2.5	0.538	0.532
6	20.1	31	15.2	88	203	5.1	0.458	0.445
2	15	22	5.2	45	203	6.3	0.556	0.550

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	207	4.3	0.469	0.426
5	17.2	41	9.1	71	207	2.5	0.538	0.532
6	20.1	31	15.2	88	207	5.1	0.458	0.445
2	15	22	5.2	45	207	6.3	0.556	0.550

Use the model to generate predictions for this simulated data.

Partial Dependence Plots

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	dis mean	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	mean
3	12.1	13	22.6	90	200	4.3	0.469	0.426	3	12.1	13	22.6	90	204	4.3	0.469	0.510	0.516
5	17.2	41	9.1	71	200	2.5	0.538	0.532	3	17.2	41	9.1	71	204	2.5	0.538	0.532	0.516
6	20.1	31	15.2	88	200	5.1	0.458	0.445	6	20.1	31	15.2	88	204	5.1	0.458	0.495	0.516
2	15	22	5.2	45	200	6.3	0.556	0.550	2	15	22	5.2	45	204	6.3	0.556	0.532	0.516

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	dis mean	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	mean
3	12.1	13	22.6	90	201	4.3	0.469	0.436	3	12.1	13	22.6	90	205	4.3	0.469	0.476	0.521
5	17.2	41	9.1	71	201	2.5	0.538	0.542	3	17.2	41	9.1	71	205	2.5	0.538	0.532	0.521
6	20.1	31	15.2	88	201	5.1	0.458	0.445	6	20.1	31	15.2	88	205	5.1	0.458	0.485	0.521
2	15	22	5.2	45	201	6.3	0.556	0.523	2	15	22	5.2	45	205	6.3	0.556	0.561	0.521

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	dis mean	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	mean
3	12.1	13	22.6	90	202	4.3	0.469	0.526	3	12.1	13	22.6	90	206	4.3	0.469	0.499	0.541
5	17.2	41	9.1	71	202	2.5	0.538	0.532	3	17.2	41	9.1	71	206	2.5	0.538	0.532	0.541
6	20.1	31	15.2	88	202	5.1	0.458	0.475	6	20.1	31	15.2	88	206	5.1	0.458	0.495	0.541
2	15	22	5.2	45	202	6.3	0.556	0.561	2	15	22	5.2	45	206	6.3	0.556	0.532	0.541

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	dis mean	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y	mean
3	12.1	13	22.6	90	203	4.3	0.469	0.426	3	12.1	13	22.6	90	207	4.3	0.469	0.426	0.551
5	17.2	41	9.1	71	203	2.5	0.538	0.532	3	17.2	41	9.1	71	207	2.5	0.538	0.532	0.551
6	20.1	31	15.2	88	203	5.1	0.458	0.512	6	20.1	31	15.2	88	207	5.1	0.458	0.575	0.551
2	15	22	5.2	45	203	6.3	0.556	0.550	2	15	22	5.2	45	207	6.3	0.556	0.550	0.551

Average the predictions for each replicated dataset
(i.e. for each possible value of the variable of interest.)

Partial Dependence Plots

dis	ptratio	medv	indus	age	tax tax 200 200	rad	Actual Y	Pred Y	dis mean 3 0.488 6 2	ptratio	medv	indus	age	tax tax 204 204	rad	Actual Y	Pred Y	mean 0.516
3	12.1	13	22.6	90	200	4.3	0.469	0.426	3	12.1	13	22.6	90	204	4.3	0.469	0.510	0.516
5	17.2	41	9.1	71	200	2.5	0.538	0.532	3	17.2	41	9.1	71	204	2.5	0.538	0.532	0.516
6	20.1	31	15.2	88	200	5.1	0.458	0.445	6	20.1	31	15.2	88	204	5.1	0.458	0.495	0.516
2	15	22	5.2	45	200	6.3	0.556	0.550	2	15	22	5.2	45	204	6.3	0.556	0.532	0.516

dis	ptratio	medv	indus	age	tax tax 201 201	rad	Actual Y	Pred Y	dis mean 3 0.489 6 2	ptratio	medv	indus	age	tax tax 205 205	rad	Actual Y	Pred Y	mean 0.521
3	12.1	13	22.6	90	201	4.3	0.469	0.436	3	12.1	13	22.6	90	205	4.3	0.469	0.476	0.521
5	17.2	41	9.1	71	201	2.5	0.538	0.542	3	17.2	41	9.1	71	205	2.5	0.538	0.532	0.521
6	20.1	31	15.2	88	201	5.1	0.458	0.445	6	20.1	31	15.2	88	205	5.1	0.458	0.485	0.521
2	15	22	5.2	45	201	6.3	0.556	0.523	2	15	22	5.2	45	205	6.3	0.556	0.561	0.521

dis	ptratio	medv	indus	age	tax tax 202 202	rad	Actual Y	Pred Y	dis mean 3 0.512 6 2	ptratio	medv	indus	age	tax tax 206 206	rad	Actual Y	Pred Y	mean 0.541
3	12.1	13	22.6	90	202	4.3	0.469	0.526	3	12.1	13	22.6	90	206	4.3	0.469	0.499	0.541
5	17.2	41	9.1	71	202	2.5	0.538	0.532	3	17.2	41	9.1	71	206	2.5	0.538	0.532	0.541
6	20.1	31	15.2	88	202	5.1	0.458	0.475	6	20.1	31	15.2	88	206	5.1	0.458	0.495	0.541
2	15	22	5.2	45	202	6.3	0.556	0.561	2	15	22	5.2	45	206	6.3	0.556	0.532	0.541

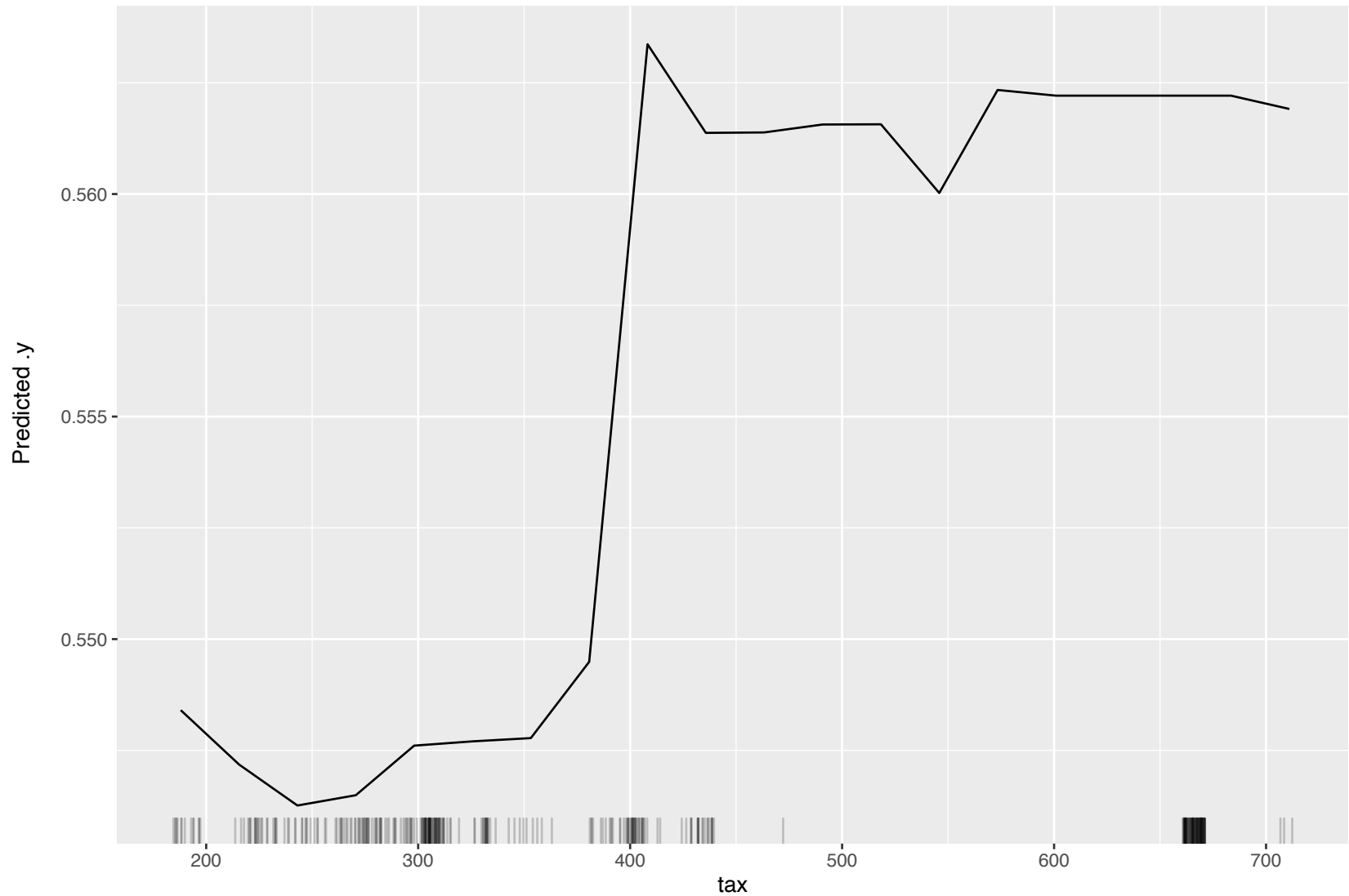
dis	ptratio	medv	indus	age	tax tax 203 203	rad	Actual Y	Pred Y	dis mean 3 0.513 6 2	ptratio	medv	indus	age	tax tax 207 207	rad	Actual Y	Pred Y	mean 0.551
3	12.1	13	22.6	90	203	4.3	0.469	0.426	3	12.1	13	22.6	90	207	4.3	0.469	0.426	0.551
5	17.2	41	9.1	71	203	2.5	0.538	0.532	3	17.2	41	9.1	71	207	2.5	0.538	0.532	0.551
6	20.1	31	15.2	88	203	5.1	0.458	0.512	6	20.1	31	15.2	88	207	5.1	0.458	0.575	0.551
2	15	22	5.2	45	203	6.3	0.556	0.550	2	15	22	5.2	45	207	6.3	0.556	0.550	0.551

The points on the partial dependence plots are shown in orange

Partial Dependence Plot

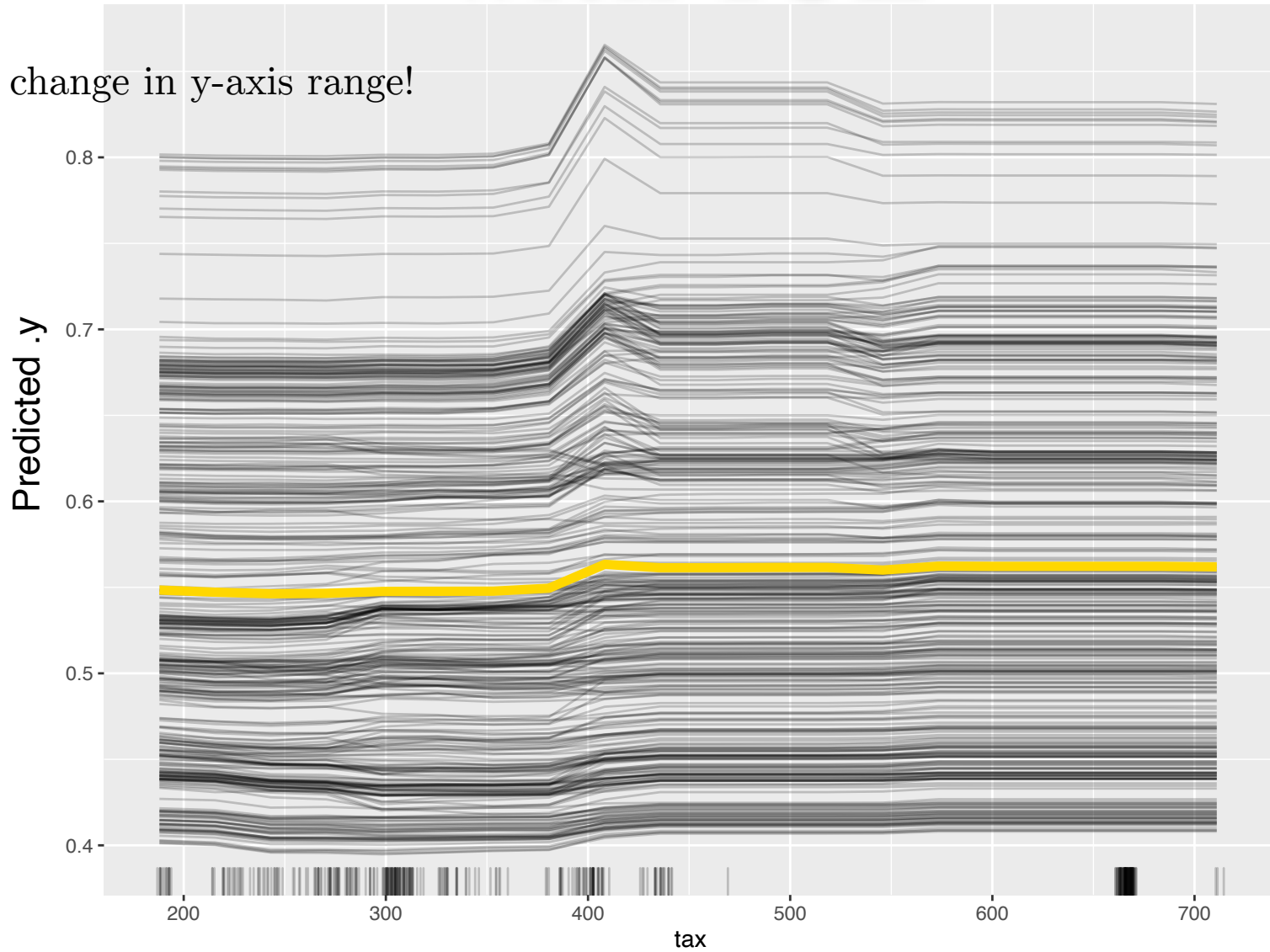
```
#####  
# ' Partial Dependence Plots - Forest Model  
# ' Calculate for all effects in model with FeatureEffects()  
# ' Calculate for a single effects in model with FeatureEffect()  
#####  
set.seed(11)  
pdps = FeatureEffects$new(forest_predictor, method='pdp')  
pdps$plot() #All charts  
pdps$plot(c("tax")) #Subset of Charts  
#####
```

Partial Dependence Plot

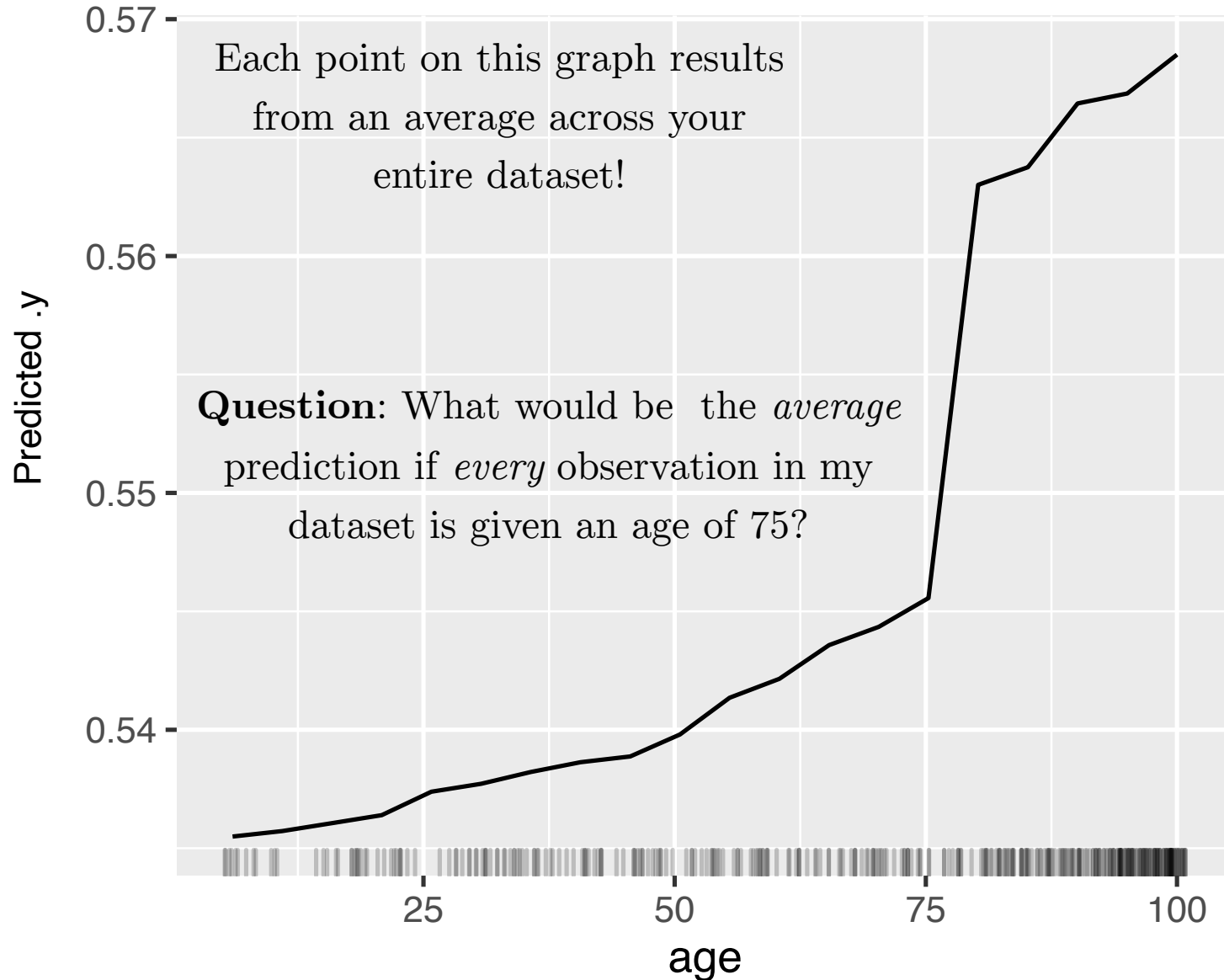


Partial Dependence Plot with ICE

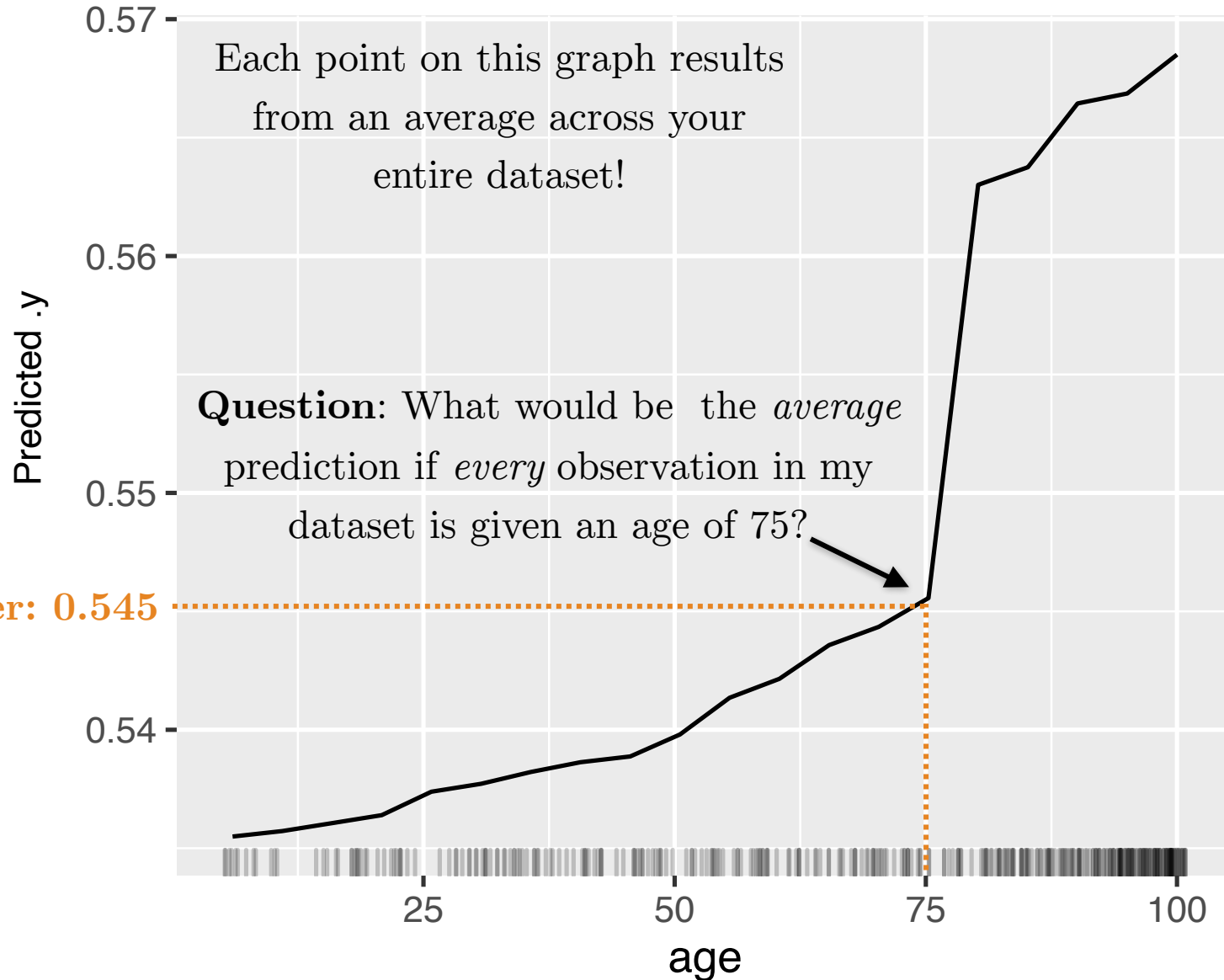
Note: change in y-axis range!



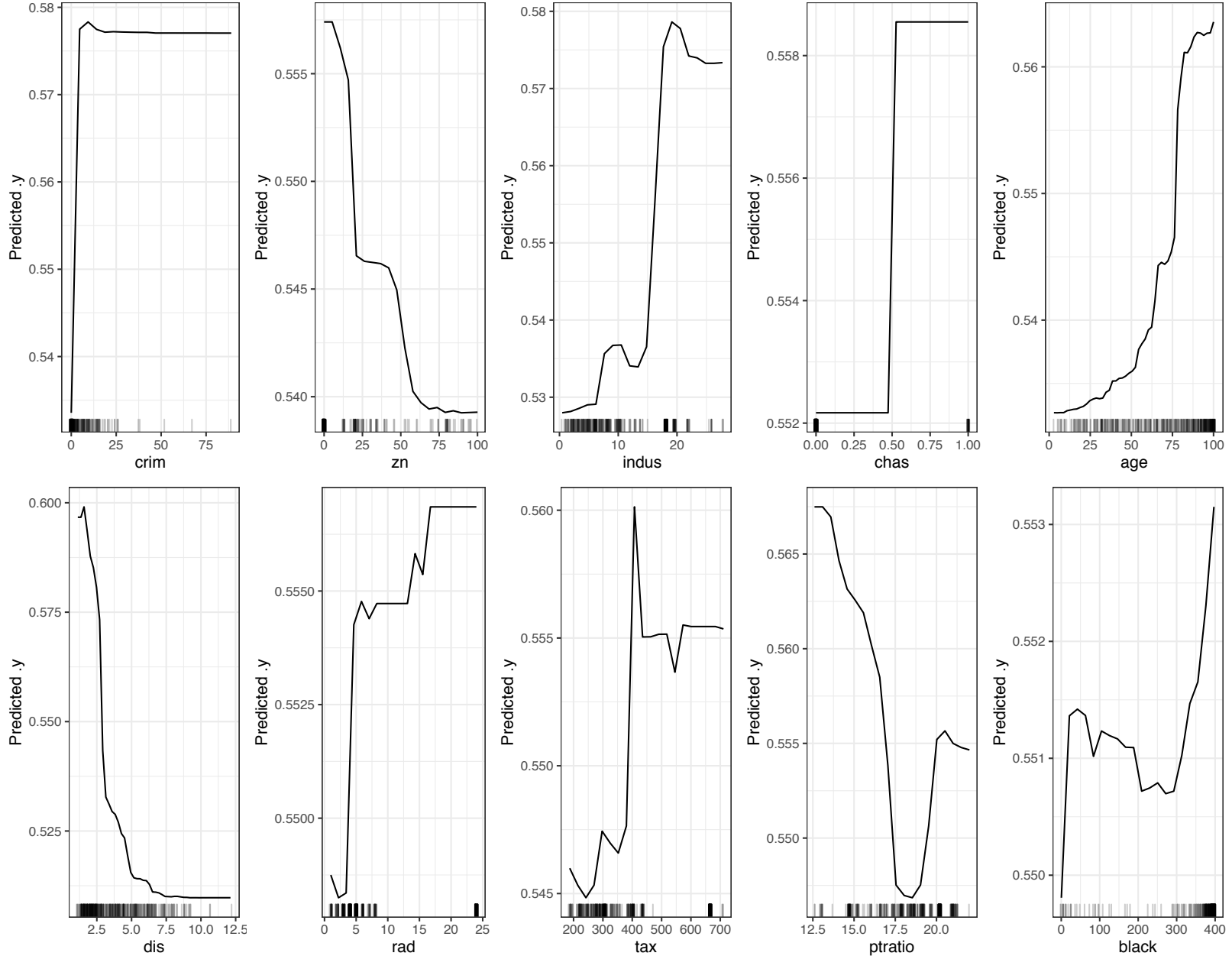
Partial Dependence Plots



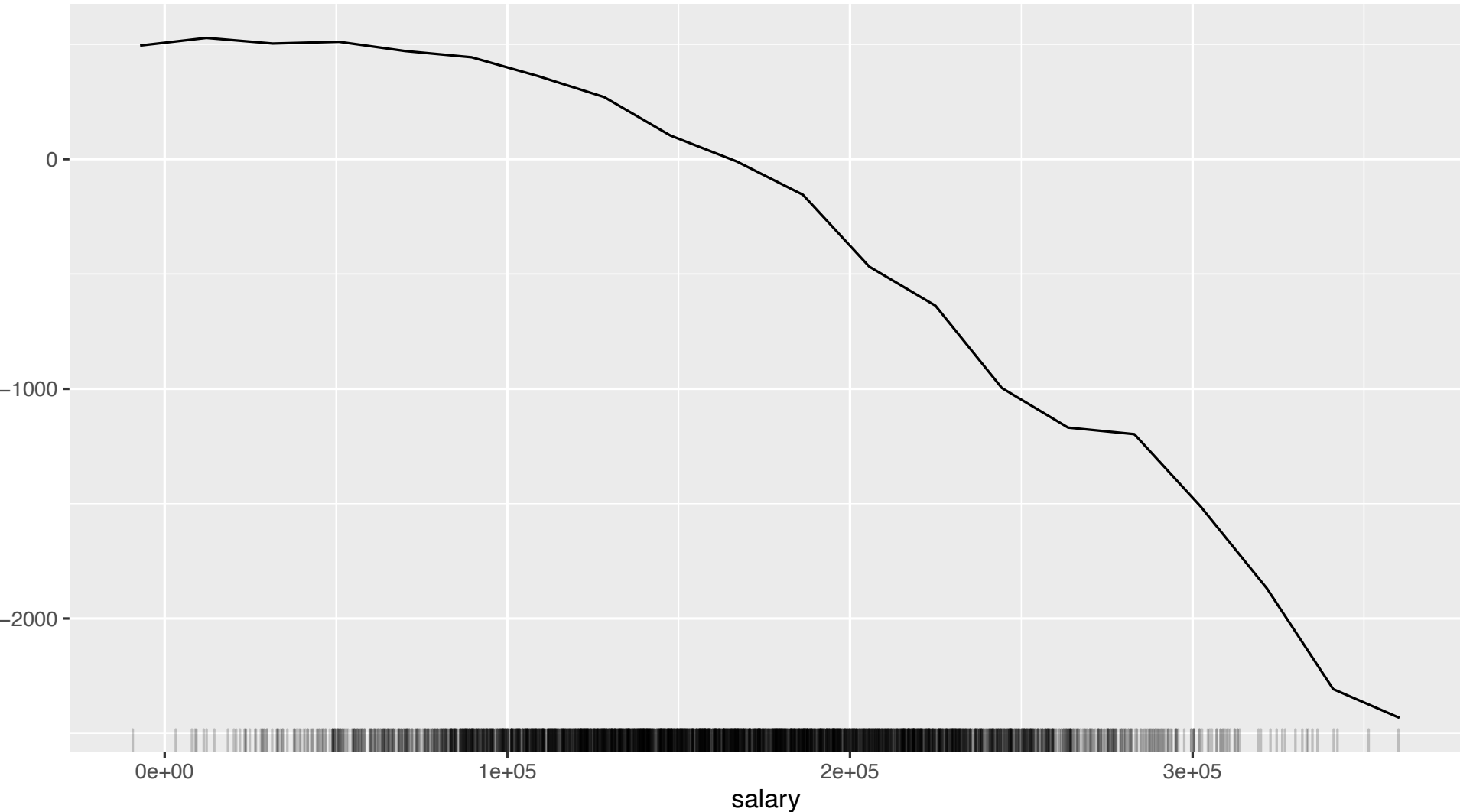
Partial Dependence Plots



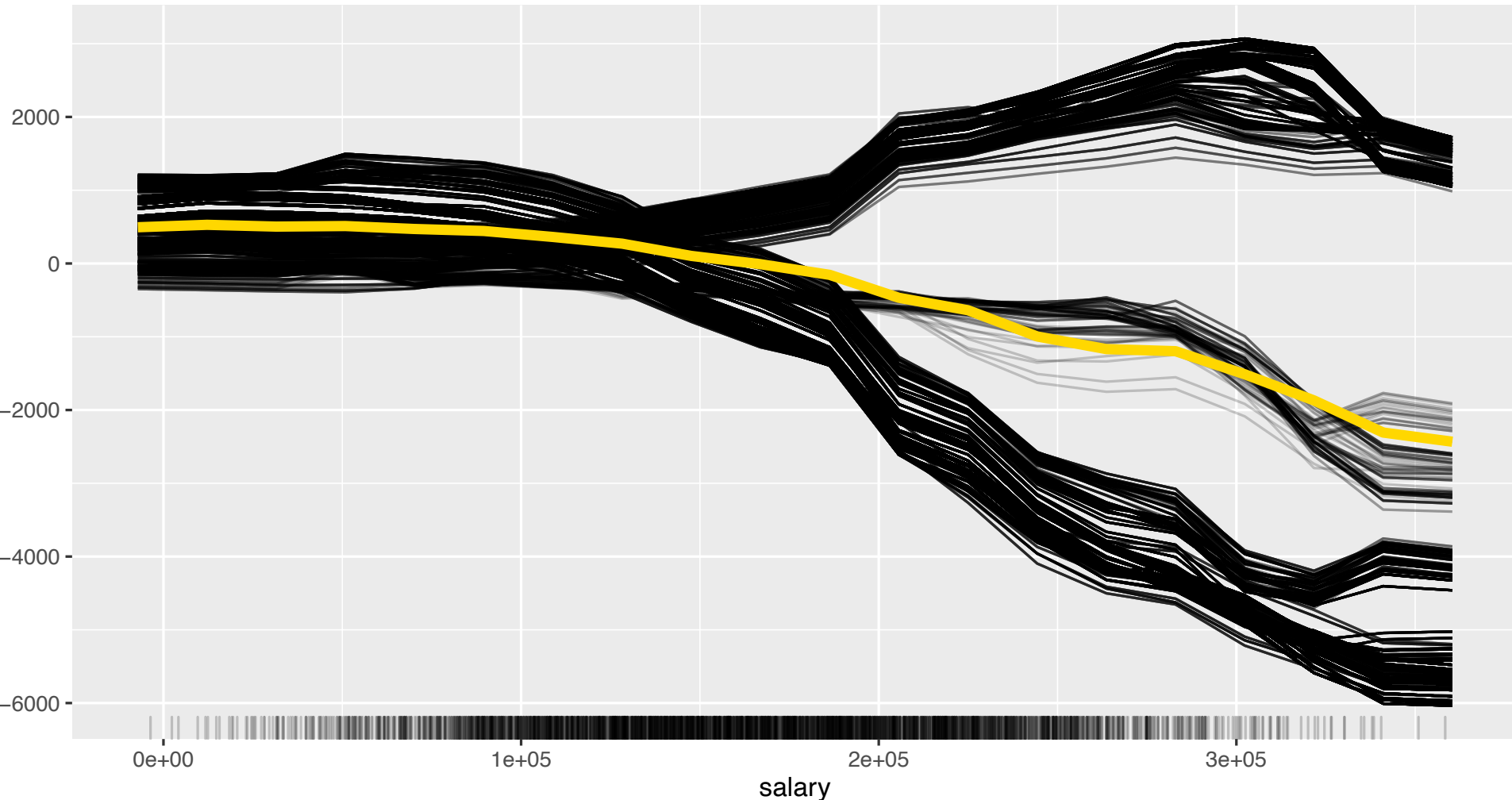
Partial Dependence Plots



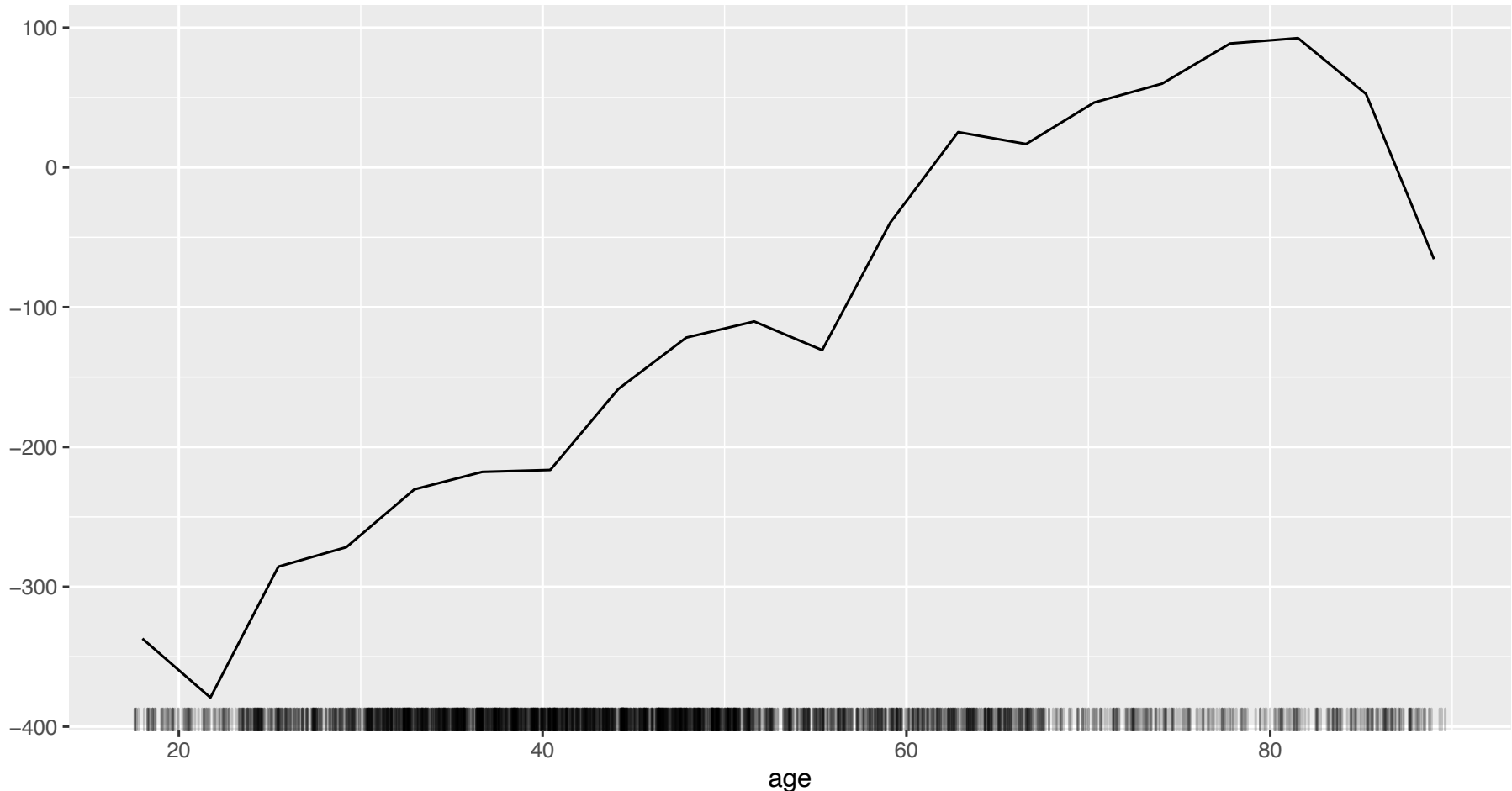
Beware of Relying on ONE Technique for this analysis!



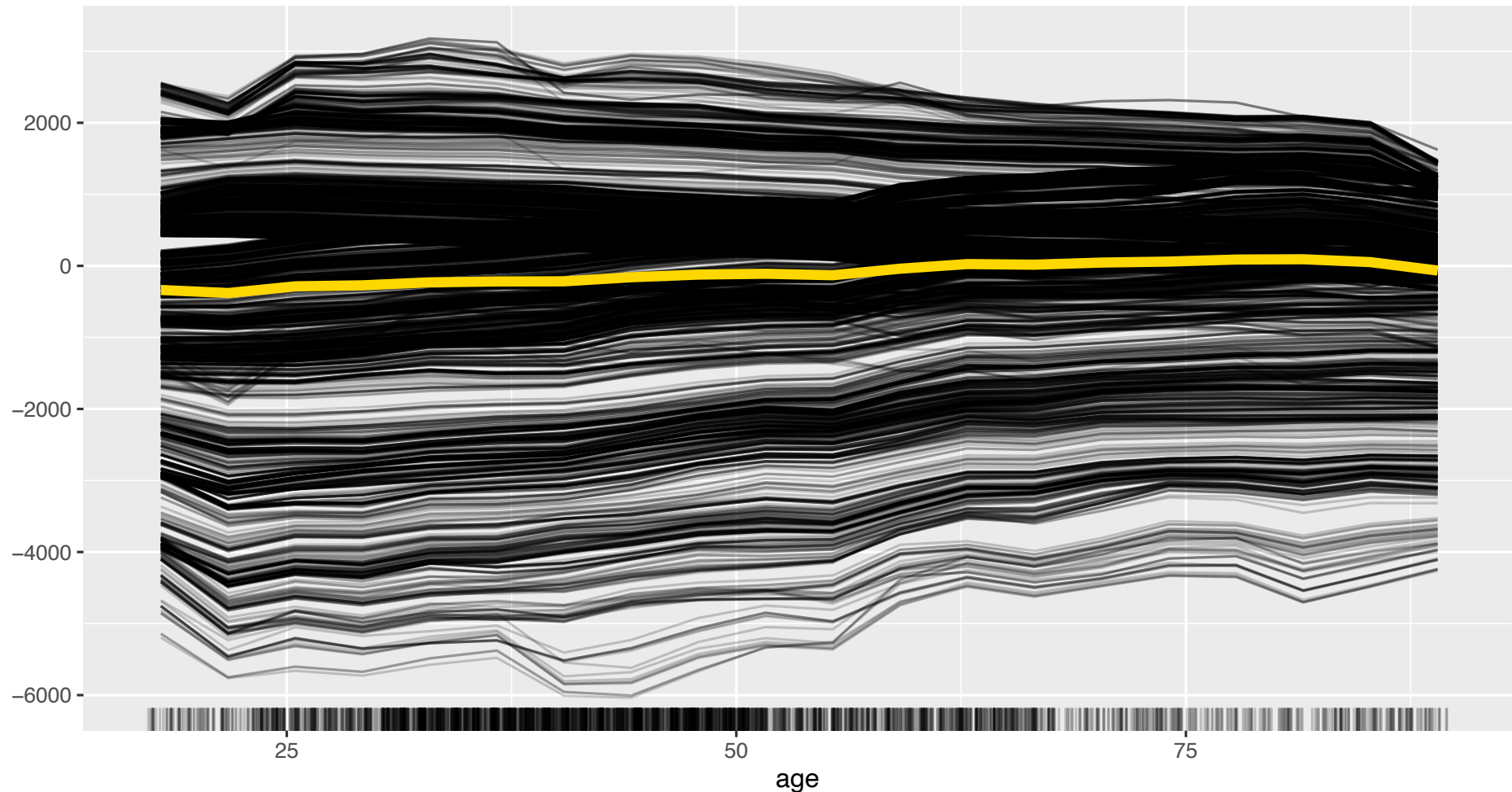
Beware of Relying on ONE Technique for this analysis!



Beware of Relying on ONE Technique for this analysis!



Beware of Relying on ONE Technique for this analysis!



Types of Model Interpretability

	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE ✓ LIME Shapley Values	Permutation Importance ✓ Partial Dependence ✓ ALE

Accumulated Local Effects (ALE)

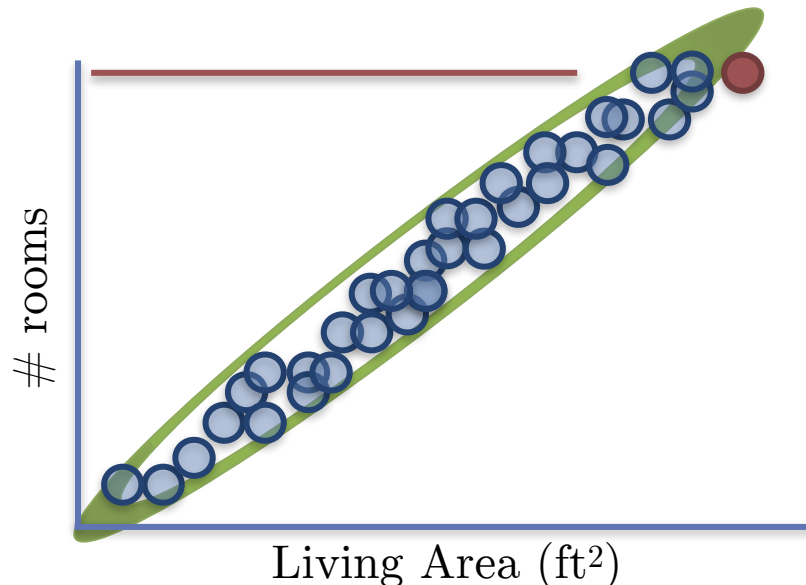
...

“Let me show you how the model predictions change when I change the variable of interest to values within a small interval around their current values.”

The Primary Problem of Partial Dependence:

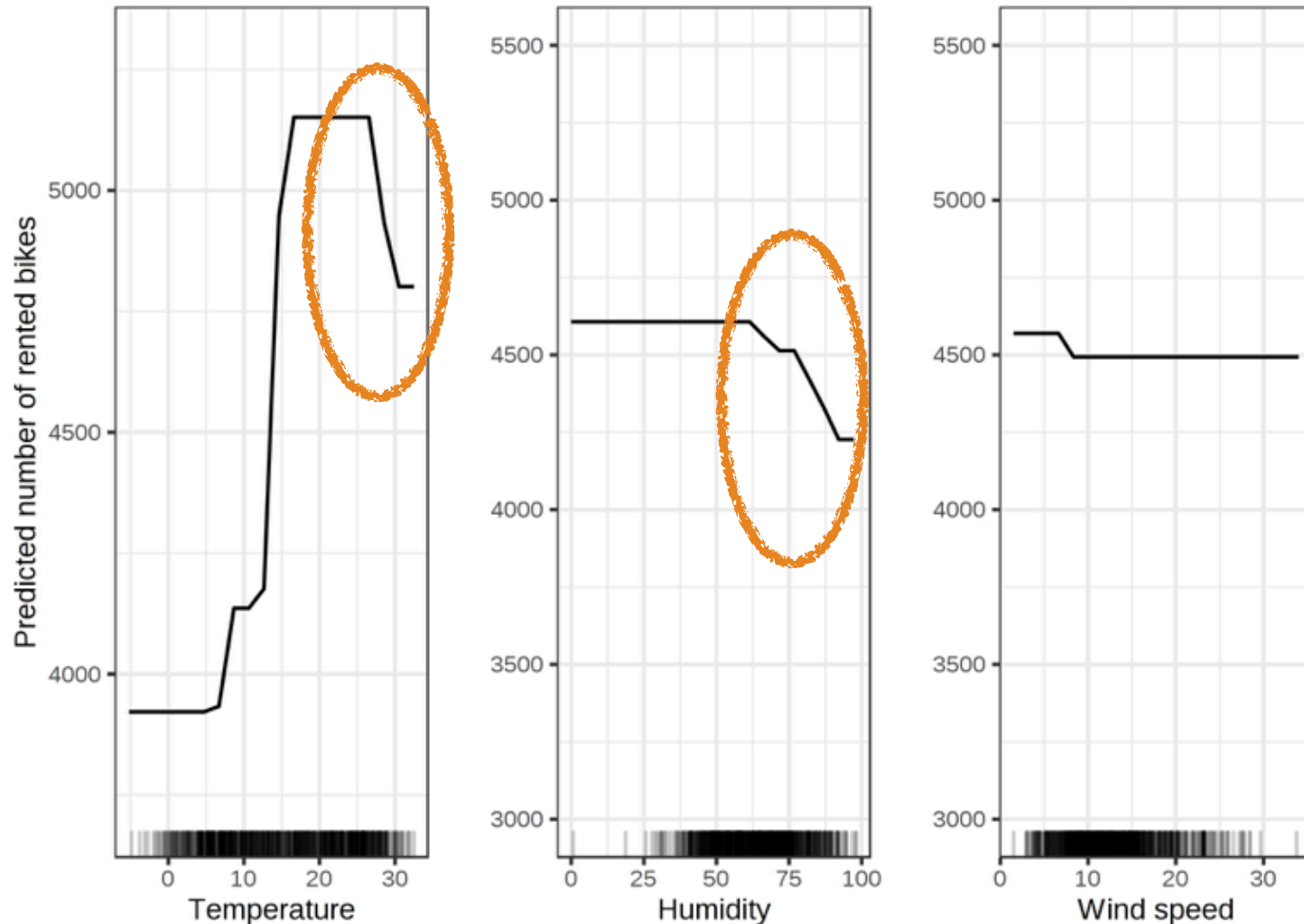
Our visualization (and therefore our conclusions) involve many simulated data points with impossible data values.

Example: *A house with 35 rooms will be given all possible values of living area - So we will have <nonsensical> houses with 35 rooms but 500 square feet of area participating in the average value we view on PDP!!*



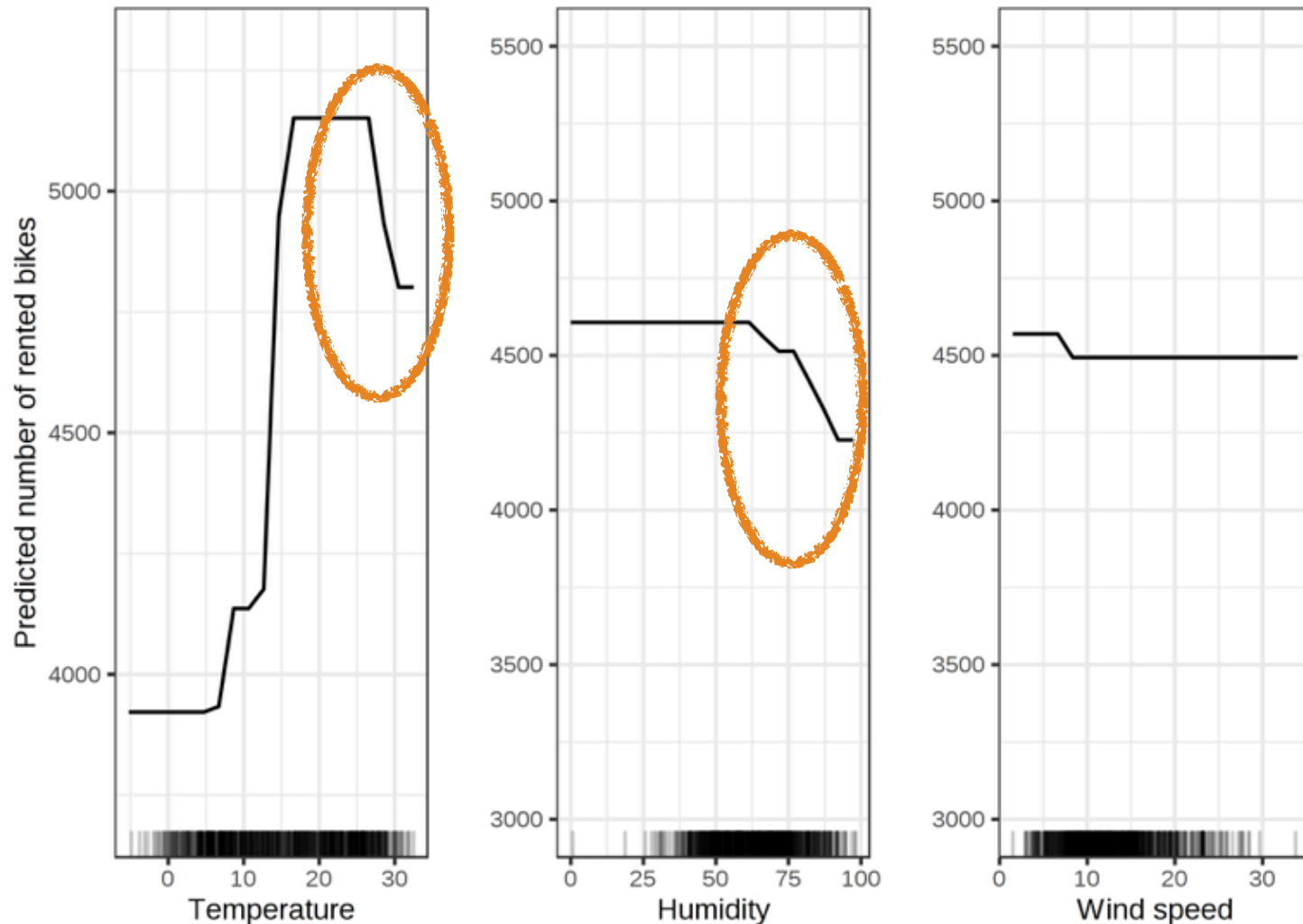
Problem with PDP

As temperature, humidity get too high,
bike rentals go down. Makes sense.



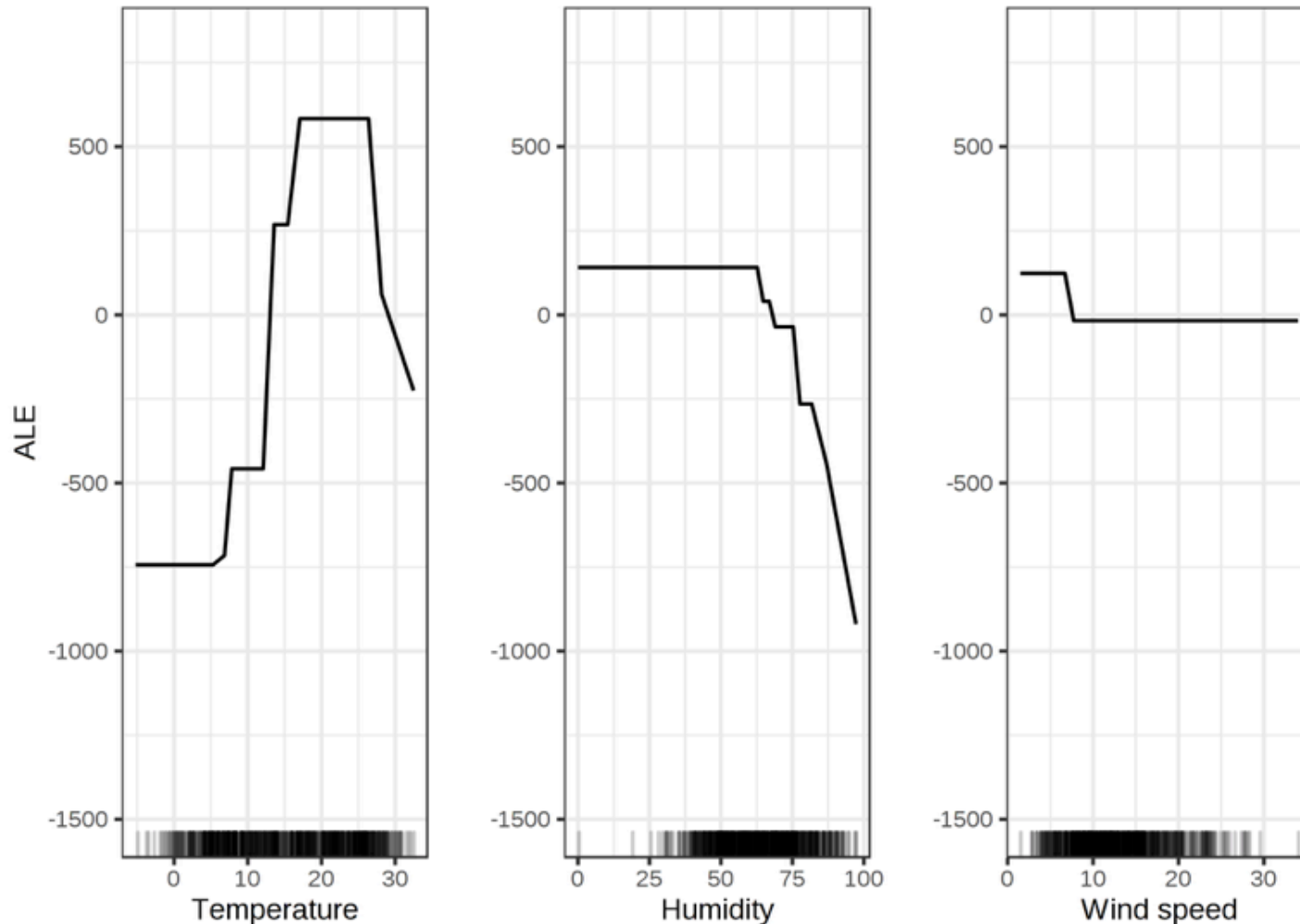
Problem with PDP

BUT, these average predictions involve EVERY observation altered to have high temperature/high humidity - even our observations from November-March! Not trustworthy!



Problem with PDP, Meet Solution with ALE

Using only reasonably contrived data, we get a more clear picture of how temperature and humidity effect bike rentals



Define a Grid

- We left a detail out from our last discussion.
- Remember when we talked about “filling in values across the range of the variable of interest” (ICE slide 25, PDP slide 46)?

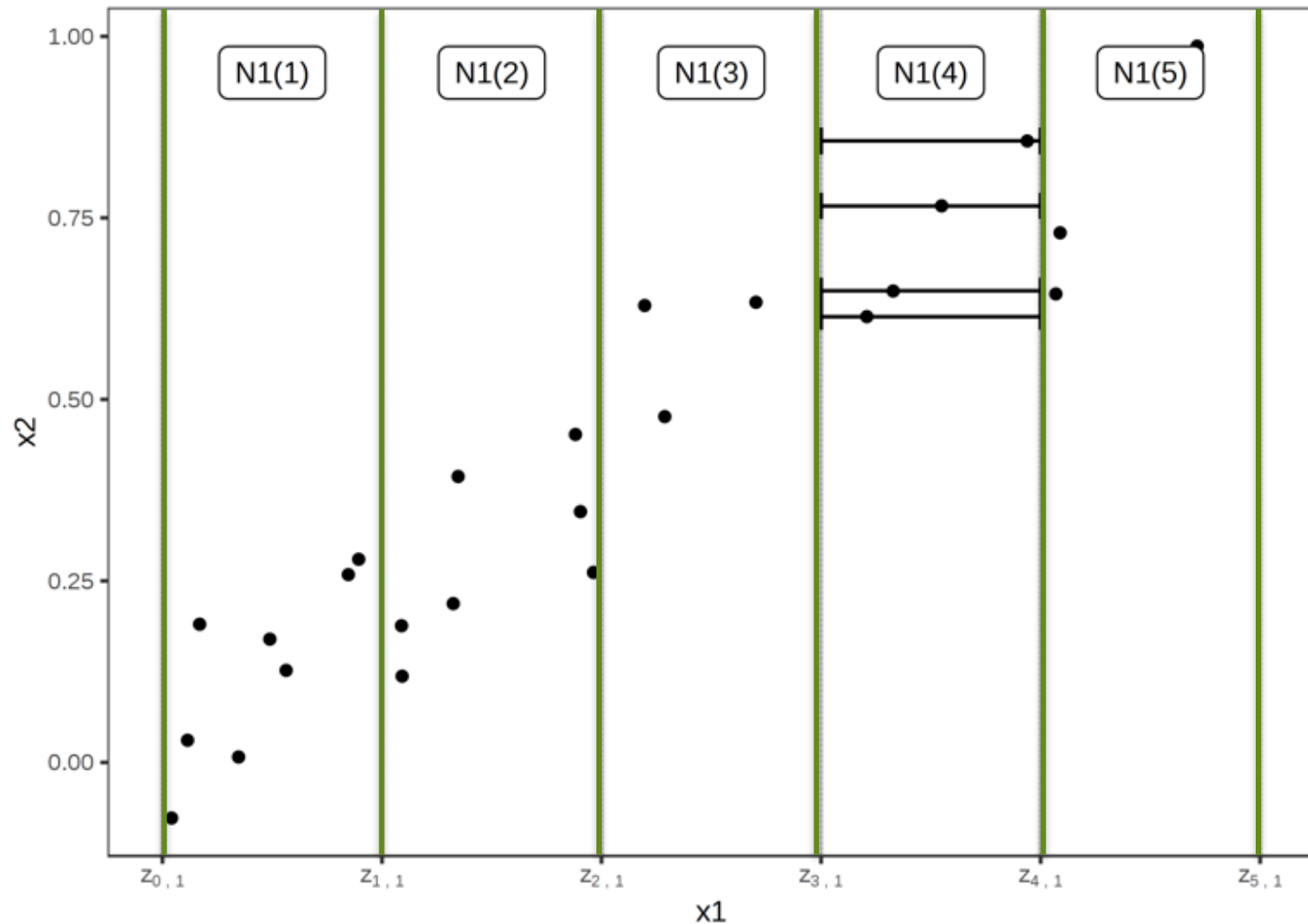
dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	
3	12.1	13	22.6	90		4.3	0.469	

dis	ptratio	medv	indus	age	tax	rad	Actual Y	Pred Y
3	12.1	13	22.6	90	200	4.3	0.469	
3	12.1	13	22.6	90	201	4.3	0.469	
3	12.1	13	22.6	90	202	4.3	0.469	
3	12.1	13	22.6	90	203	4.3	0.469	
3	12.1	13	22.6	90	204	4.3	0.469	
3	12.1	13	22.6	90	205	4.3	0.469	
3	12.1	13	22.6	90	206	4.3	0.469	

- How did I know exactly what values to fill in for tax?
- Well, we have to define this grid before hand. This is generally not a major decision. Equally spaced intervals is common.

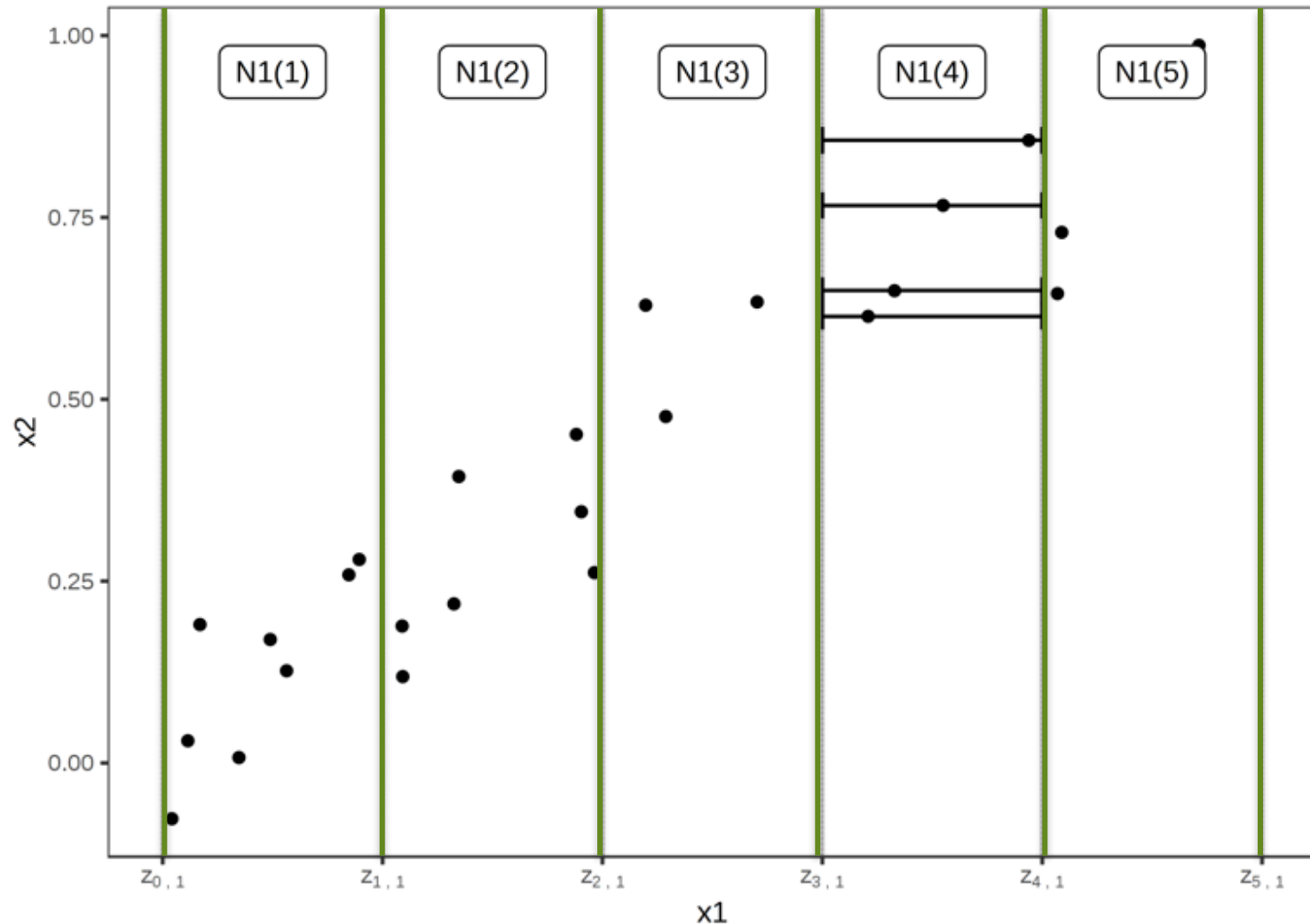
Accumulated Local Effects

Define a Grid: Most commonly uses quantiles of your data so that the same number of observations fall in each interval. (That gets weird for skewed data because intervals can have very different lengths - try it!)



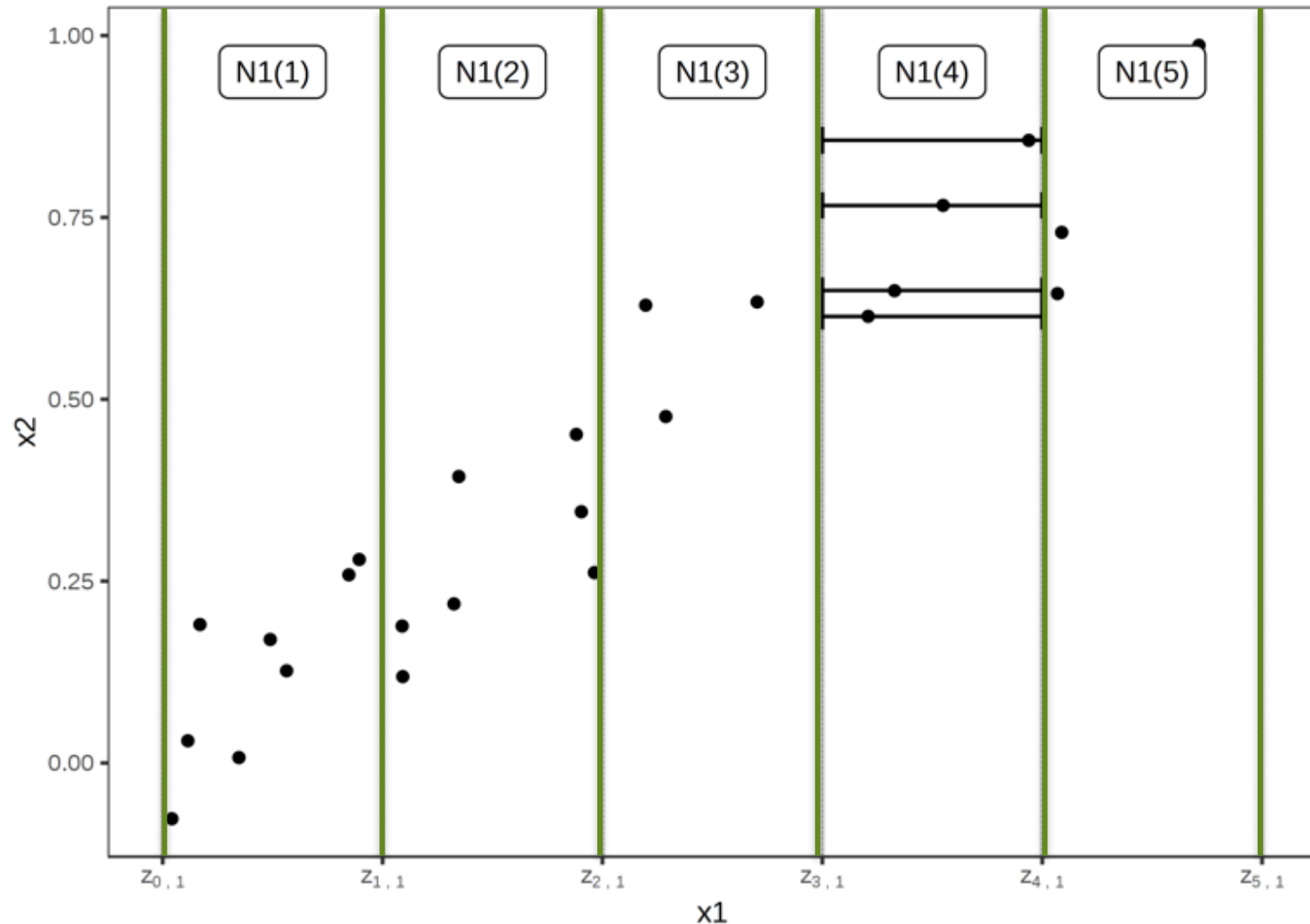
Accumulated Local Effects

For points in each interval, determine how much their prediction would change if we replace the feature of interest with the upper and lower limits of the interval (keeping all other inputs constant)



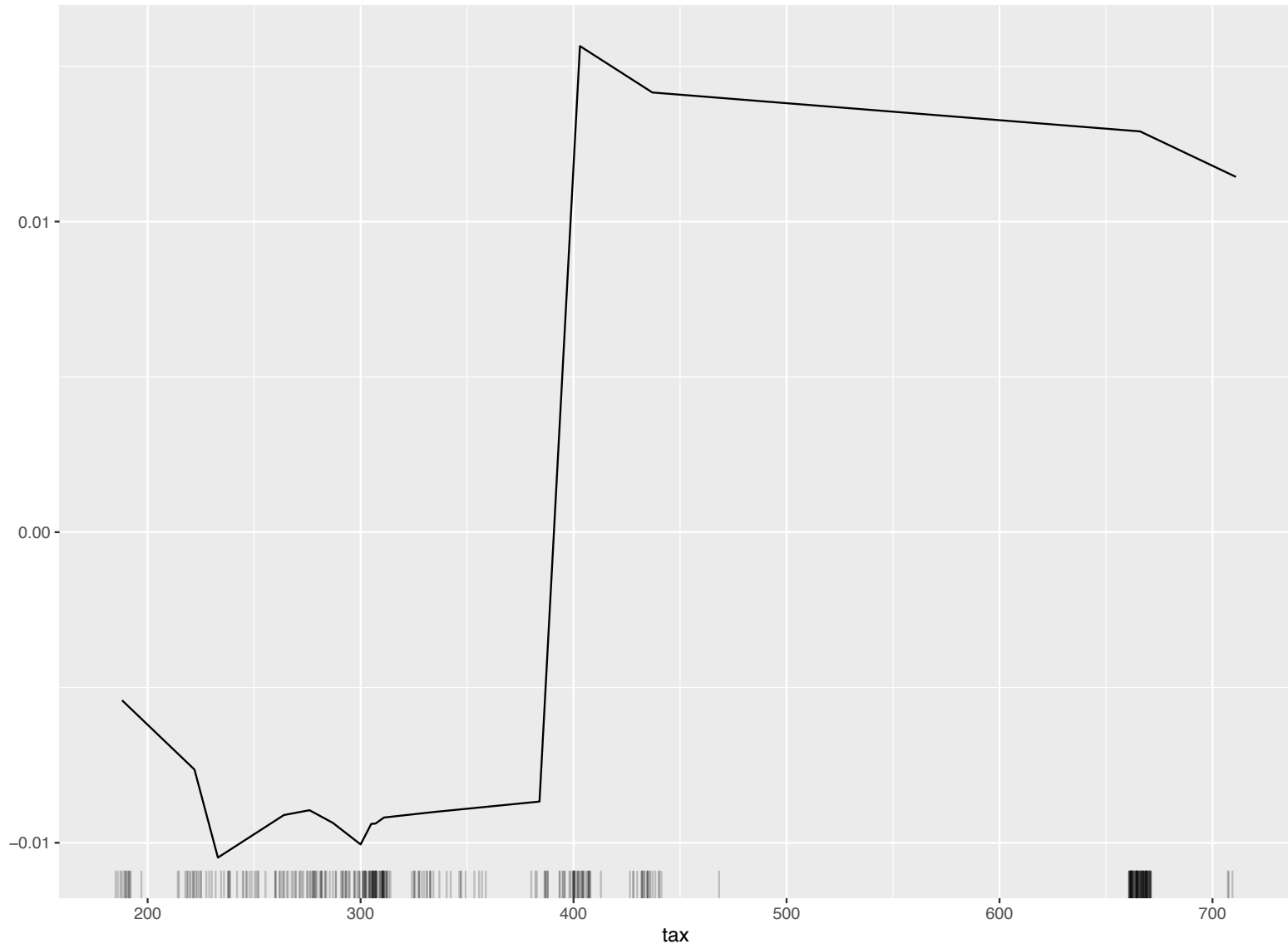
Accumulated Local Effects

These differences are later accumulated and centered, resulting in the ALE curve



Accumulated Local Effects

These differences are later accumulated and centered, resulting in the ALE curve



Accumulated Local Effects: Some Math 🤯

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

(Uncentered) ALE

Accumulated Local Effects: Some Math

Average over all points in
the interval

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

All points in the interval
(Neighborhood)

Fixing all other variables, for each
observation, difference in prediction
if variable of interest is set at the
upper vs. lower limit of the interval.

(Uncentered) ALE

Accumulated Local Effects: Some Math

ACCUMULATED: The net effect of being in, say the 3rd interval, is the sum of the effects in the 1st, 2nd, 3rd intervals.

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

LOCAL: Average difference in predictions over points in each interval.

(Uncentered) ALE

ALE after Centering

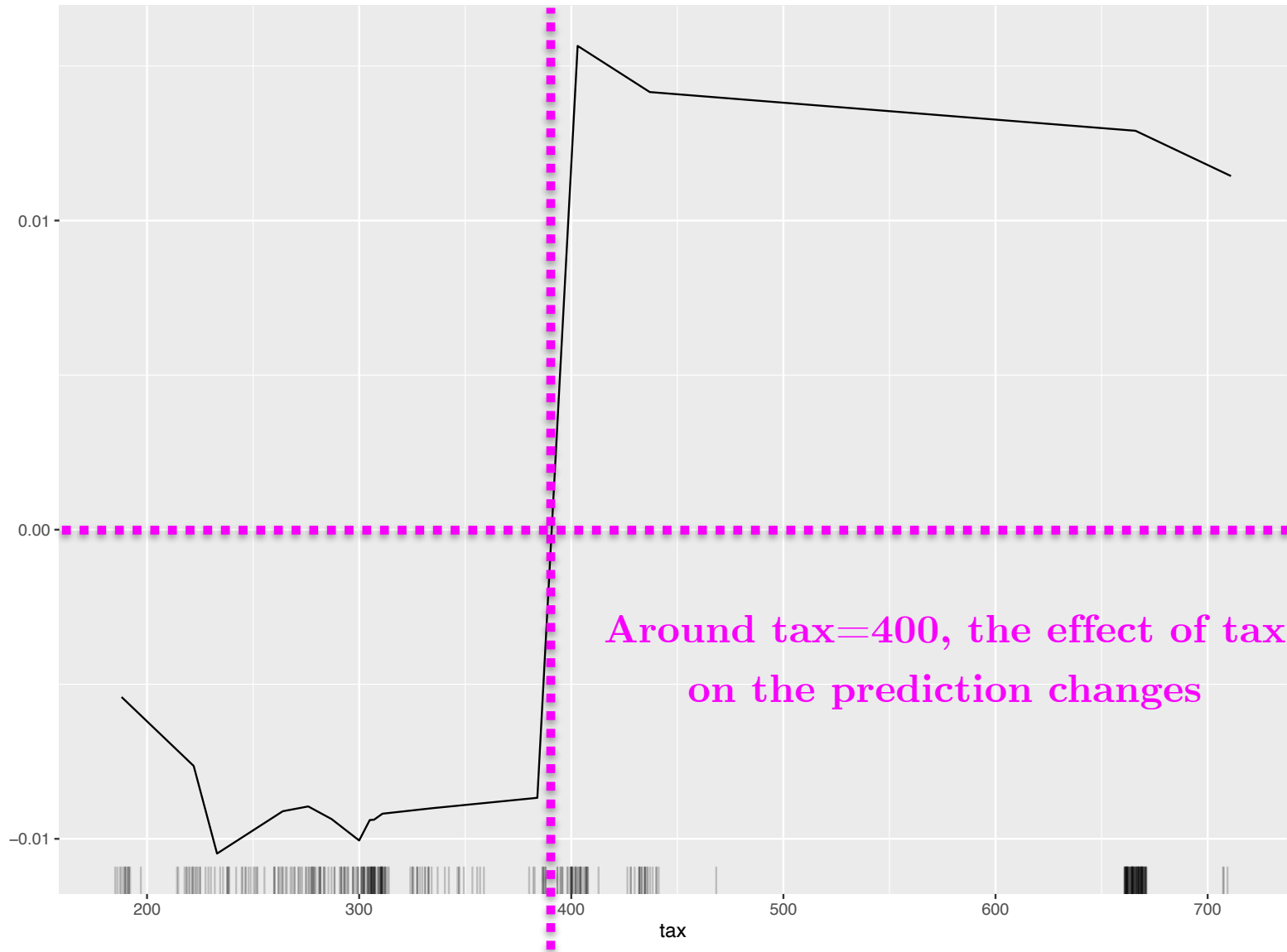
- The final math step is to center the ALE so that the mean effect is zero.

$$\hat{f}_{j,ALE}(x) = \hat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^n \hat{\tilde{f}}_{j,ALE}(x_j^{(i)})$$

This final value of ALE describes the main effect of the input variable compared to the data's average prediction.

Accumulated Local Effects

These differences are later accumulated and centered, resulting in the ALE curve



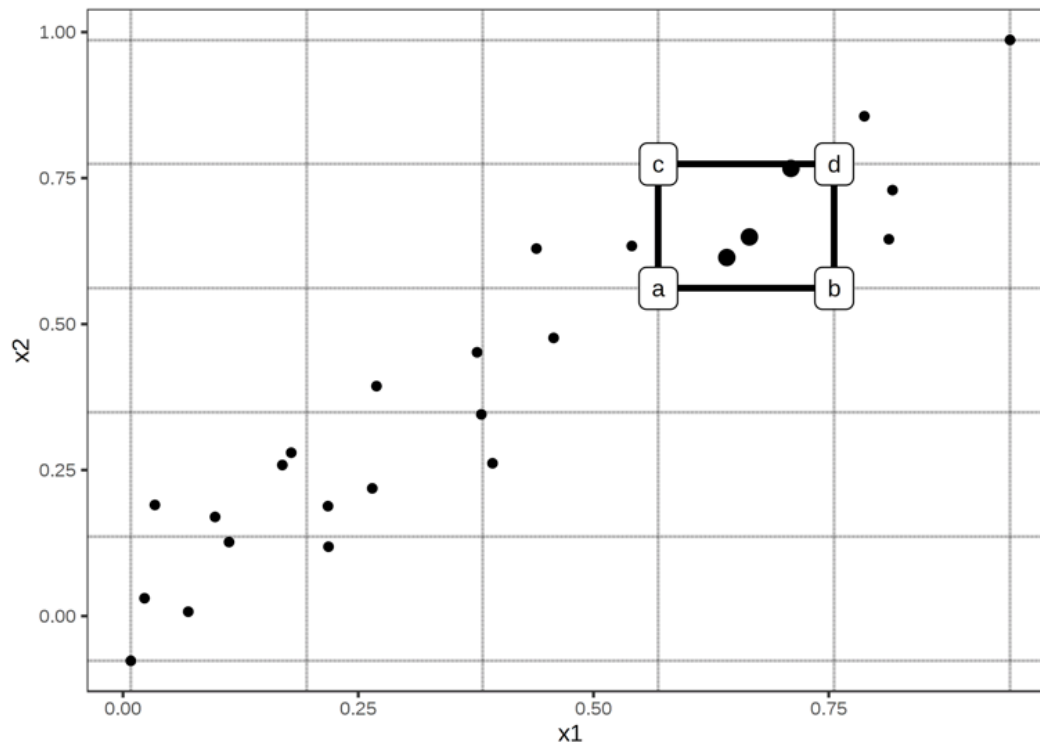
ALE Interpretation

The final value of ALE describes the main effect of the input variable compared to the data's average prediction.

Example: if the ALE for tax is 0.01 when $\text{tax}=400$, this means that when $\text{tax}=400$, the prediction is higher by 0.01 compared to the average prediction.

Second-Order ALE

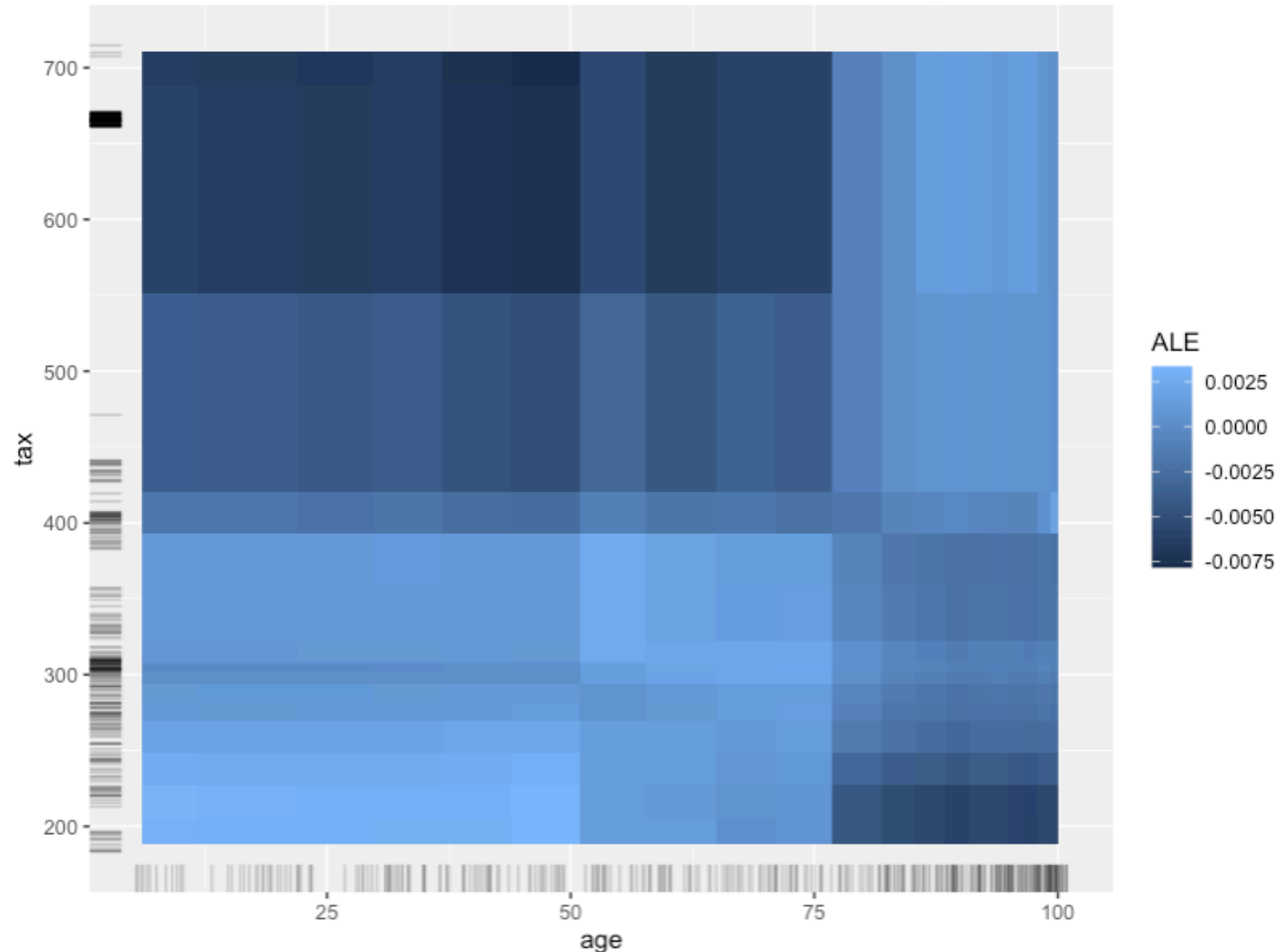
- We can do the same procedure for two input variables simultaneously, defining a rectangular grid to accumulate effects across the 2 dimensions.



ALE plots for the Interaction of Two Features

- We can do the same procedure for two input variables simultaneously, defining a rectangular grid to accumulate effects across the 2 dimensions.
- HOWEVER. This shows us only the second-order interaction effect ($a*b$) of the two variables, *after the main effects have been accounted for*.
- Example: $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$
If the interaction term is not really significant, the 2-d ALE plot for the pair (x_1, x_2) would be constant at 0 because the main effects were already accounted for.

ALE plots for the Interaction of Two Features



Types of Model Interpretability

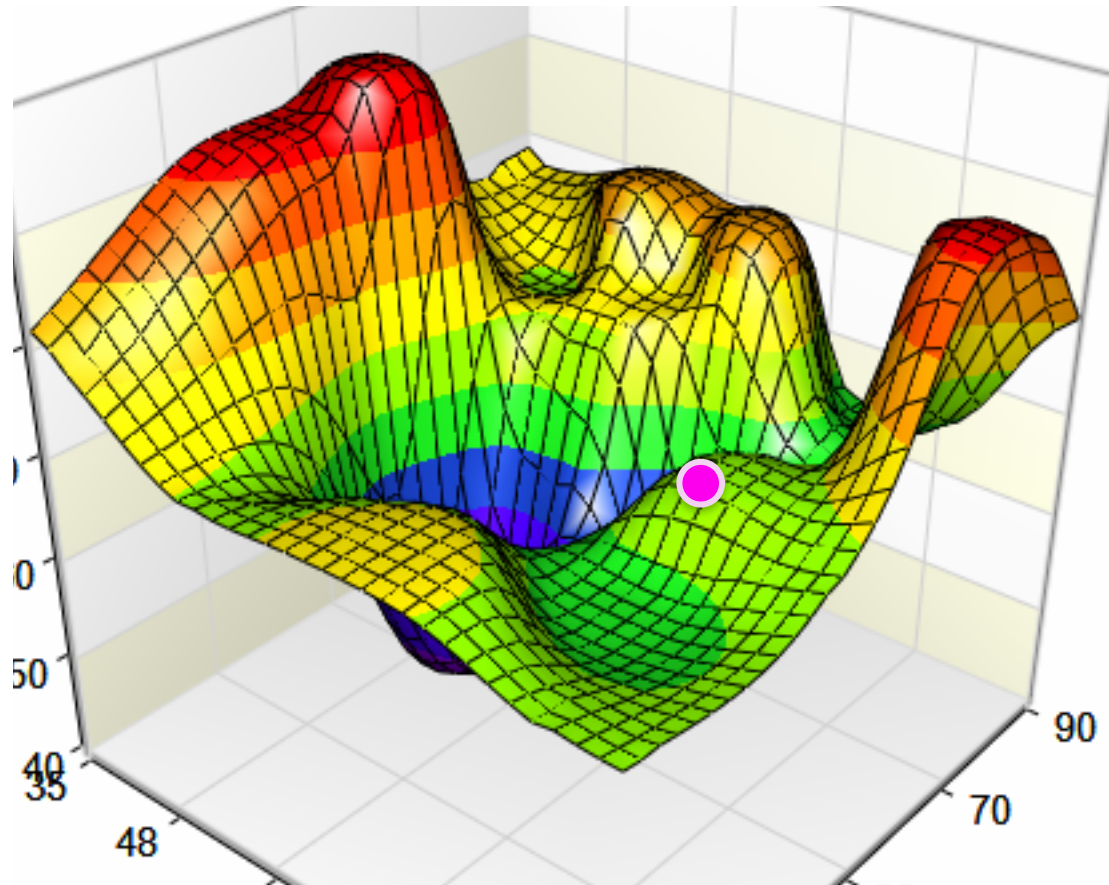
	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE ✓ LIME Shapley Values	Permutation Importance ✓ Partial Dependence ✓ ALE ✓

LIME

• • •

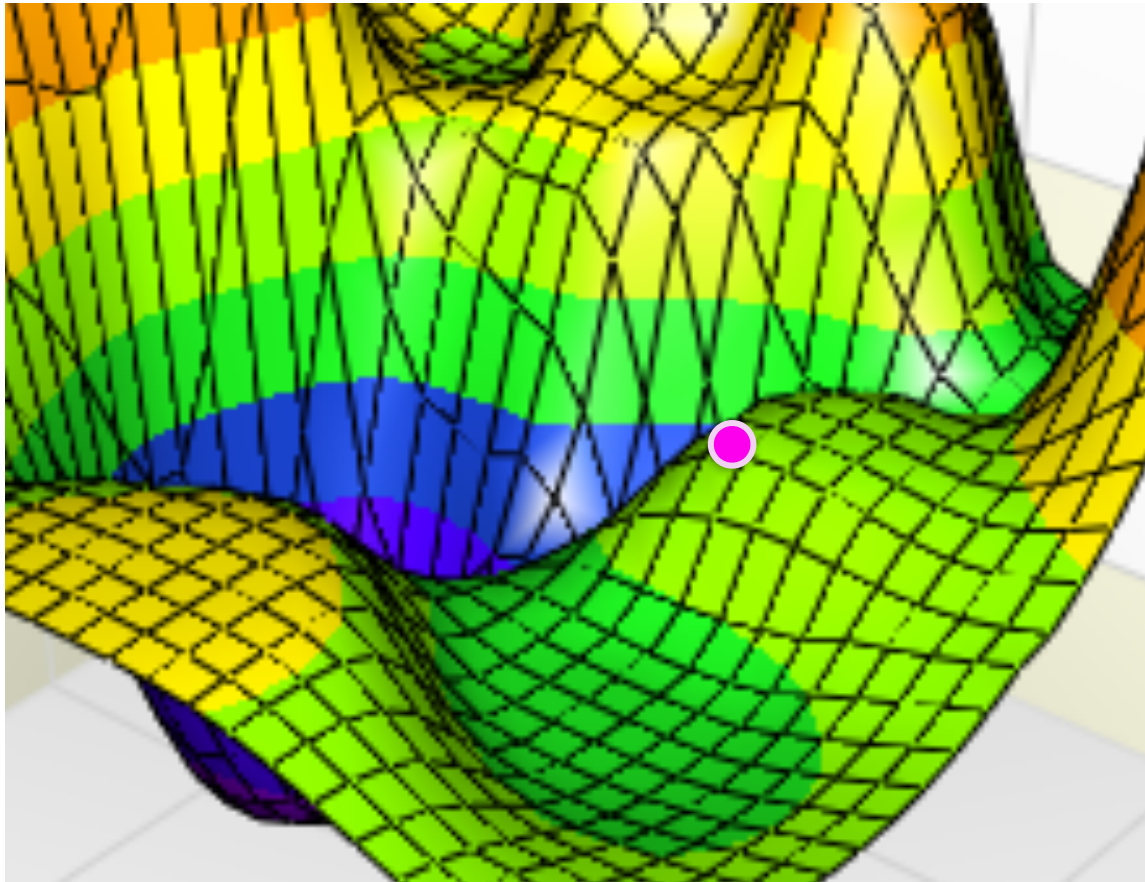
Local Interpretable Model-agnostic Explanations

LIME: Intuition



Zoom in....

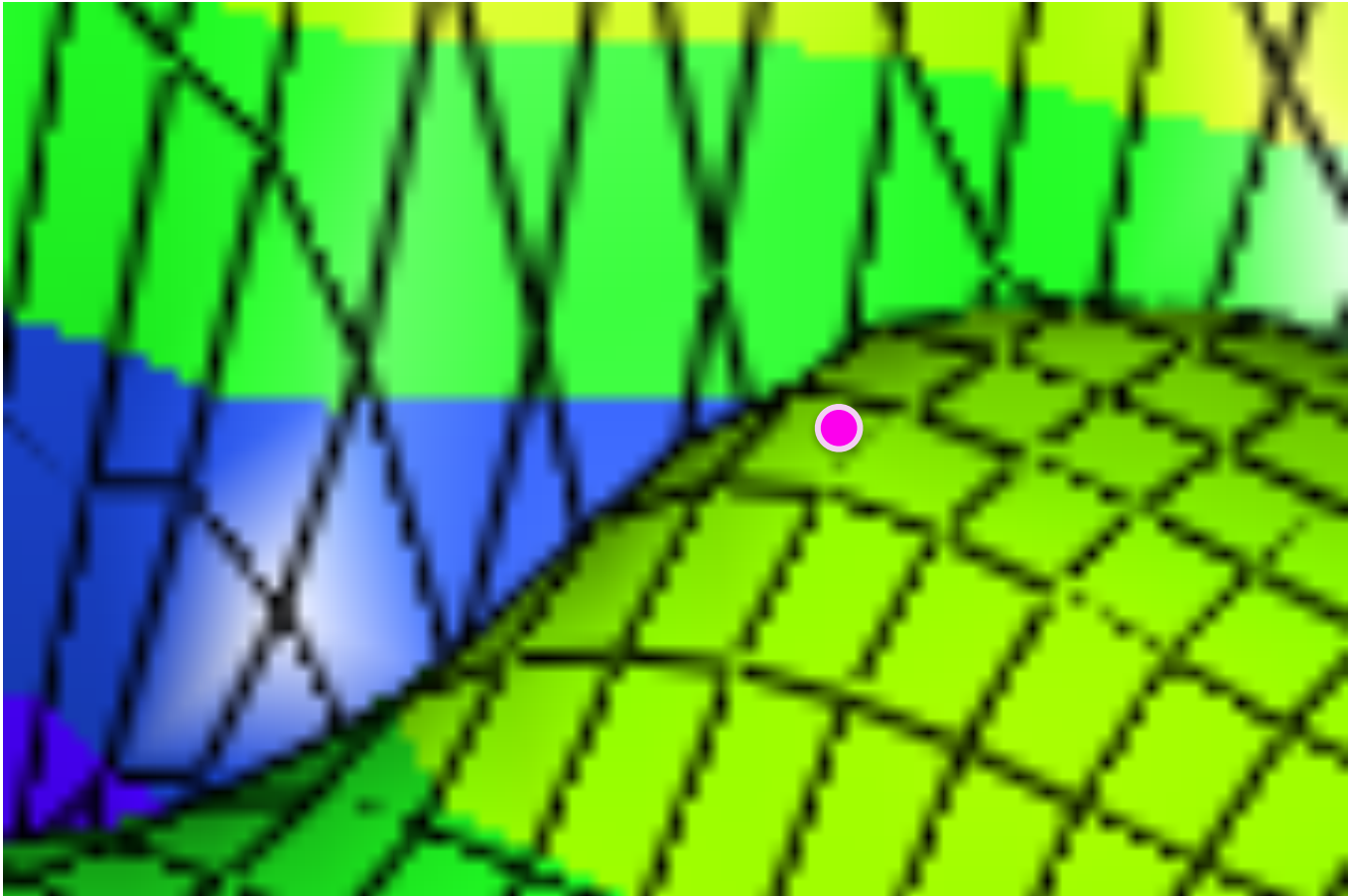
LIME: Intuition



Zoom in....

...closer

LIME: Intuition

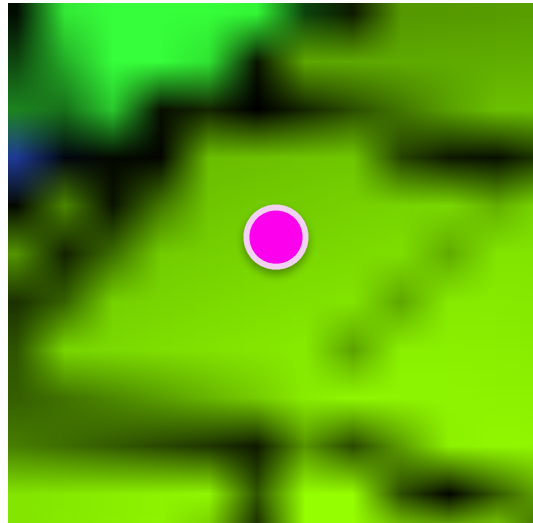


Zoom in....

...closer

LIME: Intuition

Zoom in....

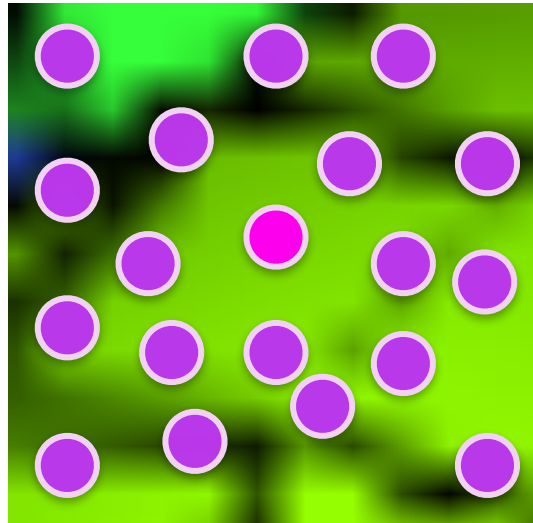


...closer

Just a dot on a flat surface!

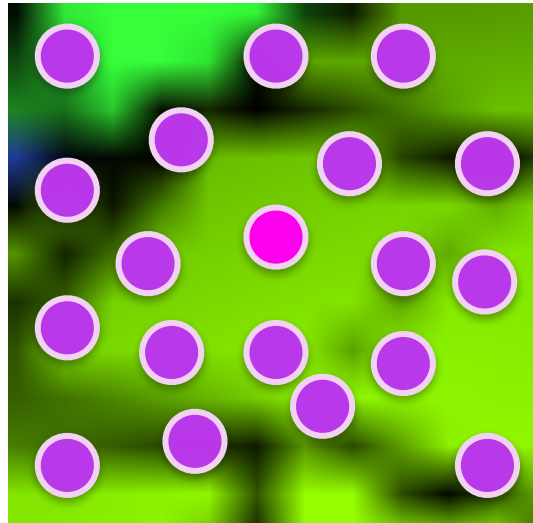
We can model that with linear regression!

LIME: Intuition



Put a bunch more data points on the function near this pink point of interest, then create a linear model using the purple and pink points as input data.

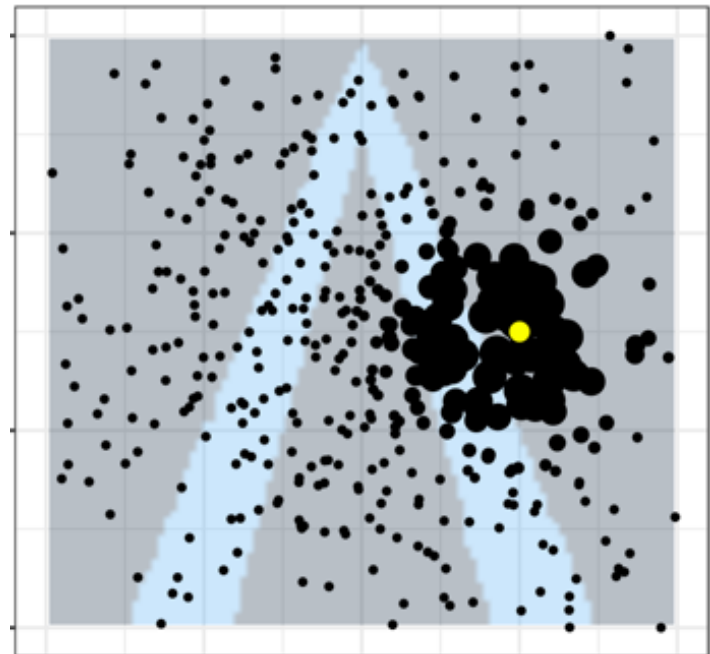
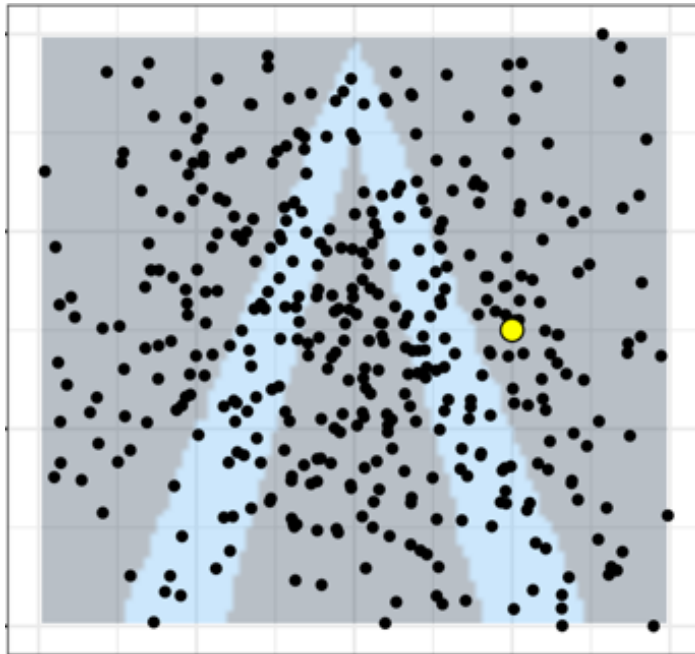
LIME: Intuition



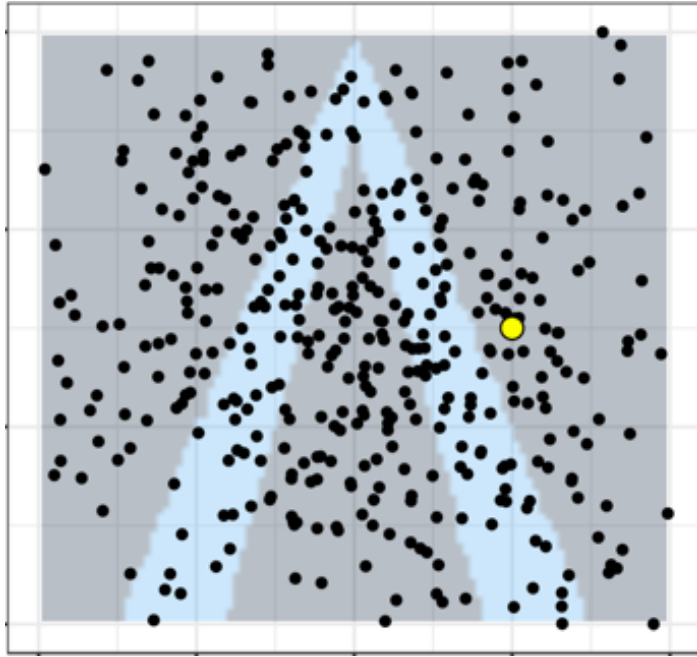
The resulting linear model could explain the exact orientation of the predictive model *at the pink point*.

LIME: Details

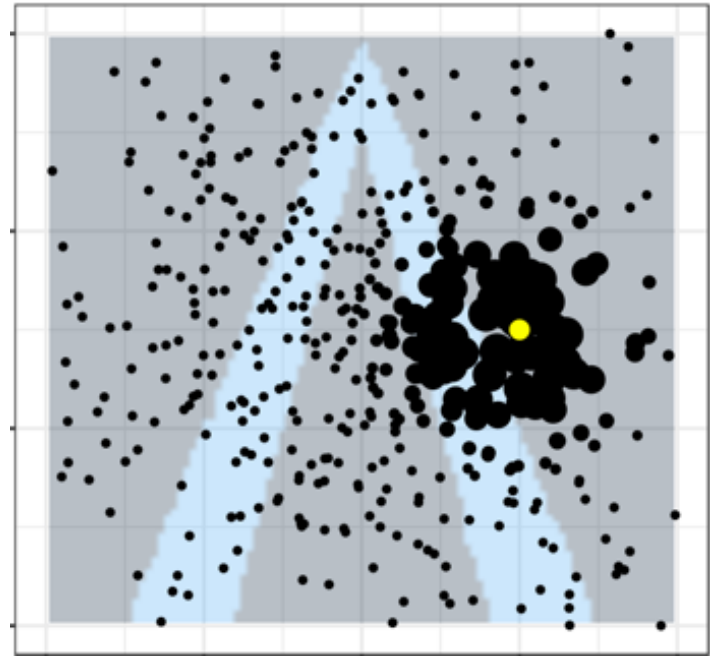
- LIME doesn't completely ignore the rest of the input space as we showed with the purple dots.
- The points it generates are normally distributed around the data's mean.
- Then the local model *weights* observations by proximity to the point of interest.



LIME: Details



Normally distributed
points sampled across
the input domain



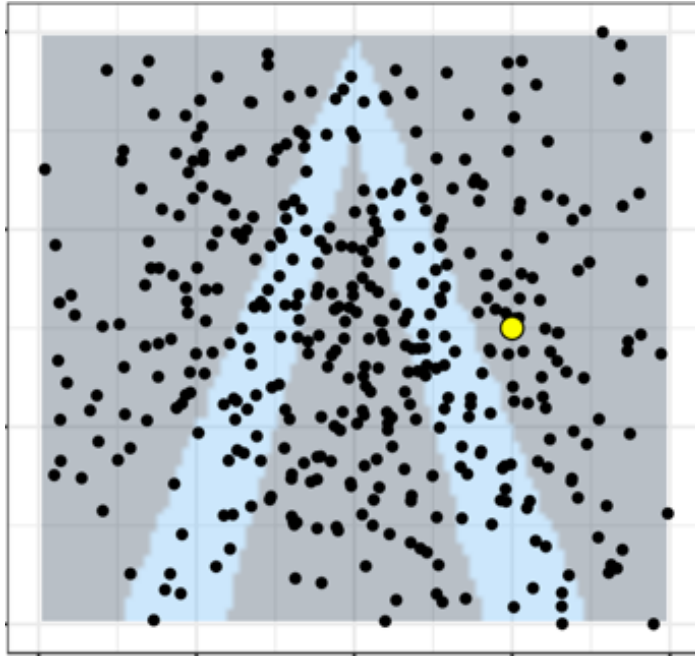
Weights of samples for
OLS determined by
proximity to yellow point.
(i.e. *weighted* least squares)

Weighted Least Squares

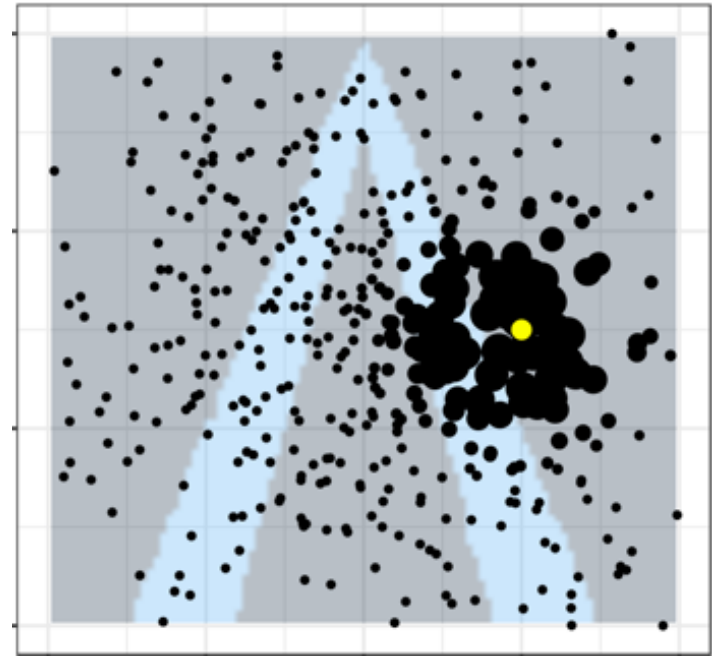
Weighted least squares gives higher weight to certain observations' residuals in the objective function of OLS.

Observations with higher weight will be more important for the linear model to predict accurately, while the model will not put much emphasis on observations with lower weight.

LIME: Details



Normally distributed
points sampled across
the input domain



Weights of samples for
OLS determined by
proximity to yellow point.

As measured by a Kernel!
Gaussian Radial Basis Function Common

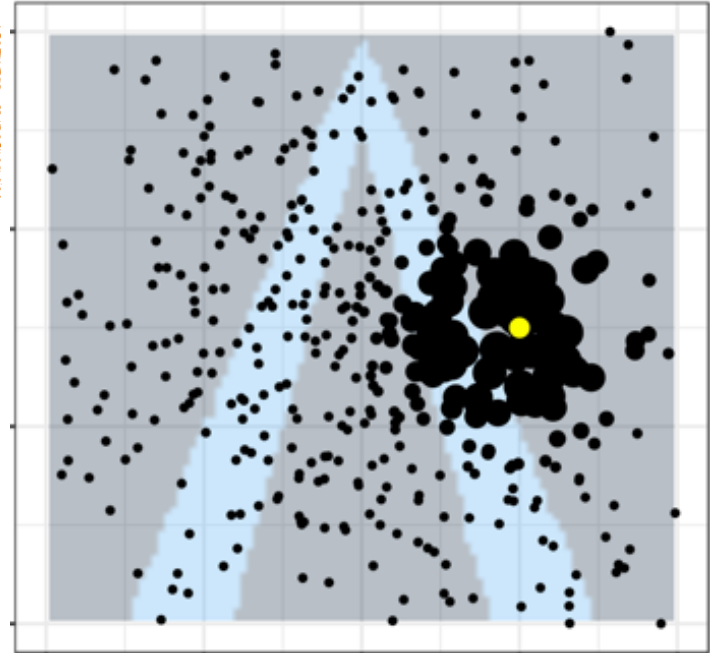
i.e. *weighted* least squares)

LIME: Devils in Details

As measured by a Kernel!
Gaussian Radial Basis Function Common

$$e^{\frac{-\|x_i - x\|_2}{2\sigma^2}}$$

How do we determine the width of our
kernel, σ ?



Weights of samples for
OLS determined by
proximity to yellow point.
(i.e. *weighted* least squares)

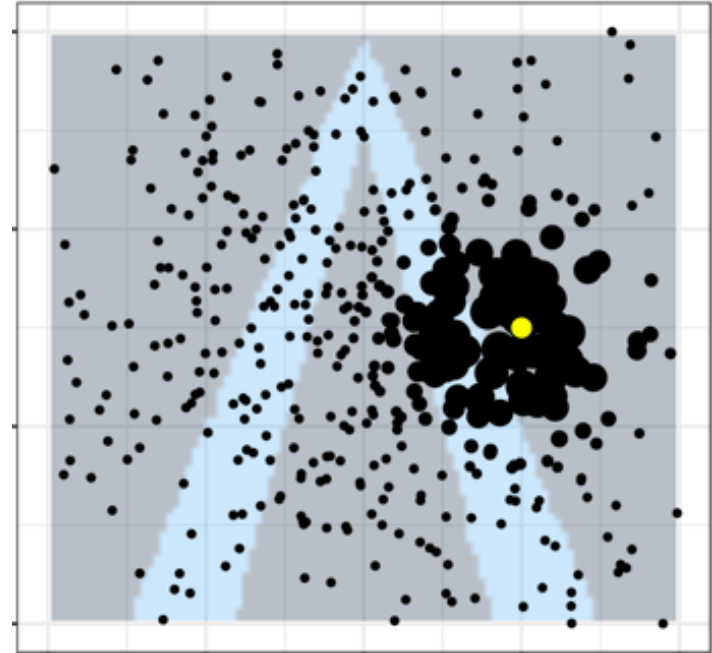
LIME: Devils in Details

As measured by a Kernel!

Gaussian Radial Basis Function Common

$$e^{\frac{-\|x_i - x\|_2}{2\sigma^2}}$$

How do we determine the width of our kernel, σ ?



In [lime/lime_tabular.py](#) you'll get some random default:

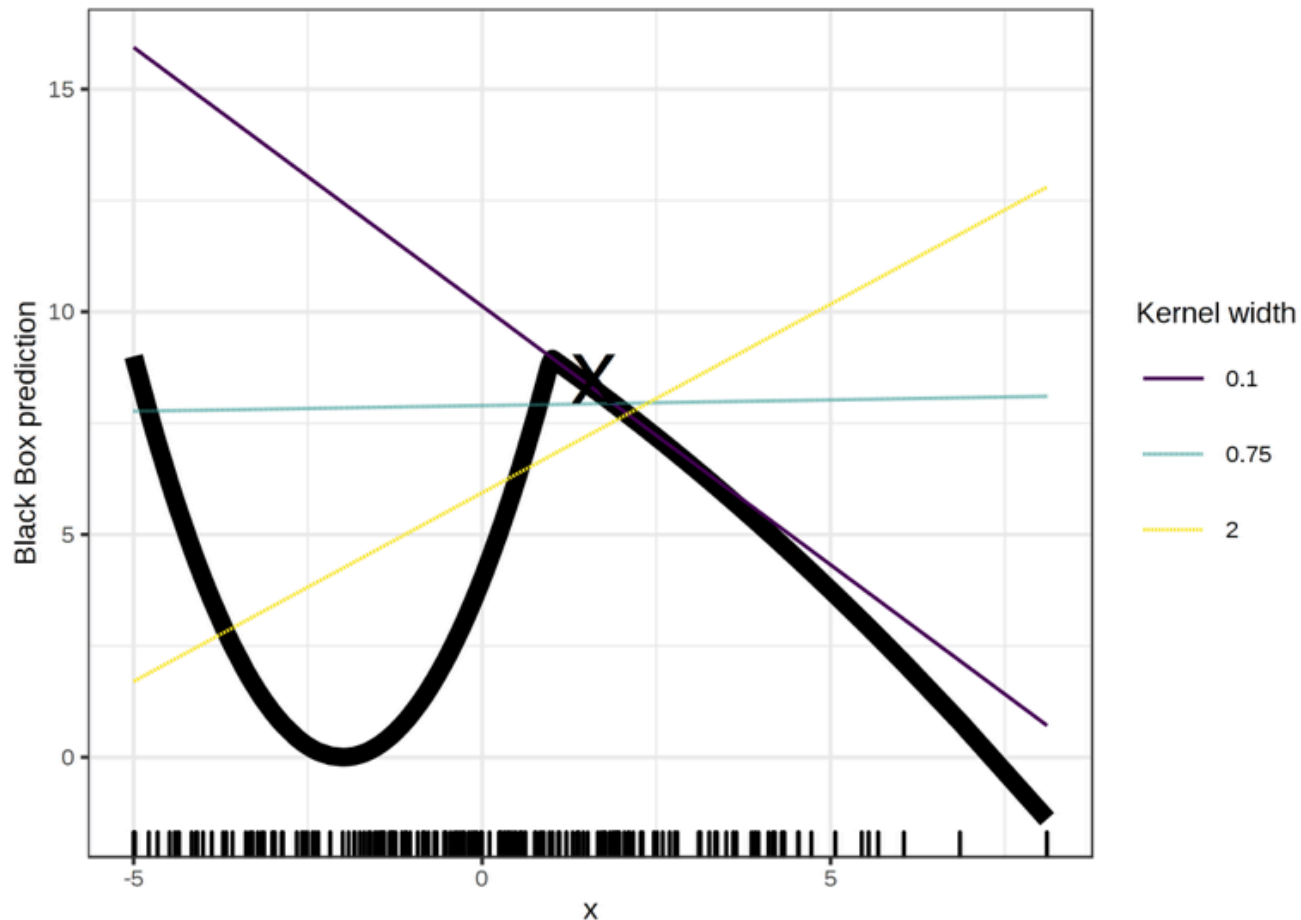
```
if kernel_width is None:
    kernel_width = np.sqrt(training_data.shape[1]) * .75
kernel_width = float(kernel_width)
```



not like this,
please

LIME: Devils in Details

And yes, it matters. Using the wrong kernel width for the wrong point can drastically alter your interpretations 😞



LIME: Devils in Details

- You have flexibility to choose any interpretable model at the local level
 - Decision trees
 - Lasso
 - OLS w/ stepwise selection
- You have to determine *how much* explanation you want: **decide how complex you want the model** to be!
 - Specify lambda
 - Specify sigma in kernel
 - Specify number of variables to use in model
- LIME **commonly used for short** (few variables in local model) **explanations and text data**
- **Uncommon in compliance** scenarios where a *full* explanation is required.

LIME: Devils in Details

- The correct definition of a “neighborhood” is a **very big, unsolved problem** when using LIME with tabular data.
- My recommendation? **Only use LIME with text data** where the simulated data is created differently:
 - From the document of interest, new documents generated by randomly removing words from original text.
 - Generated data is binary (1 if word is in document, 0 otherwise)
- Data points for local model are sampled from normal distribution, ignoring correlation between features => Unlikely or impossible generated points contribute to conclusions
- If you repeat the sampling procedure, explanations can come out different. Instability => Untrustworthy.

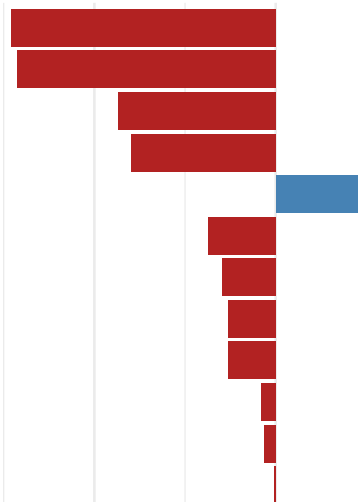
Lime Explanations

Case: 1

Prediction: 0.510662200000001

Explanation Fit: 0.22

crim <= 0.0826
3.28 < dis <= 5.14
12.5 < zn
indus <= 5.19
ptratio <= 17.4
44.2 < age <= 76.6
278 < tax <= 330
lstat <= 6.89
rad <= 4
21.2 < medv <= 25.0
6.21 < rm <= 6.58
396 < black

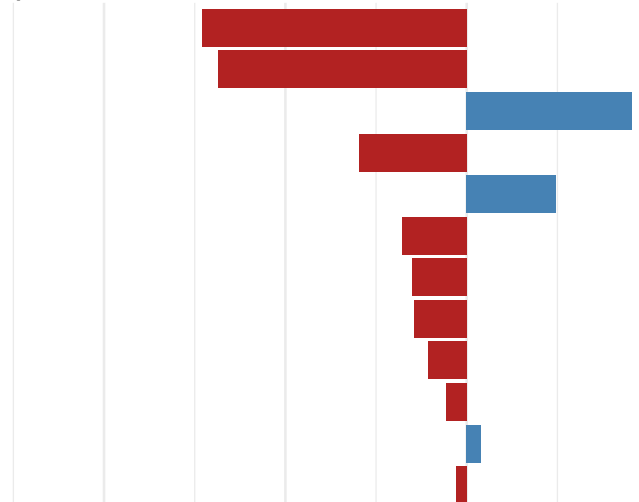


Case: 2

Prediction: 0.469658836666665

Explanation Fit: 0.21

crim <= 0.0826
3.28 < dis <= 5.14
zn <= 12.5
5.19 < indus <= 9.12
76.6 < age <= 94.5
tax <= 278
rad <= 4
17.4 < ptratio <= 19.0
6.21 < rm <= 6.58
21.2 < medv <= 25.0
396 < black
6.89 < lstat <= 11.04

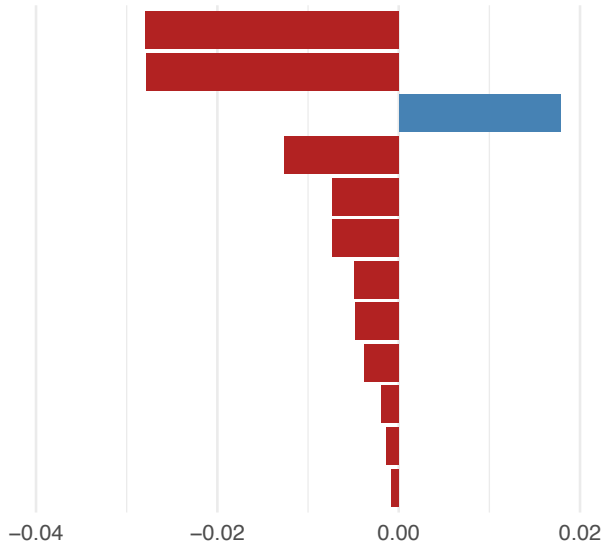


Case: 3

Prediction: 0.465765839999999

Explanation Fit: 0.20

3.28 < dis <= 5.14
crim <= 0.0826
zn <= 12.5
5.19 < indus <= 9.12
44.2 < age <= 76.6
17.4 < ptratio <= 19.0
tax <= 278
rad <= 4
lstat <= 6.89
25.0 < medv
391 < black <= 396
6.58 < rm

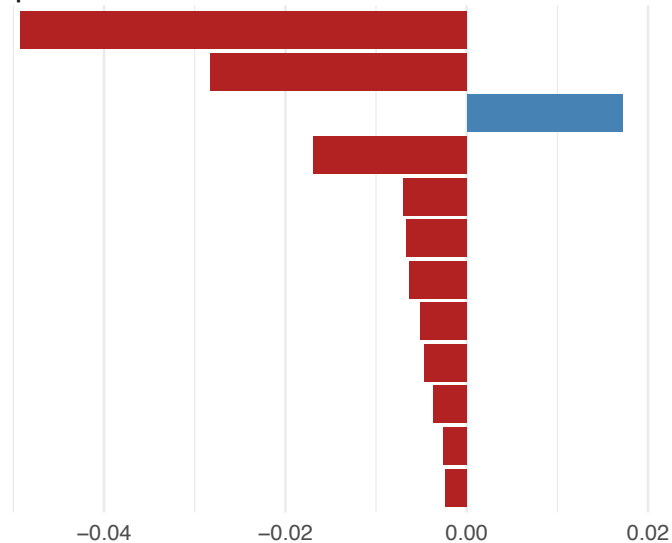


Case: 4

Prediction: 0.456098319999999

Explanation Fit: 0.36

5.14 < dis
crim <= 0.0826
zn <= 12.5
indus <= 5.19
17.4 < ptratio <= 19.0
tax <= 278
44.2 < age <= 76.6
rad <= 4
lstat <= 6.89
25.0 < medv
6.58 < rm
391 < black <= 396



Feature

Weight

Types of Model Interpretability

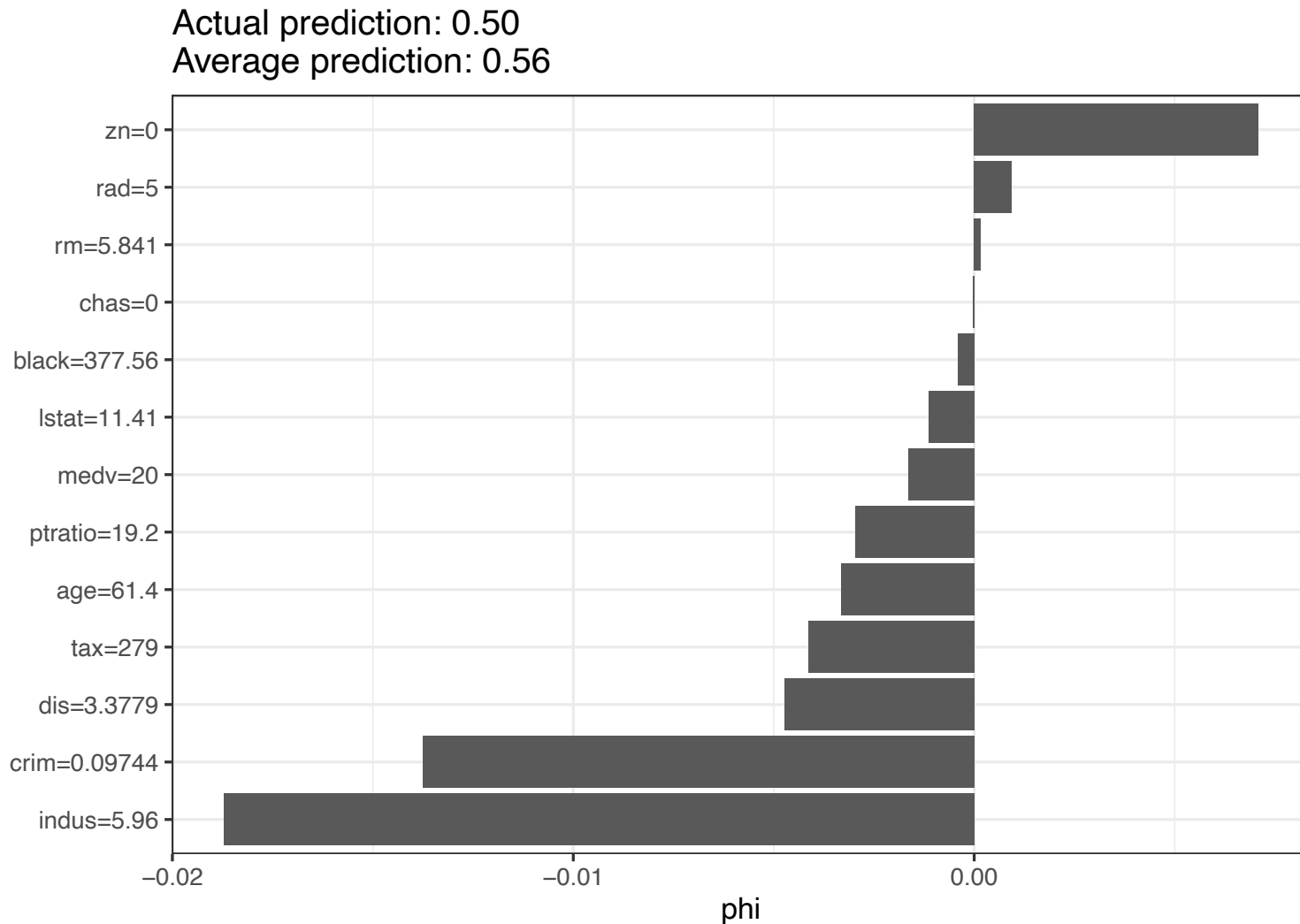
	Local	Global
Model Specific	Saliency Maps TreeSHAP	Tree Variable Importance
Model Agnostic	ICE ✓ LIME ✓ Shapley Values	Permutation Importance ✓ Partial Dependence ✓ ALE ✓

Shapley Values

Interpretation: The value of the j^{th} feature contributed ϕ_j to the prediction of this particular instance compared to the average prediction for the dataset.

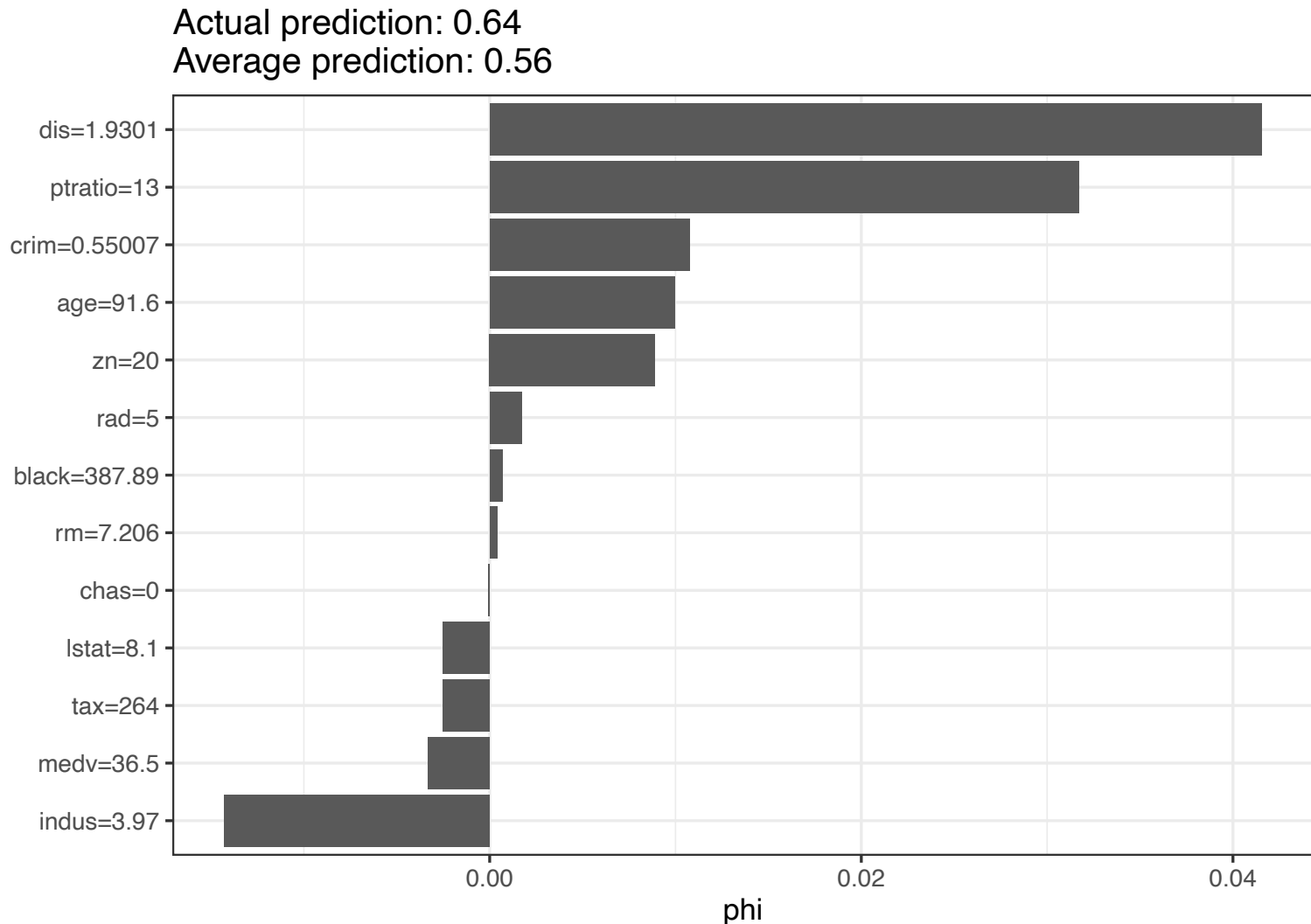
Shapley Values

Interpretation: The value of the j^{th} feature contributed ϕ_j to the prediction of this particular instance compared to the average prediction for the dataset.



Shapley Values

Interpretation: The value of the j^{th} feature contributed ϕ_j to the prediction of this particular instance compared to the average prediction for the dataset.



Shapley Values

Interpretation: The value of the j^{th} feature contributed ϕ_j to the prediction of this particular instance compared to the average prediction for the dataset.

Best Part: Shapley Values for a given instance *sum to the difference* between the given prediction and the average prediction.

Shapley Values

- Exact definition of Shapley Values is complicated and infeasible to solve for many variables
- Many feasible approximation techniques (more every day) to estimate them.
- Ideas are all similar in spirit to previous methods — Monte Carlo simulations to estimate how model changes with and without this precise feature value.

Shapley Values: Intuition

Explains a prediction as a game played by the features, each contributing a portion of points to an overall sum that yields the difference between the prediction of interest and the average prediction.

(variable of interest)

tax

chas

indus

zn

rm

medv

dis

age

black

ptratio

crime



Shapley Values: Intuition

Explains a prediction as a game played by the features, each contributing a portion of points to an overall sum that yields the difference between the prediction of interest and the average prediction.

Observation of Interest:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.55007	20	3.97	0	0.647	7.206	91.6	1.9301	5	264	13	387.89	8.1	36.5

tax

chas

indus

ptratio

zn

rm

black

medv

age

dis

crime



Shapley Values: Intuition

Explains a prediction as a game played by the features, each contributing a portion of points to an overall sum that yields the difference between the prediction of interest and the average prediction.

Observation of Interest:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
0.55007	20	3.97	0	0.647	7.206	91.6	1.9301	5	264	13	387.89	8.1	36.5

tax
264

chas
0

indus
3.97

ptratio
13

zn
20

rm
7.206

black
387.89

medv
36.5

age
91.6

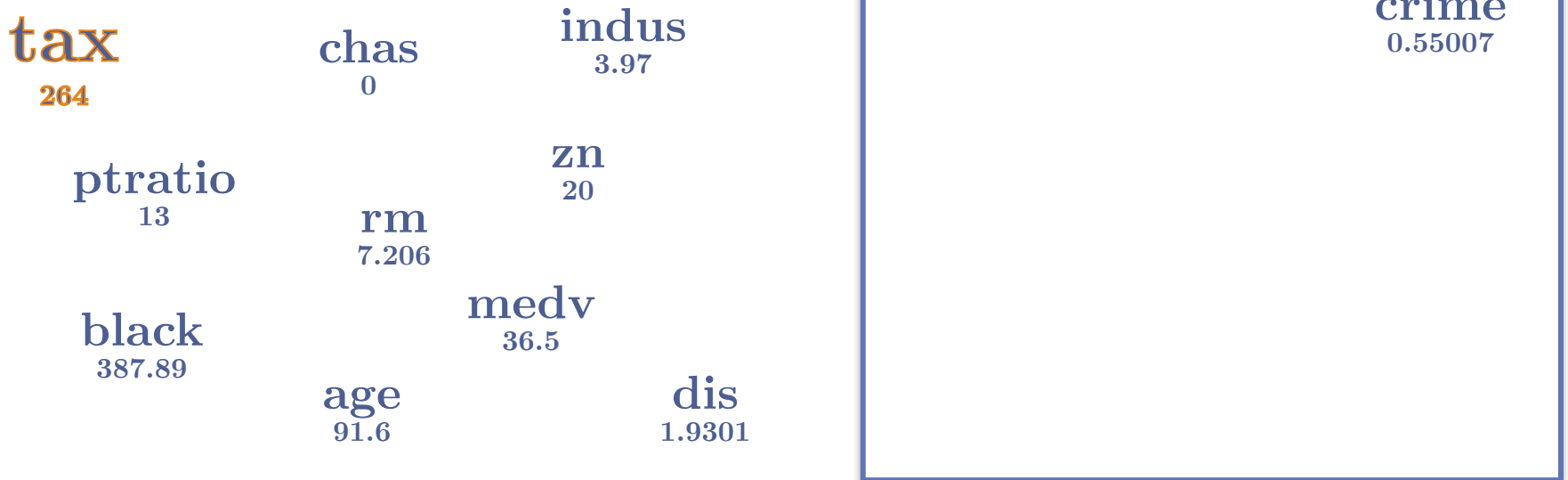
dis
1.9301

crime
0.55007



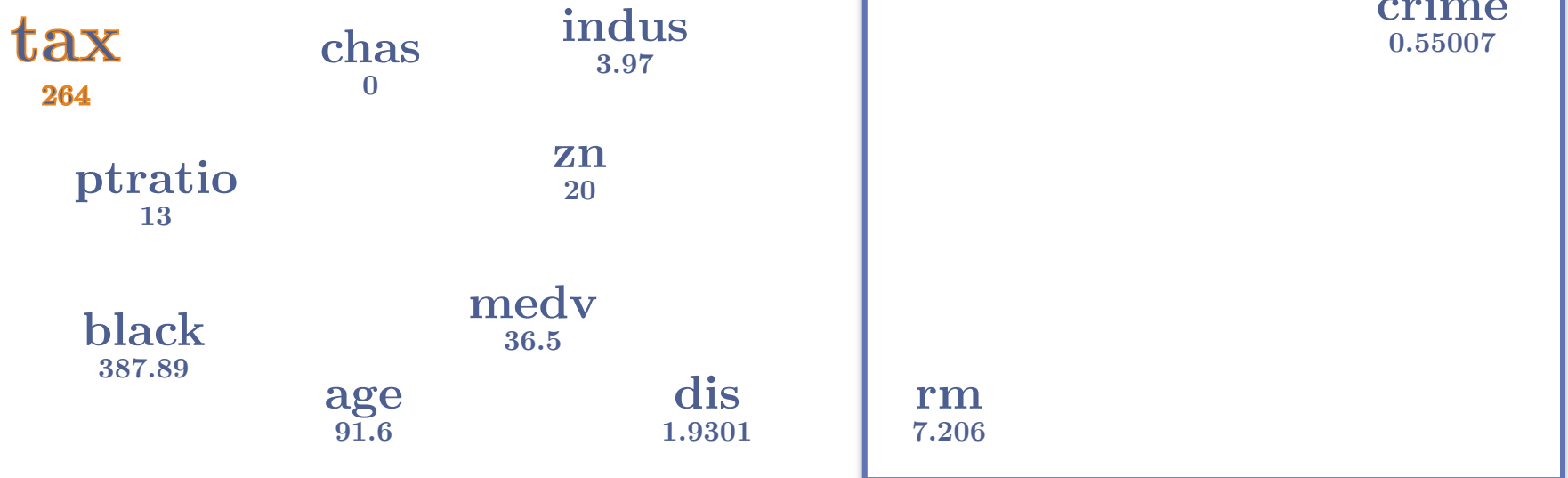
Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value



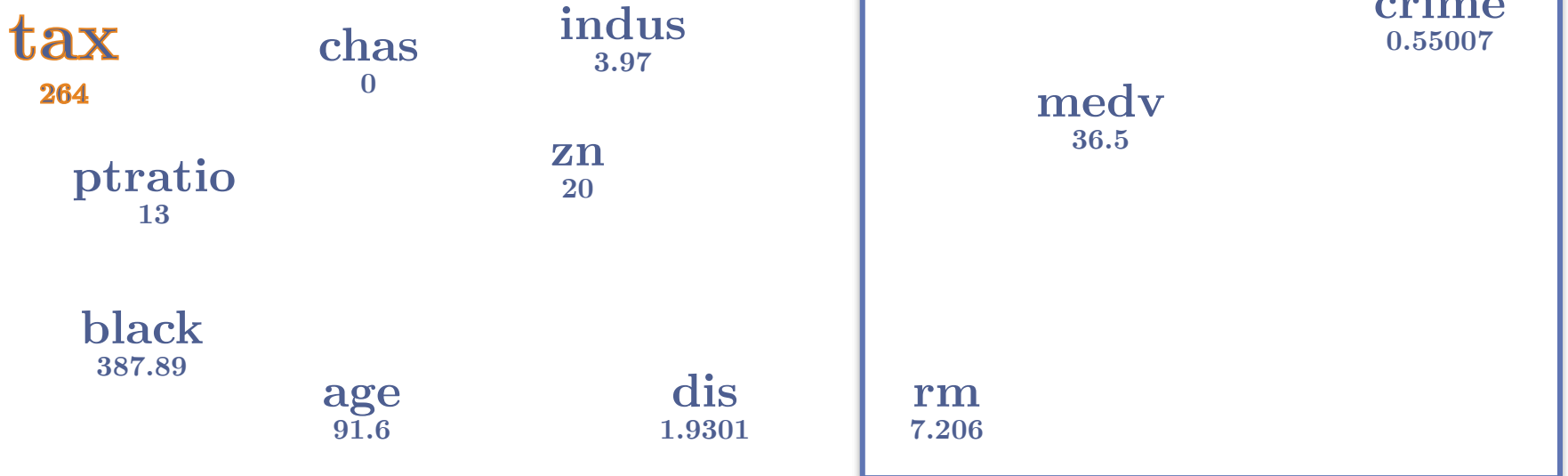
Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value



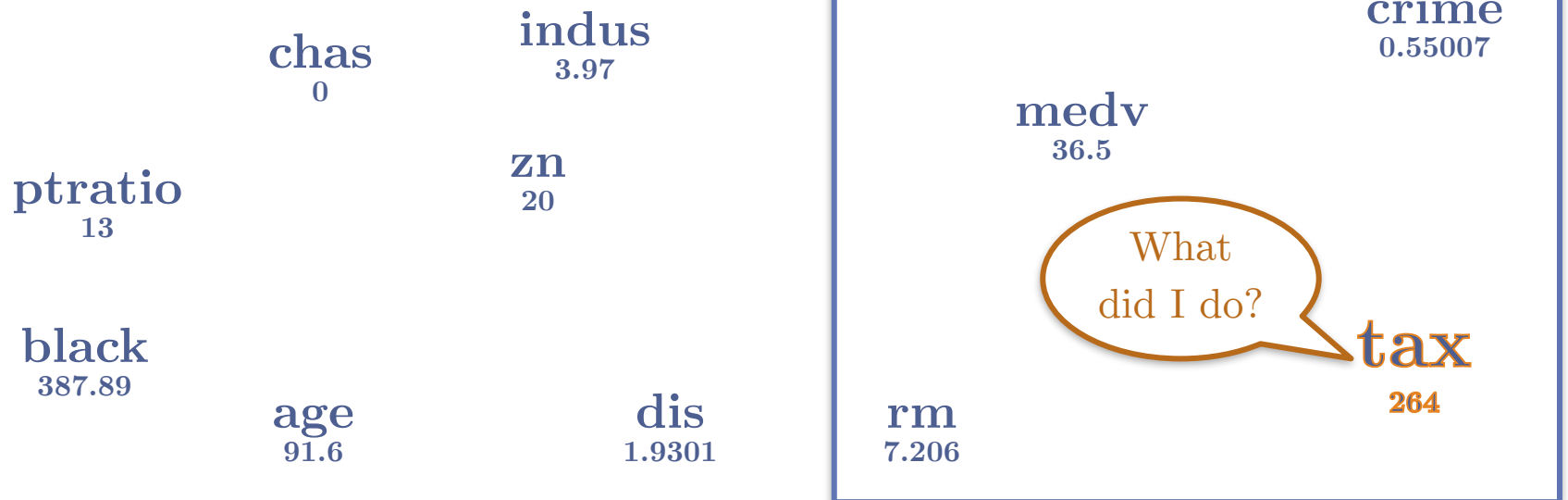
Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value



Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value



Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value

tax
264

chas
0

indus
3.97

age
91.6

ptratio
13

zn
20

rm
7.206

black
387.89

medv
36.5

dis
1.9301

crime
0.55007

Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value

tax
264

chas
0

indus
3.97

age
91.6

ptratio
13

zn
20

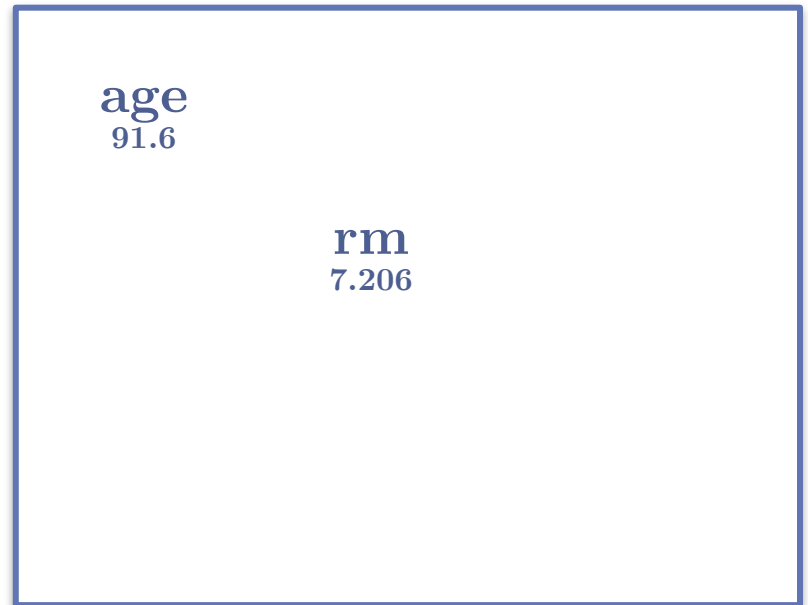
rm
7.206

black
387.89

medv
36.5

dis
1.9301

crime
0.55007



Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value

tax
264

chas
0

indus
3.97

age
91.6

ptratio
13

zn
20

rm
7.206

black
387.89

medv
36.5

dis
1.9301

crime
0.55007

Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value

tax
264

indus
3.97

zn
20

black
387.89

medv
36.5

age
91.6

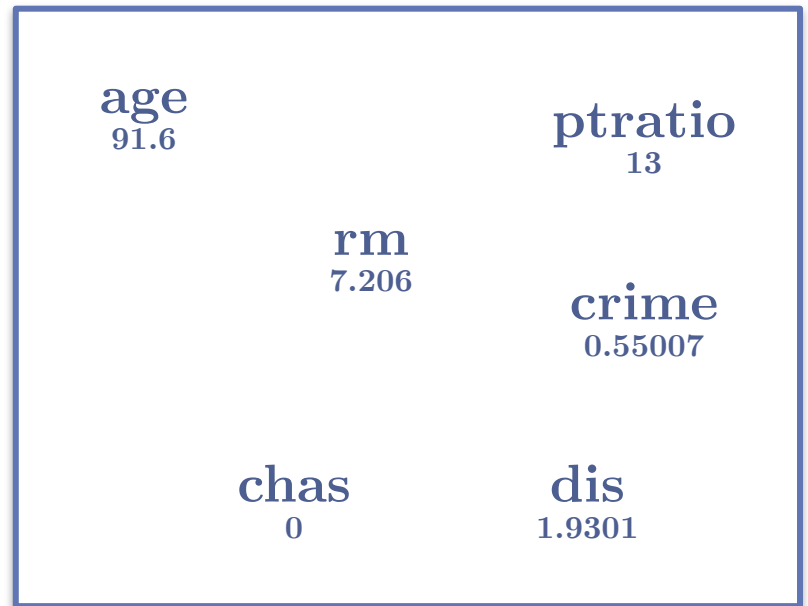
ptratio
13

rm
7.206

crime
0.55007

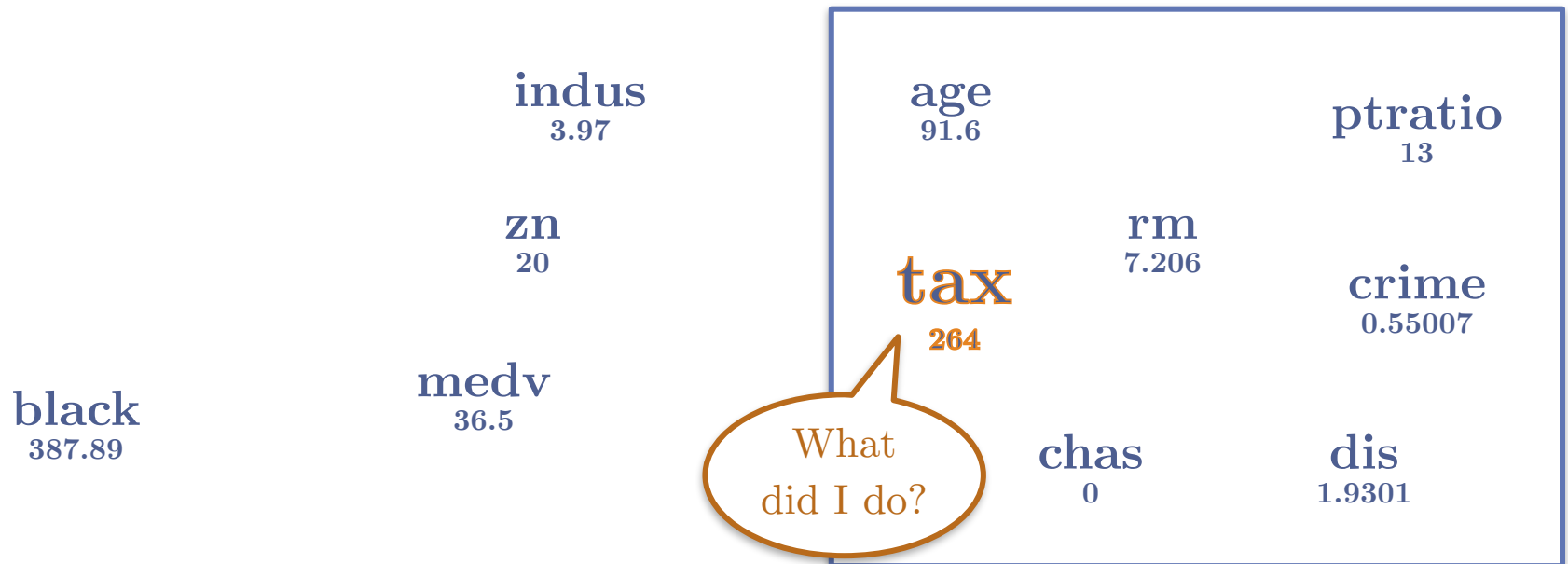
chas
0

dis
1.9301



Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value



Shapley Values: Intuition

Game play: Features enter a room (black box model) in a random order. All features in the room make contribution to the prediction. The Shapley value of a given feature value is the average change in the prediction that the “coalition” of features already in the room experiences upon being joined by that specific feature value

Features outside of the “room” take on values from a random observation rather than the observation of interest. => explanations built on “Frankenstein” observations

black
387.89

medv
36.5

indus
3.97

zn
20

age
91.6

ptratio
13

tax
264

rm
7.206

crime
0.55007

chas
0

dis
1.9301

Approximation of Shapley Values

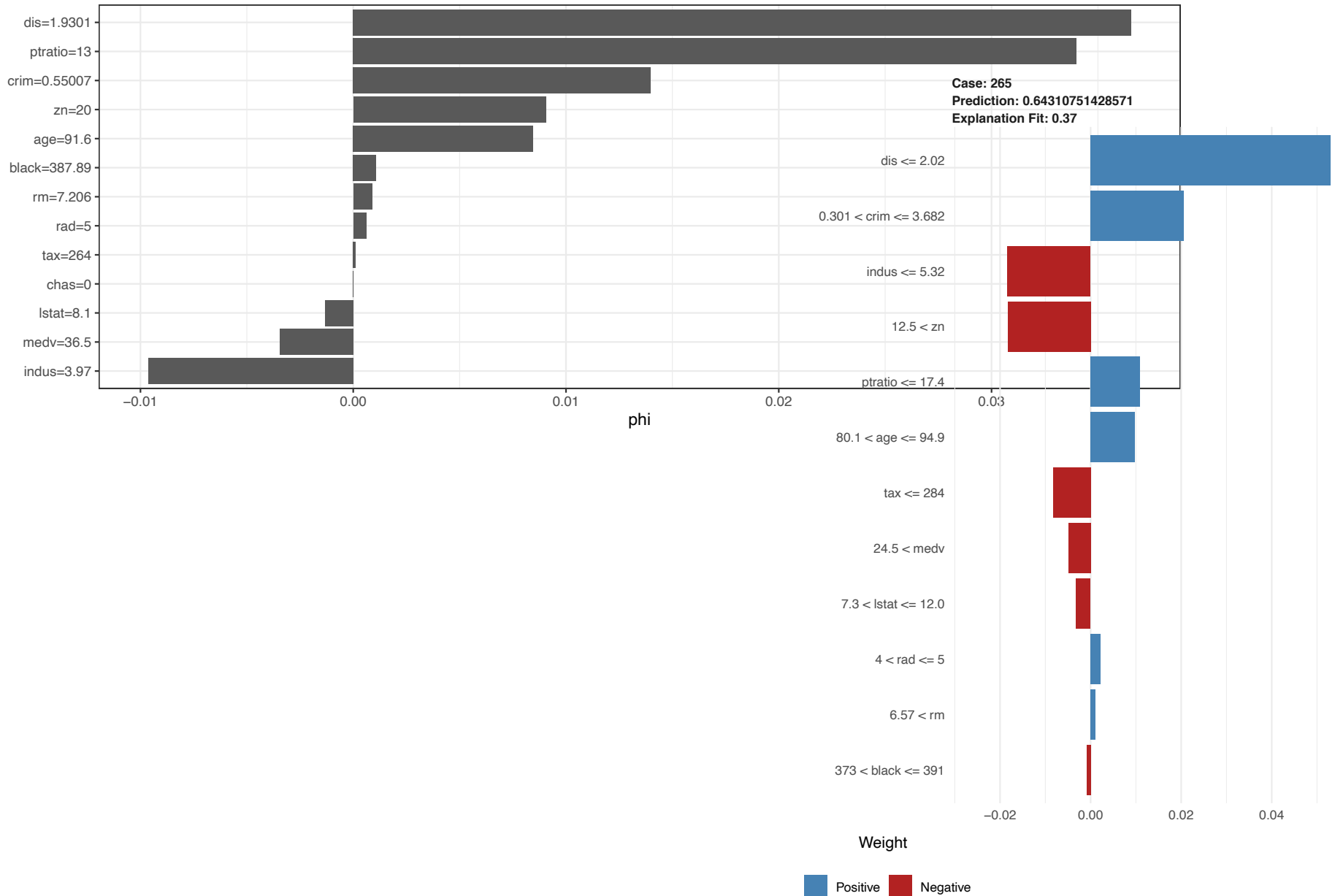
Approximate Shapley estimation for single feature value:

- Output: Shapley value for the value of the j-th feature
- Required: Number of iterations M, instance of interest x, feature index j, data matrix X, and machine learning model f
- For all $m = 1, \dots, M$:
 - Draw random instance z from the data matrix X
 - Choose a random permutation o of the feature values
 - Order instance x: $x_o = (x_{(1)}, \dots, x_{(j)}, \dots, x_{(p)})$
 - Order instance z: $z_o = (z_{(1)}, \dots, z_{(j)}, \dots, z_{(p)})$
 - Construct two new instances
 - With feature j: $x_{+j} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots, z_{(p)})$
 - Without feature j: $x_{-j} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots, z_{(p)})$
 - Compute marginal contribution: $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$
- Compute Shapley value as the average: $\phi_j(x) = \frac{1}{M} \sum_{m=1}^M \phi_j^m$

(Strumbelj et al 2014)

LIME vs. Shapley

Observation 265

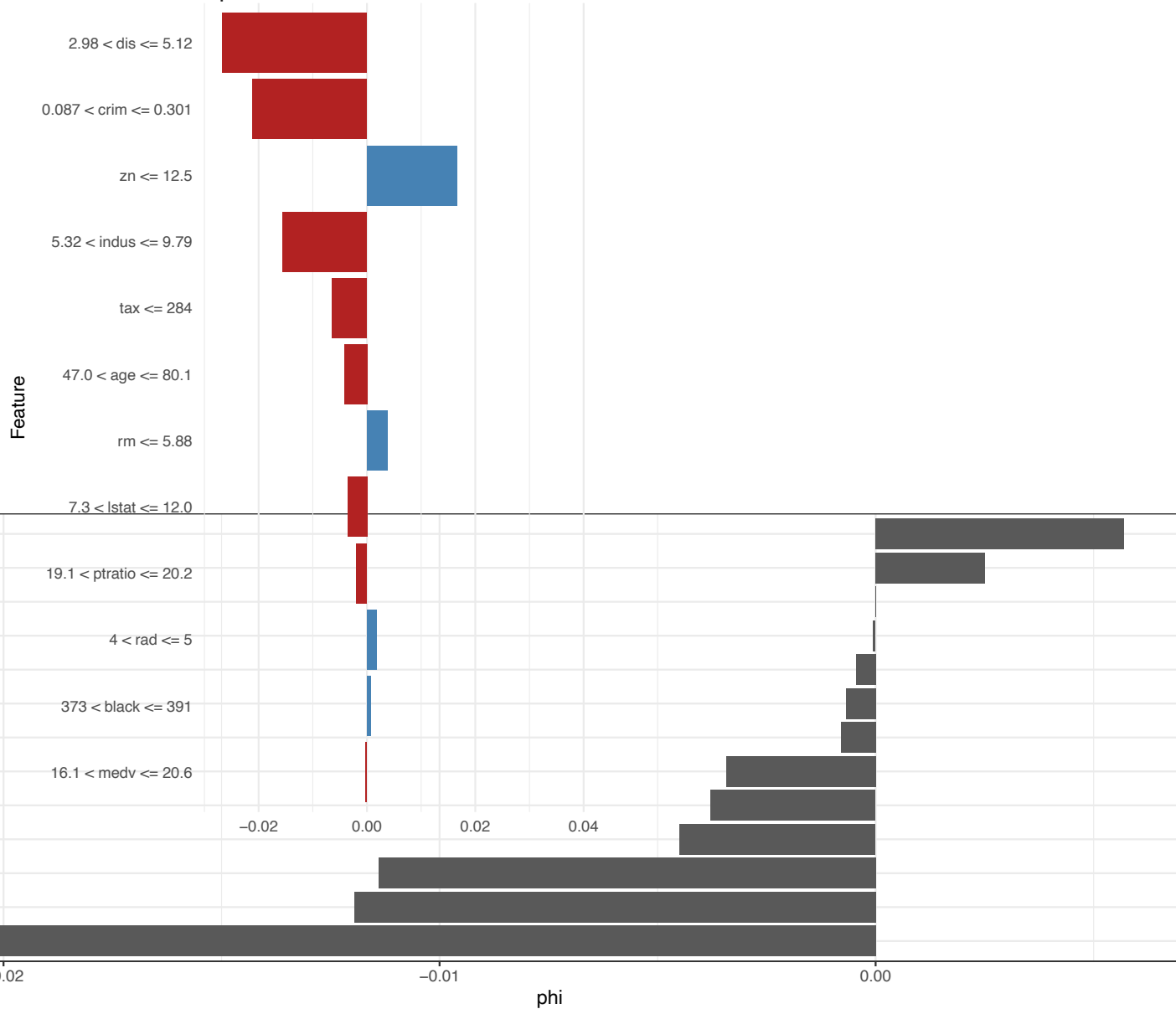


LIME vs. Shapley

Case: 37

Prediction: 0.497615083333332

Explanation Fit: 0.17



Shapley Values

The Shapley value is the *only* current explanation method with a **solid theory, proposed in 4 axioms.**

- *Efficiency*: Feature contributions must add up to the difference of prediction for point of interest vs. the average
- *Symmetry*: Contributions of two features j and k should be the same if they contribute equally to all possible coalitions
- *Dummy*: A feature that does not change the predicted value, for any of the coalitions, should have a Shapley value of 0.
- *Additivity*: For a forest of trees, the Shapley value of the forest for a given point should be the average of the Shapley Values for each tree and that given point.