Intro to Explainability in ML

PDX ML

Julio Barros

Me:

- @JulioBarros
- Machine Learning Engineer (Consultant)
- E-String.com

Explainability

The idea of how well a human can understand the decision is often called *interpretability* or *explainability*.

Confession

I pulled a bait and switch.

Explainability/Interpretability related to:

- Fairness - (socially) unbiased and not discriminating
- Safety / Reliability - errors are not catastrophic
- Privacy - protecting sensitive information
- Justification - why a decision is good

You might be asked: *What*

What features were used to make this prediction/decision?

- Feature selection and engineering
- Model selection

Or maybe: *How*

- How does the algorithm work?
- How does the input affect the output?
- How are features and outputs correlated?

But *Why* is special

Why was this prediction made?

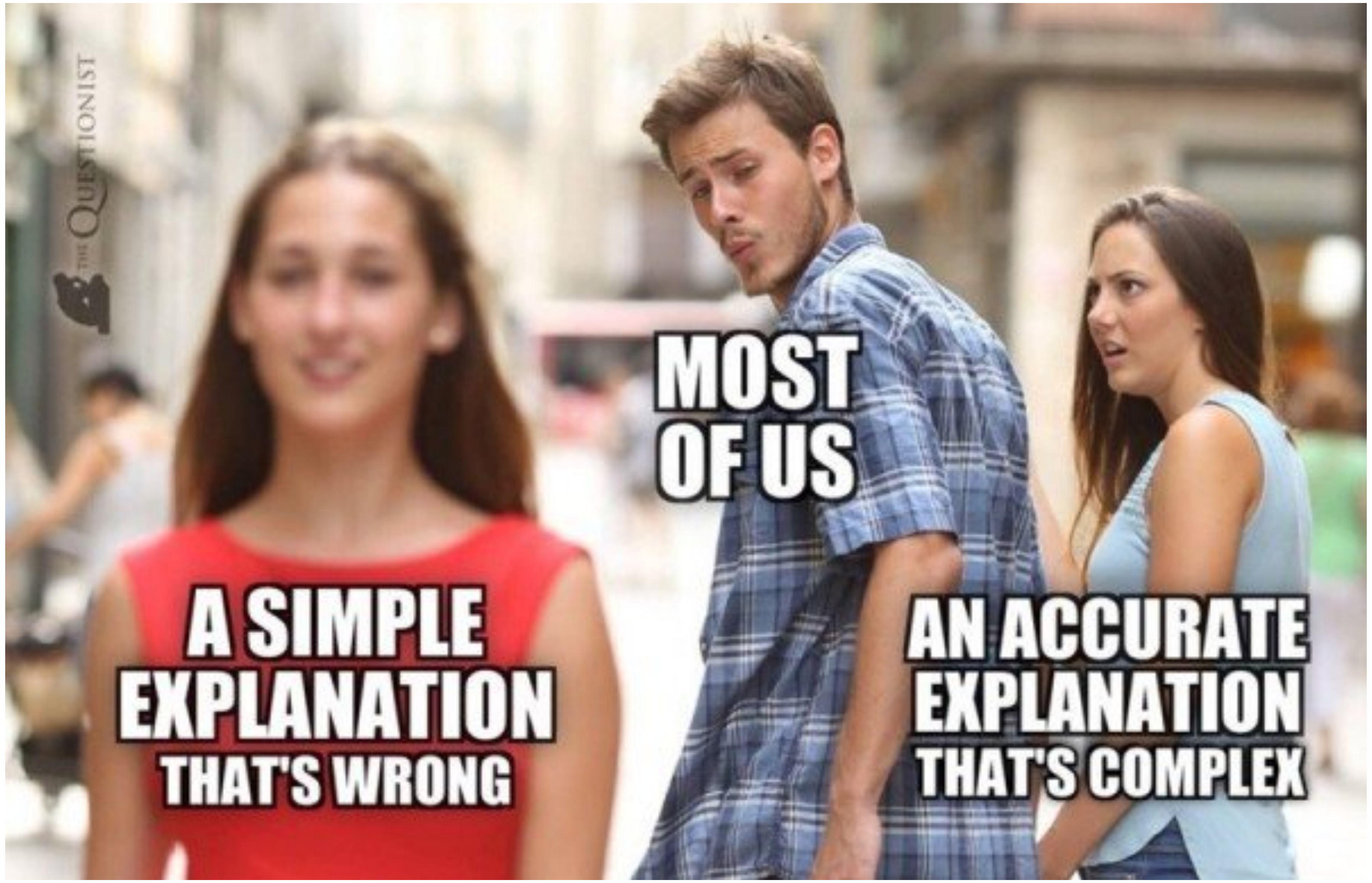
Why should I trust your prediction?

Good (understandable?) explanations

Deep down your boss/client/user wants the explanation to be:

- monotonic
- homoscedastic
- not probabilistic
- contrastive
- selective
- perscriptive
- conformant to social expectations

Almost everyone


THE QUESTIONIST
MOST OF US
A SIMPLE EXPLANATION THAT'S WRONG
AN ACCURATE EXPLANATION THAT'S COMPLEX

Radical Thesis

Explainability and interpretability are two different things.

- Explainability is *why*
- Interpretability is *how*

Unfortunately

Humans want meaning (*why*).

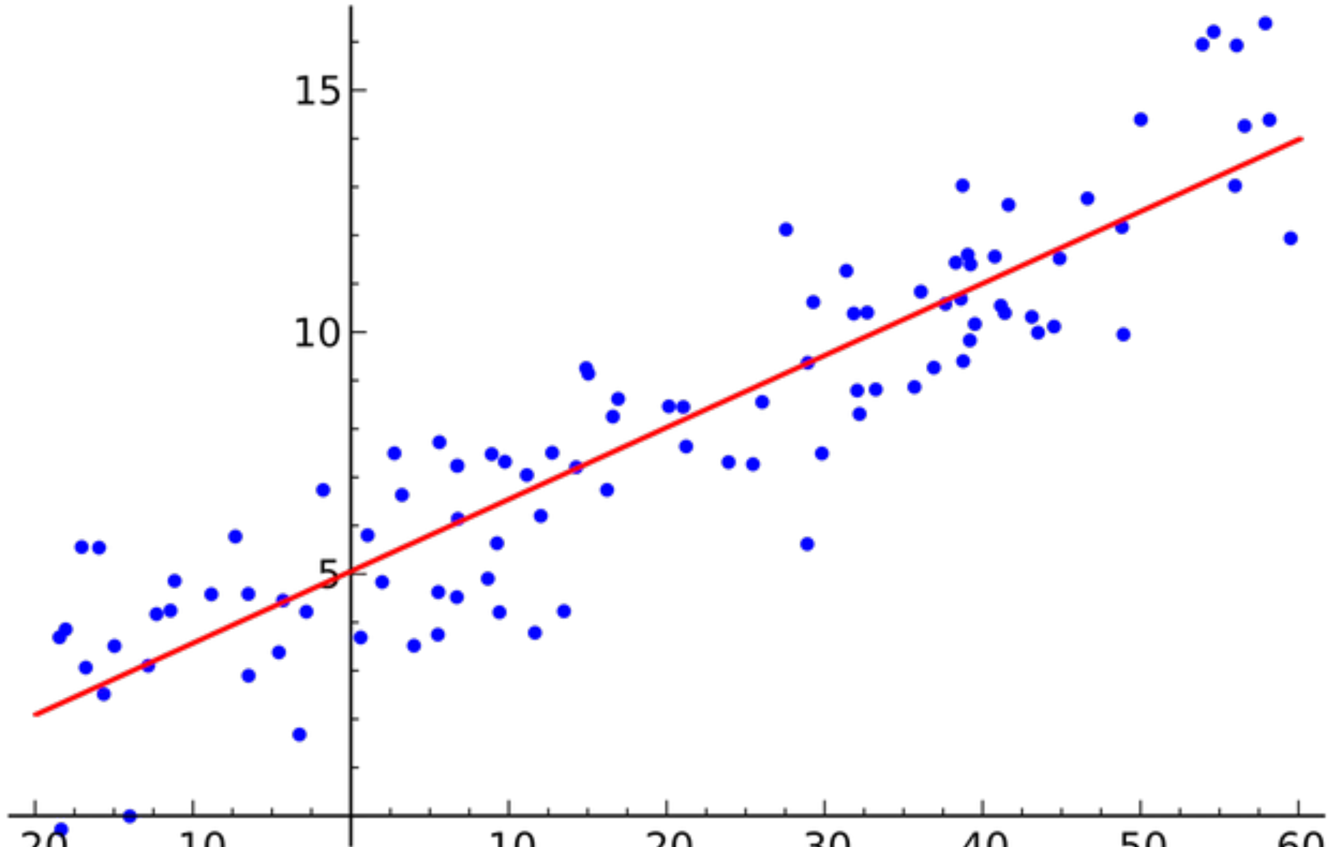
Explainability is understanding the *why* (causation).

DS/ML deals in correlations.

Interpretability is understanding the *how* (correlations).

Correlation is not Causation.

Example



So, what can we do?

Well …

- Run randomized control trials (?)
- Build structural/causal models (?)
- Understand the correlations the best we can …

Great power / responsibility

Keep in mind the human (mind's) tendencies to:

- want causation
- want to compare and contrast
- want *perscriptive* insight

Understanding the correlations in our models

Approach	Tool	Area
Linear models	coefficients	global
Decision trees / Rules	nodes	global
Tree ensemble	feature importance	global
Feature exploration	permutation importance	global
	Partial Dependence Plot (PDP)	single feature* / global
	Indiv. Cond. Expectation (ICE)	single feature / subsample
Surrogate Models	Local Inter. Model-agnostic Exp. (LIME)	multi feature / local
	Shapley values	multi feature / local

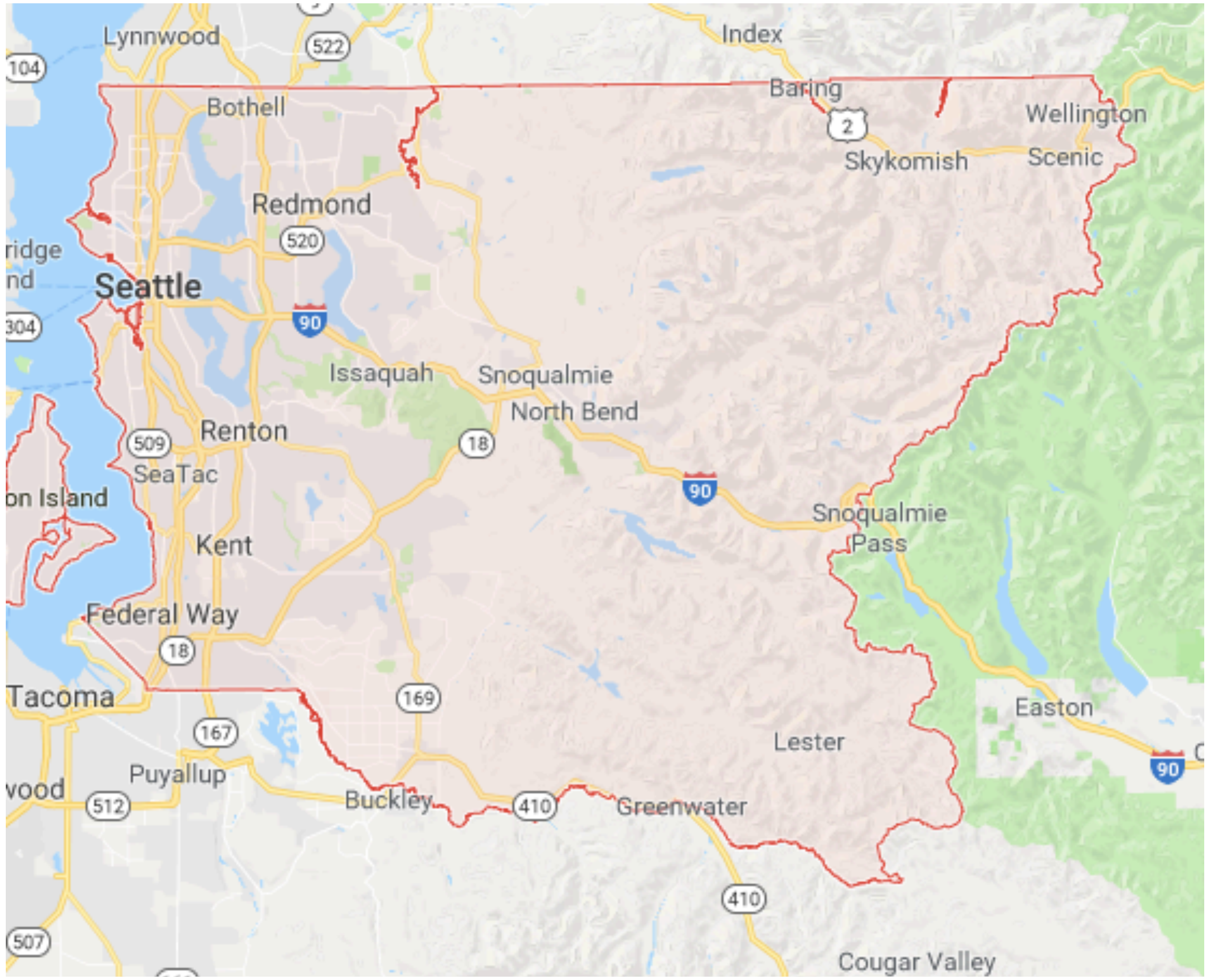
Data: King County Washington home sales

- May 2014 and May 2015
- Kaggle https://www.kaggle.com/harlfoxem/housesalesprediction
- 19 features
- 21,613 observations

Features

id - a notation for a house	date - Date house was sold
price - Price is prediction target	bedrooms - Number of Bedrooms/House
bathrooms - Number of bathrooms/bedrooms	sqft_living - square footage of the home
sqft_lot - square footage of the lot	floors - Total floors (levels) in house
waterfront - House which has a view to a waterfront	view - Has been viewed
condition - How good the condition is (Overall)	grade - overall grade given to the housing unit, based on King County grading system
sqft_above - square footage of house apart from basement	sqft_basement - square footage of the basement
yr_built - Built Year	yr_renovated - Year when house was renovated
zipcode - zip	lat - Latitude coordinate
long - Longitude coordinate	sqft_living - 15Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
sqft_lot15 - lotSize area in 2015(implies-- some renovations)	

King County


Lynnwood
522
Index
104
Bothell
Baring
2
Wellington
Skykomish
Scenic
Redmond
520
ridge
nd
Seattle
304
90
Issaquah
Snoqualmie
North Bend
509
Renton
18
SeaTac
90
on Island
Snoqualmie
Pass
Kent
Federal Way
18
Tacoma
169
167
Easton
Lester
90
Puyallup
vood
512
Buckley
410
Greenwater
410
507
Cougar Valley

Switch to notebook

Summary

- Explainability vs Interpretability
- Feature permutation - Eli5
- Partial Dependece Plot - pdpbox
- Shapley - shap
- IML for R

Resources

A Survey Of Methods For Explaining Black Box Models
Explanation in Artificial Intelligence
Consistent Individualized Feature Attribution for Tree Ensembles
https://christophm.github.io/interpretable-ml-book/

https://github.com/SauceCat/PDPbox
https://github.com/TeamHG-Memex/eli5
https://github.com/marcotcr/lime
https://github.com/slundberg/shap

https://www.kaggle.com/harlfoxem/housesalesprediction

Thank You!

@JulioBarros / E-String.com / Julio@E-String.com

Questions?