A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models [work in progress]



Josua Krause*, Adam Perer**, Enrico Bertini*

Mon, August 20th 2018









"Why Should I Trust You?" Explaining the Predictions of Any Classifier Marco Riberio, Sameer Singh, Carlos Guestrin International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2016)

Instance Explanations



From: pauld@verdix.com (Paul Durbin) Subject: Re: DAVID CORESH IS! GOD! Nntp-Posting-Host: sarge.hq.verdix.com Organization: Verdix Corp Lines: 8



Finding Data Biases



"Why Should I Trust You?" Explaining the Predictions of Any Classifier Marco Riberio, Sameer Singh, Carlos Guestrin International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2016)

Lines: 8



Problem: Inspecting single instances does not scale well



Solution: Aggregating data and explanations







Solution: Aggregating data and explanations









Solution: Aggregating data and explanations

Living Area (numeric)

7











-	\sim
	()
	• • •
	~

Sorted by Importance Living Area (nu...) Diverall Quality...







What is the impact of instance-level explanations?

How do those settings affect the ability to detect biases in the data?

What is the impact of aggregation?

12

Table

Νο
Explanation

Foundation Garage House Style Living Area Mon Poured Attac One sto 1795 Poured Attac One sto 1704 One sto 1704 1704 Cinder Attac One sto 1704 Poured Attac One sto 1700 Poured Attac One sto 1700 Poured Attac One sto 1751 Poured Attac One sto 1561 Poured Attac One sto 1656 Poured Attac One sto 1262 Poured Attac One sto 1262 N/A Attac One and 1362 N/A Detac One and 1774 Brick Attac One and 1077					
Poured Attac One sto 1795 Poured Attac One sto 1704 Cinder Attac One sto 1700 Poured Attac One sto 1700 Poured Attac One sto 1700 Poured Attac One sto 1761 Poured Attac One sto 1752 Poured Attac One sto 1752 Poured Attac. One sto 1752 Poured Attac One sto 1656 Foundation Garage House Style Living Area Mon Cinder Attac One sto 1262 N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077	Foundation	Garage	House Style	Living Area	Mon
PouredAttacOne sto1795PouredAttacOne sto1704CinderAttacOne sto1700PouredAttacOne sto1561PouredAttacOne sto1752PouredAttac.One sto1656PouredAttacOne sto1656PouredAttac.One sto1262FoundationGarageHouse StyleLiving AreaMonCinderAttacOne sto1262N/AAttacOne and1362BrickDetacOne and1774BrickAttacOne and1077				20.0	
PouredAttacOne sto1704CinderAttacOne sto1700PouredAttacOne sto1561PouredAttacOne sto1752PouredAttacOne sto1656PouredAttacOne sto1656FoundationGarageHouse StyleLiving AreaMonCinderAttacOne sto1262N/AAttacOne and1362BrickDetacOne and1774BrickAttacOne and1077	Poured	Attac	One sto	1795	
CinderAttacOne sto1700PouredAttacOne sto1561PouredAttacOne sto1752PouredAttacOne sto1656PouredAttacOne sto1656FoundationGarageHouse StyleLiving AreaMonCinderAttacOne sto1262N/AAttacOne and1362BrickDetacOne and1774	Poured	Attac	One sto	1704	
 Poured Attac One sto 1561 Poured Attac One sto 1752 Poured Attac One sto 1656 Foundation Garage House Style Living Area Mon Cinder Attac One sto 1262 N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077 	Cinder	Attac	One sto	1700	
 Poured Attac One sto 1752 Poured Attac One sto 1656 Foundation Garage House Style Living Area Mon Cinder Attac One sto 1262 N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077 	Poured	Attac	One sto	1561	
Poured Attac: One sto 1656 Foundation Garage House Style Living Area Mon Cinder Attac One sto 1262 N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077	Poured	Attac	One sto	1752	
Foundation Garage House Style Living Area Mon Cinder Attac One sto 1262 N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077					
FoundationGarageHouse StyleLiving AreaMonCinderAttacOne sto1262N/AAttacOne and1362BrickDetacOne and1774BrickAttacOne and1077	Poured	Attac	One sto	1656	
Cinder Attac One sto 1262 N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077	Poured	Attac	One sto	1656	
N/A Attac One and 1362 Brick Detac One and 1774 Brick Attac One and 1077	Poured Foundation	Attac Garage	One sto House Style	1656 Living Area	Mon
Brick Detac One and 1774 Brick Attac One and 1077	Foundation Cinder	Attac Garage Attac	One sto House Style One sto	1656 Living Area 1262	Mon
Brick Attac One and 1077	Foundation Cinder N/A	Attac Garage Attac Attac	One sto House Style One sto One and	1656 Living Area 1262 1362	Mon
	Poured Foundation Cinder N/A Brick	Attac Garage Attac Attac Detac	One sto House Style One sto One and	1656 Living Area 1262 1362 1774	Mon

Overall Quality	Room Count	Foundation	Living Area	
,	J		2010	
Very Good	7	Poured	1795	
Very Good	7	Poured	1704	
Average	6	Cinder	1700	
Excellent	6	Poured	1561	
Excellent	6	Poured	1752	
Verv Good	7	Poured	1656	
Very Good	7	Poured	1656	
Very Good Overall Quality	7 Room Count	Poured Foundation	1656 Living Area	
Very Good Overall Quality Above Avera	7 Room Count 6	Foundation Cinder	1656 Living Area 1262	
Verv Good Overall Quality Above Avera Average	7 Room Count 6 5	Foundation Cinder N/A	1656 Living Area 1262 1362	
Very Good Overall Quality Above Avera Average Good	7 Room Count 6 5 8	Poured Foundation Cinder N/A Brick	1656 Living Area 1262 1362 1774	
Verv Good Overall Quality Above Avera Average Good Average	7 Room Count 6 5 8 5	Poured Foundation Cinder N/A Brick Brick	1656 Living Area 1262 1362 1774 1077	

Explanation

Histogram





13

Table

Νο
Explanation

Foundation	Garage	House Style	Living Area	Mon
			20.0	
Poured	Attac	One sto	1795	
Poured	Attac	One sto	1704	
Cinder	Attac	One sto	1700	
Poured	Attac	One sto	1561	
Poured	Attac	One sto	1752	
Poured	Attac	One sto	1656	
Poured	Attac	One sto	1656	
Poured Foundation	Attac Garage	One sto House Style	1656 Living Area	Mon
Poured Foundation Cinder	Attac Garage Attac	One sto House Style One sto	1656 Living Area 1262	Mon
Poured Foundation Cinder N/A	Attac Garage Attac Attac	One sto House Style One sto One and	1656 Living Area 1262 1362	Mon
Poured Foundation Cinder N/A Brick	Attac Garage Attac Attac Detac	One sto House Style One sto One and	1656 Living Area 1262 1362 1774	Mon

Living Area	Foundation	Room Count	Overall Quality
2010		÷	,
1795	Poured	7	Very Good
1704	Poured	7	Very Good
1700	Cinder	6	Average
1561	Poured	6	Excellent
1752	Poured	6	Excellent
1656	Poured	7	Verv Good
1656	Poured	7	Very Good
1656 Living Area	Foundation	7 Room Count	Very Good Overall Quality
1656 Living Area 1262	Foundation Cinder	7 Room Count 6	Very Good Overall Quality Above Avera
1656 Living Area 1262 1362	Foundation Cinder N/A	7 Room Count 6 5	Very Good Overall Quality Above Avera Average
1656 Living Area 1262 1362 1774	Foundation Cinder N/A Brick	7 Room Count 6 5 8	Very Good Overall Quality Above Avera Average Good
1656 Living Area 1262 1362 1774 1077	Poured Foundation Cinder N/A Brick Brick	7 Room Count 6 5 8 5	Verv Good Overall Quality Above Avera Average Good Average

Explanation

Histogram





1	4

Table

Νο
Explanation

Foundation	Garage	House Style	Living Area	Mon
			20.0	
Poured	Attac	One sto	1795	
Poured	Attac	One sto	1704	
Cinder	Attac	One sto	1700	
Poured	Attac	One sto	1561	
Poured	Attac	One sto	1752	
Poured	Attac	One sto	1656	
Poured	Attac	One sto	1656	
Poured Foundation	Attac Garage	One sto House Style	1656 Living Area	Mon
Poured Foundation Cinder	Attac Garage Attac	One sto House Style One sto	1656 Living Area 1262	Mon
Poured Foundation Cinder N/A	Attac Garage Attac Attac	One sto House Style One sto One and	1656 Living Area 1262 1362	Mon
Poured Foundation Cinder N/A Brick	Attac Garage Attac Attac Detac	One sto House Style One sto One and	1656 Living Area 1262 1362 1774	Mon
Poured Foundation Cinder N/A Brick Brick	Attac Garage Attac Attac Detac Attac	One sto House Style One sto One and One and	1656 Living Area 1262 1362 1774 1077	Mon

Overall Quality	Room Count	Foundation	Living Area	
,	J		20.0	
Very Good	7	Poured	1795	
Very Good	7	Poured	1704	
Average	6	Cinder	1700	
Excellent	6	Poured	1561	
Excollent	6	Poured	1752	
Excellent				_
Very Good	7	Poured	1656	
Very Good	7	Poured	1656	
Very Good Overall Quality	7 Room Count	Poured Foundation	1656 Living Area	
Very Good Overall Quality Above Avera	7 Room Count 6	Poured Foundation Cinder	1656 Living Area 1262	
Very Good Overall Quality Above Avera Average	7 Room Count 6 5	Poured Foundation Cinder N/A	1656 Living Area 1262 1362	
Very Good Overall Quality Above Avera Average Good	7 Room Count 6 5 8	Poured Foundation Cinder N/A Brick	1656 Living Area 1262 1362 1774	
Very Good Overall Quality Above Avera Average Good Average	7 Room Count 6 5 8	Poured Foundation Cinder N/A Brick Brick	1656 Living Area 1262 1362 1774 1077	

Explanation

House Style (cat)	Living Area (num)
Month Sold (num)	Neighborhood (cat)
	hill the filt has a second
Living Area (nu	Foundation (cat)
Living Area (nu	Foundation (cat)
Living Area (nu)	Foundation (cat)
Living Area (nu)	Foundation (cat)

Histogram Foundation (cat) Garage (cat)

House Style (c... Garage (cat)

15

Table

Νο
Explanation

Foundation	Garage	House Style	Living Area	Mon
			20.0	
Poured	Attac	One sto	1795	
Poured	Attac	One sto	1704	
Cinder	Attac	One sto	1700	
Poured	Attac	One sto	1561	
Poured	Attac	One sto	1752	
Poured	Attac	One sto	1656	
Poured	Attac	One sto	1656	
Poured Foundation	Attac Garage	One sto House Style	1656 Living Area	Mon
Poured Foundation Cinder	Attac Garage Attac	One sto House Style One sto	1656 Living Area 1262	Mon
Poured Foundation Cinder N/A	Attac Garage Attac Attac	One sto House Style One sto One and	1656 Living Area 1262 1362	Mon
Poured Foundation Cinder N/A Brick	Attac Garage Attac Attac Detac	One sto House Style One sto One and	1656 Living Area 1262 1362 1774	Mon

Topic distant	TO SO SO STATE	a de la compañía de l	er har and the street
Living Area	Foundation	Room Count	Overall Quality
2010		v	,
1795	Poured	7	Very Good
1704	Poured	7	Very Good
1700	Cinder	6	Average
1561	Poured	6	Excellent
1752	Poured	6	Excellent
1656	Poured	7	Verv Good
Living Area	Foundation	Room Count	Overall Quality
1262	Cinder	6	Above Avera
1262 1362	Cinder N/A	6 5	Above Avera Average
1262 1362 1774	Cinder N/A Brick	6 5 8	Above Avera Average Good
1262 1362 1774 1077	Cinder N/A Brick Brick	6 5 8 5	Above Avera Average Good Average
1262 1362 1774 1077 1040	Cinder N/A Brick Brick Cinder	6 5 8 5 5	Above Avera Average Good Average Average

Explanation

Histogram





16

Table

Foundation	Garage	House Style	Living Area	Mon
	-		2010	
Poured	Attac	One sto	1795	
Poured	Attac	One sto	1704	
Cinder	Attac	One sto	1700	
Poured	Attac	One sto	1561	
Poured	Attac	One sto	1752	
Poured	Attac	One sto	1656	
Foundation	Garage	House Style	Living Area	Mon
Cinder	Attac	One sto	1262	
N/A	Attac	One and	1362	
Brick	Detac	One and	1774	

Overall Quality	Room Count	Foundation	Living Area	
,	v		20.0	
Very Good	7	Poured	1795	
Very Good	7	Poured	1704	
Average	6	Cinder	1700	
Excellent	6	Poured	1561	
Excellent	6	Poured	1752	
Verv Good	7	Poured	1656	
Verv Good	7	Poured	1656	
Verv Good Overall Quality	7 Room Count	Poured Foundation	1656 Living Area	
Very Good Overall Quality Above Avera	7 Room Count 6	Foundation Cinder	1656 Living Area 1262	
Very Good Overall Quality Above Avera Average	7 Room Count 6 5	Foundation Cinder N/A	1656 Living Area 1262 1362	
Very Good Overall Quality Above Avera Average Good	7 Room Count 6 5 8	Poured Foundation Cinder N/A Brick	1656 Living Area 1262 1362 1774	
Verv Good Overall Quality Above Avera Average Good Average	7 Room Count 6 5 8 5	Poured Foundation Cinder N/A Brick Brick	1656 Living Area 1262 1362 1774 1077	

Νο Explanation

Explanation

Histogram











Two Data Sets







Confusi	on Mat	rix:		
	high	low	← Pred.	🚃 Li
high	422	69	491	
low	53	501	554	
1 Label	475	570	1045	
Мос	lel Acci	uracy:	88.325%	💻 Fo
Label	high 🗌	vs.	low	
Pred.	🗌 high	VS.	low	
	Corr.	VS.	Incorr.	



18

Two Data Sets









19

Two Data Sets







20

Questions

Individual models:

- How much do you trust the model?
- Why do you trust or not trust this model?

Summary: Which model do you prefer?

 Do you think the predictions of the model make sense? 5 point Likert scale (Not at all – Very much) How well does the model perform in terms of accuracy? 5 point Likert scale (Not much – Very well) 5 point Likert scale (Not at all – Very much)

Free text answer

Multiple choice and text answer



100 participants

- 4 conditions (25 each): • Table without Explanations (**T/N**) • Table with Explanations (**T/E**) • Histogram without Explanations (H/N) • Histogram with Explanations (H/E)

Random model order

Evaluation metrics:

Model preference (trust) **Bias detection**

Study

- Correctly identified more accurate model



Participants Who Trusted the Correct Model





Participants Who Trusted the Correct Model





Participants Who Trusted the Correct Model





30%

10%

0%-

"It is accurate, yet the predictions do not make much sense. Higher quality houses having a larger amount of low priced houses, percentage-wise? More rooms, area, or stories resulting in lower prices? The logic does not work out."

T/E

"larger houses are valued lower than others which are smaller"

T: Table H: Histogram E: Explanation N: No Explanation



"It has higher accuracy so should be more trustworthy than the other one. However some of the results don't make sense to me. Maybe this is just an atypical property market."









"If the data says it's true, then it's true I suppose and it's more trustworthy than my

20%

10%

0%-

"I feel like the results of [the biased model] where strange even though they where correct according to the dataset."

"I'm drawn to trusting the model which was more accurate even though it didn't entirely make sense to me."

T/E

T: Table H: Histogram E: Explanation N: No Explanation



25% of the participants who found the bias did not change their mind!

H/N

H/E







Participants Who Detected the Bias



28





Participants Who Detected the Bias





Bootstrapped 95% Confidence Intervals







Number of Hovered Cells



Benefitting InfoVis with Visual Difficulties

Jessica Hullman, Student Member, IEEE, Eytan Adar, and Priti Shah Abstract-Many well-cited theories for visualization design state that a visual representation should be optimized for quick and immediate interpretation by a user. Distracting elements like decorative "chartjunk" or extraneous information are avoided so as not to slow comprehension. Yet several recent studies in visualization research provide evidence that non-efficient visual elements may benefit comprehension and recall on the part of users. Similarly, findings from studies related to learning from visual displays in various subfields of psychology suggest that introducing cognitive difficulties to visualization interaction can improve a user's understanding of important information. In this paper, we synthesize empirical results from cross-disciplinary research on visual information representations, providing a counterpoint to efficiency-based design theory with guidelines that describe how visual difficulties can be introduced to benefit comprehension and recall. We identify conditions under which the application of visual difficulties is appropriate based on underlying factors in visualization interaction like active processing and engagement. We characterize effective graph design as a trade-off between efficiency and learning difficulties in order to provide Information Visualization (InfoVis) researchers and practitioners with a framework for organizing explorations of graphs for which d recall are crucial. We identify implications of this view for the design and evaluation of information

Explanations Considered Harmful? User Interactions with Machine Learning Systems

Abstract

It has been suggested that the intelligibility of machine learning system behavior is an important factor in ensuring that users can identify that the system has erred, understand how the system operates and that thereby they are better able to provide appropriate feedback to the machine learning system to improve its accuracy. There has been increasing research into how to make machine learning intelligible to users without a background in AI, and it has been shown that providing explanations of a system's reasoning has many benefits. In this paper we review recent work in this area but also point to instances when explanations might have less desirable effects. Further work is warranted to understand how best to expose the reasoning of machine learning systems to improve their

Author Keywords

Machine learning; explanations; reliability; intelligibility.



Participants Who Detected the Bias





Foundation	Garage	House Style	Living Area	Mon
Poured	Attac	One sto	1795	
Poured	Attac	One sto	1704	
Cinder	Attac	One sto	1700	
Poured	Attac	One sto	1561	
Poured	Attac	One sto	1752	
Poured	Attac	One sto	1656	
Foundation	Garage	House Style	Living Area	Mon
Cinder	Attac	One sto	1262	
N/A	Attac	One and	1362	
Brick	Detac	One and	1774	
Brick	Attac	One and	1077	
Cinder	Detac	One sto	1040	
Cinder	Attac	One sto	1253	

Histograms scale better to larger data sets or more complex errors in the data. In tables you have to extrapolate...



Note that the task was chosen in a way that under all conditions it was possible to find the bias.

33

Lessons Learned

People trust accuracy (too much).

Aggregating instance-level explanations significantly helps detecting biases compared to individual explanations.

Individual instance-level explanations may hurt performance.



More targeted studies to confirm hypotheses

Different results for expert users?

Further Work

35

Foundation	Garage	House Style	Living Area	Mon
Poured	Attac	One sto	1795	
Poured	Attac	One sto	1704	
Cinder	Attac	One sto	1700	
Poured	Attac	One sto	1561	
Poured	Attac	One sto	1752	
Poured	Attac	One sto	1656	
Poured	Attac	One sto	1656	
Poured Foundation	Attac Garage	One sto House Style	1656 Living Area	Mon
Poured Foundation Cinder	Attac Garage Attac	One sto House Style One sto	1656 Living Area 1262	Mon
Poured Foundation Cinder N/A	Attac Garage Attac Attac	One sto House Style One sto One and	1656 Living Area 1262 1362	Mon
Poured Foundation Cinder N/A Brick	Attac Garage Attac Attac Detac	One sto House Style One sto One and One and	1656 Living Area 1262 1362 1774	Mon

A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models [work in progress]

Josua Krause*, Adam Perer**, Enrico Bertini*

Thank You!









