Runtime operation count can serve as a proxy metric for local interpretability.

Assessing the Local Interpretability of Machine Learning Models

Dylan Slack, Sorelle Friedler, Carlos Scheidegger, Chitradeep Dutta Roy

INTRODUCTION

- How do we provide user grounded metrics for motions of model interpretability?
- We focus on *simulatibility* (ability to trace computation of input) and *"what if"* local explainability (determine local changes on input).

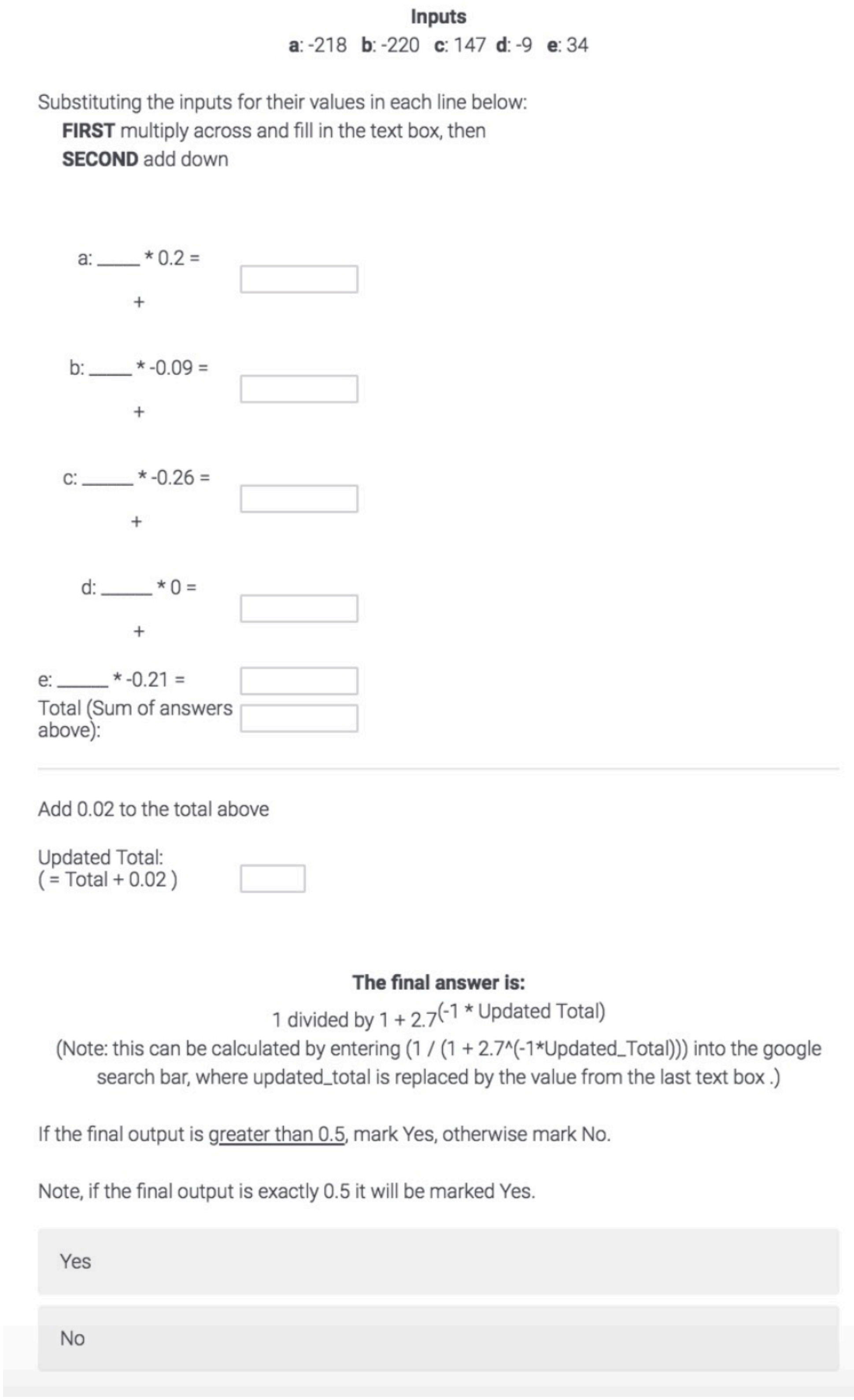
METHODS

- We assess *runtime operation count* as a proxy metric for our proposed notions of interpretability in *decision trees, logistic regression, and (small) feedforward neural networks* using a 1,000 person user study.

EXACT BINOMIAL TEST WRT RANDOM GUESSING

		Simulatability	"What If" Local Explainability
DT	Correct	717 / 930	719 / 930
	p-Value	5.9×10^{-63}	5.16×10^{-64}
	95% CI	$[0.73, 0.81]$	$[0.73, 0.82]$
LR	Correct	592 / 930	579 / 930
	p-Value	1.94×10^{-15}	2.07×10^{-12}
	95% CI	$[0.59, 0.69]$	$[0.57, 0.67]$
NN	Correct	556 / 930	499 / 930
	p-Value	7.34×5.5^{-8}	0.78
	95% CI	$[0.55, 0.65]$	$[0.49, 0.59]$

EXAMPLE LOGISTIC REGRESSION SURVEY QUESTION



Inputs

a: -218 **b**: -220 **c**: 147 **d**: -9 **e**: 34

Substituting the inputs for their values in each line below:
FIRST multiply across and fill in the text box, then
SECOND add down

a: _____ * 0.2 =

+

b: _____ * -0.09 =

+

c: _____ * -0.26 =

+

d: _____ * 0 =

+

e: _____ * -0.21 =

Total (Sum of answers above):

Add 0.02 to the total above

Updated Total:
(= Total + 0.02)

The final answer is:

1 divided by 1 + 2.7$^{(-1 * \text{Updated Total})}$

(Note: this can be calculated by entering (1 / (1 + 2.7^(-1*Updated_Total))) into the google search bar, where updated_total is replaced by the value from the last text box .)

If the final output is greater than 0.5, mark Yes, otherwise mark No.

Note, if the final output is exactly 0.5 it will be marked Yes.

MODEL COMPARISON USING FISHER EXACT TEST

Relative Simulatability:

Contingency Table	DT > NN		DT > LR		LR > NN	
Correct	717	556	717	592	592	556
Incorrect	213	374	213	338	338	374
p-value, 95% CI	1.5×10^{-14}	$[1.69, \infty]$	3.7×10^{-9}	$[1.43, \infty]$	1.3	$[0.90, \infty]$

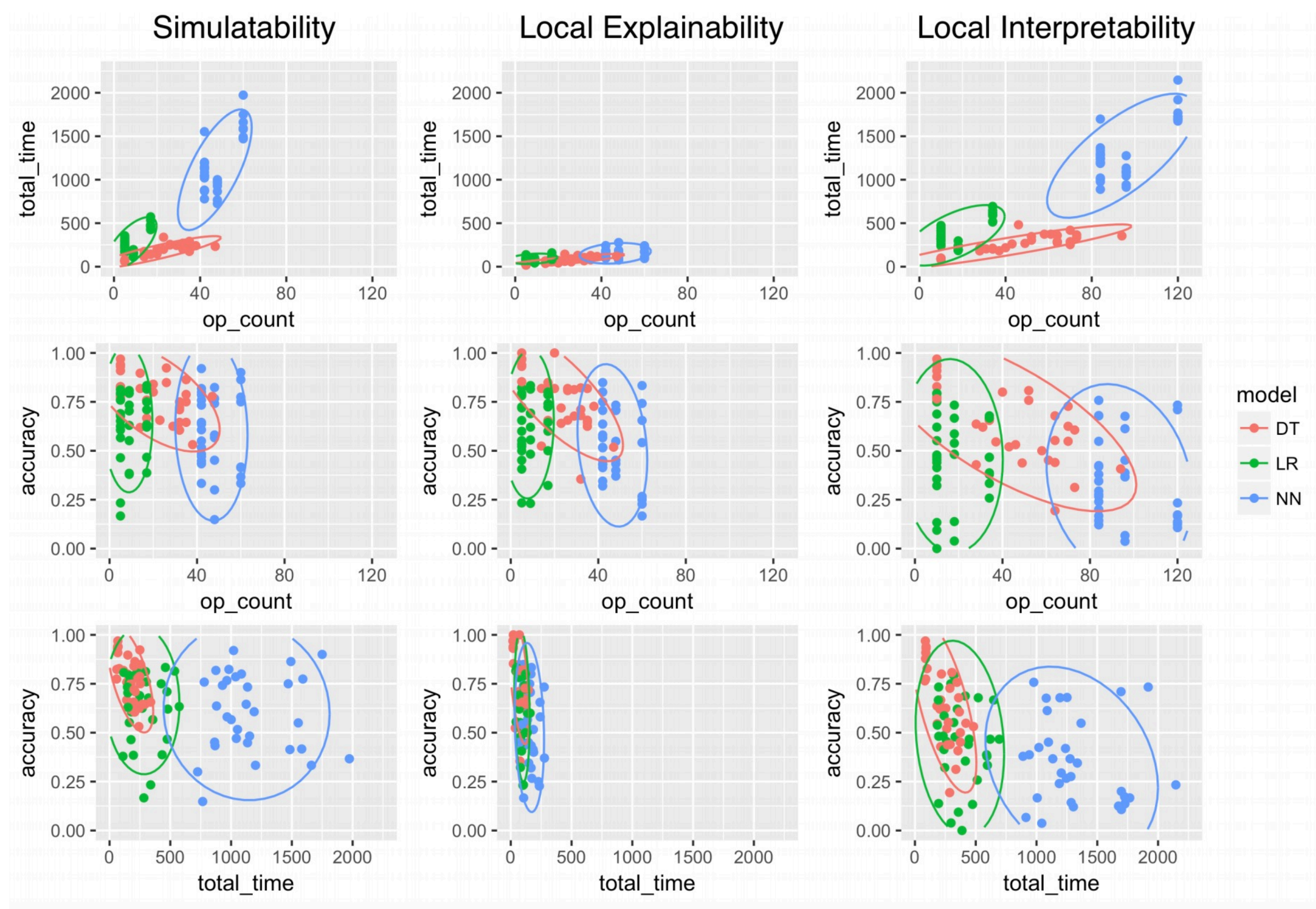
Relative "What If" Local Explainability:

Contingency Table	DT > NN		DT > LR		LR > NN	
Correct	719	499	719	579	579	499
Incorrect	211	431	211	351	351	431
p-value, 95% CI	7.3×10^{-26}	$[2.20, \infty]$	2.6×10^{-11}	$[1.54, \infty]$	2.9×10^{-3}	$[1.09, \infty]$

Relative Local Interpretability:

Contingency Table	DT > NN		DT > LR		LR > NN	
Correct	594	337	594	425	425	337
Incorrect	336	593	336	505	505	593
p-value, 95% CI	9.3×10^{-32}	$[2.36, \infty]$	5.9×10^{-14}	$[1.60, \infty]$	5.7×10^{-4}	$[1.13, \infty]$

RELATIONSHIP BETWEEN OPERATION COUNT, TIME, AND ACCURACY



Scan code for arXiv.