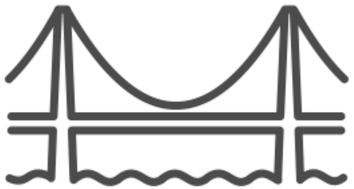
Bridge Improvement Cost



DSO 530

Project Presentation

Group Member: Cheng Chen, Xi Jiang, Yu Zhang, Yuting Zheng, Hangyu Zhou

Problem of Interest

Bridge plays a significant role in infrastructural system. Budget is limited for each year and therefore estimation on cost of each improvement is necessary. If we could build an algorithm to approximately estimate the cost, we can save tons of money for the government.

We want to come up with an ***decent estimation*** of bridge improvement.

Background

Investment Estimation Stage	Allowable Deviation
1. Project Planning Stage	>=30%
2. Project Proposal Stage	<=30%
3. Preliminary Feasibility Study Stage	<=20%
4. Detailed Feasibility Study	<=10%

Data Description

data_NBI:https://www.kaggle.com/broach/build-bridges-not-walls

Data Source:
United States Federal Highway Administration (FHWA)

Content:
Information about bridges on federal land.

Data Size:
607,070 observations
x
135 variables

Variables:
location, owner, bypass length, year built, lanes, average daily traffic, total project cost, etc.

Data Preparation

Variable Selection

- Reduce predictors from 135 to 56 (categorical & numerical)

Variable Creation and Adjustment

- Age of Bridge=Year of Improvement Cost Estimate-Year Built
- Adjusted Project Cost=Project Cost adjusted by Producer Price Index (PPI)

Scope Reduction

- Year of Improvement Cost Estimate >= 2009
- Total Project Cost >= $40,000

Training & Testing Data Sets

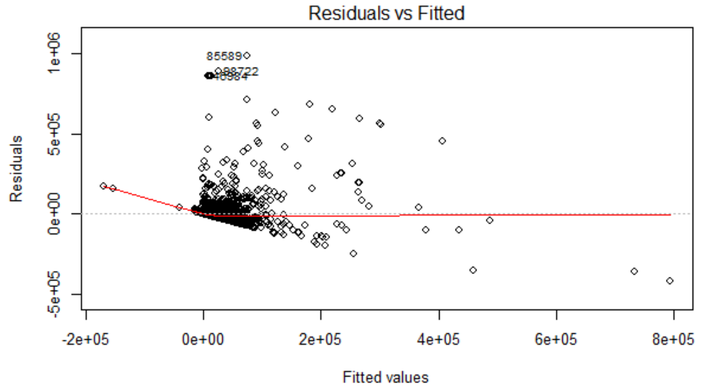
- set.seed(1)
- Train:Test=7:3

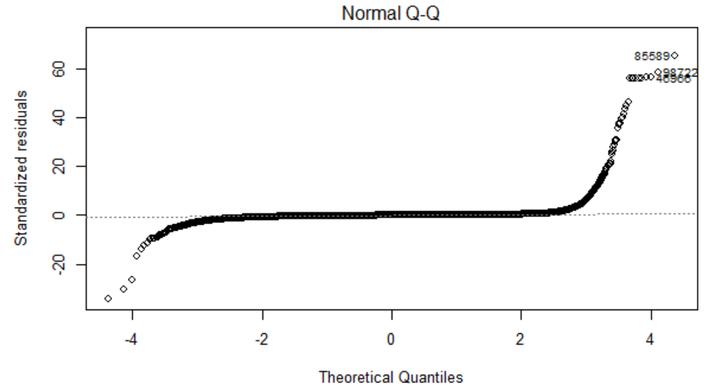
Method

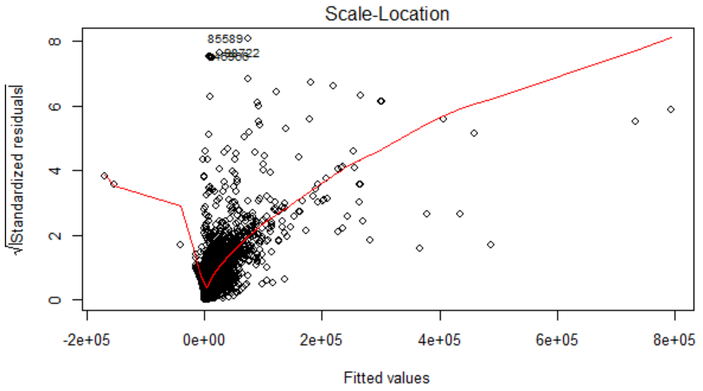
Regression

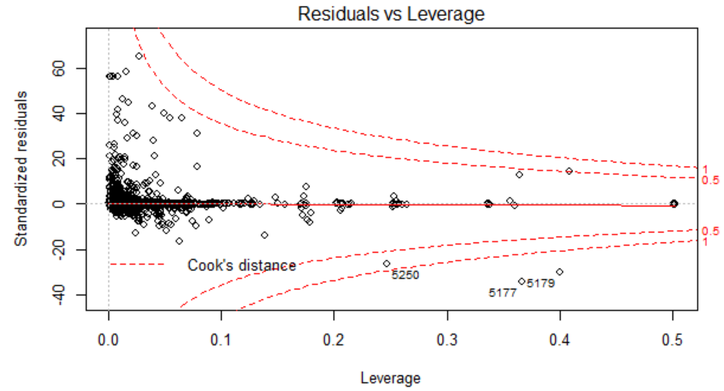
	Linear	Lasso	Ridge
Test MSE	**1.94E+08**	**1.93E+08**	**1.95E+08**

- The true relationship between response and predictors is far from LINEAR



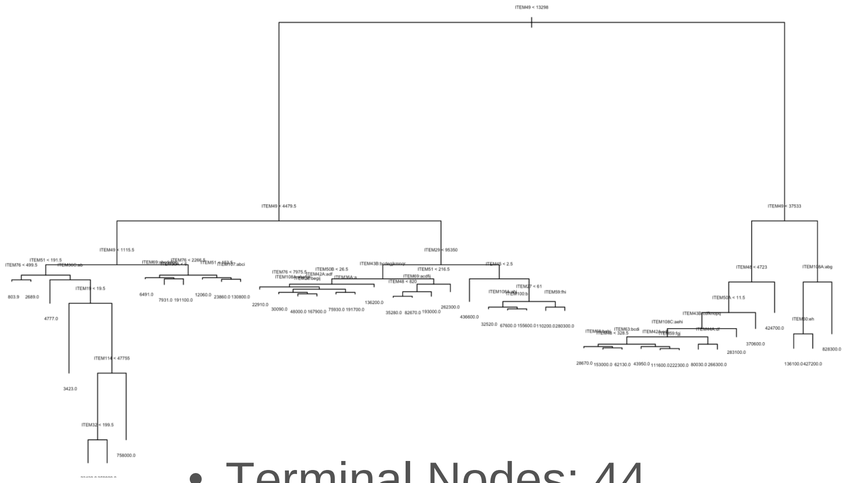




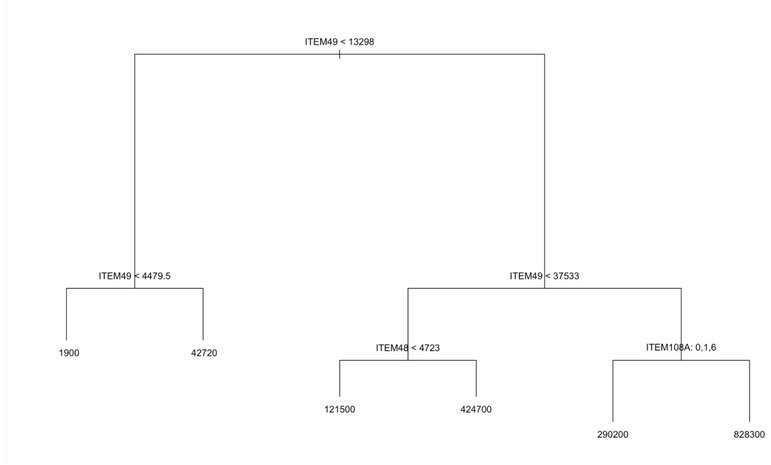


Method

Decision Trees

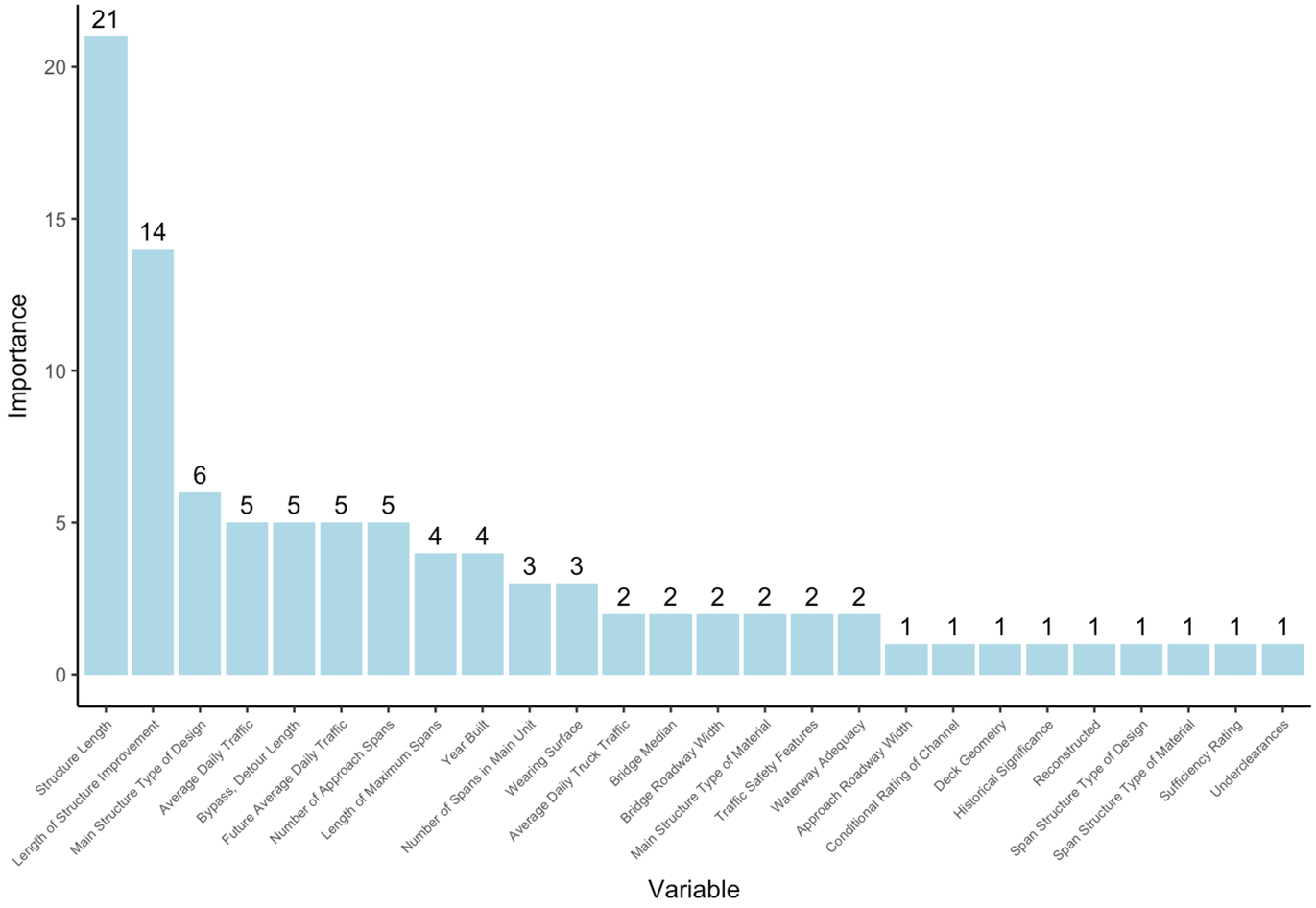


- Terminal Nodes: 44
- The test MSE: 236976719



- Terminal Node:6
- The test MSE: 194870062

Variable Importance



Method

Several Nonlinear Regression models

So far we know:
1. If there is a mapping between independent variables and improvement cost, it is nonlinear
2. Decision trees is far from accuracy

We then tried several other nonlinear approaches which normally could generate better results than decision trees. Compare the result of each method by MSE (Mean Square Error):

Neural Network Regression, MSE: 1.43E+8
Gradient Boost Model, MSE: 1.44E+8
Random Forest, MSE: 1.58E+8
(Decision Trees, MSE: 1.95E+8)

Key problem with regression:
---- It predicts negative improvement cost

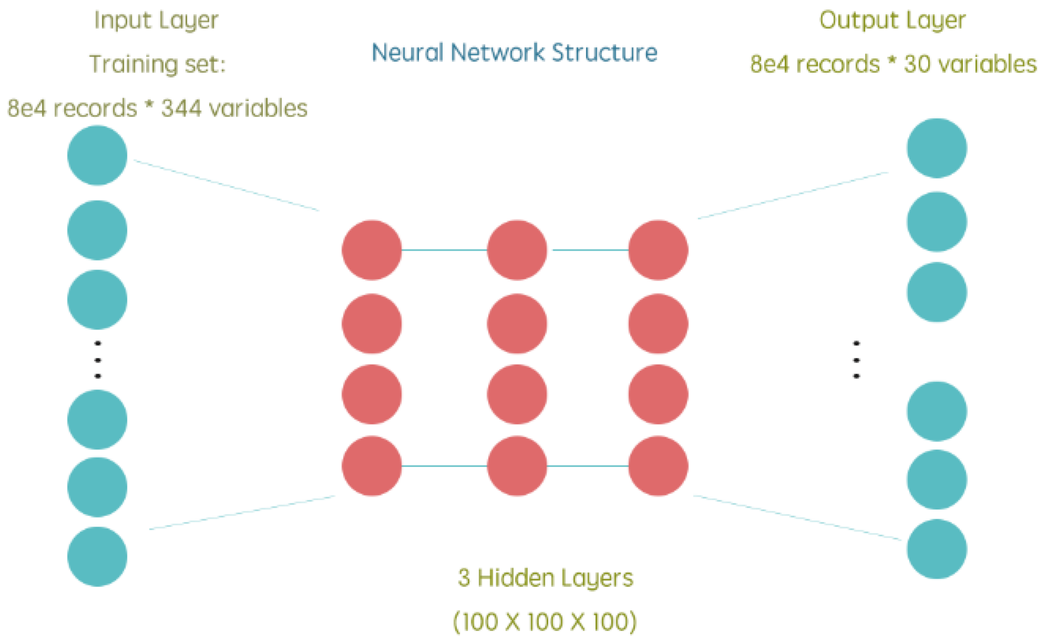
Solution:
1. Algebra transformation, force the predicted value to be positive
---- Problem: MSE would get even worse

2. Use **classification**, while sacrificing some accuracy
---- Advantage: we really do not need completely accuracy in this case
---- Approach: cr

Range of total project cost in each class i: ement cost (Max: 1061115, Min: 44)

$$[44 * 1.4^{\{i-1\}}, 44 * 1.4^{i})$$

Artificial Neural Network - Classification

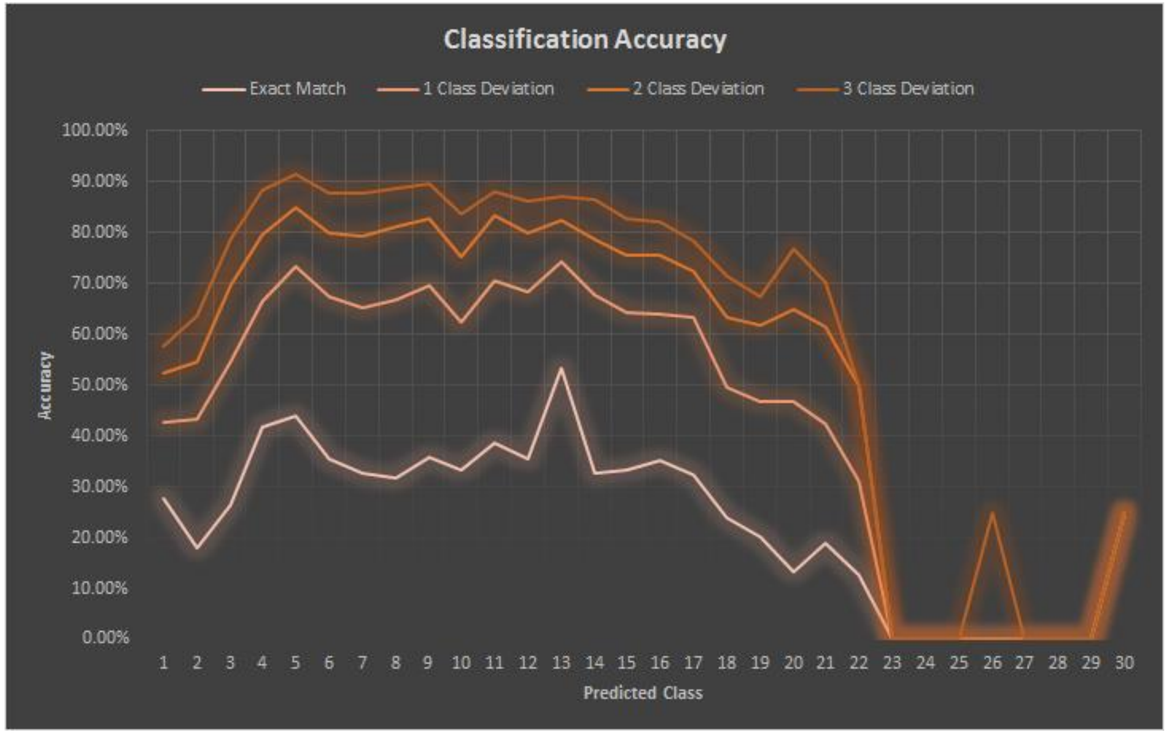


Variability from True Value within Different Classification Interval:

Classification Interval	Variability	Comment
Exact Match	8.30%	Very Satisfactory
1-Class Deviation	40.00%	Satisfactory
2-Class Deviation	100.00%	Acceptable
3-Class Deviation	180.00%	Critical Point
4-Class Deviation	300.00%	Worse than expert's intuitive estimation

Classification Result

True\Predict	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	177	89	61	67	71	47	50	47	36	21	13	8	10	13	2	1	0	1	0	0	2	0	0	0	0	0	0	0	0	0
2	96	155	145	97	84	72	76	31	27	25	11	13	4	11	7	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0
3	61	131	318	210	170	110	114	60	32	31	15	11	8	3	9	6	2	1	1	0	0	0	0	0	0	0	0	0	0	0
4	35	96	198	982	518	205	189	101	48	55	29	22	19	8	3	3	3	2	0	0	0	0	0	0	0	0	0	0	0	0
5	52	80	119	370	1495	536	302	136	67	57	30	29	21	9	6	8	4	2	1	0	0	0	0	0	0	0	0	0	0	0
6	48	69	108	211	480	1157	752	263	128	79	66	35	31	14	11	7	6	4	3	1	0	0	0	0	0	0	0	0	0	0
7	37	60	73	142	219	502	1313	659	215	122	50	47	35	17	10	6	6	4	2	0	0	0	0	0	0	0	0	0	0	0
8	34	42	52	87	138	204	557	1070	565	181	75	44	43	16	11	6	7	2	1	0	1	0	0	1	0	0	0	0	0	0
9	24	36	38	56	73	144	272	517	1088	403	142	78	50	14	13	8	9	4	1	2	2	1	1	0	0	0	0	0	0	0
10	25	35	23	42	59	90	144	231	460	856	403	144	64	35	13	18	13	10	1	0	1	0	1	0	1	1	0	0	0	1
11	21	17	22	20	35	63	75	115	191	343	813	345	106	61	37	25	20	11	2	0	0	0	0	0	0	1	0	0	0	1
12	11	11	16	23	11	45	52	55	78	151	267	653	223	84	39	21	12	19	2	1	0	0	0	0	0	0	0	0	0	0
13	9	19	10	25	30	49	66	47	57	95	125	265	1072	229	64	33	24	19	7	0	0	0	1	0	0	0	0	0	0	0
14	5	3	9	7	5	19	32	22	19	62	27	69	202	406	135	47	24	14	7	4	1	0	0	0	0	0	0	0	0	0
15	1	4	7	7	3	9	9	7	14	33	9	36	52	206	261	121	28	22	11	2	4	1	0	0	0	0	0	0	0	0
16	2	2	3	2	2	6	8	5	5	21	13	18	32	51	108	235	86	29	6	1	7	1	0	0	0	0	0	0	0	0
17	0	2	3	1	0	1	2	1	4	16	4	13	12	36	23	71	166	45	14	4	3	0	0	0	0	0	0	0	0	0
18	0	3	0	0	1	0	5	1	2	7	5	8	12	11	18	30	72	93	23	7	6	4	0	0	0	0	0	0	0	0
19	0	0	0	1	0	1	2	1	2	3	2	7	5	8	10	11	18	56	36	11	18	0	0	0	0	0	0	0	0	0
20	1	4	0	0	0	0	1	0	0	1	0	1	6	4	1	3	7	25	24	8	18	3	0	0	0	0	0	0	0	0
21	0	2	1	0	0	0	1	1	0	2	2	0	1	2	2	5	2	9	13	9	30	2	0	0	0	0	0	0	0	0
22	0	0	1	0	0	0	1	1	0	0	0	1	2	1	1	0	0	7	4	4	19	2	0	0	0	0	0	0	0	0
23	0	0	1	0	0	0	1	0	2	0	0	0	0	0	0	0	2	2	6	3	12	1	0	0	0	1	0	0	0	1
24	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1	0	5	5	0	8	0	0	0	0	0	0	0	0	0
25	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	2	0	15	0	0	0	0	0	0	0	0	0
26	0	1	0	1	0	0	2	0	0	1	0	0	0	0	0	0	0	1	2	1	3	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	1	3	0	0	2	0	0	0	0	0	0	0	1	3	0	2	0	0	0	0	0	0	0	0	0
28	0	3	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1



Range of total project cost in each class i:

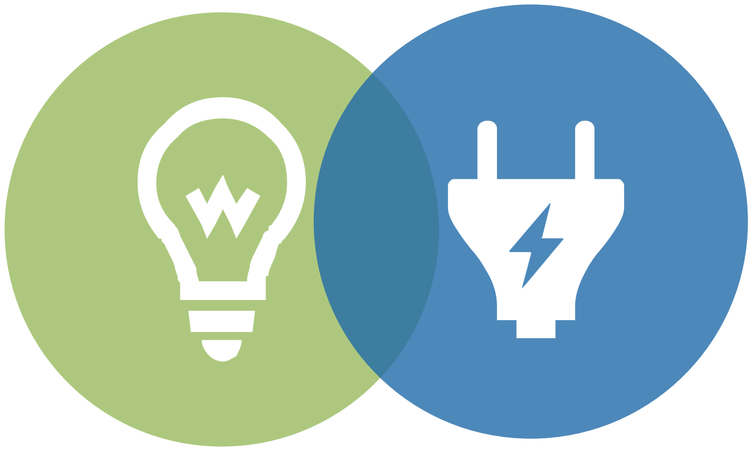
$$[44 * 1.4^{\{i-1\}}, 44 * 1.4^{i})$$

Insights & Limitation

Insights

- Our model performs well for the bridges from Class 3 to 17, whose bridge improvement costs are between $120 thousand to $19 million.
- If given better variables and quality of data, we have confidence to improve the results.
- We have confidence to meet the required accuracy of investment estimation in the **Project Planning Stage** and **Project Proposal Stage**

----Save Cost & Instant Estimation



Limitation

- The observations in the top quartile are so limited for the model to learn the data.
- In the original data, some important metrics have too many missing values and are inaccurately recorded.
- Poor reliability of dependent variable.
  - 1. Variable itself is an estimated value.
    (No evidence of accuracy is provided)
  - 2. Definition of improvement is not clear.
    (Long-term, Short-term? Overall, Local?)
- The model needs more insightful variables to generate lower bias. Otherwise, increasing data volume will not help much.

Recommendation

We seek for recommendations from engineering experts...

Ying Li

Vice Chief Engineering at Bridge Engineering Design Institute, TJAD

"The idea is quite new in this traditional industry.
Adding additional features such as price fluctuation on steel and concrete might be better. "

Yinghui Dong

Civil Engineering Graduate Student at Georgia Institute of Technology

"Government data makes things harder.
The result obviously demonstrates a certain relation between input and output.
But I believe there is discrepancy between bridge assessment of each state, or even each record."

Thank You!