Supplementary Material of 'The Effects of Mismatched Train and Test Data Cleaning Pipelines on Regression Models: Lessons for Practice'

No Author Given

No Institute Given

A Cleaning Setups - Airbnb Data

Setup number	mv_repair	outlier_detection	outlier_repair	duplicate_repair
0	delete	none	NA	NA
1	delete	none	NA	key_val
2	delete	SD	mean	NA
3	delete	SD	mean	key_val
4	delete	SD	median	NA
5	delete	SD	median	key_val
6	delete	SD	mode	NA
7	delete	SD	mode	key_val
8	delete	IQR	mean	NA
9	delete	IQR	mean	key_val
10	delete	IQR	median	NA
11	delete	IQR	median	key_val
12	delete	IQR	mode	NA
13	delete	IQR	mode	key_val
14	mean-mode	none	NA	NA
15	mean-mode	none	NA	key_val
16	mean-mode	SD	mean	NA
17	mean-mode	SD	mean	key_val
18	mean-mode	SD	median	NA
19	mean-mode	SD	median	key_val
20	mean-mode	SD	mode	NA
21	mean-mode	SD	mode	key_val
22	mean-mode	IQR	mean	NA
23	mean-mode	IQR	mean	key_val
24	mean-mode	IQR	median	NA
25	mean-mode	IQR	median	key_val
26	mean-mode	IQR	mode	NA
27	mean-mode	IQR	mode	key_val

28	median-mode	none	NA	NA
29	median-mode	none	NA	key_val
30	median-mode	SD	mean	NA
31	median-mode	SD	mean	key_val
32	median-mode	SD	median	NA
33	median-mode	SD	median	key_val
34	median-mode	SD	mode	NA
35	median-mode	SD	mode	key_val
36	median-mode	IQR	mean	NA
37	median-mode	IQR	mean	key_val
38	median-mode	IQR	median	NA
39	median-mode	IQR	median	key_val
40	median-mode	IQR	mode	NA
41	median-mode	IQR	mode	key_val
42	mode-mode	none	NA	NA
43	mode-mode	none	NA	key_val
44	mode-mode	SD	mean	NA
45	mode-mode	SD	mean	key_val
46	mode-mode	SD	median	NA
47	mode-mode	SD	median	key_val
48	mode-mode	SD	mode	NA
49	mode-mode	SD	mode	key_val
50	mode-mode	IQR	mean	NA
51	mode-mode	IQR	mean	key_val
52	mode-mode	IQR	median	NA
53	mode-mode	IQR	median	key_val
54	mode-mode	IQR	mode	NA
55	mode-mode	IQR	mode	key_val
56	mean-dummy	none	NA	NA
57	mean-dummy	none	NA	key_val
58	mean-dummy	SD	mean	NA
59	mean-dummy	SD	mean	key_val
60	mean-dummy	SD	median	NA
61	mean-dummy	SD	median	key_val
62	mean-dummy	SD	mode	NA
63	mean-dummy	SD	mode	key_val
64	mean-dummy	IQR	mean	NA
65	mean-dummy	IQR	mean	key_val
66	mean-dummy	IQR	median	NA
67	mean-dummy	IQR	median	key_val
68	mean-dummy	IQR	mode	NA
69	mean-dummy	IQR	mode	key_val
70	median-dummy	none	NA	NA
71	median-dummy	none	NA	key_val
72	median-dummy	SD	mean	NA

73	median-dummy	SD	mean	key_val
74	median-dummy	SD	median	NA
75	median-dummy	SD	median	key_val
76	median-dummy	SD	mode	NA
77	median-dummy	SD	mode	key_val
78	median-dummy	IQR	mean	NA
79	median-dummy	IQR	mean	key_val
80	median-dummy	IQR	median	NA
81	median-dummy	IQR	median	key_val
82	median-dummy	IQR	mode	NA
83	median-dummy	IQR	mode	key_val
84	mode-dummy	none	NA	NA
85	mode-dummy	none	NA	key_val
86	mode-dummy	SD	mean	NA
87	mode-dummy	SD	mean	key_val
88	mode-dummy	SD	median	NA
89	mode-dummy	SD	median	key_val
90	mode-dummy	SD	mode	NA
91	mode-dummy	SD	mode	key_val
92	mode-dummy	IQR	mean	NA
93	mode-dummy	IQR	mean	key_val
94	mode-dummy	IQR	median	NA
95	mode-dummy	IQR	median	key_val
96	mode-dummy	IQR	mode	NA
97	mode-dummy	IQR	mode	key_val

Table 1: Numbering of cleaning pipelines for Airbnb dataset.

B Airbnb Data Results Snapshot

		Test cleaning										
		0	1	2	3	4	...	93	94	95	96	97
Train cleaning	0	0.6356	0.6327	0.6332	0.6304	0.6335	...	0.5812	0.5841	0.5824	0.583	0.5813
		(0.0099)	(0.0108)	(0.0101)	(0.0112)	(0.0101)		(0.0137)	(0.0132)	(0.0137)	(0.0134)	(0.0139)
	1	0.6352	0.6325	0.633	0.6303	0.6331	...	0.582	0.585	0.5833	0.5838	0.5821
		(0.0101)	(0.0109)	(0.0101)	(0.0112)	(0.0102)		(0.0134)	(0.0131)	(0.0135)	(0.0134)	(0.0138)
	2	0.6354	0.6325	0.6334	0.6304	0.6332	...	0.5823	0.5845	0.5828	0.5833	0.5816
		(0.0094)	(0.0101)	(0.0095)	(0.0104)	(0.0096)		(0.0124)	(0.0123)	(0.0127)	(0.0122)	(0.0127)
	3	0.6345	0.6318	0.6325	0.6297	0.6324	...	0.5835	0.5858	0.584	0.5846	0.5829
		(0.0099)	(0.0104)	(0.0099)	(0.0105)	(0.0099)		(0.012)	(0.0122)	(0.0122)	(0.0124)	(0.0124)
	4	0.6351	0.6323	0.6329	0.6299	0.633	...	0.5821	0.5849	0.5833	0.5839	0.5823
		(0.0082)	(0.0094)	(0.0083)	(0.0096)	(0.0083)		(0.0129)	(0.0124)	(0.0131)	(0.0125)	(0.0132)
	...	...	...	...	...	...	...	...	...	...	...	...
	93	0.6231	0.6206	0.621	0.6185	0.6211	...	0.5887	0.5916	0.589	0.5904	0.5878
		(0.0106)	(0.0113)	(0.0106)	(0.0114)	(0.0107)		(0.0124)	(0.0123)	(0.0126)	(0.0125)	(0.0128)
	94	0.6252	0.6225	0.6228	0.62	0.6231	...	0.5895	0.5932	0.5905	0.5921	0.5894
		(0.0108)	(0.0114)	(0.0107)	(0.0114)	(0.0107)		(0.0128)	(0.0126)	(0.0128)	(0.0128)	(0.0129)
	95	0.6235	0.621	0.6211	0.6185	0.6213	...	0.5883	0.5918	0.5893	0.5908	0.5882
		(0.0104)	(0.0113)	(0.0104)	(0.0114)	(0.0103)		(0.0131)	(0.013)	(0.0132)	(0.013)	(0.0132)
	96	0.6237	0.6208	0.6212	0.6183	0.6215	...	0.5879	0.5919	0.589	0.5913	0.5884
		(0.0108)	(0.0115)	(0.0109)	(0.0117)	(0.0109)		(0.0117)	(0.0116)	(0.0118)	(0.0117)	(0.0119)
	97	0.623	0.6204	0.6207	0.6181	0.6209	...	0.5881	0.5918	0.5892	0.5913	0.5886
		(0.0106)	(0.0113)	(0.0107)	(0.0115)	(0.0107)		(0.0127)	(0.0127)	(0.0127)	(0.0129)	(0.0129)

Table 2: Snapshot of performance table for GBR models on Airbnb data. Average R^2 score (standard deviation) of a given training cleaning pipeline (row) on a test cleaning pipeline (column) is calculated over different train-test splits.