Technical Reports of CleanML

The CleanML Team

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Technical Reports of CleanML

***Abstract*—Here we report additional experimental details and results that are left out in the CleanML paper due to space limitation.**

I. HoloClean Experiment Setups and Results

- *: We do not count a label as an attribute. A numerical variable is a variable where a measurement or number has a numerical meaning. It is different from a categorical variable expressed as a number.
- tba: For marketing, if we do some conversion with the text representations, the ratio would be 3/13. These three attributes could be converted into numeric representations: age, age, person, person under 18.
- For sensor, hour and minute are used to identify points of measure from a sensor; these two variables are hence regarded as categorical.

			imputation_acc				
error_type	dataset	ratio_numeric_attributes*	epoch1	epoch2	epoch3	epoch4	epoch5
outliers	Airbnb	97.44%	95.27%	95.27%	95.27%	95.27%	95.27%
	Credit	100.00%	46.26%	46.26%	46.26%	46.26%	46.26%
	EEG	100.00%	11.09%	11.09%	11.09%	11.09%	11.09%
	Sensor	62.50%	42.96%	42.96%	42.96%	42.96%	42.96%
missing_values	Airbnb	97.44%	93.23%	93.23%	93.23%	93.23%	93.23%
	Credit	100.00%	62.73%	62.73%	62.73%	62.73%	62.73%
	Marketing	0.00%	91.04%	91.04%	91.04%	91.04%	91.04%
	Titanic	44.44%	81.23%	81.17%	81.17%	81.17%	81.17%
	USCensus	35.71%	98.31%	98.31%	98.31%	98.31%	98.31%

II. SQL Query Tables

TABLE I. Q1(E=Missing Values)

R	P	S	N
R1	46.94% (115)	32.65% (80)	20.41% (50)
R2	57.14% (20)	25.71% (9)	17.14% (6)
R3	20.00% (1)	80.00% (4)	0.00% (0)

TABLE II. Q2(E=Missing Values)

R	Scenario	P	S	N
R1	CD	46.94% (115)	32.65% (80)	20.41% (50)
R2	CD	57.14% (20)	25.71% (9)	17.14% (6)
R3	CD	20.00% (1)	80.00% (4)	0.00% (0)

TABLE III. Q3(E=Missing Values)

R	Model	P	S	N
R1	AdaBoost	62.86% (22)	20.00% (7)	17.14% (6)
	Decision Tree	45.71% (16)	40.00% (14)	14.29% (5)
	Gaussian Naive Bayes	14.29% (5)	42.86% (15)	42.86% (15)
	KNN	45.71% (16)	40.00% (14)	14.29% (5)
	Logistic Regression	62.86% (22)	11.43% (4)	25.71% (9)
	Random Forest	62.86% (22)	22.86% (8)	14.29% (5)
	XGBoost	62.86% (22)	20.00% (7)	17.14% (6)

TABLE IV. Q4.1(E=Missing Values)

R	P	S	N
R1	46.94% (115)	32.65% (80)	20.41% (50)
R2	57.14% (20)	25.71% (9)	17.14% (6)

TABLE V. Q4.2(E=Missing Values)

R	Imputation Mehtod	P	S	N
R1	HoloClean	45.71% (16)	34.29% (12)	20.00% (7)
	Mean Dummy	42.86% (15)	28.57% (10)	28.57% (10)
	Mean Mode	60.00% (21)	34.29% (12)	5.71% (2)
	Median Dummy	42.86% (15)	28.57% (10)	28.57% (10)
	Median Mode	57.14% (20)	37.14% (13)	5.71% (2)
	Mode Dummy	42.86% (15)	28.57% (10)	28.57% (10)
	Mode Mode	37.14% (13)	37.14% (13)	25.71% (9)
R2	HoloClean	60.00% (3)	20.00% (1)	20.00% (1)
	Mean Dummy	60.00% (3)	20.00% (1)	20.00% (1)
	Mean Mode	80.00% (4)	20.00% (1)	0.00% (0)
	Median Dummy	40.00% (2)	40.00% (2)	20.00% (1)
	Median Mode	80.00% (4)	20.00% (1)	0.00% (0)
	Mode Dummy	40.00% (2)	40.00% (2)	20.00% (1)
	Mode Mode	40.00% (2)	20.00% (1)	40.00% (2)

TABLE VI. Q5(E=Missing Values)

R	Dataset	P	S	N
R1	Airbnb	12.24% (6)	81.63% (40)	6.12% (3)
	Credit	42.86% (21)	51.02% (25)	6.12% (3)
	Marketing	48.98% (24)	8.16% (4)	42.86% (21)
	Titanic	65.31% (32)	0.00% (0)	34.69% (17)
	USCensus	85.71% (42)	0.00% (0)	14.29% (7)
R2	Airbnb	0.00% (0)	100.00% (7)	0.00% (0)
	Credit	57.14% (4)	28.57% (2)	14.29% (1)
	Marketing	57.14% (4)	0.00% (0)	42.86% (3)
	Titanic	71.43% (5)	0.00% (0)	28.57% (2)
	USCensus	100.00% (7)	0.00% (0)	0.00% (0)
R3	Airbnb	0.00% (0)	100.00% (1)	0.00% (0)
	Credit	0.00% (0)	100.00% (1)	0.00% (0)
	Marketing	0.00% (0)	100.00% (1)	0.00% (0)
	Titanic	0.00% (0)	100.00% (1)	0.00% (0)
	USCensus	100.00% (1)	0.00% (0)	0.00% (0)

TABLE VII. Q1(E=Outliers)

R	P	S	N
R1	31.43% (176)	60.54% (339)	8.04% (45)
R2	38.75% (31)	56.25% (45)	5.00% (4)
R3	12.50% (1)	87.50% (7)	0.00% (0)

TABLE VIII. Q2(E=Outliers)

R	Scenario	P	S	N
R1	CD	26.79% (75)	63.57% (178)	9.64% (27)
	BD	36.07% (101)	57.50% (161)	6.43% (18)
R2	CD	27.50% (11)	70.00% (28)	2.50% (1)
	BD	50.00% (20)	42.50% (17)	7.50% (3)
R3	CD	0.00% (0)	100.00% (4)	0.00% (0)
	BD	25.00% (1)	75.00% (3)	0.00% (0)

TABLE IX. Q3(E=Outliers)

R	Model	P	S	N
R1	AdaBoost	12.50% (10)	70.00% (56)	17.50% (14)
	Decision Tree	30.00% (24)	68.75% (55)	1.25% (1)
	Guassian Naive Bayes	31.25% (25)	63.75% (51)	5.00% (4)
	KNN	52.50% (42)	42.50% (34)	5.00% (4)
	Logistic Regression	22.50% (18)	60.00% (48)	17.50% (14)
	Random Forest	32.50% (26)	60.00% (48)	7.50% (6)
	XGBoost	38.75% (31)	58.75% (47)	2.50% (2)

TABLE X. Q4.1(E=OUTLIERS)

R	Detection	P	S	N
R1	IF	33.93% (57)	47.02% (79)	19.05% (32)
	IQR	58.93% (99)	38.10% (64)	2.98% (5)
	SD	7.74% (13)	89.88% (151)	2.38% (4)
R2	IF	37.50% (9)	58.33% (14)	4.17% (1)
	IQR	70.83% (17)	16.67% (4)	12.50% (3)
	SD	16.67% (4)	83.33% (20)	0.00% (0)

TABLE XI. Q4.2(E=OUTLIERS)

R	Repair	P	S	N
R1	HoloClean	12.50% (7)	80.36% (45)	7.14% (4)
	Mean	33.33% (56)	60.12% (101)	6.55% (11)
	Median	33.33% (56)	57.74% (97)	8.93% (15)
	Mode	33.93% (57)	57.14% (96)	8.93% (15)
R2	HoloClean	12.50% (1)	87.50% (7)	0.00% (0)
	Mean	41.67% (10)	54.17% (13)	4.17% (1)
	Median	45.83% (11)	50.00% (12)	4.17% (1)
	Mode	37.50% (9)	54.17% (13)	8.33% (2)

TABLE XII. Q5(E=OUTLIERS)

R	Dataset	P	S	N
R1	Airbnb	10.00% (14)	87.14% (122)	2.86% (4)
	Credit	14.29% (20)	70.00% (98)	15.71% (22)
	EEG	57.14% (80)	40.71% (57)	2.14% (3)
	Sensor	44.29% (62)	44.29% (62)	11.43% (16)
R2	Airbnb	30.00% (6)	70.00% (14)	0.00% (0)
	Credit	0.00% (0)	80.00% (16)	20.00% (4)
	EEG	75.00% (15)	25.00% (5)	0.00% (0)
	Sensor	50.00% (10)	50.00% (10)	0.00% (0)
R3	Airbnb	0.00% (0)	100.00% (2)	0.00% (0)
	Credit	0.00% (0)	100.00% (2)	0.00% (0)
	EEG	50.00% (1)	50.00% (1)	0.00% (0)
	Sensor	0.00% (0)	100.00% (2)	0.00% (0)

TABLE XIII. Q1(E=MISLABELS)

R	P	S	N
R1	39.88% (67)	54.17% (91)	5.95% (10)
R2&R3	45.83% (11)	54.17% (13)	0.00% (0)

TABLE XIV. Q2(E=MISLABELS)

R	Scenario	P	S	N
R1	BD	28.57% (24)	70.24% (59)	1.19% (1)
	CD	51.19% (43)	38.10% (32)	10.71% (9)
R2&R3	BD	33.33% (4)	66.67% (8)	0.00% (0)
	CD	58.33% (7)	41.67% (5)	0.00% (0)

TABLE XV. Q3(E=MISLABELS)

R	Model	P	S	N
R1	Adaboost	45.83% (11)	54.17% (13)	0.00% (0)
	Decision Tree	41.67% (10)	54.17% (13)	4.17% (1)
	Gaussian Naive Bayes	25.00% (6)	58.33% (14)	16.67% (4)
	KNN	37.50% (9)	58.33% (14)	4.17% (1)
	Logistic Regression	37.50% (9)	54.17% (13)	8.33% (2)
	Random Forest	41.67% (10)	54.17% (13)	4.17% (1)
	XGBoost	50.00% (12)	45.83% (11)	4.17% (1)

TABLE XVI. Q4.1(E=MISLABELS)

R	P	S	N
R1	39.88% (67)	54.17% (91)	5.95% (10)
R2	45.83% (11)	54.17% (13)	0.00% (0)

TABLE XVII. Q4.2(E=MISLABELS)

R	P	S	N
R1	39.88% (67)	54.17% (91)	5.95% (10)
R2	45.83% (11)	54.17% (13)	0.00% (0)

TABLE XVIII. Q5(E=MISLABELS)

R	Dataset	P	S	N
R1	EEG_major	71.43% (10)	21.43% (3)	7.14% (1)
	EEG_minor	78.57% (11)	21.43% (3)	0.00% (0)
	EEG_uniform	78.57% (11)	14.29% (2)	7.14% (1)
	Marketing_major	0.00% (0)	100.00% (14)	0.00% (0)
	Marketing_minor	7.14% (1)	92.86% (13)	0.00% (0)
	Marketing_uniform	0.00% (0)	100.00% (14)	0.00% (0)
	Titanic_major	0.00% (0)	57.14% (8)	42.86% (6)
	Titanic_minor	7.14% (1)	92.86% (13)	0.00% (0)
	Titanic_uniform	42.86% (6)	57.14% (8)	0.00% (0)
	USCensus_major	50.00% (7)	42.86% (6)	7.14% (1)
	USCensus_minor	71.43% (10)	28.57% (4)	0.00% (0)
	USCensus_uniform	71.43% (10)	21.43% (3)	7.14% (1)
R2	EEG_major	100.00% (2)	0.00% (0)	0.00% (0)
	EEG_minor	100.00% (2)	0.00% (0)	0.00% (0)
	EEG_uniform	100.00% (2)	0.00% (0)	0.00% (0)
	Marketing_major	0.00% (0)	100.00% (2)	0.00% (0)
	Marketing_minor	0.00% (0)	100.00% (2)	0.00% (0)
	Marketing_uniform	0.00% (0)	100.00% (2)	0.00% (0)
	Titanic_major	0.00% (0)	100.00% (2)	0.00% (0)
	Titanic_minor	0.00% (0)	100.00% (2)	0.00% (0)
	Titanic_uniform	50.00% (1)	50.00% (1)	0.00% (0)
	USCensus_major	50.00% (1)	50.00% (1)	0.00% (0)
	USCensus_minor	50.00% (1)	50.00% (1)	0.00% (0)
	USCensus_uniform	100.00% (2)	0.00% (0)	0.00% (0)
R3	EEG_major	100.00% (2)	0.00% (0)	0.00% (0)
	EEG_minor	100.00% (2)	0.00% (0)	0.00% (0)
	EEG_uniform	100.00% (2)	0.00% (0)	0.00% (0)
	Marketing_major	0.00% (0)	100.00% (2)	0.00% (0)
	Marketing_minor	0.00% (0)	100.00% (2)	0.00% (0)
	Marketing_uniform	0.00% (0)	100.00% (2)	0.00% (0)
	Titanic_major	0.00% (0)	100.00% (2)	0.00% (0)
	Titanic_minor	0.00% (0)	100.00% (2)	0.00% (0)
	Titanic_uniform	50.00% (1)	50.00% (1)	0.00% (0)
	USCensus_major	50.00% (1)	50.00% (1)	0.00% (0)
	USCensus_minor	50.00% (1)	50.00% (1)	0.00% (0)
	USCensus_uniform	100.00% (2)	0.00% (0)	0.00% (0)

*Note: uniform, major, minor c.f. mislabel distribution as defined in the CleanML paper.

TABLE XIX. Q1(E=INCONSISTENCIES)

R	P	S	N
R1	12.50% (7)	87.50% (49)	0.00% (0)
R2 & R3	25.00% (2)	75.00% (6)	0.00% (0)

TABLE XX. Q2(E=INCONSISTENCIES)

R	Scenario	P	S	N
R1	CD	17.86% (5)	82.14% (23)	0.00% (0)
	BD	7.14% (2)	92.86% (26)	0.00% (0)
R2	CD	25.00% (1)	75.00% (3)	0.00% (0)
	BD	25.00% (1)	75.00% (3)	0.00% (0)
R3	CD	25.00% (1)	75.00% (3)	0.00% (0)
	BD	25.00% (1)	75.00% (3)	0.00% (0)

TABLE XXI. Q3(E=INCONSISTENCIES)

R	Model	P	S	N
R1	Adaboost	12.50% (1)	87.50% (7)	0.00% (0)
R1	Decision Tree	0.00% (0)	100.00% (8)	0.00% (0)
R1	Gaussian Naive Bayes	12.50% (1)	87.50% (7)	0.00% (0)
R1	KNN	25.00% (2)	75.00% (6)	0.00% (0)
R1	Logistic Regression	12.50% (1)	87.50% (7)	0.00% (0)
R1	Random Forest	25.00% (2)	75.00% (6)	0.00% (0)
R1	XGBoost	0.00% (0)	100.00% (8)	0.00% (0)

TABLE XXII. Q4.1(E=INCONSISTENCIES)

R	P	S	N
R1	12.50% (7)	87.50% (49)	0.00% (0)
R2	25.00% (2)	75.00% (6)	0.00% (0)

TABLE XXIII. Q4.2(E=Inconsistencies)

R	P	S	N
R1	12.50% (7)	87.50% (49)	0.00% (0)
R2	25.00% (2)	75.00% (6)	0.00% (0)

TABLE XXIV. Q5(E=Inconsistencies)

R	Dataset	P	S	N
R1	Company	28.57% (4)	71.43% (10)	0.00% (0)
	Movie	14.29% (2)	85.71% (12)	0.00% (0)
	Restaurant	0.00% (0)	100.00% (14)	0.00% (0)
	University	7.14% (1)	92.86% (13)	0.00% (0)
R2&R3	Company	0.00% (0)	100.00% (2)	0.00% (0)
	Movie	100.00% (2)	0.00% (0)	0.00% (0)
	Restaurant	0.00% (0)	100.00% (2)	0.00% (0)
	University	0.00% (0)	100.00% (2)	0.00% (0)

TABLE XXV. Q1 (E=Duplicates)

R	P	S	N
R1	10.71% (12)	66.96% (75)	22.32% (25)
R2	12.50% (2)	56.25% (9)	31.25% (5)
R3	12.50% (1)	50.00% (4)	37.50% (3)

TABLE XXVI. Q2 (E=Duplicates)

R	Scenario	P	S	N
R1	CD	16.07% (9)	66.07% (37)	17.86% (10)
	BD	5.36% (3)	67.86% (38)	26.79% (15)
R2	CD	25.00% (2)	62.50% (5)	12.50% (1)
	BD	0.00% (0)	50.00% (4)	50.00% (4)
R3	CD	0.00% (0)	75.00% (3)	25.00% (1)
	BD	0.00% (0)	50.00% (2)	50.00% (2)

TABLE XXVII. Q3 (E=Duplicates)

R	Model	P	S	N
R1	AdaBoost	12.50% (2)	75.00% (12)	12.50% (2)
	Decision Tree	0.00% (0)	87.50% (14)	12.50% (2)
	Gussian Naive Bayes	25.00% (4)	56.25% (9)	18.75% (3)
	KNN	6.25% (1)	81.25% (13)	12.50% (2)
	Logistic Regression	18.75% (3)	62.50% (10)	18.75% (3)
	Random Forest	12.50% (2)	37.50% (6)	50.00% (8)
	XGBoost	0.00% (0)	68.75% (11)	31.25% (5)

TABLE XXVIII. Q4.1 (E=Duplicates)

R	Detection	P	S	N
R1	AutoER	5.36% (3)	60.71% (34)	33.93% (19)
	Key Collision	16.07% (9)	73.21% (41)	10.71% (6)
R2	AutoER	12.50% (1)	50.00% (4)	37.50% (3)
	Key Collision	12.50% (1)	62.50% (5)	25.00% (2)

TABLE XXIX. Q4.2 (E=Duplicates)

R	P	S	N
R1	12.50% (7)	87.50% (49)	0.00% (0)
R2	25.00% (2)	75.00% (6)	0.00% (0)

TABLE XXX. Q5 (E=Duplicates)

R	Dataset	P	S	N
R1	Airbnb	3.57% (1)	85.71% (24)	10.71% (3)
	Citation	10.71% (3)	71.43% (20)	17.86% (5)
	Movie	28.57% (8)	21.43% (6)	50.00% (14)
	Restaurant	0.00% (0)	89.29% (25)	10.71% (3)
R2	Movie	25.00% (1)	0.00% (0)	75.00% (3)
	Restaurant	0.00% (0)	25.00% (1)	75.00% (3)
	Citation	50.00% (2)	50.00% (2)	0.00% (0)
	Airbnb	0.00% (0)	75.00% (3)	25.00% (1)
R3	Movie	0.00% (0)	0.00% (0)	100.00% (2)
	Restaurant	0.00% (0)	0.00% (0)	100.00% (2)
	Citation	50.00% (1)	50.00% (1)	0.00% (0)
	Airbnb	0.00% (0)	100.00% (2)	0.00% (0)