DEALING WITH MISSING DATA USING ATTENTION AND LATENT SPACE REGULARIZATION (SUPPLEMENTARY MATERIALS)

Anonymous authors
Paper under double-blind review

1 METHODS TO CORRUPT DATASETS WITH MISSINGNESS

For corrupting data with missingness, we have three predefined patterns.

1.1 MCAR

In order to generate data MCAR, 40% of data from 40% of randomly selected columns is randomly deleted.

1.2 MAR

For generating data MAR 40% of data from 40% of columns is deleted such that a value in a row is deleted depending on the value of it's neighboring column to the left.

1.3 MNAR

Finally, for data MNAR 40% of data from 40% of columns is deleted such that a value in a row is deleted depending on it's value.

2 MODEL TRAINING

2.1 HYPER-PARAMETER OPTIMIZATION

Our approach to optimization was using Bayesian hyperparamater (Biewald 2020). For the transformer model, 4 hyperparameter ranges were optimized including learning rate (1e-3 - 1e-5), weight decay (1e-1 - 1e-7), model dimensionality (4 - 150) and optimizer (adam, adabelief, and stochastic gradient descent).

For the LightGBM model, 5 hyperparameters were optimized including number of leaves (1 - 1000), the learning rate (1e-1 - 1e-6), the minimum data per leaf (1 - 500), the maximum number of bins (100 - 1000), and the type of boosting (gbdt, rf, dart, goss).

2.2 MODEL TRAINING

Both models were trained using a cross-entropy loss objective. The LSAM included weight regularization. For LSAM model was trained for a maximum of 5000 steps with early stopping.

LightGBM models were trained for 100 iterations with early stopping.

3 PERFORMANCE TABLES

Complete results for performance on the benchmark datasets with difference missingness regimes are available in the following tables.

Table 1: Baseline accuracy with no missing data on benchmark datasets.

	LSAM	LightGBM
ozone-level-8hr	**0.92**	0.90
splice	0.82	**0.92**
pc3	0.85	**0.87**
qsar-biodeg	**0.85**	0.85
mfeat-zernike	**0.81**	0.77
mfeat-fourier	**0.84**	0.80
texture	**0.99**	0.95
kr-vs-kp	**0.99**	0.98
satimage	**0.91**	0.90
dna	0.96	**0.96**
optdigits	**0.98**	0.97
first-order-theorem-proving	0.50	**0.57**
GesturePhaseSegmentationProcessed	0.55	**0.66**
pc4	**0.89**	0.84
mfeat-karhunen	**0.97**	0.95
mfeat-pixel	**0.97**	0.85
spambase	**0.95**	0.93
mfeat-factors	**0.97**	0.95

Table 2: Change in accuracy from baseline when data is missing completely at random.

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
ozone-level-8hr	**-0.01**	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02
splice	-0.07	-0.01	-0.02	**-0.01**	-0.05	**-0.01**	-0.04	-0.02
pc3	-0.02	**-0.00**	-0.02	-0.00	-0.01	-0.01	-0.02	-0.01
qsar-biodeg	0.00	-0.01	0.00	-0.00	**0.00**	-0.00	0.00	-0.00
mfeat-zernike	-0.01	-0.01	-0.01	-0.01	**-0.00**	-0.01	-0.01	-0.01
mfeat-fourier	**-0.01**	-0.01	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
texture	-0.00	-0.01	-0.00	-0.01	**-0.00**	-0.01	-0.00	-0.01
kr-vs-kp	-0.01	-0.01	-0.03	-0.02	-0.04	-0.02	-0.02	**-0.01**
satimage	-0.00	-0.01	-0.00	-0.01	**0.00**	-0.00	-0.00	-0.01
dna	**0.00**	-0.00	-0.01	-0.00	-0.01	-0.00	-0.01	-0.00
optdigits	-0.00	-0.01	**-0.00**	-0.01	-0.01	-0.01	-0.01	-0.01
first-order-theorem-proving	-0.01	-0.02	-0.01	-0.02	**-0.00**	-0.02	-0.01	-0.02
GesturePhaseSegmentationProcessed	-0.02	-0.03	-0.02	-0.03	**-0.01**	-0.01	-0.03	-0.04
pc4	-0.00	0.01	-0.00	0.01	-0.00	**0.02**	-0.00	0.01
mfeat-karhunen	-0.01	-0.01	-0.01	-0.01	**-0.00**	-0.01	-0.01	-0.01
mfeat-pixel	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	**-0.00**
spambase	-0.01	-0.01	**-0.00**	-0.01	-0.01	-0.01	-0.01	-0.01
mfeat-factors	**-0.00**	-0.00	-0.01	-0.00	-0.01	-0.00	-0.01	-0.01

Table 3: Change in accuracy from baseline when data is missing at random.

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
ozone-level-8hr	**0.00**	-0.01	-0.00	-0.00	-0.01	-0.01	-0.00	-0.01
splice	-0.16	-0.27	-0.06	-0.05	-0.06	-0.05	**-0.02**	-0.07
pc3	-0.01	-0.01	**0.00**	-0.00	-0.01	0.00	-0.00	-0.01
qsar-biodeg	**0.01**	-0.01	-0.00	-0.01	0.01	-0.01	-0.00	-0.02
mfeat-zernike	-0.00	-0.04	-0.01	-0.01	**-0.00**	-0.01	-0.01	-0.01
mfeat-fourier	-0.02	-0.03	-0.02	-0.01	-0.01	-0.01	-0.02	**-0.01**
texture	-0.03	-0.08	-0.02	-0.08	**-0.00**	-0.02	-0.01	-0.07
kr-vs-kp	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	-0.01	**-0.01**
satimage	-0.02	-0.01	-0.02	-0.00	-0.01	**-0.00**	-0.01	-0.00
dna	-0.00	-0.00	-0.01	**0.00**	-0.01	**0.00**	-0.00	-0.00
optdigits	-0.04	-0.04	-0.02	-0.02	**-0.00**	-0.01	-0.01	-0.01
first-order-theorem-proving	-0.04	-0.04	-0.05	-0.02	**-0.02**	-0.02	-0.03	-0.02
GesturePhaseSegmentationProcessed	-0.02	-0.05	-0.02	-0.03	**-0.01**	-0.02	-0.02	-0.04
pc4	-0.01	**0.02**	-0.01	0.02	-0.02	0.02	0.00	0.01
mfeat-karhunen	-0.03	-0.09	-0.02	-0.02	-0.01	**-0.01**	-0.02	-0.01
mfeat-pixel	-0.03	-0.12	-0.01	-0.09	**-0.01**	-0.07	-0.01	-0.09
spambase	-0.01	-0.01	-0.01	-0.02	**-0.00**	-0.01	-0.01	-0.02
mfeat-factors	**-0.00**	-0.01	-0.01	-0.01	**-0.00**	-0.00	-0.01	-0.01

Table 4: Change in accuracy from baseline when data is missing not at random

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
ozone-level-8hr	-0.00	-0.00	-0.01	-0.01	-0.01	-0.02	-0.02	**-0.00**
splice	-0.40	-0.26	**-0.10**	-0.12	-0.14	-0.12	-0.27	-0.23
pc3	-0.01	**0.00**	-0.02	0.00	-0.01	-0.00	-0.01	0.00
qsar-biodeg	0.00	-0.01	-0.01	-0.01	**0.01**	-0.02	-0.01	-0.01
mfeat-zernike	**-0.01**	-0.01	-0.04	-0.02	-0.02	-0.01	-0.03	-0.02
mfeat-fourier	-0.03	-0.01	-0.06	-0.01	-0.06	**-0.00**	-0.05	-0.01
texture	-0.03	-0.04	-0.02	-0.03	**-0.00**	-0.02	-0.00	-0.02
kr-vs-kp	-0.01	**-0.01**	-0.05	-0.04	-0.05	-0.04	-0.02	-0.02
satimage	-0.01	-0.01	-0.03	-0.01	-0.02	**-0.00**	-0.01	-0.00
dna	**0.00**	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
optdigits	-0.00	**-0.00**	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
first-order-theorem-proving	-0.04	-0.06	-0.10	-0.05	**-0.01**	-0.04	-0.03	-0.06
GesturePhaseSegmentationProcessed	**-0.03**	-0.06	-0.08	-0.08	-0.06	-0.06	-0.08	-0.06
pc4	-0.01	-0.01	-0.01	0.00	-0.02	0.01	-0.02	**0.01**
mfeat-karhunen	**-0.02**	-0.02	-0.07	-0.03	-0.05	-0.02	-0.05	-0.02
mfeat-pixel	-0.01	-0.02	-0.01	-0.01	**-0.00**	-0.07	-0.01	-0.02
spambase	-0.01	**-0.00**	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01
mfeat-factors	-0.02	-0.01	-0.02	**-0.01**	-0.01	-0.01	-0.02	-0.01

Table 5: Baseline negative log-likelihood with no missing data for the benchmark datasets.

	NLL (No Missing Data)	
	LSAM	LightGBM
ozone-level-8hr	**0.51**	0.85
splice	1.47	**0.38**
pc3	0.92	**0.65**
qsar-biodeg	0.72	**0.71**
mfeat-zernike	**0.41**	0.68
mfeat-fourier	**0.41**	0.56
texture	**0.03**	0.96
kr-vs-kp	**0.09**	0.54
satimage	**0.24**	0.25
dna	0.17	**0.15**
optdigits	**0.06**	0.13
first-order-theorem-proving	1.47	**1.25**
GesturePhaseSegmentationProcessed	1.22	**0.95**
pc4	**0.57**	0.79
mfeat-karhunen	**0.12**	0.21
mfeat-pixel	**0.13**	2.27
spambase	**0.31**	0.44
mfeat-factors	**0.13**	0.21

Table 6: Negative change in NLL from baseline when data is missing completely at random (bigger is better).

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
dna	**0.00**	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
ozone-level-8hr	**0.00**	-0.02	-0.01	-0.02	-0.01	-0.03	-0.03	-0.03
mfeat-karhunen	-0.01	-0.03	-0.02	-0.04	**0.00**	-0.03	-0.03	-0.04
GesturePhaseSegmentationProcessed	-0.05	-0.06	-0.04	-0.05	**-0.02**	-0.03	-0.07	-0.07
optdigits	**-0.01**	-0.03	-0.01	-0.03	-0.04	-0.02	-0.04	-0.03
mfeat-factors	**-0.01**	-0.03	-0.03	-0.03	-0.01	-0.01	-0.02	-0.02
mfeat-fourier	**-0.03**	-0.06	-0.04	-0.07	-0.04	-0.06	-0.06	-0.07
qsar-biodeg	0.02	-0.03	0.02	-0.03	0.01	-0.01	**0.02**	-0.03
mfeat-zernike	-0.04	-0.04	-0.04	-0.04	-0.02	**-0.02**	-0.03	-0.04
kr-vs-kp	-0.07	**-0.04**	-0.18	-0.11	-0.21	-0.11	-0.15	-0.04
pc4	0.02	0.01	0.03	0.01	**0.06**	0.01	0.05	0.01
spambase	-0.03	-0.03	-0.03	-0.04	**-0.03**	-0.04	-0.05	-0.04
satimage	-0.01	-0.01	-0.01	-0.01	**0.00**	0.00	-0.01	-0.01
pc3	-0.04	**-0.01**	-0.04	-0.02	-0.04	-0.02	-0.04	-0.02
texture	-0.01	-0.02	-0.02	-0.02	**-0.01**	-0.01	-0.01	-0.02
mfeat-pixel	-0.03	-0.00	-0.03	-0.00	-0.02	-0.00	-0.04	**0.00**
splice	-1.28	-0.02	**0.98**	-0.03	-1.22	-0.03	0.96	-0.05
first-order-theorem-proving	-0.00	-0.05	-0.02	-0.04	**0.00**	-0.04	-0.01	-0.04

Table 7: Negative change in NLL from baseline when data is missing at random (bigger is better).

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
dna	0.01	-0.01	-0.03	-0.01	-0.03	-0.01	**0.01**	-0.01
ozone-level-8hr	**0.05**	-0.06	0.04	-0.01	-0.07	-0.02	0.04	-0.02
mfeat-karhunen	-0.08	-0.25	-0.08	-0.07	-0.04	**-0.04**	-0.05	-0.06
GesturePhaseSegmentationProcessed	-0.05	-0.10	-0.04	-0.06	**-0.03**	-0.05	-0.05	-0.09
optdigits	-0.14	-0.13	-0.06	-0.07	**-0.02**	-0.03	-0.05	-0.05
mfeat-factors	-0.01	-0.07	-0.02	-0.04	**-0.00**	-0.02	-0.02	-0.04
mfeat-fourier	-0.07	-0.12	-0.07	-0.05	-0.04	**-0.04**	-0.06	-0.04
qsar-biodeg	**0.01**	-0.02	-0.02	-0.03	-0.01	-0.01	-0.01	-0.05
mfeat-zernike	-0.04	-0.15	-0.06	-0.06	**-0.02**	-0.02	-0.04	-0.05
kr-vs-kp	-0.06	-0.04	-0.12	-0.06	-0.10	-0.06	-0.07	**-0.03**
pc4	-0.00	-0.01	-0.01	-0.01	-0.05	0.00	**0.02**	-0.00
spambase	-0.05	-0.09	-0.05	-0.09	**-0.03**	-0.06	-0.07	-0.08
satimage	-0.05	-0.03	-0.05	-0.01	-0.02	**-0.00**	-0.01	-0.01
pc3	0.05	-0.01	**0.06**	0.02	-0.08	0.02	0.02	-0.00
texture	-0.09	-0.18	-0.06	-0.19	**-0.01**	-0.05	-0.02	-0.16
mfeat-pixel	-0.10	-0.00	-0.03	**-0.00**	-0.02	-0.01	-0.03	-0.00
splice	0.50	-0.53	-0.14	-0.14	-0.13	-0.14	**0.91**	-0.16
first-order-theorem-proving	-0.09	-0.06	-0.08	-0.03	**-0.02**	-0.04	-0.04	-0.03

Table 8: Negative change in NLL from baseline when data is missing not at random (bigger is better).

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
dna	**0.00**	-0.00	-0.02	-0.01	-0.02	-0.01	-0.00	-0.01
ozone-level-8hr	0.02	-0.03	**0.08**	-0.03	-0.02	-0.03	0.03	-0.02
mfeat-karhunen	**-0.04**	-0.09	-0.21	-0.10	-0.13	-0.08	-0.14	-0.09
GesturePhaseSegmentationProcessed	**-0.07**	-0.13	-0.19	-0.17	-0.15	-0.13	-0.17	-0.15
optdigits	**-0.01**	-0.01	-0.02	-0.05	-0.02	-0.05	-0.04	-0.05
mfeat-factors	-0.06	-0.04	-0.05	-0.03	**-0.02**	-0.03	-0.04	-0.04
mfeat-fourier	-0.08	**-0.04**	-0.14	-0.06	-0.14	-0.06	-0.13	-0.07
qsar-biodeg	0.01	-0.02	-0.01	-0.03	**0.01**	-0.03	-0.03	-0.03
mfeat-zernike	**-0.05**	-0.07	-0.15	-0.09	-0.07	-0.05	-0.14	-0.08
kr-vs-kp	-0.05	**-0.02**	-0.18	-0.10	-0.18	-0.10	-0.08	-0.05
pc4	-0.03	-0.07	-0.05	-0.09	-0.05	**-0.02**	-0.06	-0.06
spambase	-0.04	**-0.02**	-0.09	-0.06	-0.04	-0.04	-0.05	-0.05
satimage	-0.03	-0.02	-0.07	-0.03	-0.04	**-0.01**	-0.03	-0.02
pc3	**0.05**	0.03	0.04	0.02	-0.02	0.01	0.01	0.02
texture	-0.10	-0.07	-0.06	-0.08	**-0.01**	-0.04	-0.02	-0.07
mfeat-pixel	-0.04	-0.00	-0.05	-0.00	-0.00	-0.00	-0.04	**-0.00**
splice	-1.36	-0.49	**-0.18**	-0.27	-0.22	-0.27	-2.71	-0.45
first-order-theorem-proving	-0.08	-0.11	-0.17	-0.10	**-0.02**	-0.07	-0.06	-0.11

Table 9: Baseline accuracy on datasets with pre-existing missingness

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
ipums_la_99-small	0.66	0.81	0.67	0.81	**0.66**	0.81	0.66	0.77
ipums_la_98-small	0.69	0.86	**0.68**	0.86	0.69	0.86	0.69	0.85
communities-and-crime-binary	0.84	0.84	0.84	0.84	**0.84**	0.84	0.84	0.84
jungle_chess_2pcs_endgame_rat_panther	1.00	**0.99**	1.00	0.99	1.00	0.99	1.00	0.99
jungle_chess_2pcs_endgame_rat_elephant	1.00	0.99	0.99	0.99	0.99	0.99	0.99	**0.99**
jungle_chess_2pcs_endgame_rat_lion	0.99	1.00	0.99	1.00	**0.99**	1.00	0.99	1.00
kdd_ipums_la_97-small	0.98	0.98	0.98	0.98	0.98	0.98	0.97	**0.77**
MiceProtein	0.99	0.83	0.99	0.83	0.99	**0.81**	0.99	0.82
cjs	1.00	0.99	1.00	0.99	0.99	0.99	1.00	**0.95**
SpeedDating	0.80	0.81	0.80	0.81	**0.78**	0.82	0.80	0.82
colleges_usnews	0.73	0.73	0.74	0.73	0.74	0.74	0.72	**0.72**

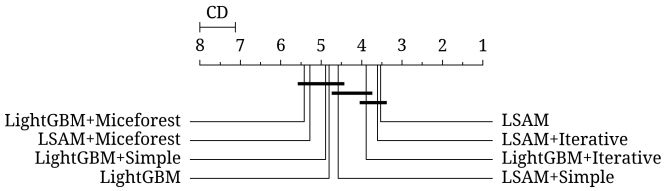
Table 10: Baseline negative log-likelihood on datasets with pre-existing missingness

	None		Simple		Iterative		MiceForest	
	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM	LSAM	LightGBM
ipums_la_99-small	5.03	0.71	4.97	**0.71**	5.03	**0.71**	5.01	0.81
ipums_la_98-small	0.88	**0.60**	0.91	0.60	0.88	0.60	0.89	0.66
communities-and-crime-binary	0.72	**0.71**	0.72	0.71	0.73	0.71	0.73	0.72
jungle_chess_2pcs_endgame_rat_panther	0.02	0.04	0.02	0.04	0.02	0.04	**0.02**	0.04
jungle_chess_2pcs_endgame_rat_elephant	0.02	0.03	0.02	0.03	**0.02**	0.03	0.03	0.05
jungle_chess_2pcs_endgame_rat_lion	0.04	**0.02**	0.04	0.02	0.04	0.02	0.04	0.03
kdd_ipums_la_97-small	0.15	0.17	0.15	0.17	**0.15**	0.17	0.22	1.29
MiceProtein	0.04	1.60	**0.03**	1.60	0.06	1.60	0.05	1.60
cjs	**0.01**	1.78	0.02	1.78	0.03	1.78	0.02	1.80
SpeedDating	**0.83**	0.97	0.85	0.97	0.89	0.96	0.84	0.96
colleges_usnews	1.03	1.08	1.04	1.09	**1.02**	1.07	1.06	1.12

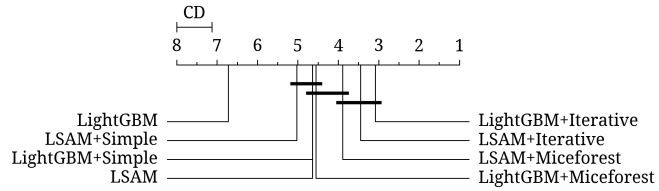
4 ACCURACY CRITICAL DIFFERENCE DIAGRAMS

In Figures 1 and 2 we report the critical difference diagrams for the outcome of accuracy.

(a) Performance when data is MCAR



(b) Performance when data is MAR



(c) Performance when data is MNAR

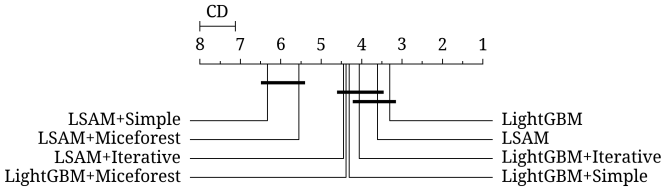


Figure 1: Critical difference diagrams comparing performance for different missingness regimes, demonstrating improved performance for the LSAM without imputation. Points are labelled by the type of model as well as the imputation strategy if used. The performance metric is the change in accuracy from baseline performance with complete data. Further right in the diagram indicates better performance. A break in the solid bar underneath demonstrates statistical significance.

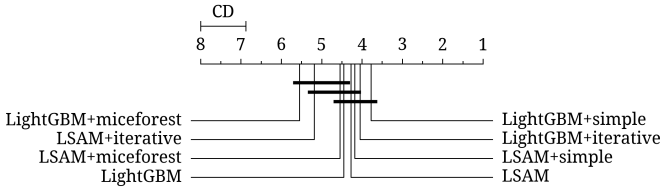


Figure 2: Critical difference diagram comparing performance with accuracy on datasets with unknown missingness pattern.

5 META-LEARNING

Meta-learning was performed to study the characteristics of the datasets where LSAM outperformed the comparator models. Random forest models were trained to predict an LSAM win based on dataset characteristics under several conditions. The first comparison looked at features importance when LSAM beat LightGBM. Then feature importance was explored when LSAM without imputation beat LSAM with imputation methods. The full feature importance rankings for the complete but corrupted datasets and incomplete datasets are presented below.

Table 11: Meta-learning approach to determine feature importance with respect to LSAM outperforming benchmarks on corrupted datasets. Analysis is stratified by overall model type as well as out-of-the-box performance of LSAM compared to imputation.

	Overall: LightGBM vs LSAM		LSAM: Simple vs None		LSAM: Iterative vs None		LSAM: Miceforest vs None	
Dataset Characteristic	Accuracy	NLL	Accuracy	NLL	Accuracy	NLL	Accuracy	NLL
NumberOfFeatures	0.08	0.06	**0.17**	0.13	**0.15**	0.11	**0.15**	0.12
NumberOfInstances	**0.21**	**0.22**	0.11	0.10	0.11	0.12	0.12	0.13
NumberOfClasses	0.15	0.15	0.07	0.06	0.08	0.07	0.05	0.08
NumberOfNumericFeatures	0.12	0.15	0.10	0.12	0.11	0.11	0.10	0.11
NumberOfSymbolicFeatures	0.06	0.12	0.04	0.03	0.04	0.04	0.04	0.03
NumericRatio	0.16	0.12	0.10	0.11	0.09	0.11	0.11	0.10
FeatureInstanceRatio	0.12	0.11	0.15	**0.13**	0.13	**0.16**	0.14	**0.14**
missingness_MAR	0.03	0.02	0.06	0.12	0.07	0.08	0.10	0.11
missingness_MCAR	0.03	0.03	0.07	0.09	0.09	0.10	0.09	0.10
missingness_MNAR	0.05	0.03	0.13	0.11	0.12	0.10	0.10	0.10

Table 12: Meta-learning approach to determine feature importance with respect to LSAM outperforming benchmarks on datasets with pre-existing missingness. Analysis is stratified by overall model type as well as out-of-the-box performance of LSAM compared to imputation.

	Overall: LightGBM vs LSAM		LSAM: Simple vs None		LSAM: Iterative vs None		LSAM: Miceforest vs None	
Dataset Characteristic	Accuracy	NLL	Accuracy	NLL	Accuracy	NLL	Accuracy	NLL
NumberOfFeatures	0.09	0.11	**0.16**	0.04	**0.19**	**0.26**	**0.22**	0.08
NumberOfInstances	0.23	0.07	0.11	**0.24**	0.11	0.13	0.12	0.15
NumberOfClasses	0.11	0.03	0.09	0.08	0.02	0.06	0.04	0.05
NumberOfNumericFeatures	0.09	**0.27**	0.10	0.16	0.14	0.09	0.17	0.18
NumberOfSymbolicFeatures	**0.25**	0.11	0.13	0.06	0.13	0.10	0.09	0.06
NumericRatio	0.09	0.23	0.16	0.17	0.18	0.12	0.07	0.11
FeatureInstanceRatio	0.06	0.14	0.11	0.21	0.07	0.08	0.15	**0.23**
FractionMissingValues	0.09	0.05	0.14	0.04	0.17	0.16	0.15	0.14

REFERENCES

Biewald L. Experiment tracking with weights and biases [Internet]. 2020. Available from: https://www.wandb.com/