Microsoft

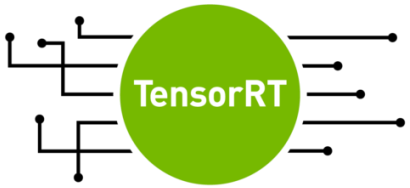
Compiling Classical ML Pipelines into Tensor Computations for OneSize-Fits-All ML Prediction Serving

Matteo Interlandi, Markus Weimer, Supun Nakandala*,
Karla Saur, Konstantinos Karanasos

GSL, Azure Data
*UCSD

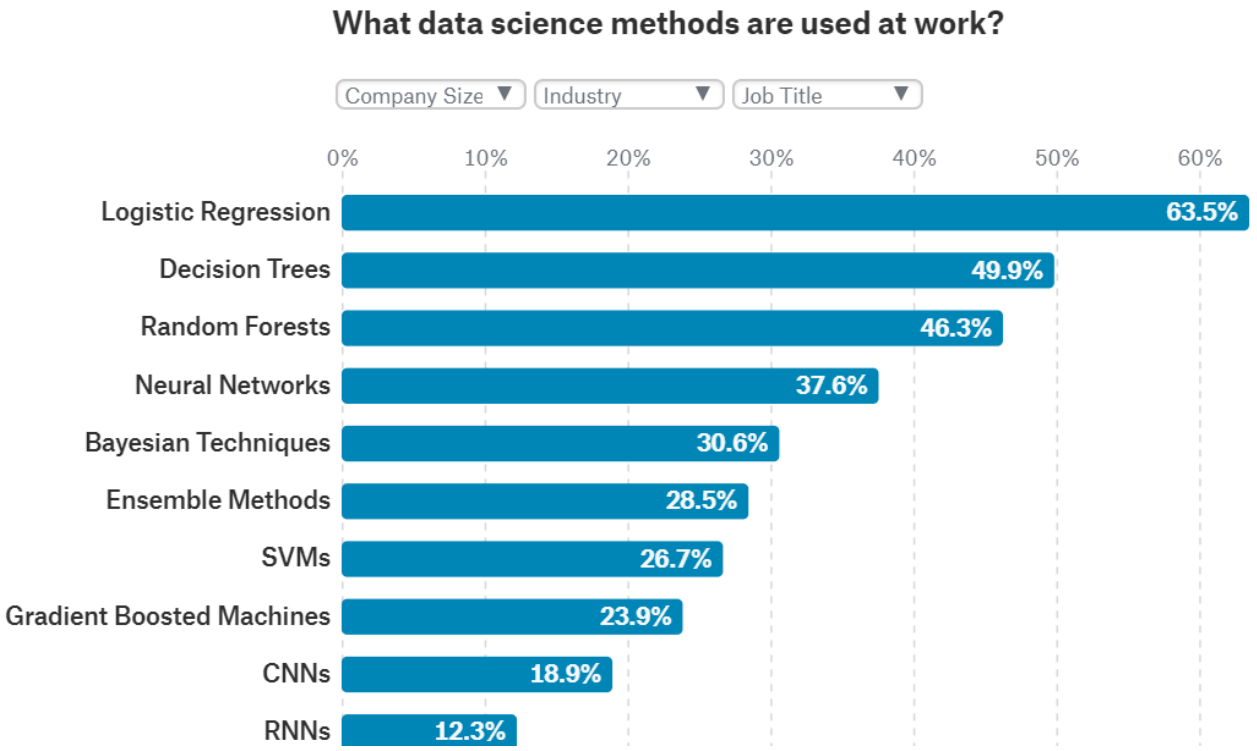
Motivation

Specialized Prediction Systems have been developed (mostly focus on neural networks)



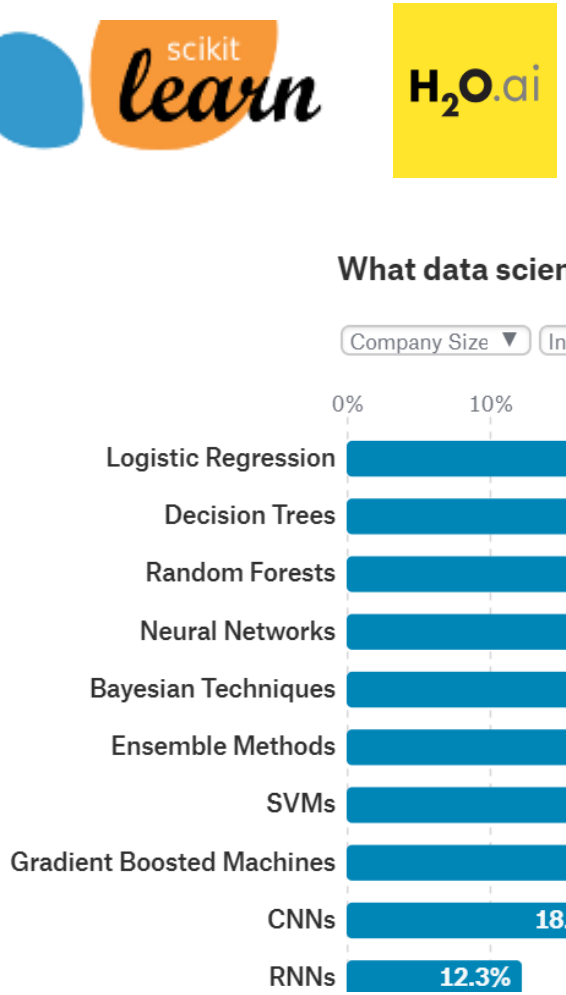
Support for classical ML methods is largely overlooked (widely used in the enterprise, scientific, and other domains)

Classical ML Models



2017 Kaggle Survey: The State of Data Science & Machine Learning

Classical ML Models



Classical
Classical
Classical
Classical
Classical
Classical

2017 Kaggle Survey: The State of Data Science & Machine Learning

Question

Can we compile classical ML pipelines into tensor computations so that we can leverage neural network systems?

Benefits:

(1) Exploit the already available optimizations
(2) Seamless hardware acceleration
(3) Significant reduction in engineering effort

Tree-models Microbenchmark: Settings

Dataset	Rows	#Features	Task
fraud	285k	28	BinaryClass
year	515k	90	Regression
covtype	581k	54	Multiclass
epsilon	500k	2000	BinaryClass

- 3 models: *RandomForest* (**rf**), *XGBoost* (**xgb**), *LightGBM* (**lgbm**)
- 80/20 train/test split
- 4 translation targets: *PyTorch* (**hb-pt**), *Torchscript* (**hb-ts**), *ONNX* (**hb-onnx**) and *TVM* (**hb-tvm**)
- Batch inferences (6 cores, batch size of 10k, w\ and w\o GPU)

Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud						
year						
covtype						
epsilon						

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud						
year						
covtype						
epsilon						

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud						
year						
covtype						
epsilon						

Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s				
year	**2.33s**	17.23s				
covtype	47.64s	**24.77s**				
epsilon	**11.22s**	26.03s				

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.01s**	6.4s				
year	**5.77s**	15.75s				
covtype	**63.45s**	173.92s				
epsilon	**14.84s**	29s				

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**3.76s**	6.59s				
year	**6.18s**	10.14s				
covtype	**67.12s**	158.3s				
epsilon	**14.13s**	26.03s				

Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s	17.18s	17.28	92.58s	
year	**2.33s**	17.23s	17.95s	17.23s	154.71s	
covtype	47.64s	**24.77s**	38.27s	38.02s	260.35s	
epsilon	**11.22s**	26.03s	48.52s	45.87s	SEGFAULT	

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.01s**	6.4s	17.23s	16.38s	89.71s	
year	**5.77s**	15.75s	17.26s	15.74s	153.96s	
covtype	**63.45s**	173.92s	295.6s	295.3s	1255s	
epsilon	**14.84s**	29s	47.38s	48.78s	SEGFAULT	

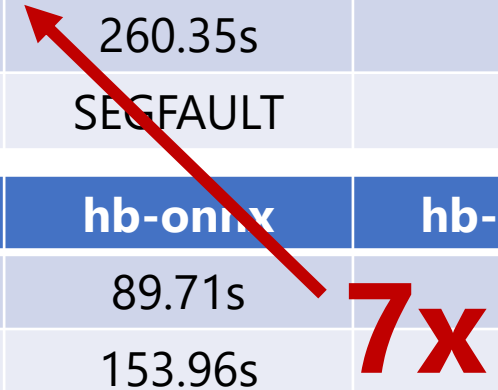
	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**3.76s**	6.59s	17.41s	16.43s	89.90s	
year	**6.18s**	10.14s	18.3s	18.01s	153.67s	
covtype	**67.12s**	158.3s	296s	294s	1256s	
epsilon	**14.13s**	26.03s	47.21s	45.89s	SEGFAULT	

Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s	17.18s	17.28	92.58s	
year	**2.33s**	17.23s	17.95s	17.23s	154.71s	
covtype	47.64s	**24.77s**	38.27s	38.02s	260.35s	
epsilon	**11.22s**	26.03s	48.52s	45.87s	SEGFAULT	

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.01s**	6.4s	17.23s	16.38s	89.71s	
year	**5.77s**	15.75s	17.26s	15.74s	153.96s	
covtype	**63.45s**	173.92s	295.6s	295.3s	1255s	
epsilon	**14.84s**	29s	47.38s	48.78s	SEGFAULT	

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**3.76s**	6.59s	17.41s	16.43s	89.90s	
year	**6.18s**	10.14s	18.3s	18.01s	153.67s	
covtype	**67.12s**	158.3s	296s	294s	1256s	
epsilon	**14.13s**	26.03s	47.21s	45.89s	SEGFAULT	

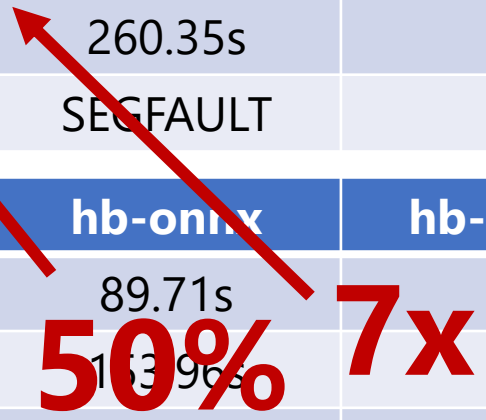


Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s	17.18s	17.28	92.58s	
year	**2.33s**	17.23s	17.95s	17.23s	154.71s	
covtype	47.64s	**24.77s**	38.27s	38.02s	260.35s	
epsilon	**11.22s**	26.03s	48.52s	45.87s	SEGFAULT	

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.01s**	6.4s	17.23s	16.38s	89.71s	
year	**5.77s**	15.75s	17.26s	15.74s	153.96s	
covtype	**63.45s**	173.92s	295.6s	295.3s	1255s	
epsilon	**14.84s**	29s	47.38s	48.78s	SEGFAULT	

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**3.76s**	6.59s	17.41s	16.43s	89.90s	
year	**6.18s**	10.14s	18.3s	18.01s	153.67s	
covtype	**67.12s**	158.3s	296s	294s	1256s	
epsilon	**14.13s**	26.03s	47.21s	45.89s	SEGFAULT	



Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s	17.18s	17.28	92.58s	3.84s
year	2.33s	17.23s	17.95s	17.23s	154.71s	**1.43s**
covtype	47.64s	24.77s	38.27s	38.02s	260.35s	**20.18s**
epsilon	11.22s	26.03s	48.52s	45.87s	SEGFAULT	**8.17s**

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	17.23s	16.38s	89.71s	**1.93s**
year	5.77s	15.75s	17.26s	15.74s	153.96s	**1.77s**
covtype	63.45s	173.92s	295.6s	295.3s	1255s	**28.53s**
epsilon	14.84s	29s	47.38s	48.78s	SEGFAULT	**4.43s**

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	17.41s	16.43s	89.90s	**1.97s**
year	6.18s	10.14s	18.3s	18.01s	153.67s	**1.78s**
covtype	67.12s	158.3s	296s	294s	1256s	**29.19s**
epsilon	14.13s	26.03s	47.21s	45.89s	SEGFAULT	**4.41s**

Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s	17.18s	17.28	92.58s	3.84s
year	2.33s	17.23s	17.95s	17.23s	154.71s	**1.43s**
covtype	47.64s	24.77s	38.27s	38.02s	260.35s	**20.18s**
epsilon	11.22s	26.03s	48.52s	45.87s	SEGFAULT	**8.17s**

50%

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	17.23s	16.38s	89.71s	**1.93s**
year	5.77s	15.75s	17.26s	15.74s	153.96s	**1.77s**
covtype	63.45s	173.92s	295.6s	295.3s	1255s	**28.53s**
epsilon	14.84s	29s	47.38s	48.78s	SEGFAULT	**4.43s**

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	17.41s	16.43s	89.90s	**1.97s**
year	6.18s	10.14s	18.3s	18.01s	153.67s	**1.78s**
covtype	67.12s	158.3s	296s	294s	1256s	**29.19s**
epsilon	14.13s	26.03s	47.21s	45.89s	SEGFAULT	**4.41s**

Tree-models Microbenchmark: Batch Inference on CPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	**2.52s**	8.1s	17.18s	17.28	92.58s	3.84s
year	2.33s	17.23s	17.95s	17.23s	154.71s	**1.43s**
covtype	47.64s	24.77s	38.27s	38.02s	260.35s	**20.18s**
epsilon	11.22s	26.03s	48.52s	45.87s	SEGFAULT	**8.17s**

50%

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	17.23s	16.38s	89.71s	**1.93s**
year	5.77s	15.75s	17.26s	15.74s	153.96s	**1.77s**
covtype	63.45s	173.92s	295.6s	295.3s	1255s	**28.53s**
epsilon	14.84s	29s	47.38s	48.78s	SEGFAULT	**4.43s**

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	17.41s	16.43s	89.90s	**1.97s**
year	6.18s	10.14s	18.3s	18.01s	153.67s	**1.78s**
covtype	67.12s	158.3s	296s	294s	1256s	**29.19s**
epsilon	14.13s	26.03s	47.21s	44.89s	SEGFAULT	**4.41s**

3x

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	8.1s				
year	2.33s	17.23s				
covtype	47.64s	24.77s				
epsilon	11.22s	26.03s				

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s				
year	5.77s	15.75s				
covtype	63.45s	173.92s				
epsilon	14.84s	29s				

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s				
year	6.18s	10.14s				
covtype	67.12s	158.3s				
epsilon	14.13s	26.03s				

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	8.1s	0.15s	**0.11s**	17.55s	
year	2.33s	17.23s	0.15s	**0.10s**	45.12s	
covtype	47.64s	24.77s	0.32s	**0.26s**	62.81s	
epsilon	11.22s	26.03s	0.36s	**0.29s**	OOM	

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	0.16s	**0.11s**	16.89s	
year	5.77s	15.75s	0.14s	**0.1s**	44.82s	
covtype	63.45s	173.92s	1.47s	**1.29s**	445.89s	
epsilon	14.84s	29s	0.37s	**0.28s**	OOM	

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	0.16s	**0.11s**	17.70s	
year	6.18s	10.14s	0.14s	**0.10s**	OOM	
covtype	67.12s	158.3s	1.47s	**1.29s**	446s	
epsilon	14.13s	26.03s	0.36s	**0.28s**	OOM	

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	8.1s	0.15s	**0.11s**	17.55s	
year	2.33s	17.23s	0.15s	**0.10s**	45.12s	
covtype	47.64s	24.77s	0.32s	**0.26s**	62.81s	
epsilon	11.22s	26.03s	0.36s	**0.29s**	OOM	

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	0.16s	**0.11s**	16.89s	
year	5.77s	15.75s	0.14s	**0.1s**	44.82s	
covtype	63.45s	173.92s	1.47s	**1.29s**	445.89s	
epsilon	14.84s	29s	0.37s	**0.28s**	OOM	

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	0.16s	**0.11s**	17.70s	
year	6.18s	10.14s	0.14s	**0.10s**	OOM	
covtype	67.12s	158.3s	1.47s	**1.29s**	446s	
epsilon	14.13s	26.03s	0.36s	**0.28s**	OOM	

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	[illegible]s	0.15s	**0.11s**	17.55s	
year	2.33s	17.23s	[illegible]s	**0.10s**	45.12s	
covtype	47.64s	24.77s	0.32s	**0.26s**	62.81s	
epsilon	11.22s	26.03s	0.36s	**0.29s**	OOM	

100x

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	0.16s	**0.11s**	16.89s	
year	5.77s	15.75s	0.14s	**0.1s**	44.82s	
covtype	63.45s	173.92s	[illegible]s	**1.29s**	445.89s	
epsilon	14.84s	29s	0.37s	**0.28s**	OOM	

20x

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	0.16s	**0.11s**	17.70s	
year	6.18s	10.14s	0.14s	**0.10s**	OOM	
covtype	67.12s	158.3s	1.47s	**1.29s**	446s	
epsilon	14.13s	26.03s	0.36s	**0.28s**	OOM	

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	8.1s	0.15s	0.11s	17.55s	**0.02s**
year	2.33s	17.23s	0.15s	0.10s	45.12s	**0.03s**
covtype	47.64s	24.77s	0.32s	0.26s	62.81s	**0.06s**
epsilon	11.22s	26.03s	0.36s	0.29s	OOM	**0.14s**

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	0.16s	0.11s	16.89s	**0.02s**
year	5.77s	15.75s	0.14s	0.1s	44.82s	**0.03s**
covtype	63.45s	173.92s	1.47s	1.29s	445.89s	**0.25s**
epsilon	14.84s	29s	0.37s	0.28s	OOM	**0.14s**

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	0.16s	0.11s	17.70s	**0.02s**
year	6.18s	10.14s	0.14s	0.10s	OOM	**0.03s**
covtype	67.12s	158.3s	1.47s	1.29s	446s	**0.25s**
epsilon	14.13s	26.03s	0.36s	0.28s	OOM	**0.14s**

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	8.1s	0.15s	0.11s	17.55s	**0.02s**
year	2.33s	17.23s	0.15s	0.10s	45.12s	**0.03s**
covtype	47.64s	24.77s	0.32s	0.26s	62.81s	**0.06s**
epsilon	11.22s	26.03s	0.36s	0.28s	OOM	**0.14s**

60x

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	0.16s	0.11s	16.89s	**0.02s**
year	5.77s	15.75s	0.14s	0.1s	44.82s	**0.03s**
covtype	63.45s	173.92s	1.47s	1.29s	445.89s	**0.25s**
epsilon	14.84s	29s	0.37s	0.28s	OOM	**0.14s**

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	0.16s	0.11s	17.70s	**0.02s**
year	6.18s	10.14s	0.14s	0.10s	OOM	**0.03s**
covtype	67.12s	158.3s	1.47s	1.29s	446s	**0.25s**
epsilon	14.13s	26.03s	0.36s	0.28s	OOM	**0.14s**

Tree-models Microbenchmark: Batch w/ GPU

	rf	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	8.1s	0.15s	0.11s	17.55s	**0.02s**
year	2.33s	17.23s	0.15s	0.10s	45.12s	**0.03s**
covtype	47.64s	24.77s	0.32s	0.26s	62.81s	**0.06s**
epsilon	11.22s	26.03s	0.36s	0.28s	OOM	**0.14s**

60x

	xgb	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	6.4s	0.16s	0.11s	16.89s	**0.02s**
year	5.77s	15.75s	0.14s	0.1s	44.82s	**0.03s**
covtype	63.45s	173.92s	1.47s	1.29s	445.89s	**0.25s**
epsilon	14.84s	29s	0.37s	0.28s	OOM	**0.14s**

	lgbm	onnx-ml	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	6.59s	0.16s	0.11s	17.70s	**0.02s**
year	6.18s	10.14s	0.14s	0.10s	OOM	**0.03s**
covtype	67.12s	158.3s	1.47s	1.29s	446s	**0.25s**
epsilon	14.13s	26.03s	0.36s	0.28s	OOM	**0.14s**

200x

Tree-models Microbenchmark: Batch w/ GPU

	rf	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	!SUPPORTED	0.15s	0.11s	17.55s	**0.02s**
year	2.33s	!SUPPORTED	0.15s	0.10s	45.12s	**0.03s**
covtype	47.64s	!SUPPORTED	0.32s	0.26s	OOM	**0.06s**
epsilon	11.22s	!SUPPORTED	0.36s	0.29s	OOM	**0.14s**

	xgb	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	0.08s	0.16s	0.11s	16.89s	**0.02s**
year	5.77s	0.09s	0.14s	0.1s	44.82s	**0.03s**
covtype	63.45s	!SUPPORTED	1.47s	1.29s	OOM	**0.25s**
epsilon	14.84s	0.29s	0.37s	0.28s	OOM	**0.14s**

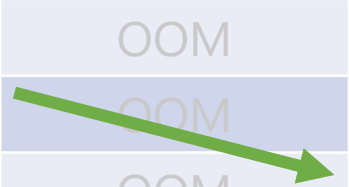
	lgbm	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	0.15s	0.16s	0.11s	17.70s	**0.02s**
year	6.18s	0.09s	0.14s	0.10s	OOM	**0.03s**
covtype	67.12s	!SUPPORTED	1.47s	1.29s	OOM	**0.25s**
epsilon	14.13s	0.29s	0.36s	0.28s	OOM	**0.14s**

Tree-models Microbenchmark: Batch w/ GPU

	rf	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	!SUPPORTED	0.15s	0.11s	17.55s	**0.02s**
year	2.33s	!SUPPORTED	0.15s	0.10s	45.12s	**0.03s**
covtype	47.64s	!SUPPORTED	0.32s	0.26s	OOM	**0.06s**
epsilon	11.22s	!SUPPORTED	0.36s	0.29s	OOM	**0.14s**

	xgb	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	0.08s	0.16s	0.11s	16.89s	**0.02s**
year	5.77s	0.09s	0.14s	0.1s	44.82s	**0.03s**
covtype	63.45s	!SUPPORTED	1.47s	1.29s	OOM	**0.25s**
epsilon	14.84s	0.29s	0.37s	0.28s	OOM	**0.14s**

	lgbm	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	0.15s	0.16s	0.11s	17.70s	**0.02s**
year	6.18s	0.09s	0.14s	0.10s	OOM	**0.03s**
covtype	67.12s	!SUPPORTED	1.47s	1.29s	OOM	**0.25s**
epsilon	14.13s	0.29s	0.36s	0.28s	OOM	**0.14s**

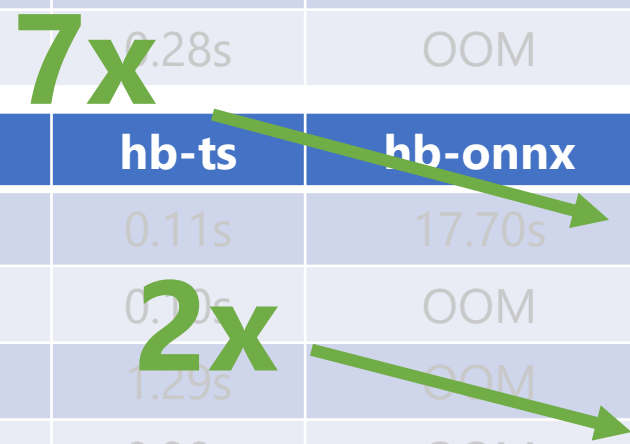


Tree-models Microbenchmark: Batch w/ GPU

	rf	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.52s	!SUPPORTED	0.15s	0.11s	17.55s	**0.02s**
year	2.33s	!SUPPORTED	0.15s	0.10s	45.12s	**0.03s**
covtype	47.64s	!SUPPORTED	0.32s	0.26s	OOM	**0.06s**
epsilon	11.22s	!SUPPORTED	0.36s	0.29s	OOM	**0.14s**

	xgb	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	2.01s	0.08s	0.16s	0.11s	16.89s	**0.02s**
year	5.77s	0.09s	0.14s	0.1s	44.82s	**0.03s**
covtype	63.45s	!SUPPORTED	1.47s	1.29s	OOM	**0.25s**
epsilon	14.84s	0.29s	0.37s	0.28s	OOM	**0.14s**

	lgbm	RAPIDS	hb-pt	hb-ts	hb-onnx	hb-tvm
fraud	3.76s	0.15s	0.16s	0.11s	17.70s	**0.02s**
year	6.18s	0.09s	0.14s	0.10s	OOM	**0.03s**
covtype	67.12s	!SUPPORTED	1.47s	1.29s	OOM	**0.25s**
epsilon	14.13s	0.29s	0.36s	0.28s	OOM	**0.14s**



Currently Supported Operators

Operator Group	Supported Operators
Linear Classifiers	Logistic Regression, Linear SVC, SVC, NuSVC, SGDClassifier, LogisticRegressionCV
Tree Methods	DecisionTreeClassifier, RandomForestClassifier/Regressor, GradientBoostingClassifier/Regressor, XGBClassifier/Regressor, LGBMClassifier/Regressor
Neural Networks	MLPClassifier
Others	BernouliNB, KMeans
Feature Selectors	SelectKBest
Decomposition	PCA, TruncatedSVD
Feature Pre-Processing	SimpleImputer, Imputer, ColumnTransformer, RobustScaler, MaxAbsScaler, MinMaxScaler, StandardScaler, Binarizer, KBinsDiscretizer, Normalizer, PolynomialFeatures, OneHotEncoder, LabelEncoder, FeatureHasher
Text Feature Extractor	CountVectorizer

Conclusions

- **Hummingbird**: Compiles ML Pipelines into tensor operations for better inference performance

- **Idea**: use Neural Network frameworks to solve classical ML system problems

- **Results**: faster than current custom implementations (e.g., C++ and CUDA)
  1. With higher flexibility (run both on CPU and GPU), and
  2. Less engineering effort

Microsoft

Thank you!

mainterl@microsoft.com