TreeFlow: Going Beyond Tree-based Parametric Probabilistic Regression

Patryk Wielopolski, Maciej Zięba

Motivation

- **Fact #1:** Tree-based ensembles excel in classification and regression with mixed-type variable tabular data, e.g., CatBoost.
- **Fact #2:** Current approaches use Gaussian or parametric distributions for uncertainty modeling, e.g., CatBoost, NGBoost, PGBM.
- **Fact #3:** Existing methods struggle to handle multi-modal distributions and do not support high-dimensional probabilistic predictions.

Here comes the TreeFlow!

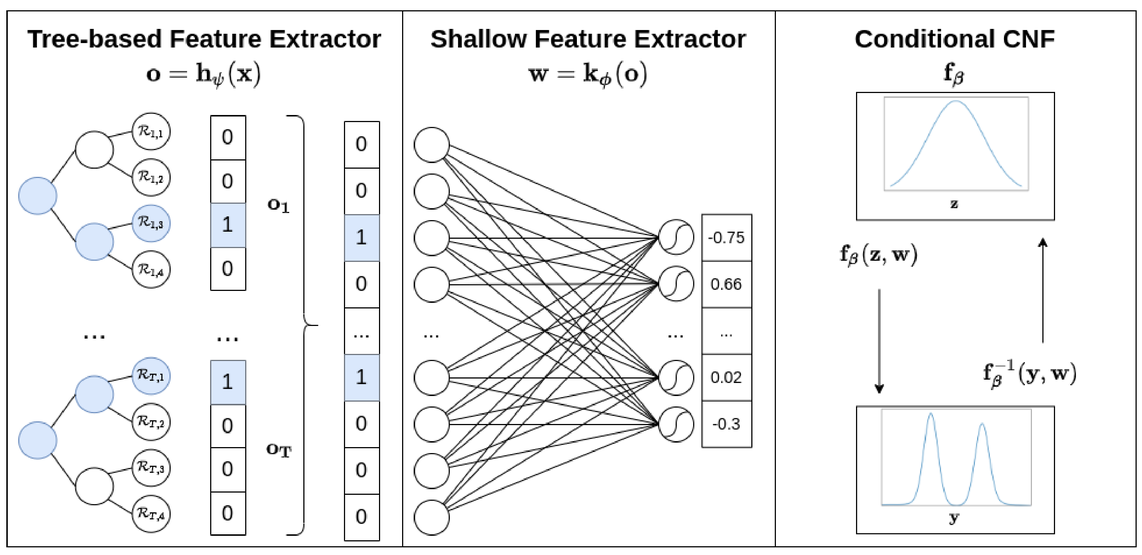
- **Regression model** for **tabular data**
- **Numerical** and **categorical** data
- **Univariate** and **multivariate** targets
- **Non-Gaussian**, **non-parametric** distributions
- **Probabilistic** and **deterministic** predictions

Paper

Code

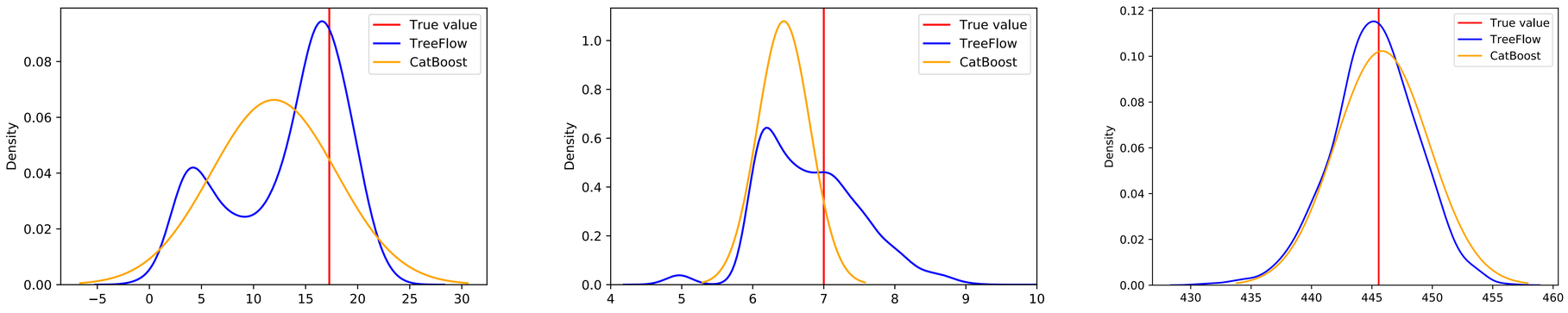
Connect

Method

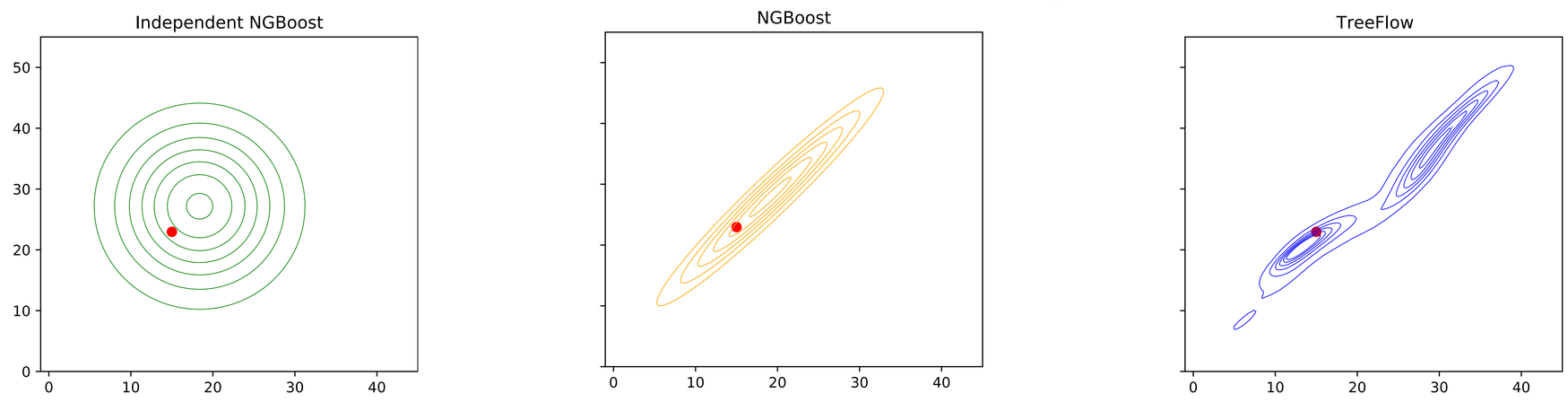


Tree-based Feature Extractor - extract the vector of binary features from the structure of the tree-based ensemble model.
Shallow Feature Extractor - a shallow neural network; maps high-dimensional binary vectors to low-dimensional feature space.
Conditional Continuous Normalizing Flow - takes previous vector as a conditioning factor; models complex probability distribution.

Univariate Flexible Probabilistic Regression



Multivariate Flexible Probabilistic Regression



Univariate Regression on Mixed-type Data

Probabilistic Regression

Dataset	NLL			CRPS		
	CatBoost	PGBM	TreeFlow	CatBoost	PGBM	TreeFlow
Avocado	-0.40 ± 0.01	-0.45 ± 0.01	**-0.47 ± 0.03**	0.0992 ± 0.0018	0.0870 ± 0.0013	**0.0854 ± 0.0024**
BigMart	-0.05 ± 0.02	**-0.10 ± 0.02**	-0.08 ± 0.02	0.1270 ± 0.0021	**0.1259 ± 0.0023**	0.1294 ± 0.0027
Diamonds	-1.80 ± 0.02	-1.41 ± 0.76	**-1.94 ± 0.03**	0.0222 ± 0.0002	0.0447 ± 0.0474	**0.0210 ± 0.0005**
Diamonds 2	-1.89 ± 0.02	-1.24 ± 0.83	**-2.14 ± 0.05**	0.0217 ± 0.0002	0.0461 ± 0.0504	**0.0197 ± 0.0005**
Laptop	-0.89 ± 0.08	**-0.97 ± 0.09**	-0.74 ± 0.13	0.0572 ± 0.0049	**0.0474 ± 0.0034**	0.0563 ± 0.0043
Pak Wheel	-1.40 ± 0.05	-0.53 ± 0.02	**-1.60 ± 0.03**	0.0362 ± 0.0006	0.0813 ± 0.0009	**0.0327 ± 0.0007**
Sydney	-0.54 ± 0.04	0.20 ± 1.02	**-0.66 ± 0.01**	0.0726 ± 0.0011	0.2383 ± 0.2646	**0.0721 ± 0.0008**

Univariate Regression on Mixed-type Data

Deterministic Regression

Dataset	RMSE				
	CatBoost	PGBM	TreeFlow(Avg)	TreeFlow(@1)	TreeFlow(@2)
Avocado	0.1939 ± 0.0043	**0.1624 ± 0.0024**	0.1676 ± 0.0058	0.1769 ± 0.0087	0.1713 ± 0.0066
BigMart	0.2284 ± 0.0039	**0.2274 ± 0.0040**	0.2335 ± 0.0045	0.2514 ± 0.0087	0.2480 ± 0.0083
Diamonds	0.0419 ± 0.0007	0.0403 ± 0.0006	0.0407 ± 0.0009	0.0445 ± 0.0015	**0.0343 ± 0.0017**
Diamonds 2	0.0421 ± 0.0006	0.0492 ± 0.0010	0.0398 ± 0.0006	0.0460 ± 0.0014	**0.0364 ± 0.0004**
Laptop	0.1028 ± 0.0092	**0.0848 ± 0.0063**	0.1014 ± 0.0082	0.1015 ± 0.0076	0.0958 ± 0.0058
Pak Wheel	0.0783 ± 0.0009	0.1630 ± 0.0018	0.0729 ± 0.0018	0.0796 ± 0.0021	**0.0654 ± 0.0047**
Sydney	0.1528 ± 0.0057	0.1561 ± 0.0047	0.1518 ± 0.0051	0.1721 ± 0.0041	**0.1361 ± 0.0066**

Univariate Regression on Numeric Data

Probabilistic Regression

Dataset	Deep. Ens.	CatBoost	NGBoost	RoNGBa	PGBM	TreeFlow
Concrete	3.06 ± 0.18	3.06 ± 0.13	3.04 ± 0.17	2.94 ± 0.18	**2.75 ± 0.21**	3.02 ± 0.15
Energy	1.38 ± 0.22	1.24 ± 1.28	0.60 ± 0.45	**0.37 ± 0.28**	1.74 ± 0.04	0.85 ± 0.35
Kin8nm	**-1.20 ± 0.02**	- 0.63 ± 0.02	-0.49 ± 0.02	-0.60 ± 0.03	-0.54 ± 0.04	-1.03 ± 0.06
Naval	**-5.63 ± 0.05**	-5.39 ± 0.04	-5.34 ± 0.04	-5.49 ± 0.04	-3.44 ± 0.04	-5.54 ± 0.16
Power	2.79 ± 0.04	2.72 ± 0.12	2.79 ± 0.11	2.65 ± 0.08	**2.60 ± 0.02**	2.65 ± 0.06
Protein	2.83 ± 0.02	2.73 ± 0.07	2.81 ± 0.03	2.76 ± 0.03	2.79 ± 0.01	**2.02 ± 0.02**
Wine	0.94 ± 0.12	0.93 ± 0.08	0.91 ± 0.06	0.91 ± 0.08	0.97 ± 0.20	**-0.56 ± 0.62**
Yacht	1.18 ± 0.21	0.41 ± 0.39	0.20 ± 0.26	1.03 ± 0.44	**0.05 ± 0.28**	0.72 ± 0.40
Year MSD	3.35 ± NA	3.43 ± NA	3.43 ± NA	3.46 ± NA	3.61 ± NA	**3.27 ± NA**

Multivariate Regression on Numeric Data

Probabilistic Regression

Dataset	Ind NGBoost	NGBoost	TreeFlow
Parkinsons	6.86	5.85	**5.26**
scm20d	94.40	94.81	**93.41**
WindTurbine	-0.65	-0.67	**-2.57**
Energy	**166.90**	175.80	180.00
usFlight	9.56	8.57	**7.49**
Oceanographic	7.74±0.02	**7.73±0.02**	7.84±0.01