Differential Privacy for Machine Learning:

Accuracy, Interpretability, and Privacy in Explainable Boosting

Efstathios Chatziloizos, Xichen Zhang, Aymeric Behaegel

March 12, 2025

Table of Contents

Introduction

- Model interpretability is a handy feature to have on a model, it allows for human correction and thus better results.
- But sometimes, those models deal with sensitive data, like in medicine or finance. In that case, we would also like to have privacy guarantees on the data.
- This paper introduces a new type of model for differential privacy on explainable models based on boosting trees.

Definition of Differential Privacy

(ϵ, δ)-Differential Privacy (DP) Dwork, Roth, et al. (2014)

A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if for any neighboring databases d, d',

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(d') \in S] + \delta.$$

Drawbacks:

- (ϵ, δ)-DP lacks clear interpretability, making privacy guarantees hard for users and regulators to understand.
- It also has weak composition properties, causing overly loose bounds when combining multiple DP algorithms.

Composition Theorem

Composition of Differential Privacy

If $\mathcal{M}_1, \ldots, \mathcal{M}_n$ are (ϵ_i, δ_i)-DP mechanisms, their composition satisfies:

$$(\sum_i \epsilon_i, \sum_i \delta_i)\text{-DP.}$$

Implications:

- Privacy degrades when applying multiple DP mechanisms
- Tight bounds on ϵ are necessary for practical applications.

Gaussian Differential Privacy (GDP)

Gaussian Differential Privacy (GDP) Dong, Roth, and Su (2022)

Define the Gaussian mechanism M as

$$M(D) = \theta(D) + \xi, \quad \text{where } \xi \sim \mathcal{N}\Big(0, \frac{\Delta^2}{\mu^2}\Big).$$

Then, M is μ-GDP.

- Ensures μ-GDP, where single parameter μ controls the privacy level.

k-fold Composition of GDP Mechanisms

Given $M_1, M_2, \ldots, M_k$ with μ_i-GDP, their composition satisfies:

$$\sqrt{\textstyle\sum_{i=1}^{k} \mu_i^2}\text{-GDP}.$$

- Provides a tighter bound than standard composition.

Conversion from GDP to (ϵ, δ)-DP

GDP to DP Conversion

A mechanism is μ-GDP if and only if it satisfies (ϵ, δ)-DP, where:

$$\delta = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^{\epsilon}\,\Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right).$$

Key Takeaways:

- Provides a direct link between Gaussian DP and traditional DP.
- Enables more refined privacy analysis in practical settings.

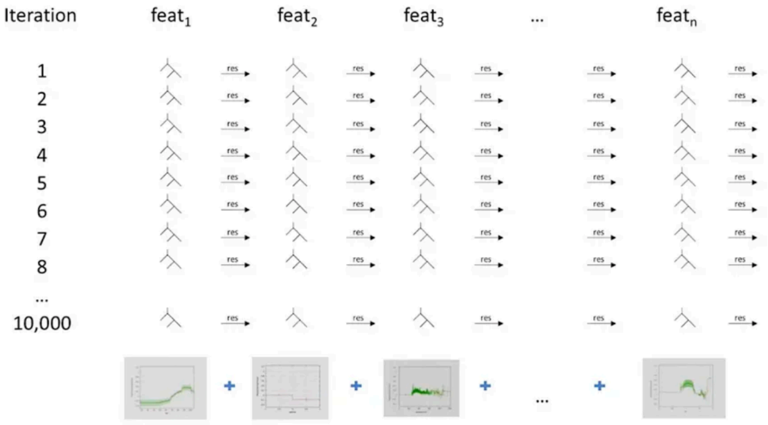
Explainable Boosting Machines (EBM): Overview

- **Glass-box models** (linear regression, decision trees) offer interpretability but typically sacrifice accuracy.
- **Explainable Boosting Machines (EBMs)** Nori et al. 2019 address this trade-off by integrating:
  - Generalized Additive Models (GAM)
  - Gradient Boosting Decision Trees (GBDT)
- EBMs represent predictions as additive combinations of shape functions:

$$g(\mathbb{E}[Y]) = \beta + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik}),$$

  where:
  - Y is the response, β an intercept.
  - f_k are feature-specific shape functions.
  - g is a link function (e.g., identity for regression, logit for classification).

EBM Training: Gradient Boosting and Splits



- EBMs use **gradient boosting** to iteratively refine shape
- **Key idea:** Each feature is updated sequentially using shallow decision trees restricted to individual features (cyclic boosting).

Differentially Private EBM (DP-EBM): Overview

- DP-EBM integrates differential privacy into Explainable Boosting Machines.
- Privacy budget (ϵ, δ) is split into:
  - **Histogram:** $((1-\tau)\epsilon, \delta/2)$
  - **Tree:** $(\tau\epsilon, \delta/2)$
- Converted to GDP
- Two key steps:
  - DP Binning
  - Noisy residual updates

Algorithm 2 Differentially Private Explainable Boosting

Input: $X, y, E, \eta, m, R, \epsilon, \delta$

Output: $\{f_k : H_k \to \mathbb{R}\}_{k=1}^{K}$

Initialization:

- For $i = 1, \dots, n$: $r_i^0 \leftarrow y_i$.
- For each feature $k = 1, \dots, K$:
  - Compute $a_k = \min_i X_{i,k}$ and $b_k = \max_i X_{i,k}$.
  - Privately bin data: $\hat{H}_k = DPBin(X[:, k], \epsilon_{\text{bin}})$
  - Set $f_k^0(b) \leftarrow 0$ for every $b \in \hat{H}_k$.

Main Loop:

for $e = 1, \dots, E$ **do**
    for $k = 1, \dots, K$ **do**
        Select Random splits $S_0, S_1, \dots, S_m \subseteq H_k$
        for $\ell = 0, \dots, m$ **do**
            $T \leftarrow \eta \cdot \sum_{b \in S_\ell} \sum_{i \in \mathcal{I}_k(b)} r_i^t$
            $\hat{T} \leftarrow T + \sigma \cdot \eta R \cdot \mathcal{N}(0, 1)$
            $\mu \leftarrow \frac{\hat{T}}{\sum_{b \in S_\ell} \hat{H}_k(b)}$
            for each $b \in S_\ell$ **do**
                $f_k^t(b) \leftarrow f_k^t(b) + \mu$
            end for
        end for
        for $i = 1, \dots, n$ **do**
            $r_i^{t+1} \leftarrow y_i - \sum_{k=1}^{K} f_k^t(\rho(H_k, X_{i,k}))$
        end for

DP-EBM: Noise Injection and Advantages

- **Noise Injection Strategy:**
  - **DP Binning:** Gaussian noise added to histograms based on privacy budget ϵ_{bin}.
  - **Tree Construction:** Aggregated residual sums perturbed with Gaussian noise:

$$\hat{T} = T + \sigma \cdot \eta R \cdot \mathcal{N}(0, 1)$$

- **Shape Function and Tree Splitting:**
  - Splits randomly selected within histogram bins.
  - Leaf nodes updated iteratively with noisy aggregates, preserving privacy.
- **Advantages of GDP Analysis:**
  - Tighter privacy composition.
  - easy to track our budget

Comparison of Non-DP Models (Test AUC)

Dataset	Model	Test AUC	Std
Breast-cancer	LR	**0.994**	0.006
	RF-100	0.992	0.009
	XGB	0.992	0.010
	APLR	0.993	0.006
	EBM	**0.994**	0.009
Telco-churn	LR	0.808	0.014
	RF-100	0.824	0.002
	XGB	0.822	0.004
	APLR	0.849	0.003
	EBM	**0.853**	0.004
Adult	LR	0.907	0.003
	RF-100	0.903	0.002
	XGB	0.928	0.001
	APLR	0.927	0.002
	EBM	**0.929**	0.002
Credit-fraud	LR	0.980	0.003
	RF-100	0.950	0.007
	XGB	0.983	0.002
	APLR	0.979	0.007
	EBM	**0.982**	0.005

Table: Test AUC (mean $\pm$ std) for Standard (non-DP) models.

DP-EBM on Adult Dataset: Claimed vs. Recreated

ϵ	**Claimed DP-EBM**		**Recreated DP-EBM**	
	Classic	GDP	Classic	GDP
0.5	0.875 ± 0.005	0.875 ± 0.005	0.826 ± 0.003	0.871 ± 0.003
1.0	0.880 ± 0.006	0.883 ± 0.005	0.859 ± 0.005	0.878 ± 0.003
2.0	0.886 ± 0.005	0.887 ± 0.004	0.877 ± 0.005	0.883 ± 0.004
4.0	0.889 ± 0.004	0.889 ± 0.004	0.875 ± 0.004	0.888 ± 0.004
8.0	0.890 ± 0.004	0.890 ± 0.004	0.887 ± 0.004	0.893 ± 0.005

Table: DP-EBM test AUC on the Adult dataset for various ϵ values. (Claimed values are from the paper and recreated results are our implementation.)

Our Custom DP-EBM Implementation

- **Data Loading:** Uses standard data loaders for Adult, Telco Churn, and Credit Card Fraud.
- **Binning Strategy:**
  - *Numeric features:* Quantile-based binning.
  - *Categorical features:* Direct mapping of unique values.
- **Cyclic Boosting:**
  - Iterates over features for a fixed number of epochs.
  - Updates each feature's *shape function* by computing the mean residual over each bin.
  - Injects Gaussian noise into the residual average update.
- **Privacy Parameters:**
  - Noise scale is computed as $\sigma = \sqrt{2\ln(1.25/\delta)}/\epsilon$.
  - A single global noise scale is used (without explicit per-round budget partitioning).
- **Prediction:** The additive predictions (across features) are passed through a sigmoid to obtain probabilities.

Differences from Algorithm 2 of the Paper

DP Binning:

- *Paper:* DPBin adds noise to bin counts/boundaries.
- *Ours:* Standard quantile binning, no DP noise.

Split Selection:

- *Paper:* Randomized splits to save privacy budget.
- *Ours:* Fixed bin edges, no randomness.

Noise Injection:

- *Paper:* Noise in aggregate residuals, budget split.
- *Ours:* Gaussian noise to mean residual update.

Ensembling:

- *Paper:* Outer/inner bagging for variance reduction.
- *Ours:* Single-model boosting, no bagging.

Complexity:

- *Paper:* Monotonicity, strict privacy accounting.
- *Ours:* A simplified toy model (cyclic boosting + noise injection).

Custom DP-EBM vs Official DP-EBM ($\epsilon = 4.0$)

- **Adult Dataset:**
  - **Custom DP-EBM Test AUC:** 0.8435
  - **Official DP-EBM (GDP) Test AUC:** 0.889 ± 0.004
- **Telco Churn Dataset:**
  - **Custom DP-EBM Test AUC:** 0.8225
  - **Official DP-EBM (GDP) Test AUC:** 0.839 ± 0.011
- **Credit Card Fraud Dataset:**
  - **Custom DP-EBM Test AUC:** 0.8754
  - **Official DP-EBM (GDP) Test AUC:** 0.969 ± 0.011

Our contribution

We tested the algorithm on two new datasets :

- Parkinson for regression
- Phishing for classification

Dataset	Domain	N	K	Task
Parkinson	Medicine	5,875	19	Reg.
Phishing	Security	11.055	30	Class.

Table: Statistics of datasets

Our contribution

Dataset	ϵ	custom	classic	gdb
Parkinson	0.1	29.71	30.4 ± 1.673	17.2 ± 0.749
	0.5	29.74	12.6 ± 0.168	11.4 ± 0.017
	2	29.71	10.6 ± 0.045	9.6 ± 0.044
	4	29.71	9.7 ± 0.119	9.4 ± 0.071
	8	29.71	9.4 ± 0.033	9.1 ± 0.044

Table: RMSE algorithm comparison on new dataset

Our contribution

Dataset	ϵ	custom	classic	gdb	logistic	rdm forest
Phishing	0.1	0.765	0.673 ± 0.005	0.914 ± 0.009	0.873 ± 0.008	0.757 ± 0.003
	0.5	0.540	0.948 ± 0.004	0.968 ± 0.003	0.888 ± 0.002	0.757 ± 0.003
	2	0.784	0.972 ± 0.003	0.977 ± 0.003	0.896 ± 0.005	0.757 ± 0.003
	4	0.943	0.975 ± 0.003	0.978 ± 0.002	0.897 ± 0.005	0.758 ± 0.003
	8	0.950	0.977 ± 0.003	0.979 ± 0.002	0.897 ± 0.005	0.758 ± 0.003

Table: AUROC algorithm comparison on new dataset

Critical analysis of results

Even though the paper shows very promising results, some critics can be made :

- limited to tabular data
- no comparison to deep DP models
- regularization through noise adding not always good

References I

Dong, Jinshuo, Aaron Roth, and Weijie J Su (2022). "Gaussian differential privacy". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84.1, pp. 3–37.

Dwork, Cynthia, Aaron Roth, et al. (2014). "The algorithmic foundations of differential privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4, pp. 211–407.

Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana (2019). "Interpretml: A unified framework for machine learning interpretability". In: *arXiv preprint arXiv:1909.09223*.