Auditing Black-Box Prediction Models for Data Minimization Compliance

by Bashir Rastegarpanah, Krishna Gummadi, Mark Crovella

Dhruv Shah

February 7, 2025

Table of contents

Introduction

Basic Definitions

Data Minimisation Principle: *"Personal data shall be adequate, relevant and limited to what is necessary for the purposes for which they are processed"*

Instability-Based Operationalization: Here the idea is that the auditor can test the need for a particular input feature, by checking the extent to which the outcomes change.

Key Features

1. Show how simple imputations can be leveraged for limiting data inputs at deployment time.
2. Define a data minimization guarantee that is based on a metric of model instability under different feasible simple imputations.
3. Define a probabilistic audit that provides a data minimization guarantee at certain confidence levels.
4. Auditing under the constraint on the number of queries to the prediction system.

Problem Setup

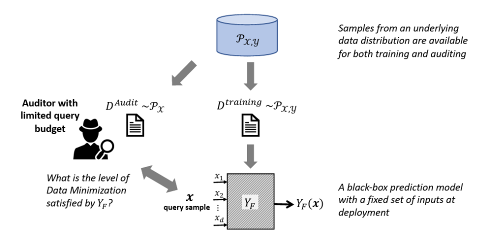

Figure 1: Problem Setup (Source: Appendix A of [1])

Formal Definitions

Notation:

- Prediction model: $\hat{Y}_F$ over features $F = \{f_1, \ldots, f_d\}$.
- Input space: $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$, Output space: $\mathcal{Y}$.
- Data follows distribution $P_{\mathcal{X}}$.
- **Goal:** Test if $\hat{Y}_F$ satisfies **data minimization** over $P_{\mathcal{X}}$.

Defining Instability:

- A feature f_j is unstable if its imputation alters predictions.
- Indicator function:

$$I_{\hat{Y}_F}(x, f_j, b) = \begin{cases} 1, & \text{if } \hat{Y}_F(x) \neq \hat{Y}_F(\tau_{f_j,b}(x)) \\ 0, & \text{otherwise} \end{cases}$$

Instability Metric:

- Probability that imputing f_j changes the prediction:

$$\beta_j^b = \mathbb{E}_{X \sim P_{\mathcal{X}}}[I_{\hat{Y}_F}(X, f_j, b)]$$

- Higher β_j^b means f_j is **more important** for the model.

Data Minimization Guarantee

Definition: A prediction model $\hat{Y}_F(x)$ satisfies data minimization at level β if there does not exist any feature f_j and any imputation value $b \in X_j$ such that $\beta_j^b < \beta$. The highest level β at which data minimization is satisfied constitutes the best data minimization guarantee an auditor can offer for $\hat{Y}_F(x)$.

Intuitively, this is to ensure that every input feature is necessary, hence the auditor would want to find the largest β.

Probabilistic Data Minimization Guarantee

Idea: Audit the prediction model by observing the outputs for a limited number of query data points and provide a probabilistic guarantee about the data minimization level.

Definition: A prediction model $\hat{Y}_F(x)$ satisfies data minimization at level β with α percent confidence if the probability that $\beta_j^b < \beta$ for atleast one feature f_j and one imputation value $b \in X_j$ is less than or equal to $1 - \alpha$.

Audit Mechanisms For Data Minimisation Guarantees

Probabilistic Audit

Key Idea:

- A probabilistic approach using **Bayesian updating** to measure **uncertainty in model instability** under different imputations.

Method:

1. Assume a **prior distribution** for each instability measure β_j^b.
2. Treat the model instability as a **Bernoulli variable** $I_{\hat{Y}_F}(X, f_j, b)$.
3. **Bayesian Update Rule**:
   - Observations of how imputations affect predictions refine our belief about β_j^b.

Mathematical Formulation:

- Let:
  - $S_j^b = \#$ times imputation $\tau_{f_j,b}()$ **changes** the model's prediction.
  - $F_j^b = \#$ times the prediction **remains unchanged**.
- **Posterior Distribution Update:**

$$\beta_j^b \sim Beta(a + S_j^b, c + F_j^b)$$

  when the prior belief is $Beta(a, c)$.

Key Takeaways:

- Provides a probabilistic **uncertainty quantification** for model instability.
- Bayesian inference **improves estimates** of how much features affect predictions.

Auditing With a Limited Query Budget

In many/most real world cases, we might not have infinite number of queries to audit the model. Moreover, we would want to minimize the total number of queries used.
Idea: Cast the problem of allocating a query budget into a bandit framework.
The paper focusses on the following two tasks

- measuring the greatest data minimization level satisfied by a prediction model given a fixed query budget, and
- deciding whether or not data minimization is satisfied at a given level using the minimum number of system queries.

Multi-Armed Bandit Framework

Framework

Formulation:

- Each **arm** corresponds to a pair (f_j, b), where:
    - f_j is an input feature.
    - b is a feasible imputation value.
- The auditor **chooses an arm** and observes a binary reward:
    - Reward is sampled from a **Bernoulli distribution** with success probability β_j^b.
    - Success means the model's prediction **changed** after imputation.
- With each observation of an arm, evaluate $I_{\hat{Y}_F}(X, f_j, b)$ at a data point x drawn randomly from $P_{\mathcal{X}}$

Decision Problem: Fixed Confidence and Fixed Level

The auditor's goal is to guarantee with a given confidence α that whether or not a prediction system satisfies data minimization at a given level β.

Good Strategy: Tries to use a small number of queries to provide this guarantee.

Measurement Problem: Fixed Confidence and Fixed Budget

Here the auditor is given a fixed query budget and the goal is to measure β, the highest level of data minimization that the prediction system is guaranteed to satisfy, with a given confidence α.

References i

B. Rastegarpanah, K. P. Gummadi, and M. Crovella.
Auditing black-box prediction models for data minimization compliance.
In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

THANK YOU!