PLUTUS: Understanding Distribution Tailoring for Machine Learning

Jiwon Chang[1], Christina Dionysio[2], Fatemeh Nargesian[1], Matthias Boehm[2]

[1] University of Rochester [2] Technische Universität Berlin

Motivation

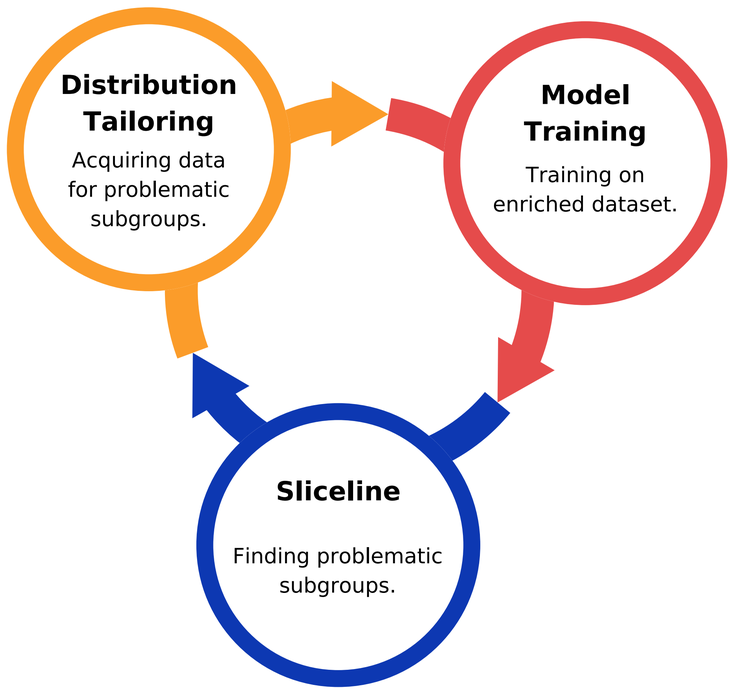
Our ML **model** is performing **poorly** on some **subgroups.**

Let's **diagnose** the model and **enrich** the train set.

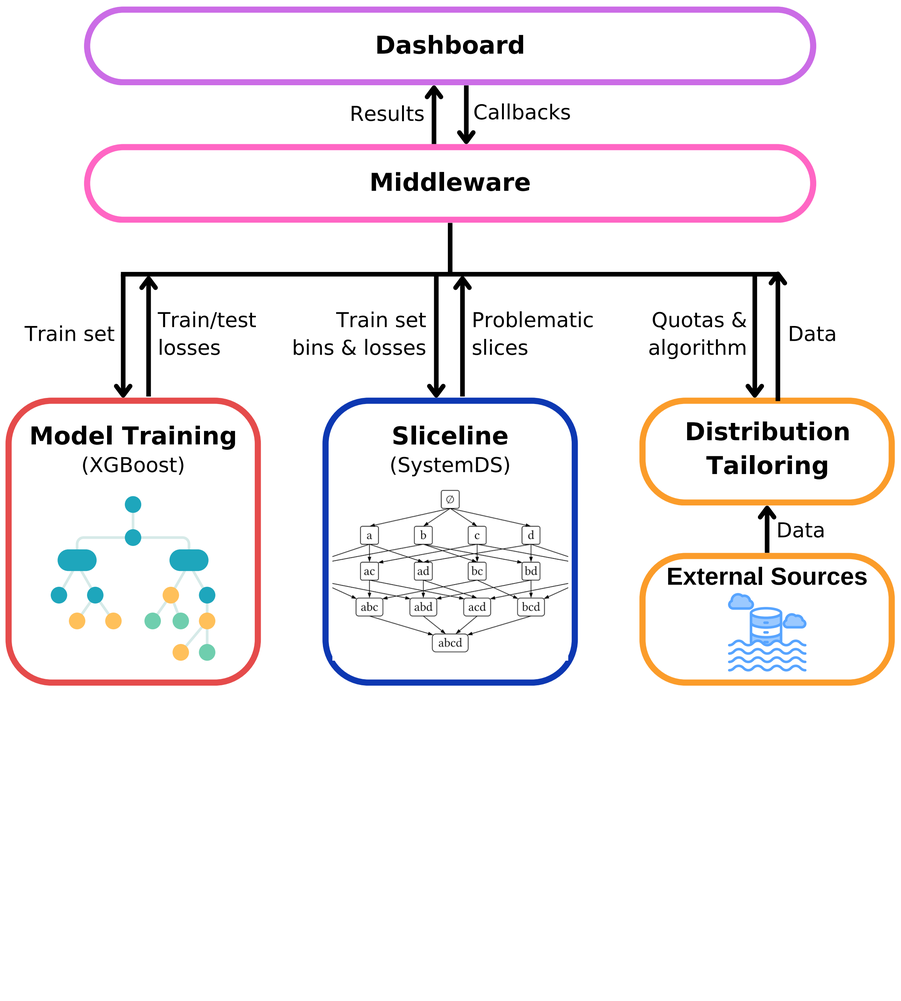
We can focus **data acquisition** on the **worst subgroups**!

I'd like the whole thing to be **interactive...**

Solution Sketch



Architecture



Key Concepts

A **subgroup** is a subset of subjects.

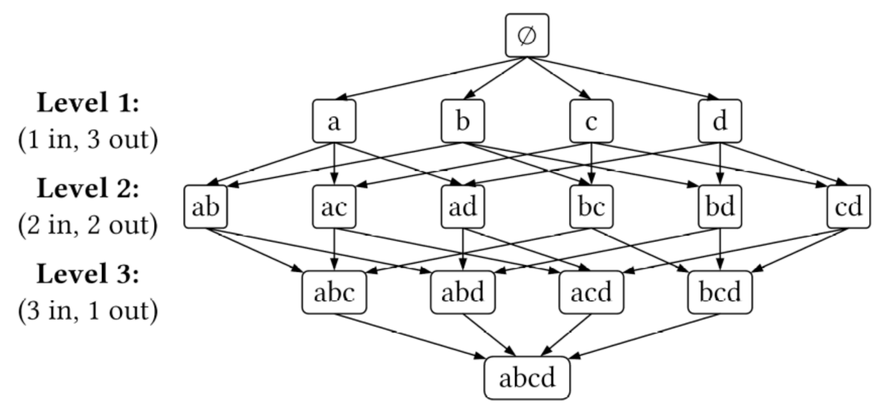
A **slice** is a subgroup specified by bins for a subset of columns in a table.
Example: $\text{Gender} = \text{Male} \wedge \text{Age} \in [0, 20)$

A **quota sample** is a non-probability sampling method where several subgroups have count requirements.
Example: 100 samples each from subgroups young male, old male, young female, old female.

Sliceline

We wish to find top K slices with **high contribution to overall error.**

$$\text{score} = \underbrace{\alpha\left(\frac{\text{slice error}}{\text{total error}} - 1\right)}_{\text{high error}} - \underbrace{(1-\alpha)\left(\frac{\text{slice size}}{\text{total size}} - 1\right)}_{\text{large size}}$$



Optimizations:
1. **Pruning** through monotonicity.
2. **Sparse matrix** multiplications.

Distribution Tailoring

Given a **quota for each slice** and a collection of external tables **unionable with train set**, we want to satisfy the quota with **minimal samples.**

PLUTUS supports:
- An adaptive sampling scheme (RatioColl).
- A zero-prior multi-armed bandit (ExploreExploit).
- Uniform random sampling baseline.

We prioritize slices that:
1. Have high **remaining quota.**
2. Are **rare** across all external sources.

References

1. Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2021. **Tailoring Data Source Distributions for Fairness-aware Data Integration.** PVLDB 14, 11.
2. Jiwon Chang, Bohan Cui, Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2024. **Data distribution tailoring revisited: cost-efficient integration of representative data.** The VLDB Journal.
3. Matthias Boehm et al. 2020. **SystemDS: A Declarative Machine Learning System for the End-to-End Data Science Lifecycle.** CIDR 2020.
4. Svetlana Sagadeeva and Matthias Boehm. 2021. **SliceLine: Fast, Linear-Algebra-based Slice Finding for ML Model Debugging.** In SIGMOD. ACM, 2290–2299.