Effective Methods to Estimate Machine Learning Performance

Yujie Wang, Shreya Shankar, Aditya Parameswaran

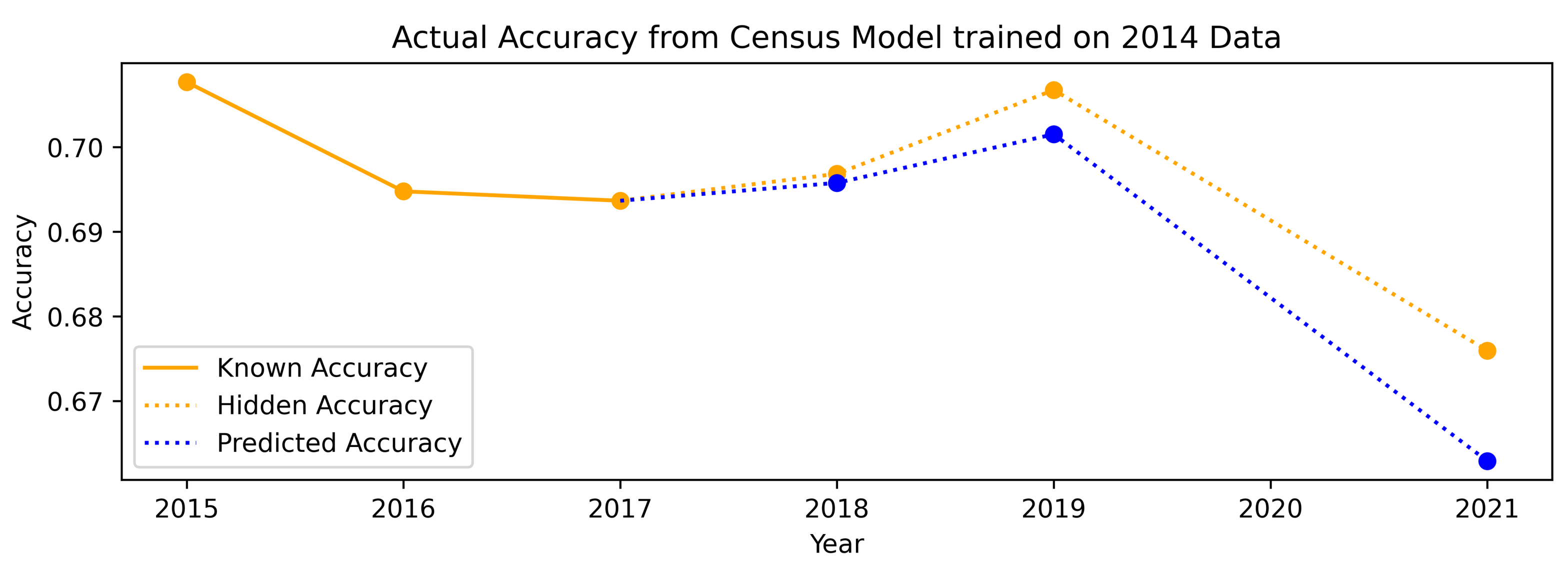
UC Berkeley

EPIC DATA lab

Motivation

- ML model accuracy, precision, and recall typically change during production due to data shift
- ML Practitioners often have limited access to ground truth labels, preventing true performance tracking
- Performance estimation techniques approximate performance metrics, but there is an estimation gap due to reducible and irreducible error.

Can we improve the current tools used to estimate ML performance?

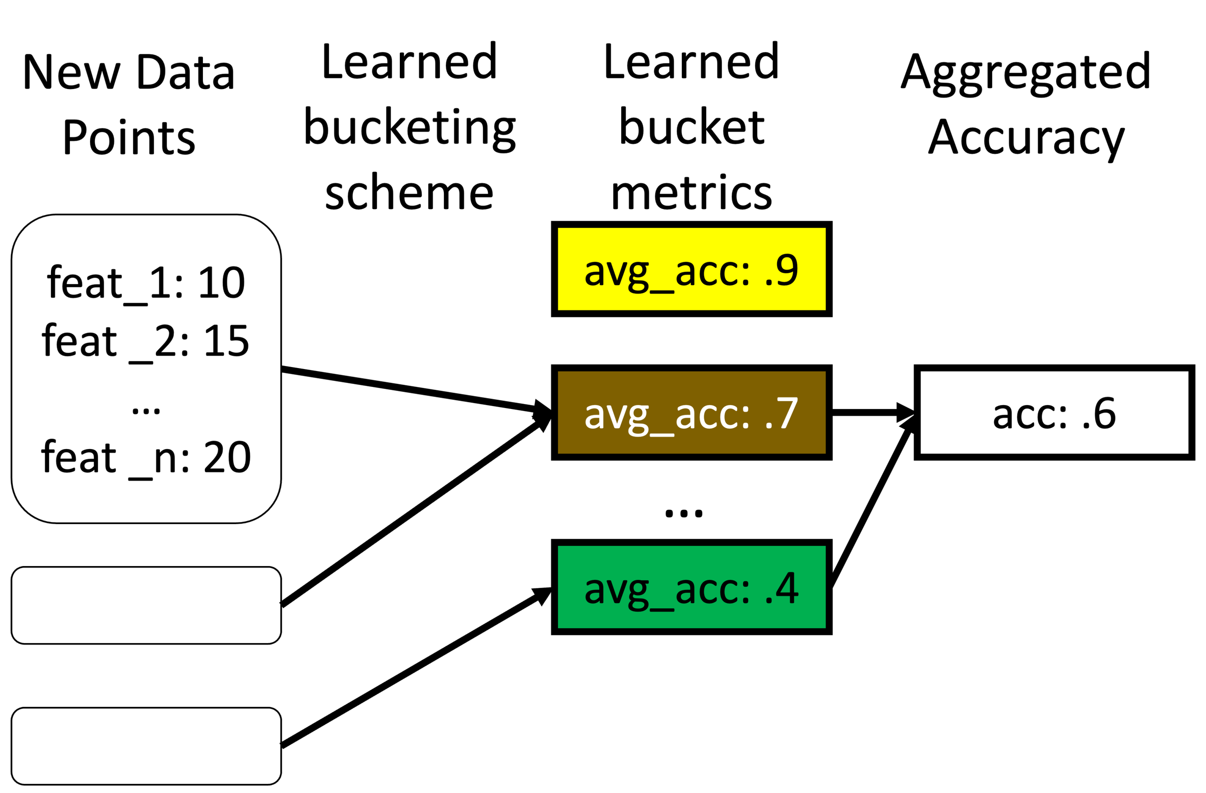


Introduction

Performance estimation for ML models broadly follows 2 types of techniques.

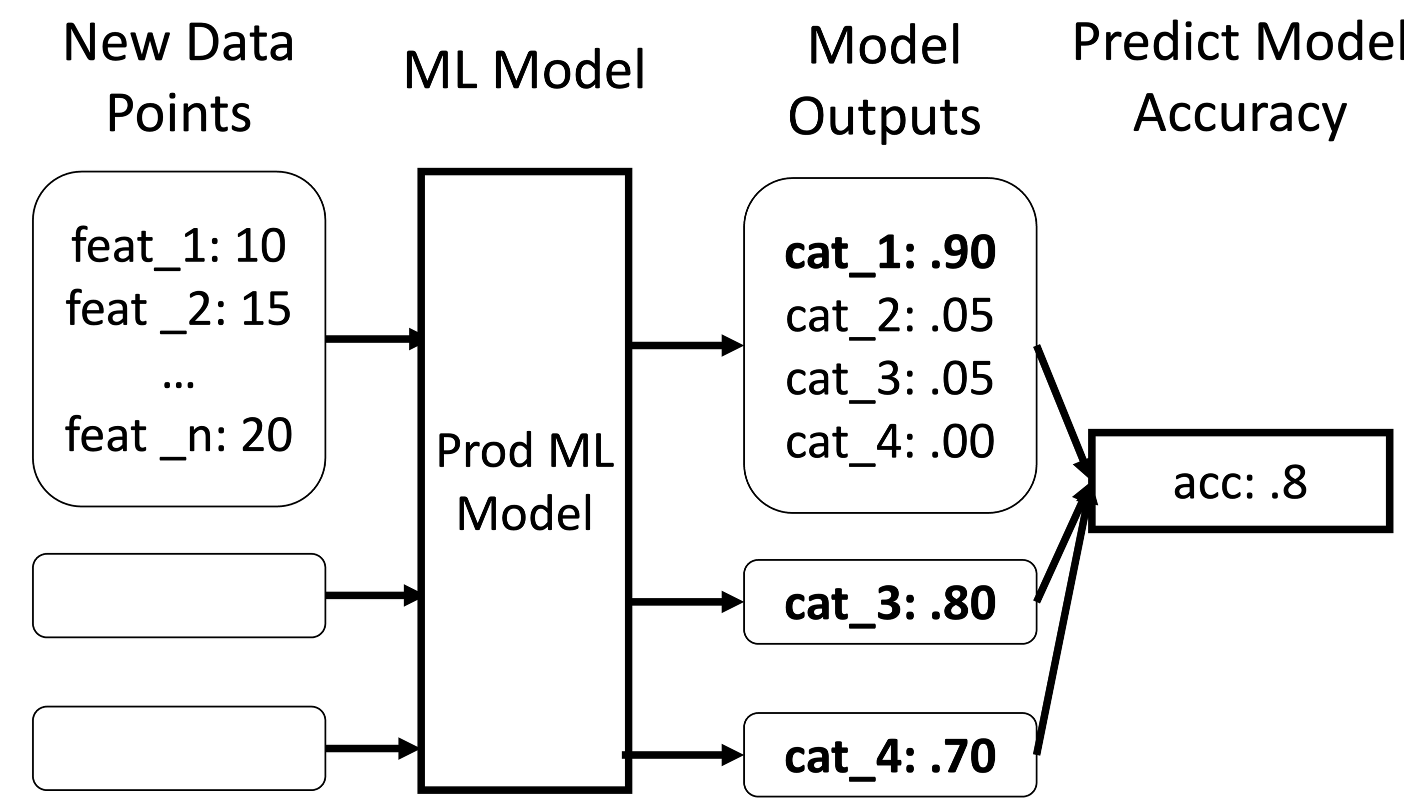
Type 1: Importance-Weighted

- Separate data into buckets by features, e.g., feat_1 < 10, feat_1 >= 10.
- Learn accuracy of validation dataset for each bucket



Problem: buckets based on individual features scale poorly with high-dimensional inputs (see Evaluation)

Type 2: Model Confidence

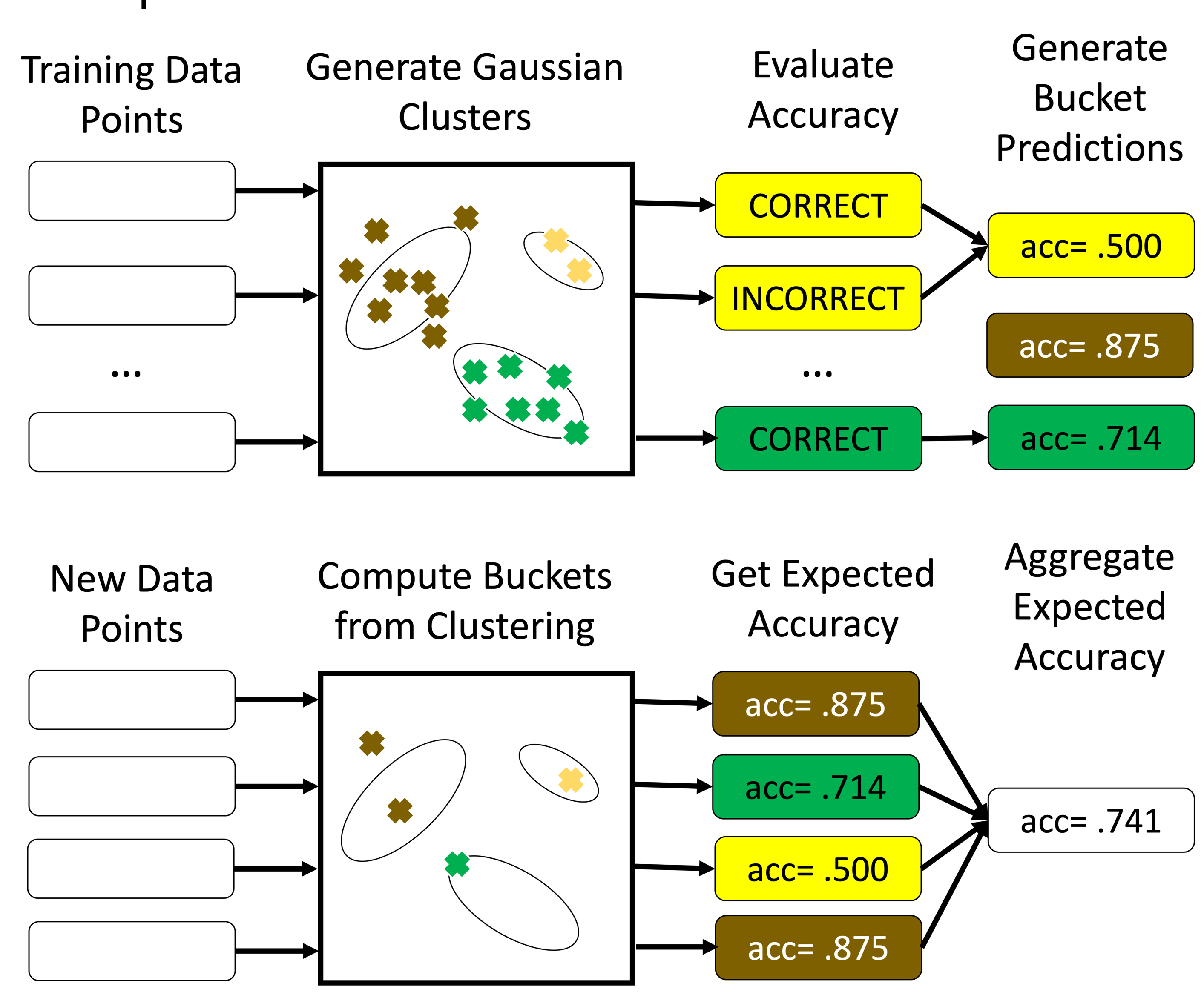


Problem: model confidence performs erratically when out of distribution

Methods Under Research

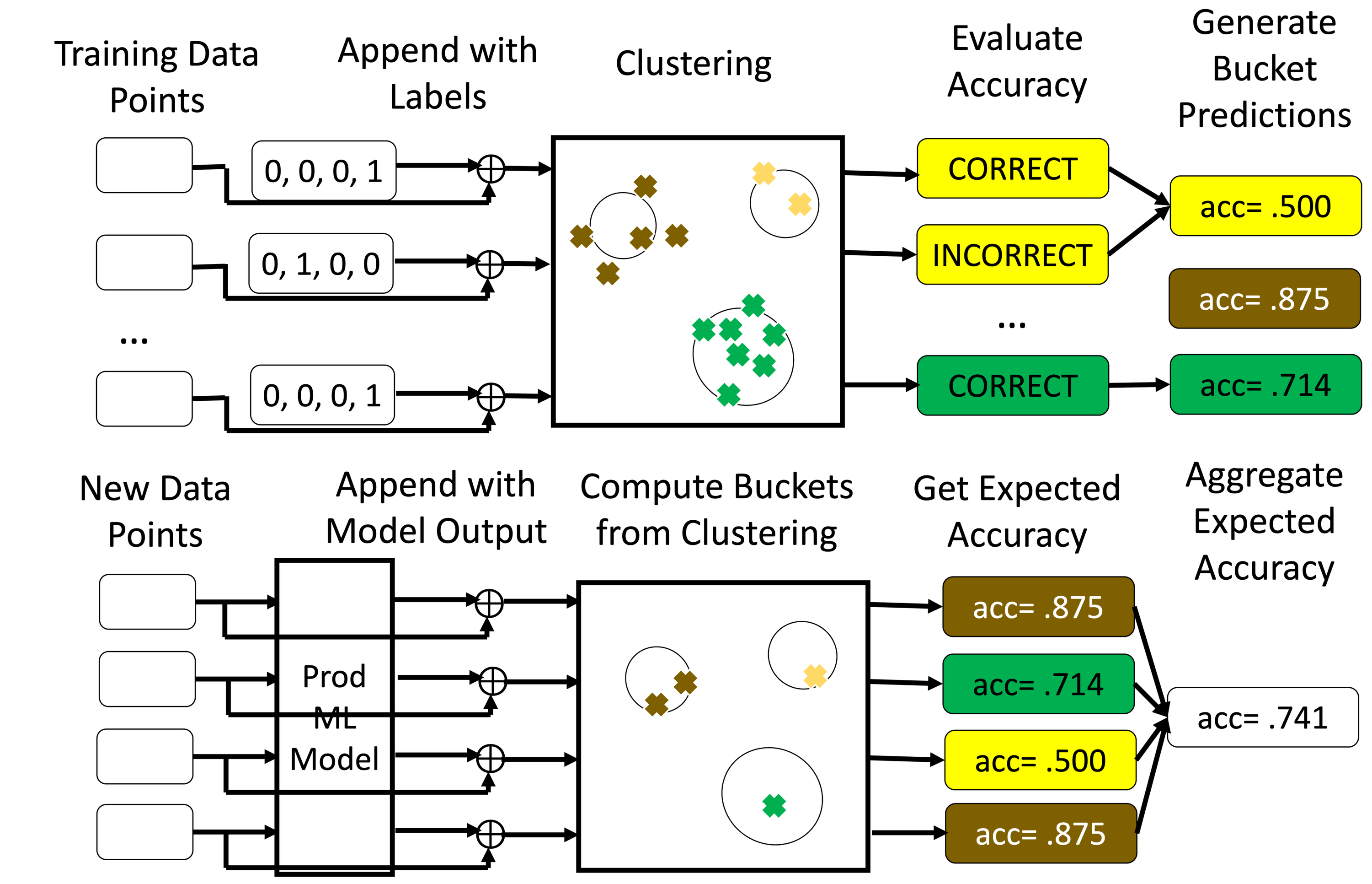
Novel Method 1: Gaussian Mixture Importance-Weighting

A type of Importance-Weighted Performance Prediction (Type 1) whose bucketing scheme is across multiple features.



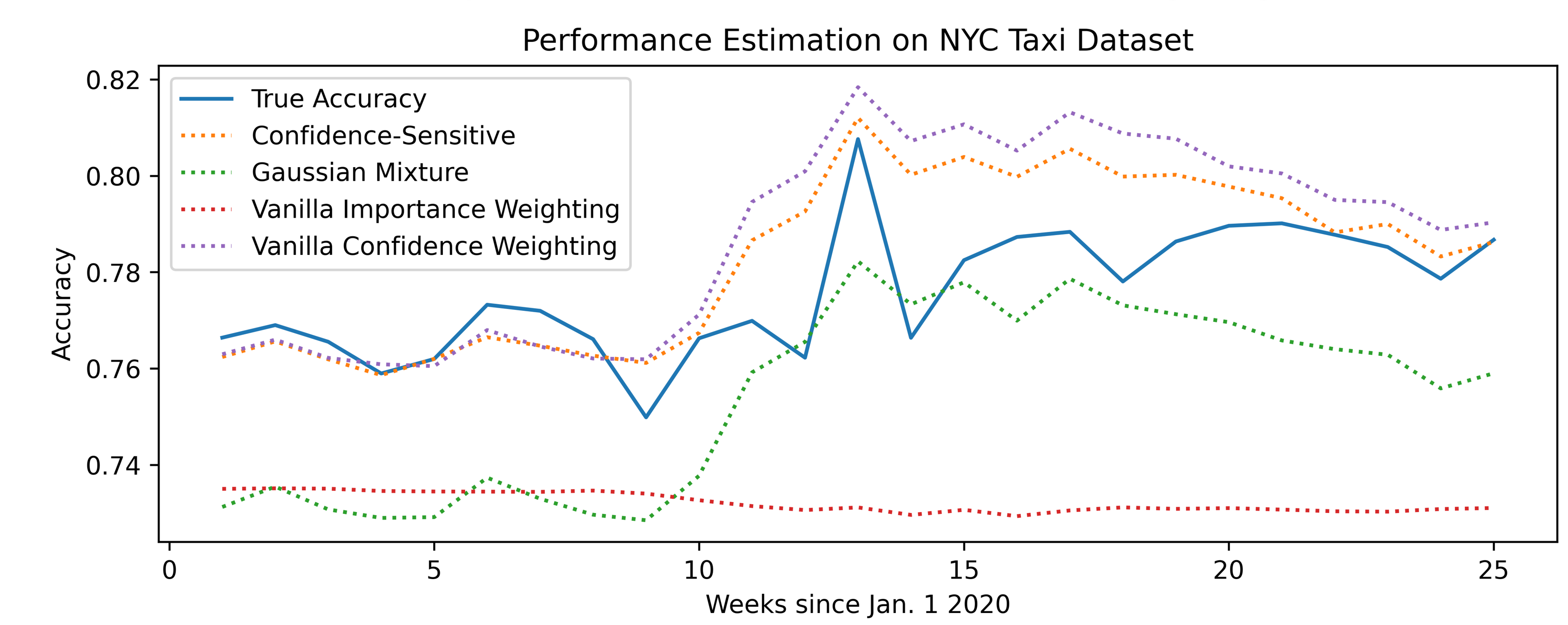
Contribution: Allows importance-weighted technique (Type 1) to generalize to higher dimensions.

Novel Method 2: Confidence-Sensitive Clustering

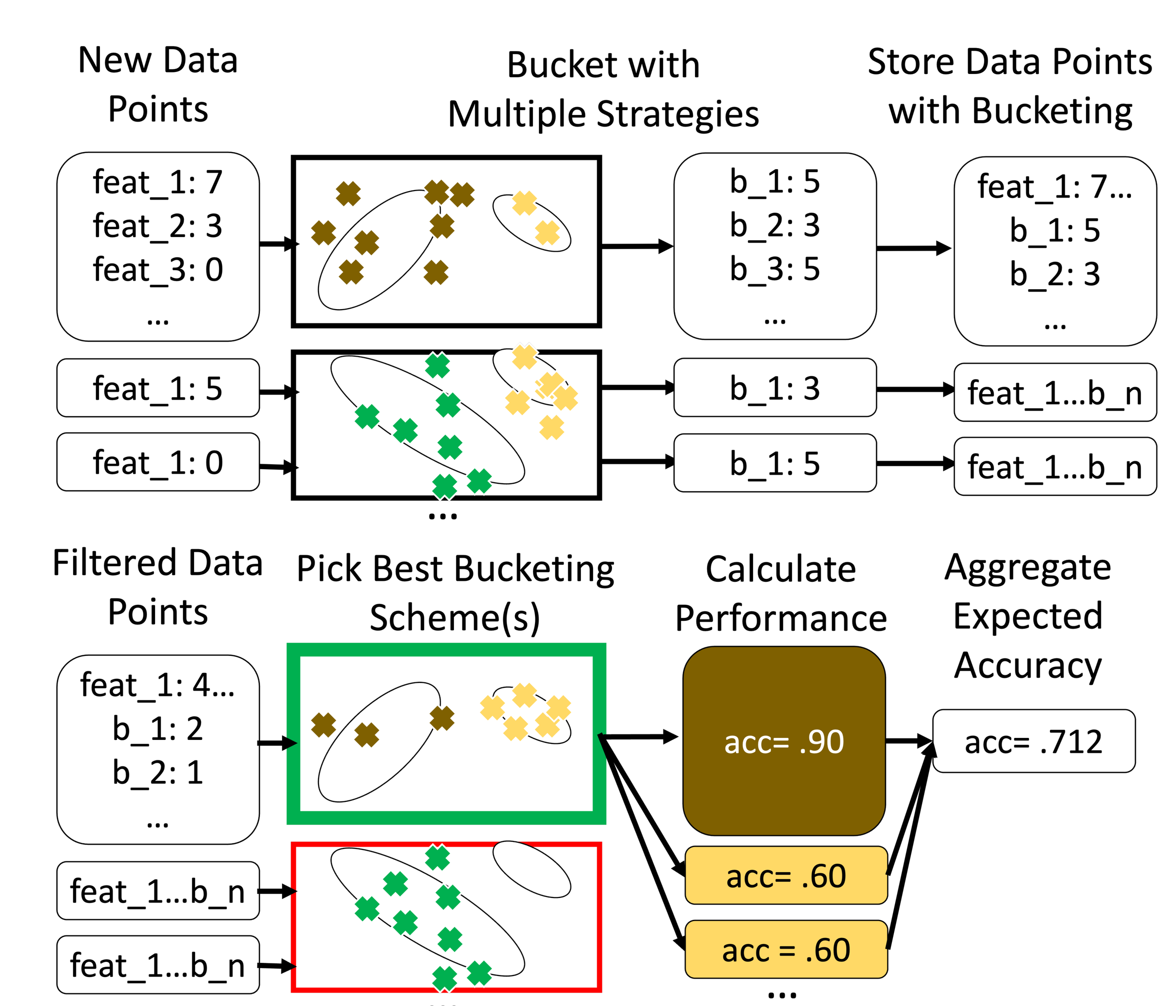


Contribution: Merges Importance-Weighting (Type 1) and Model Confidence (Type 2), allowing for confidence-aware clusters

Preliminary Evaluation on Challenging Dataset



Roadmap Vision: Dynamic Bucket Strategy



- Ensemble buckets enable better importance-weighting coverage
- Storage of bucket information allows for fast accuracy estimates for arbitrary data, enabling faster lookup of accuracy drop cause