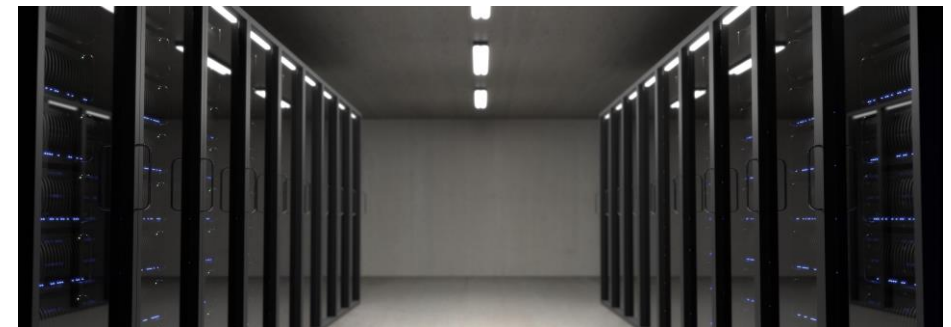


Intro

- Robust failure prediction can prevent downtime in many contexts.
- Modern hard-drives periodically measure and report usage and health statistics (called SMART metrics).
- This project looks at whether hard-drive failure can be predicted ahead of time based on these metrics.



Backblaze Hard Drive Data

- Backblaze has been publishing failure data for hard drives in their datacenters since 2014.
- For each drive, they record values of SMART metrics once per day along with whether the drive failed that day.
- There are ~20 different drive models and a total of 187 metrics, only a subset of which are available for a given drive model.
- We focus on 2021 data for one drive model (ST4000DM000).
 - After removing features with no or constant data, there are **22 SMART metrics** for this drive model.
 - There are 18,611 drives and a total of **324 failures**

<https://www.backblaze.com/b2/hard-drive-test-data.html>

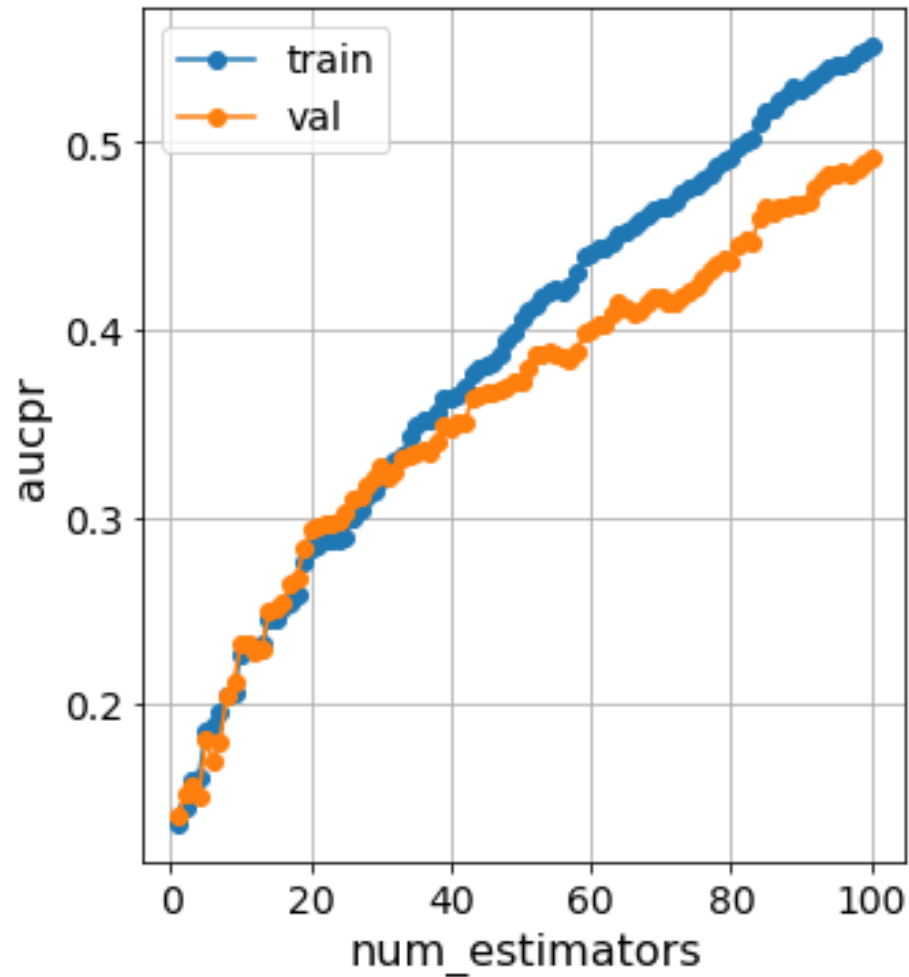
Modeling Approach

- **Problem Definition:** Predict whether a given drive will fail in the next **7 days**.
- **Feature Selection:** Use all 22 available features.
- **Model type:** Train a gradient-boosted decision tree model (**xgboost**).
- **Data sampling:** Use the latest 90 days for each drive. Use all failed drives + as many (324) randomly selected healthy drives. => 59,320 rows of data, **3.9%** of which belongs to the positive class.
- **Data splits:** **90/10** train/val split.
- **Accounting for Class Imbalance:** **Weight the failure rows** by the inverse of their relative frequency:

$$\text{failed rows weight scale} = \frac{\text{number of healthy rows}}{\text{number of failed rows}}$$

First Run with Default Parameters

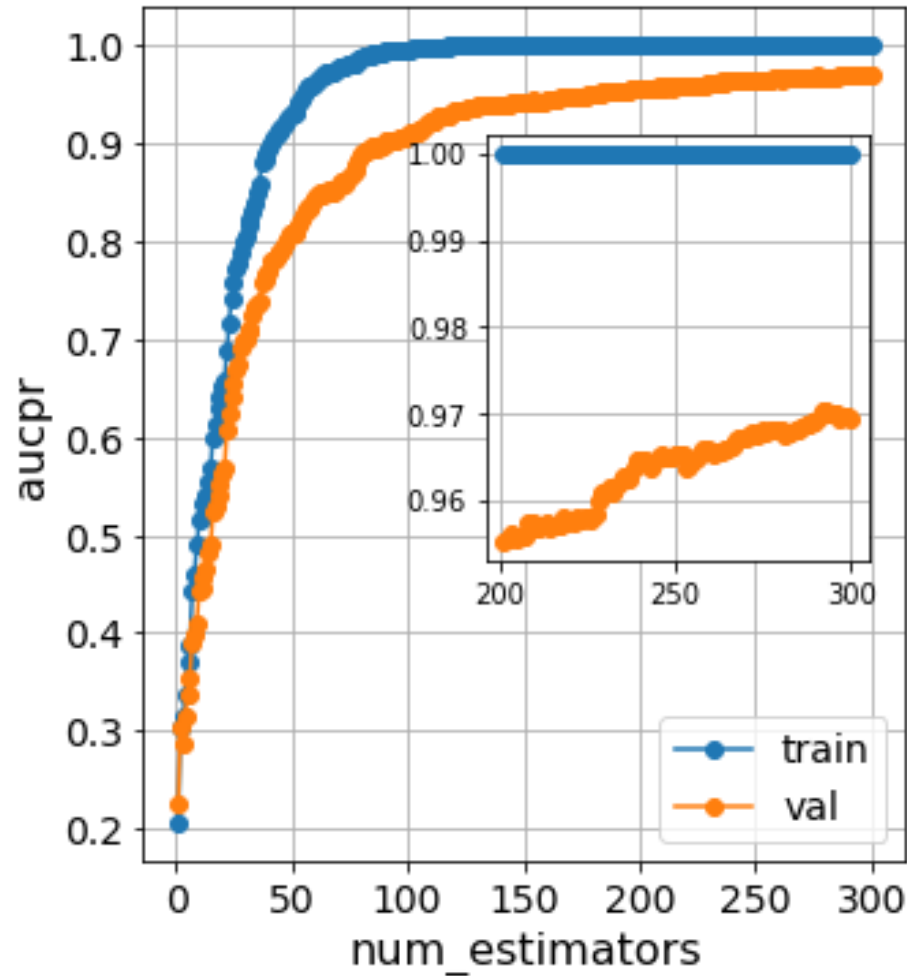
n_est=100, lr=0.3, ESR=None, gamma=0.0, max_depth=3
subsmp1=1.0, colssmpl_tree=1.0, reg_alpha=0.0, reg_lambda=0.0



ACC	87.2%	86.6%
RECALL	92.4%	87.2%
PRECISION	22.3%	20.9%
F1	35.9%	33.7%
	TRAIN	VAL

Increase max_depth and add more trees

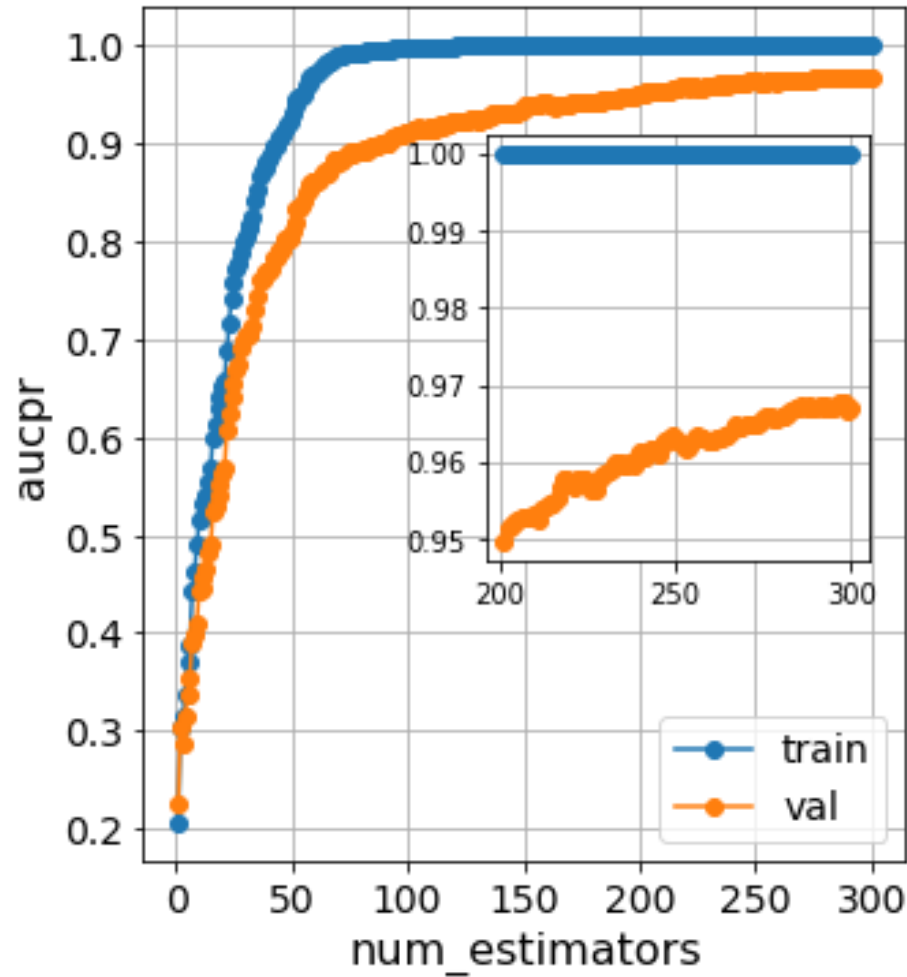
n_est=300, lr=0.3, ESR=None, gamma=0.0, max_depth=6
subsmpl=1.0, colssmpl_tree=1.0, reg_alpha=0.0, reg_lambda=0.0



ACC	100.0%	99.2%
RECALL	100.0%	90.7%
PRECISION	100.0%	90.0%
F1	100.0%	90.4%
	TRAIN	VAL

Add gamma

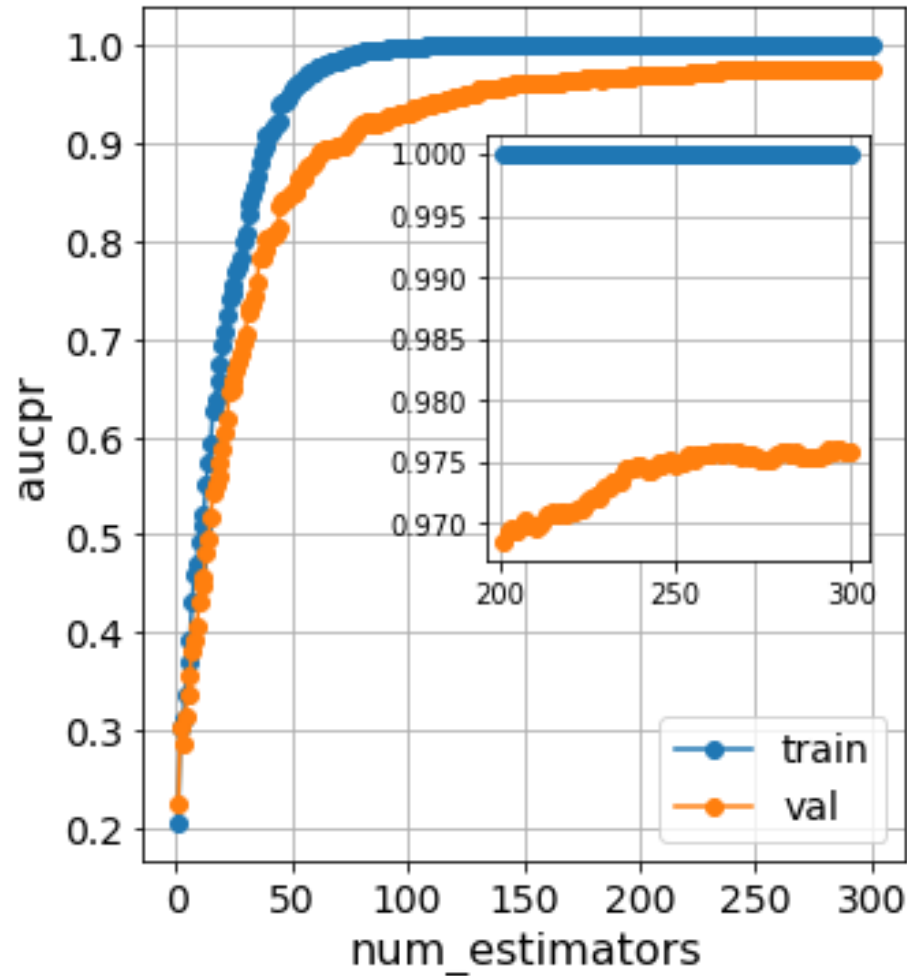
n_est=300, lr=0.3, ESR=None, gamma=0.05, max_depth=6
subsmpl=1.0, colssmpl_tree=1.0, reg_alpha=0.0, reg_lambda=0.0



ACC	100.0%	99.2%
RECALL	100.0%	88.5%
PRECISION	100.0%	90.5%
F1	100.0%	89.5%
	TRAIN	VAL

Add L1 regularization

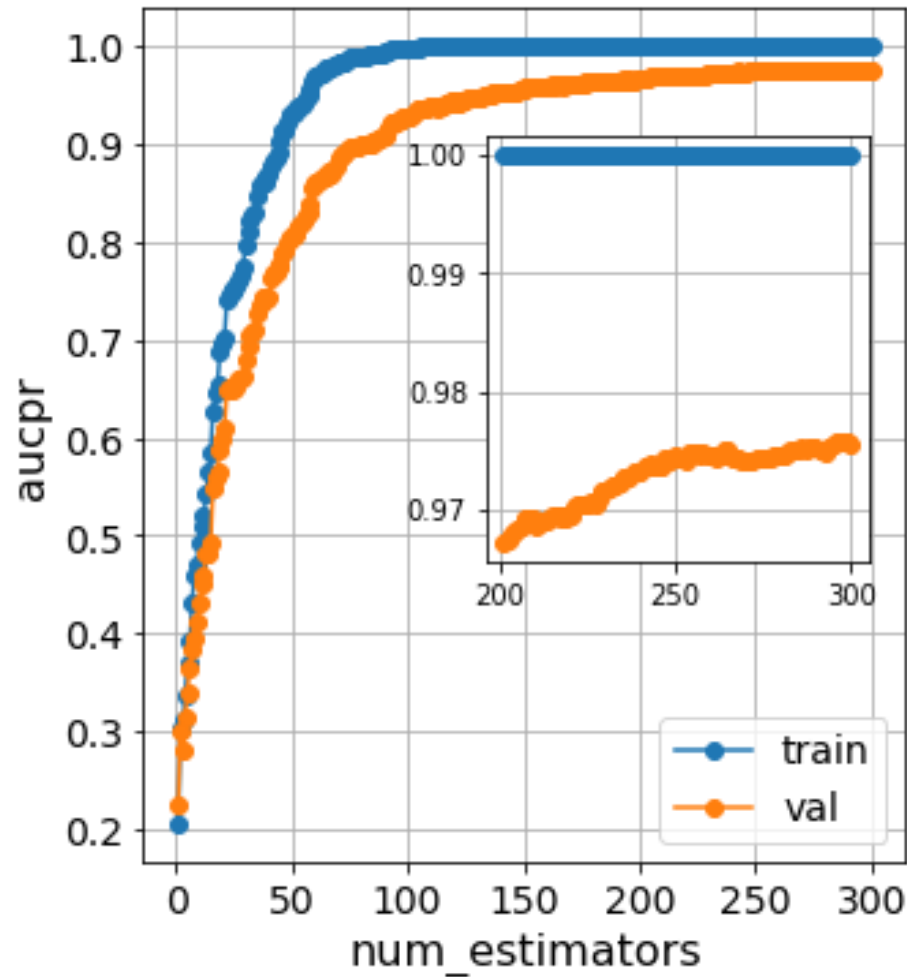
n_est=300, lr=0.3, ESR=None, gamma=0.0, max_depth=6
subsmpl=1.0, colssmpl_tree=1.0, reg_alpha=0.5, reg_lambda=0.0



ACC	100.0%	99.1%
RECALL	100.0%	89.9%
PRECISION	100.0%	88.3%
F1	100.0%	89.1%
	TRAIN	VAL

Add L2 regularization

n_est=300, lr=0.3, ESR=None, gamma=0.0, max_depth=6
subsmpl=1.0, colssmpl_tree=1.0, reg_alpha=0.0, reg_lambda=0.5



ACC	100.0%	99.2%
RECALL	100.0%	90.3%
PRECISION	99.9%	89.5%
F1	100.0%	89.9%
	TRAIN	VAL

Hyperparameter Exploration: Grid Search

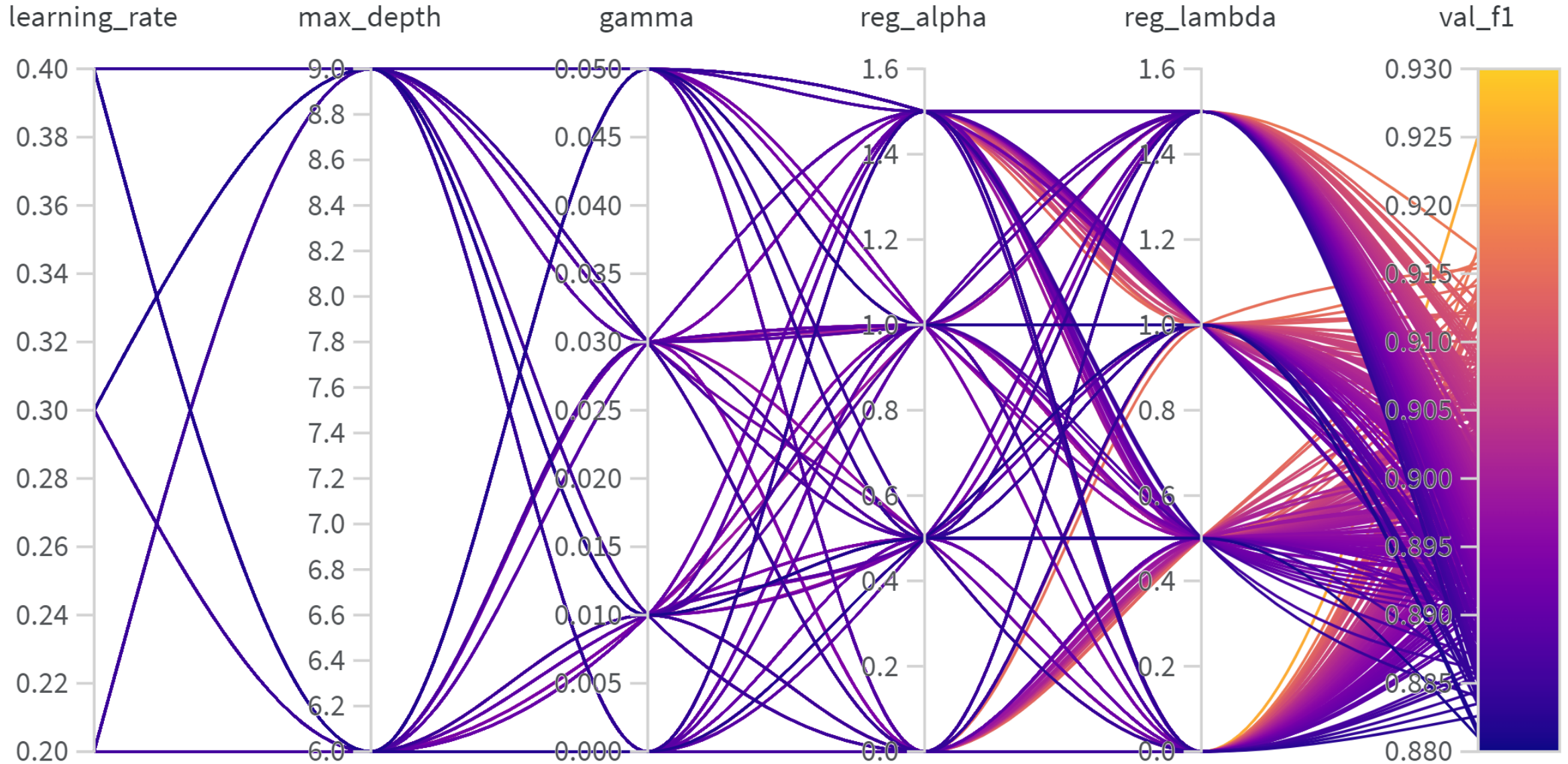
Hyperparameter Values Explored

eta (learning_rate)	0.2, 0.3, 0.4
num_boost_round (n_estimators)	$\frac{\text{default_num_boost_round} \times \text{default_eta}}{\text{eta}} \times \frac{\text{default_max_depth}}{\text{max_depth}}$
max_depth	6, 9
gamma	0.0, 0.1, 0.03, 0.05
reg_alpha (L1)	0.0, 0.5, 1.0, 1.5
reg_lambda (L2)	0.0, 0.5, 1.0, 1.5
Total Combinations	384

Fixed Parameters

objective	binary::logistic
default_eta	0.3
default_num_boost_round	300
default_max_depth	6
subsample	1
colsample_by*	1
min_child_weight	1
Early_stopping	None

Hyperparameter Exploration: Results



Hyperparameter Exploration: **Best HP Values**

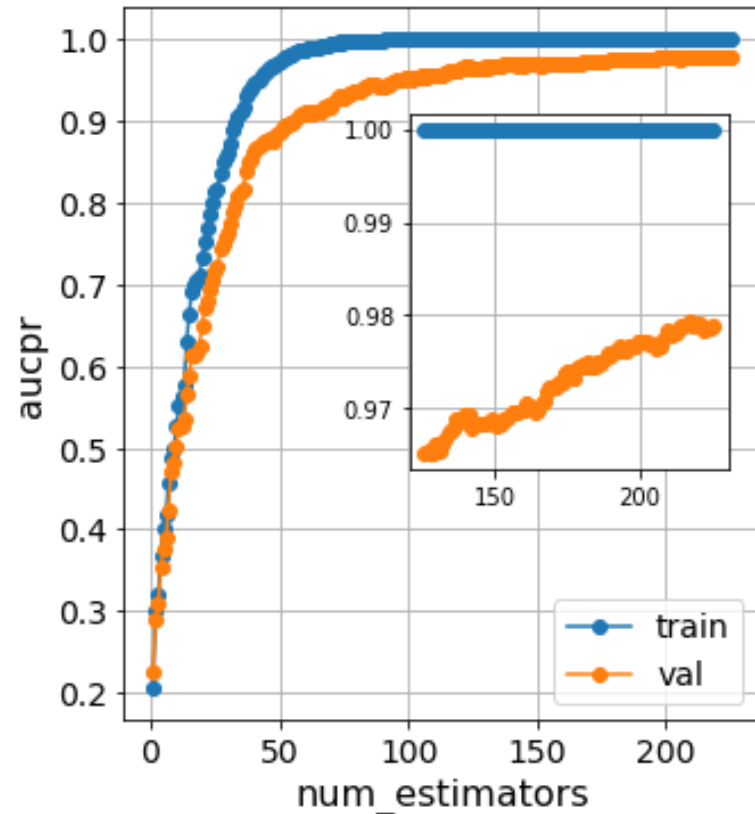
	Best F1-Score & Best Precision	Best Recall
eta (learning_rate)	0.4	0.2
num_boost_round (n_estimators)	225	450
max_depth	6	6
gamma	0	0.03
reg_alpha (L1)	0	1
reg_lambda (L2)	0	1.5
VAL. RECALL	92.5%	94.7%
VAL. PRECISION	92.5%	88.1%
VAL. F1-Score	0.925	0.913

**Selected
This One**

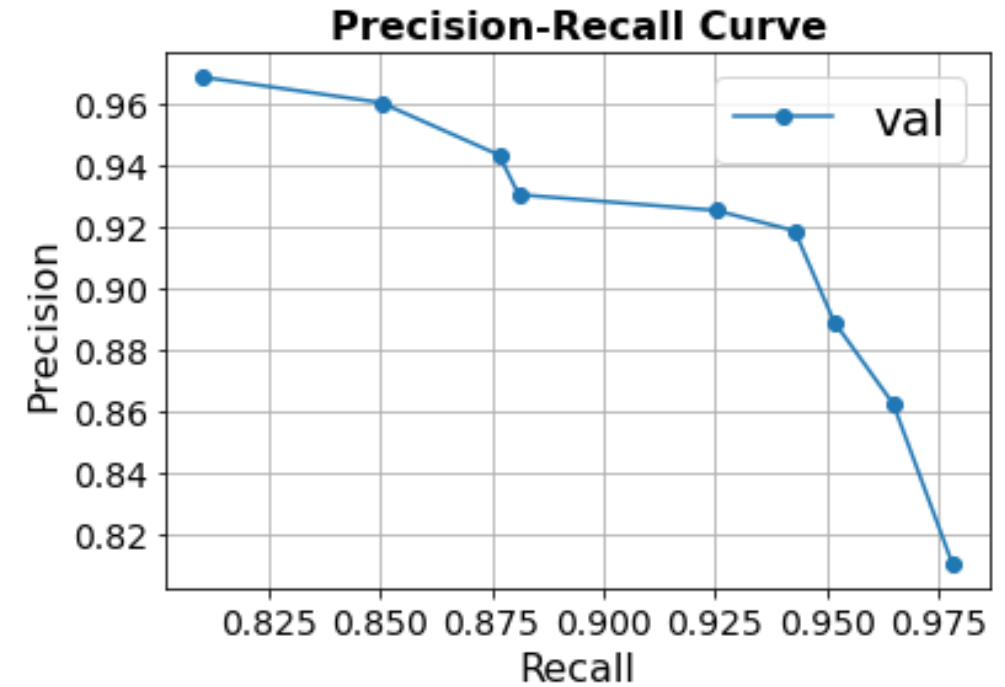


Final Model: Training + Precision-Recall Curve

n_est=225, lr=0.4, ESR=None, gamma=0, max_depth=6
subsmp1=1.0, colssmpl_tree=1.0, reg_alpha=0, reg_lambda=0



ACC	100.0%	99.4%
RECALL	100.0%	92.5%
PRECISION	100.0%	92.5%
F1	100.0%	92.5%
	TRAIN	VAL



Final Model: Feature Importance

