Predicting Injuries in Runners

Why?

OLYMPICS

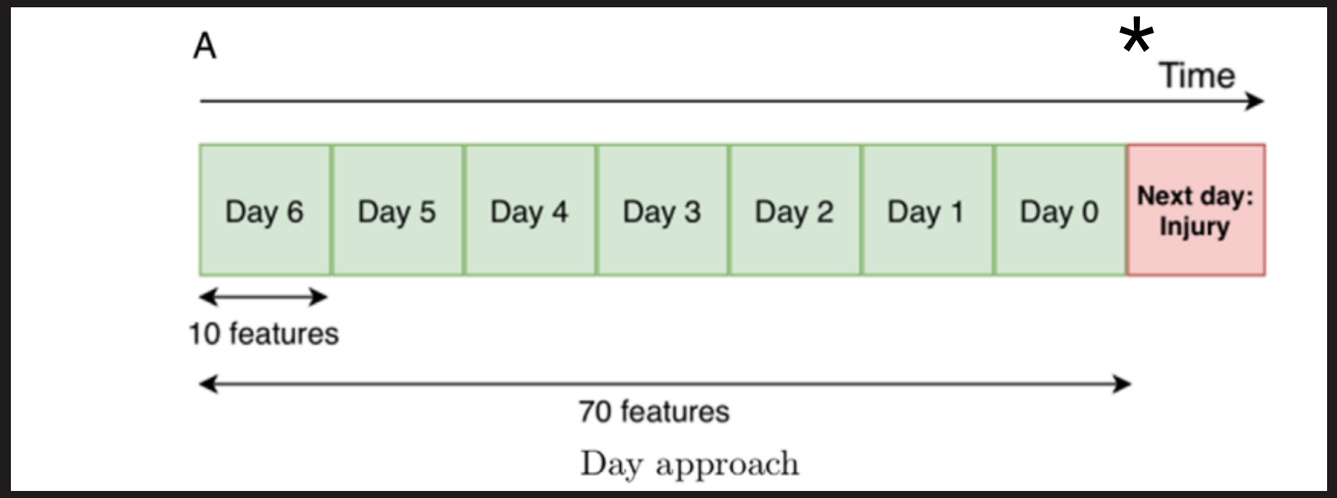
Donavan Brazier, reigning world champion, fails to qualify for Tokyo Olympics at 800 meters

Tyler Dragon and Tom Schad USA TODAY

Published 9:08 p.m. ET June 21, 2021 | Updated 11:43 a.m. ET June 22, 2021

How can we tell if we'll get injured?

Data Set



*** 10 repeating features, for 7 days**

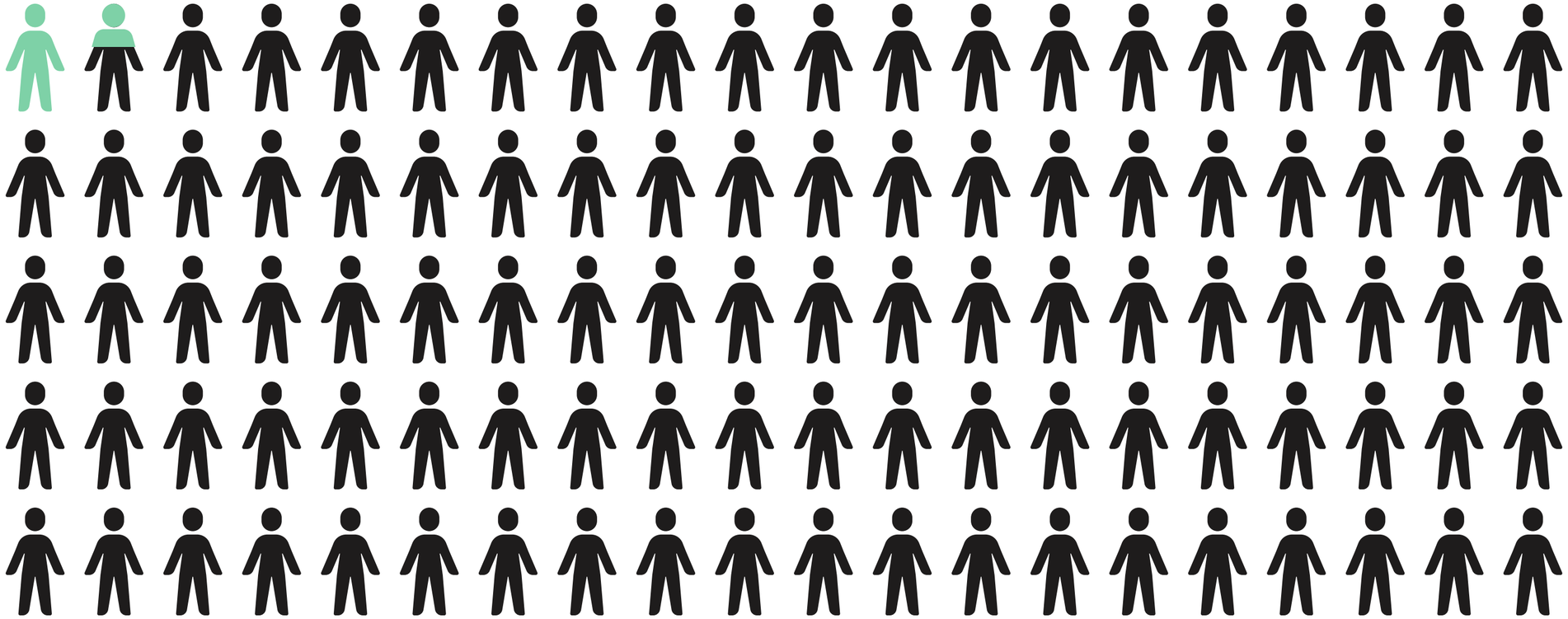
(i.e. Total km, Strength Training, Perceived exertion, etc)

70
features

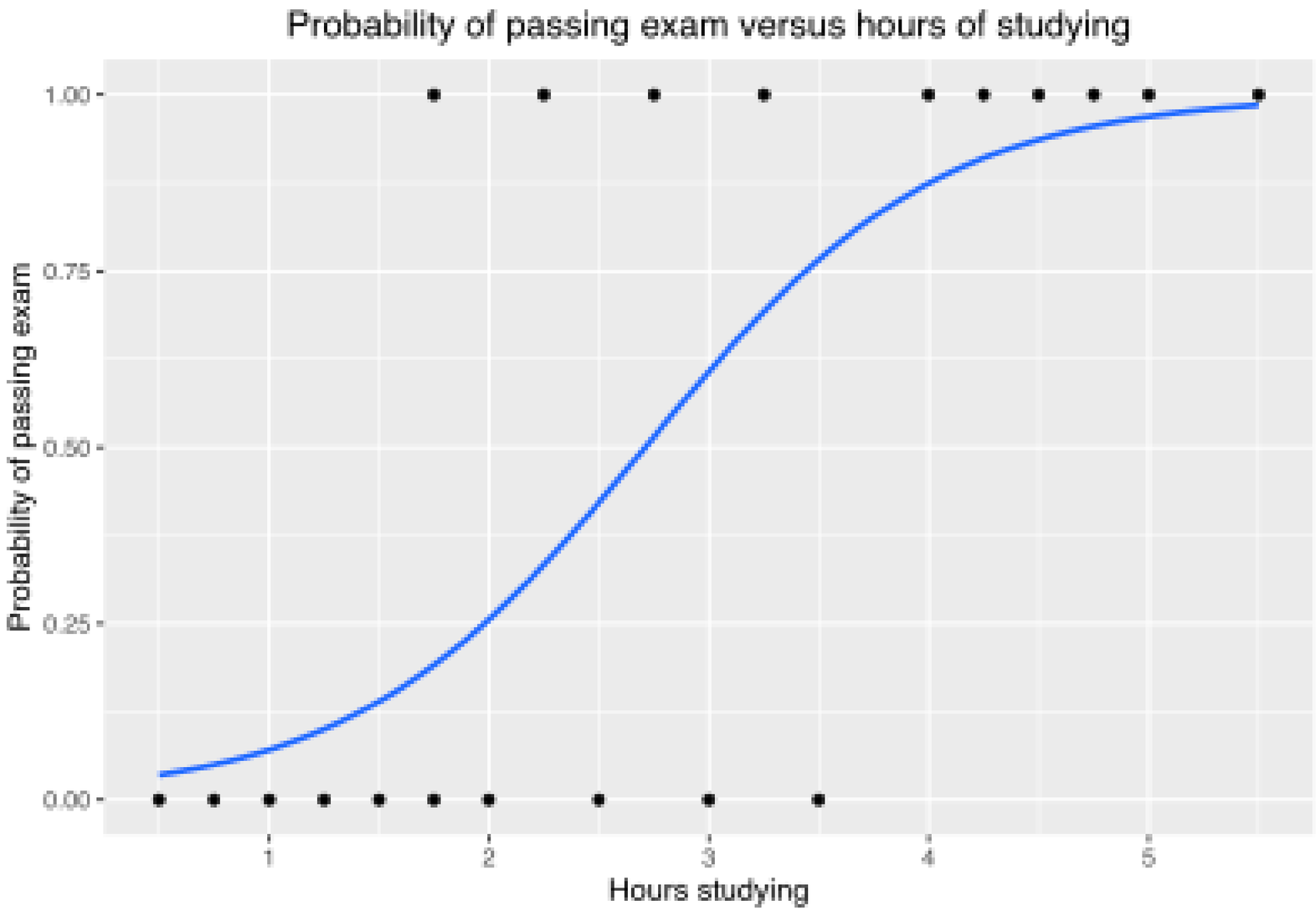
+42, 700
Rows

Severely imbalanced dataset

Only 1.3% of data points are injuries



Logistic Regression



Baseline Logistic Regression:

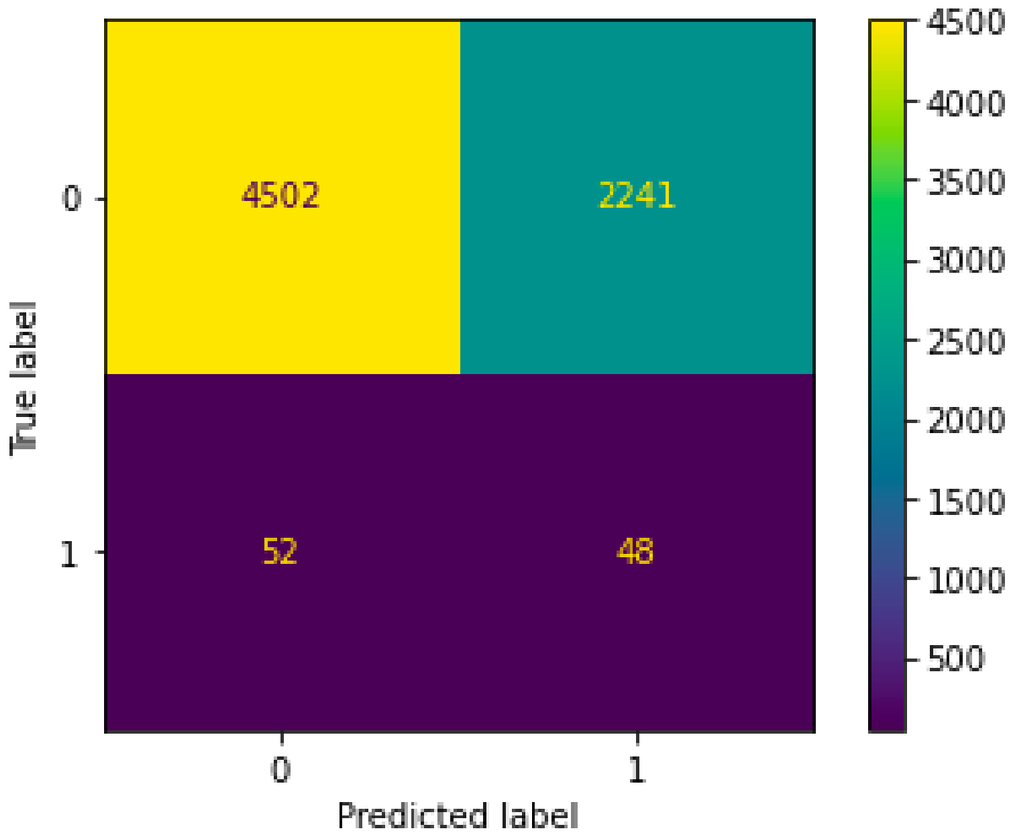
Validation score: 98%*

	Predicted Non-Injury	Predicted Injury
True Non-Injury	6743	0
True Injury	100	0

Optimizing and Resampling: LogReg

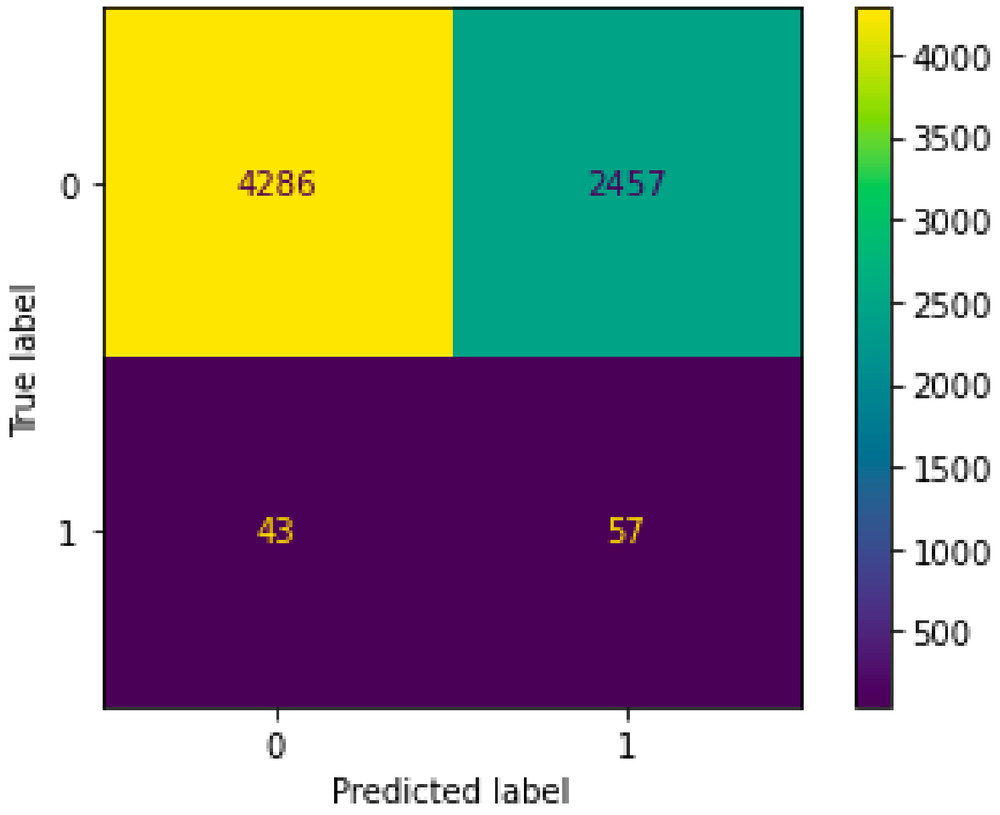
Optimized hyperparams

f1-score= **0.04**
recall=0.48
percision= 0.02

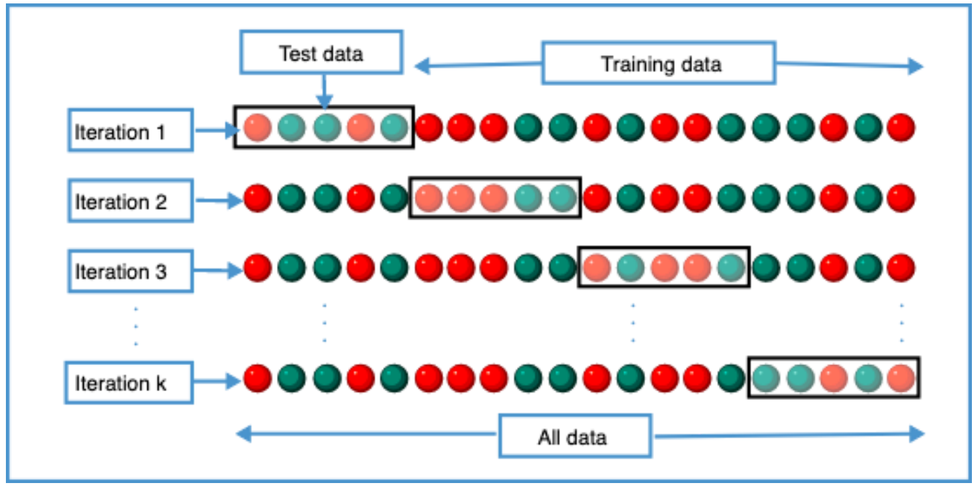


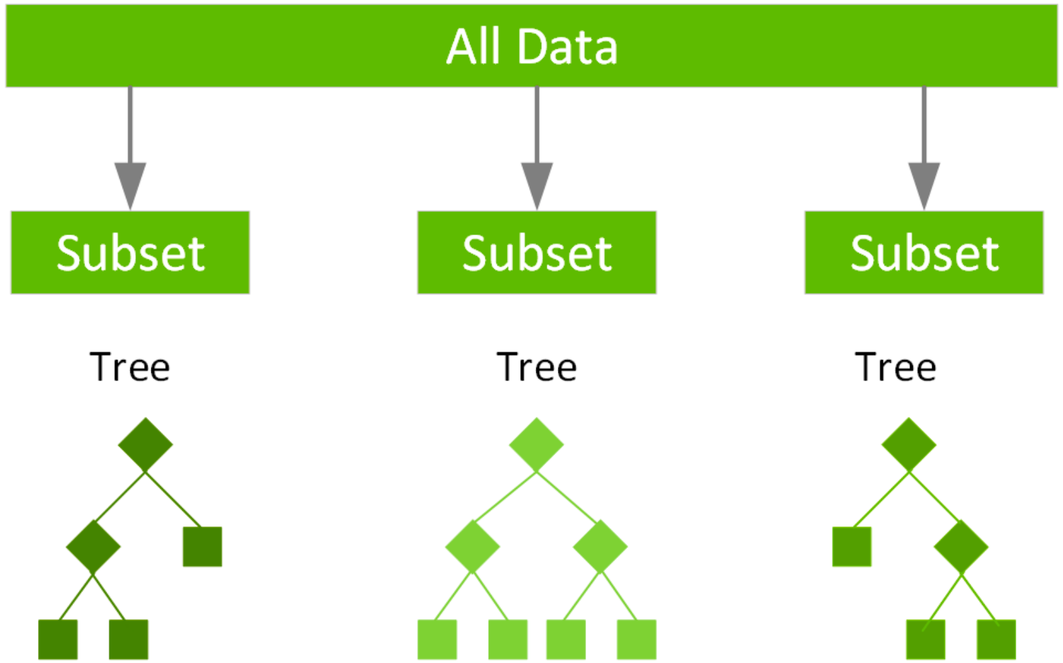
Down sampled 50-50

-f1-score= **0.04**
recall=.057,
Percsion= 0.02

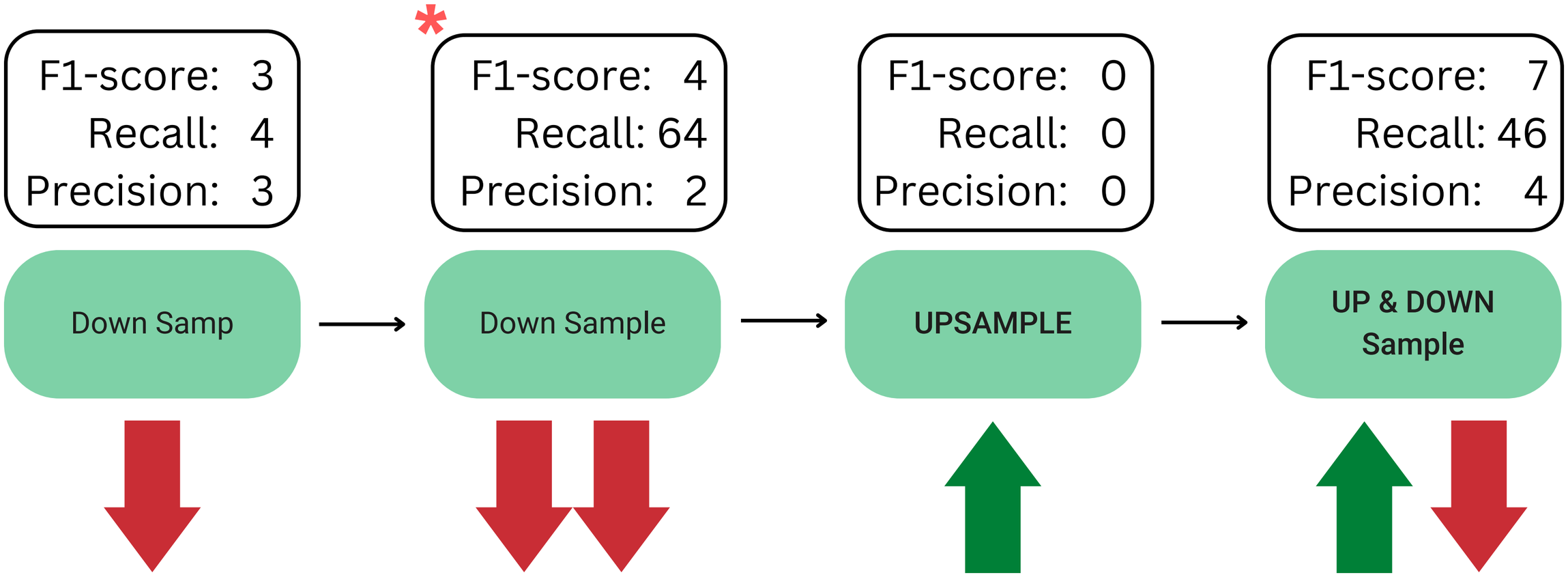


XGBoost with Cross Validation

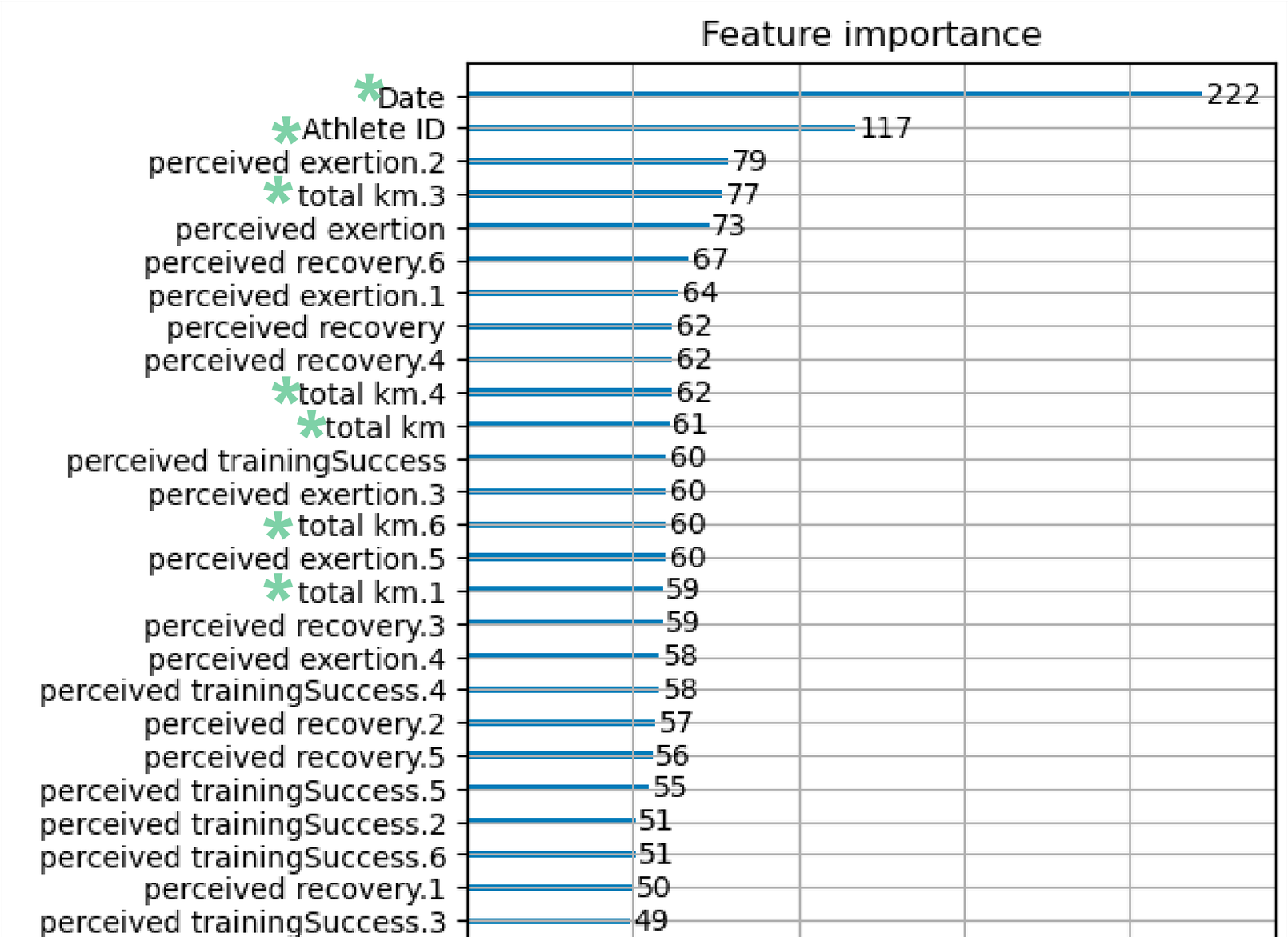




Iterations of ReSampling



Feature Importance



Most important:

- Date
- Athlete ID
- Perceived-

Least important:

- Number of Sessions
- Strength Training
- Alternative Exercise

Adjusted Datasets

	F1-score	Recall	Precision
1 Day Prior	4	77	2
1 Day Prior Key Features*	5	81	2
Feature Avg.	5	79	3
Key Features Avg.	5	75	2

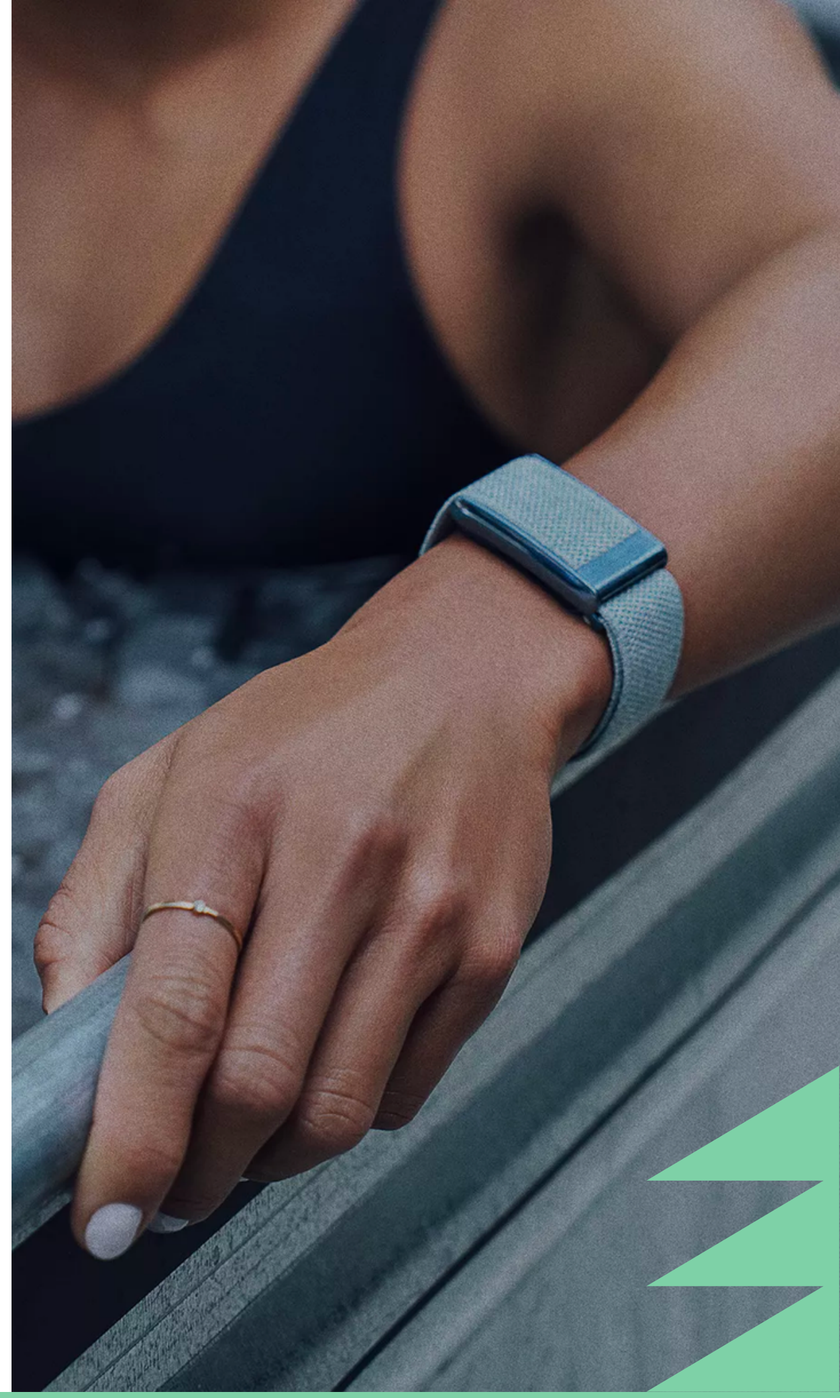
*Using Optimized Param grid from previous random search

Take Aways

- Athlete specific*
- Good Recall
- Lacking Precision

Expanding

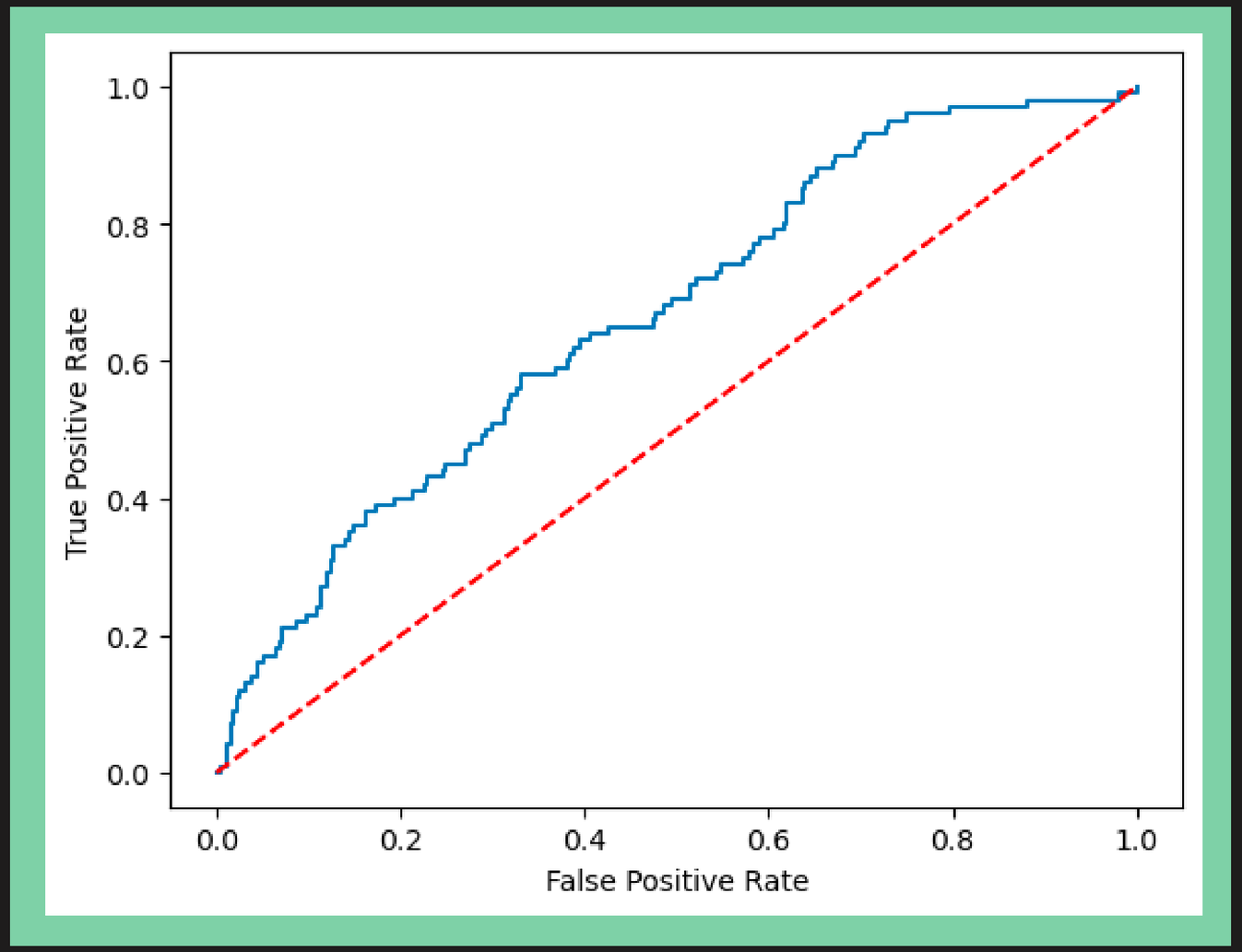
- Include Sleep/ Nutrition
- Specific Injuries
- Other sports



Extra

Back to Agenda

Scoring with ROC AUC



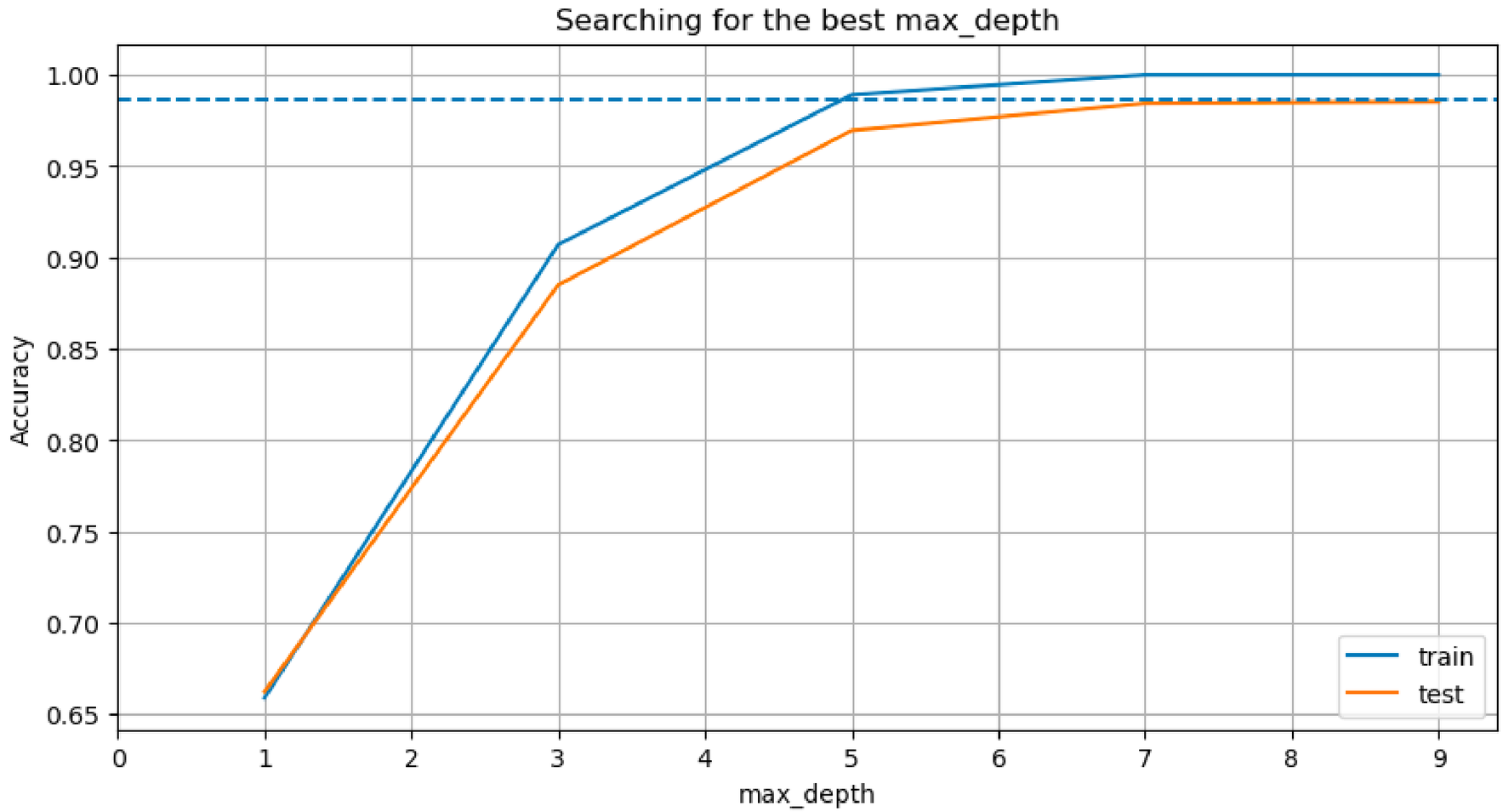
My Score:

66%

Study's Score:

72%

XGBoost: Tuning



Parameters:

- **subsample**
- **n_estimator**
- **max_depth**

- scoring = **'f1-score'***
  - **Recall**