Predicting Math PhD Student Advising

Scott Atkinson

Introduction

What makes a vibrant mathematics department?

- Active faculty
- Fresh ideas
- Prolific publication
- **PhD students**

General scenario

A PhD-granting department is hiring.

We provide two models to predict the progeny of a job candidate:

- **Screen**: ≥ 2 students in twenty years.
- **Highlighter**: ≥ 5 students in twenty years.

Screen ($\geq$ 2 students) use

- Filter out weak candidates.
- Reduce search committee workload by over 81%.

Highlighter (≥ 5 students) use

- Identify exceptionally strong (above 3^{rd} quartile) candidates.
- Can be used for novel recommendations for a department expanding research scope.

Data sources

Mathematics Genealogy Project: Online database of PhD mathematicians.

- School
- Year graduated
- Advisor
- Thesis title and classification
- Students

Data sources

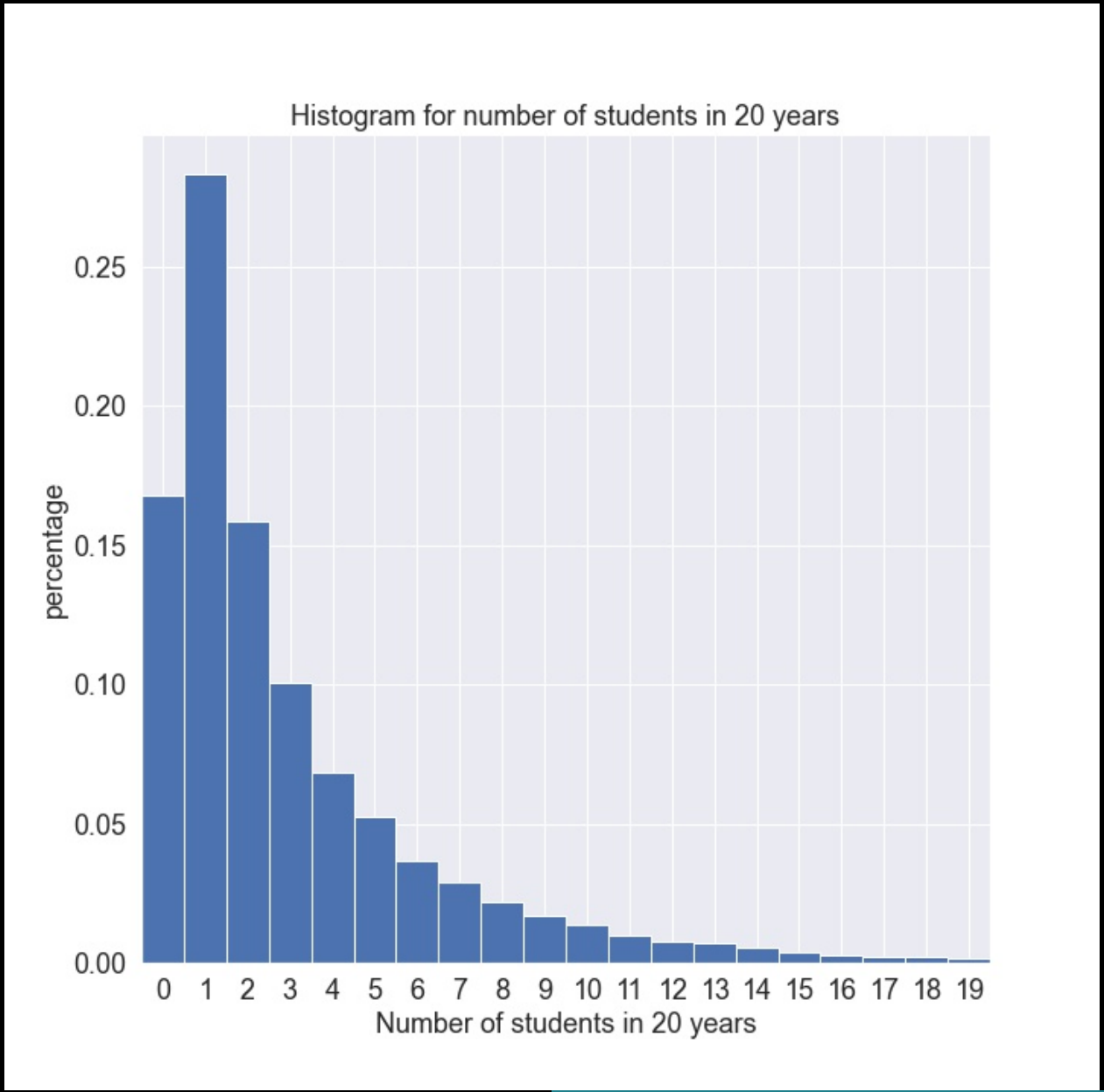
MathSciNet: Online database of publications in mathematics. Includes author summary pages.

- Earliest publication
- Total publications
- Total citations
- Collaborators
- Subject classifications

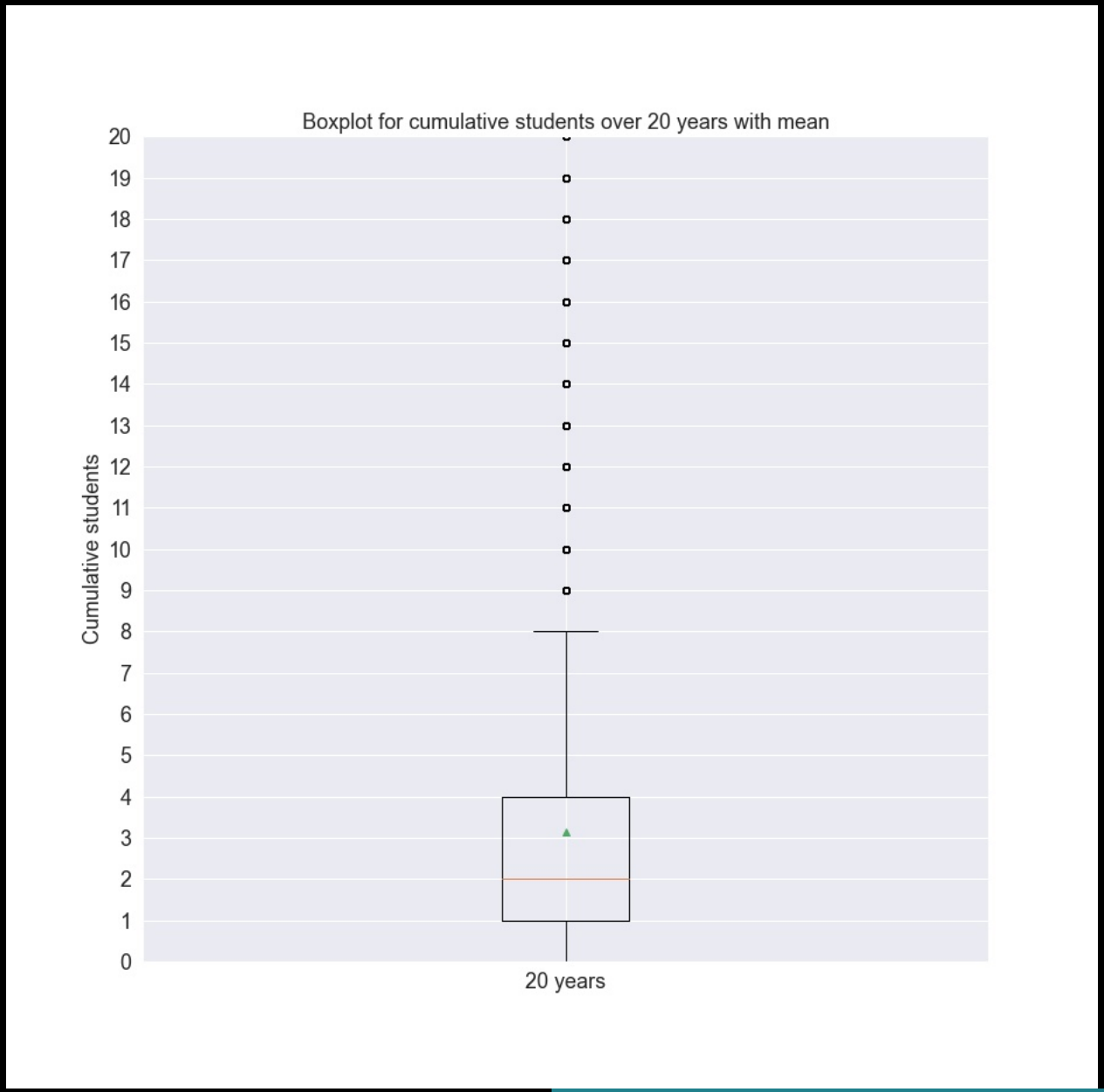
Why *20* years?

- Minimize interval to maximize records included.
- Long enough to ensure a distribution similar to unlimited data (sufficient variance).

Target histogram



Why 2 and 5?


Boxplot for cumulative students over 20 years with mean
Cumulative students
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
20 years

Features

Continuous features:

- Publications per year
- Citations per publication
- Collaborators per publication
- Advisor students at graduation

Features

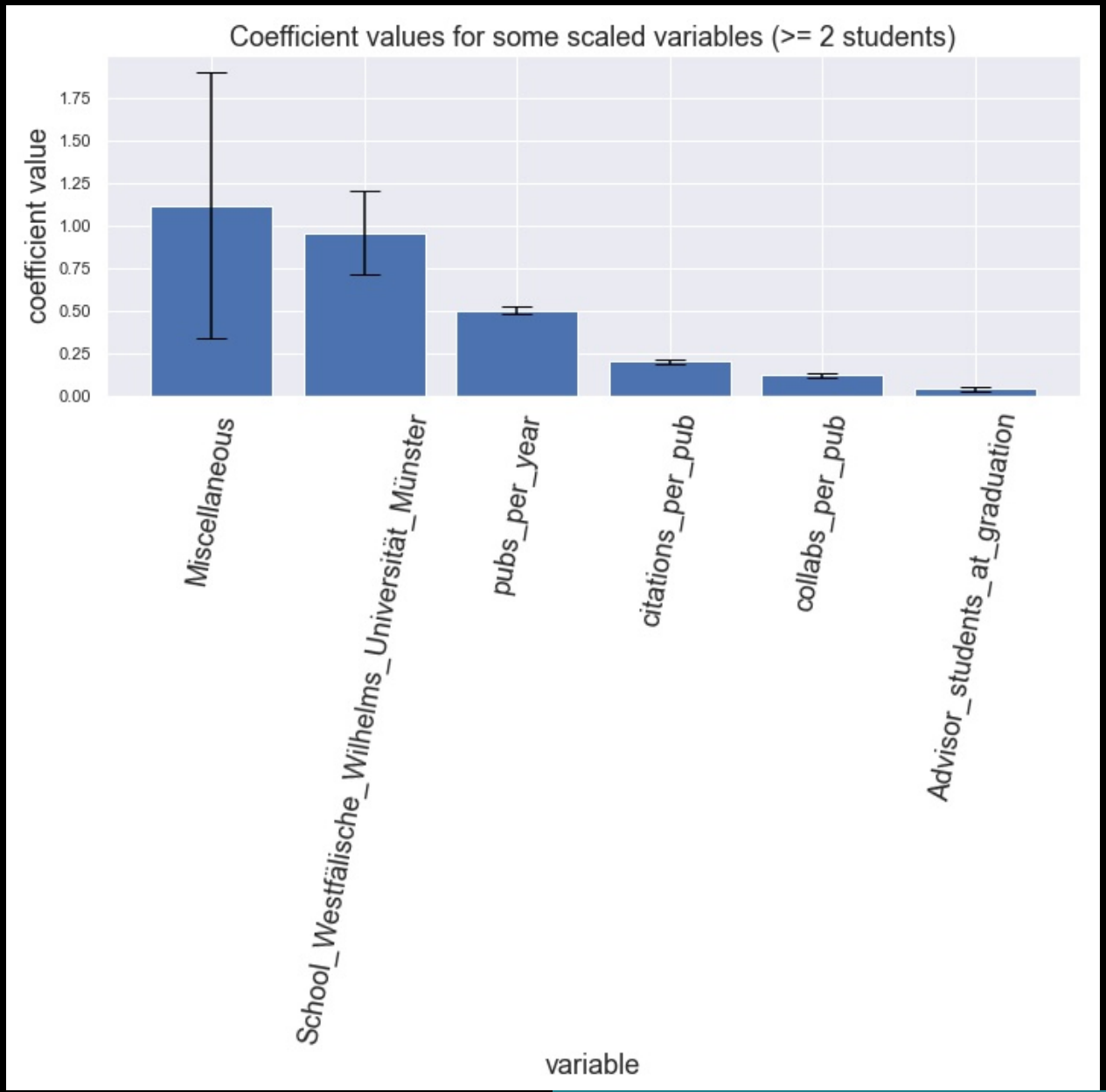
Categorical features:

- School
- Subjects

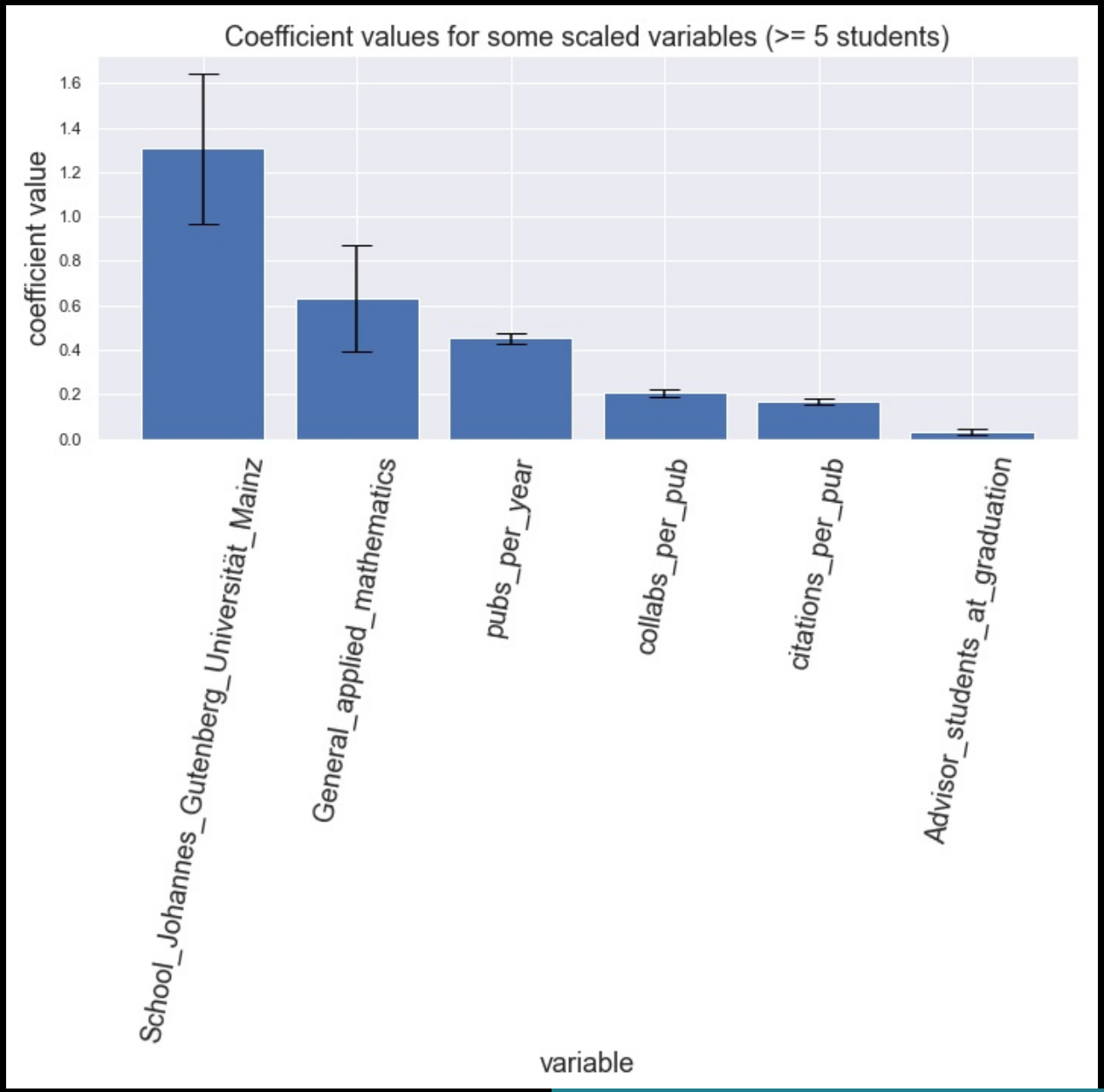
Data leakage

The data from MathSciNet are summary in nature, so there is some data leakage in the publication data. In particular, we have to use overall publication data to estimate the variable values for each record at the time before they advised students.

Feature importances



Feature importances



Model selection for Screen

Estimator	Hyperparameters	Mean AUC ROC
RandomForestClassifier	criterion='gini', max_depth=23, max_features='auto', n_estimators=233	0.6748
AdaBoostClassifier	learning_rate=1, n_estimators=292	0.6799
MLPClassifier pipelined with RobustScaler	alpha=0.1, max_iter=500, solver='sgd'	0.6719
GradientBoosting Classifier	learning_rate=0.01, max_depth=7, n_estimators=379	0.6716

Threshold tuning for Screen

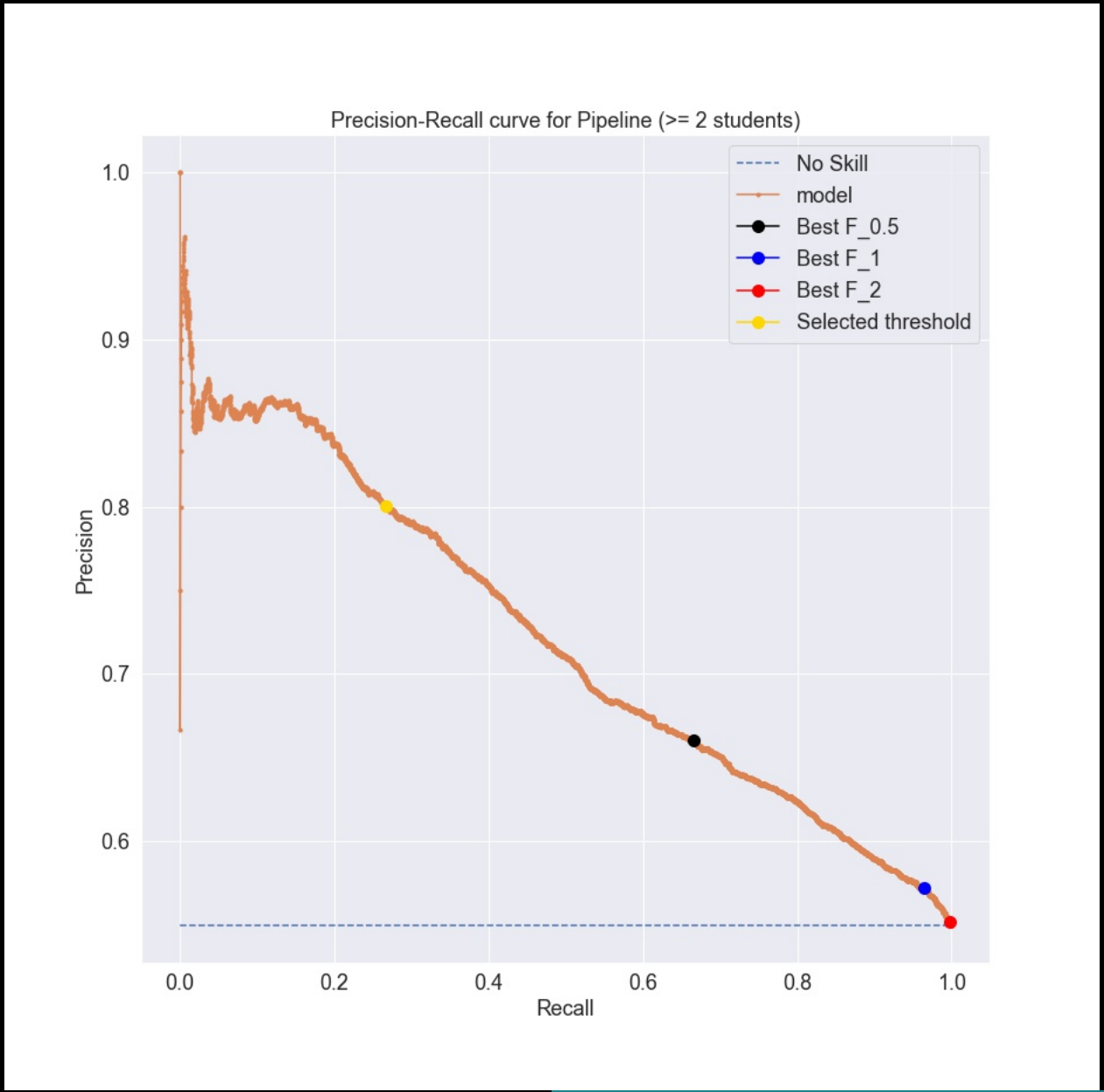
To filter out the bad: emphasize positive precision.

Selected model: `MLPClassifier` with `RobustScaler` with threshold 0.711.

Positive precision: 80%

Positive recall: 27%

Screen precision-recall curve



Screen confusion matrix

Table: Confusion matrix for `MLPClassifier` with `RobustScaler` with threshold 0.711 predicting ≥ 2 students

	Predicted False	Predicted True
Actual False	41.4%	3.6%
Actual True	40.3%	14.7%

Note: 92% negative recall.

Model selection for Highlighter

Estimator	Hyperparameters	Mean AUC ROC
`RandomForestClassifier`	`criterion='entropy', max_depth=31, max_features='auto', n_estimators=1000`	0.7120
`AdaBoostClassifier`	`learning_rate=0.1, n_estimators=1000`	0.7063
`MLPClassifier`	`alpha=0.1, max_iter=500, solver='sgd'`	0.7067
`GradientBoosting Classifier`	`learning_rate=0.1, max_depth=3, n_estimators=316`	0.7098

Threshold tuning for Highlighter

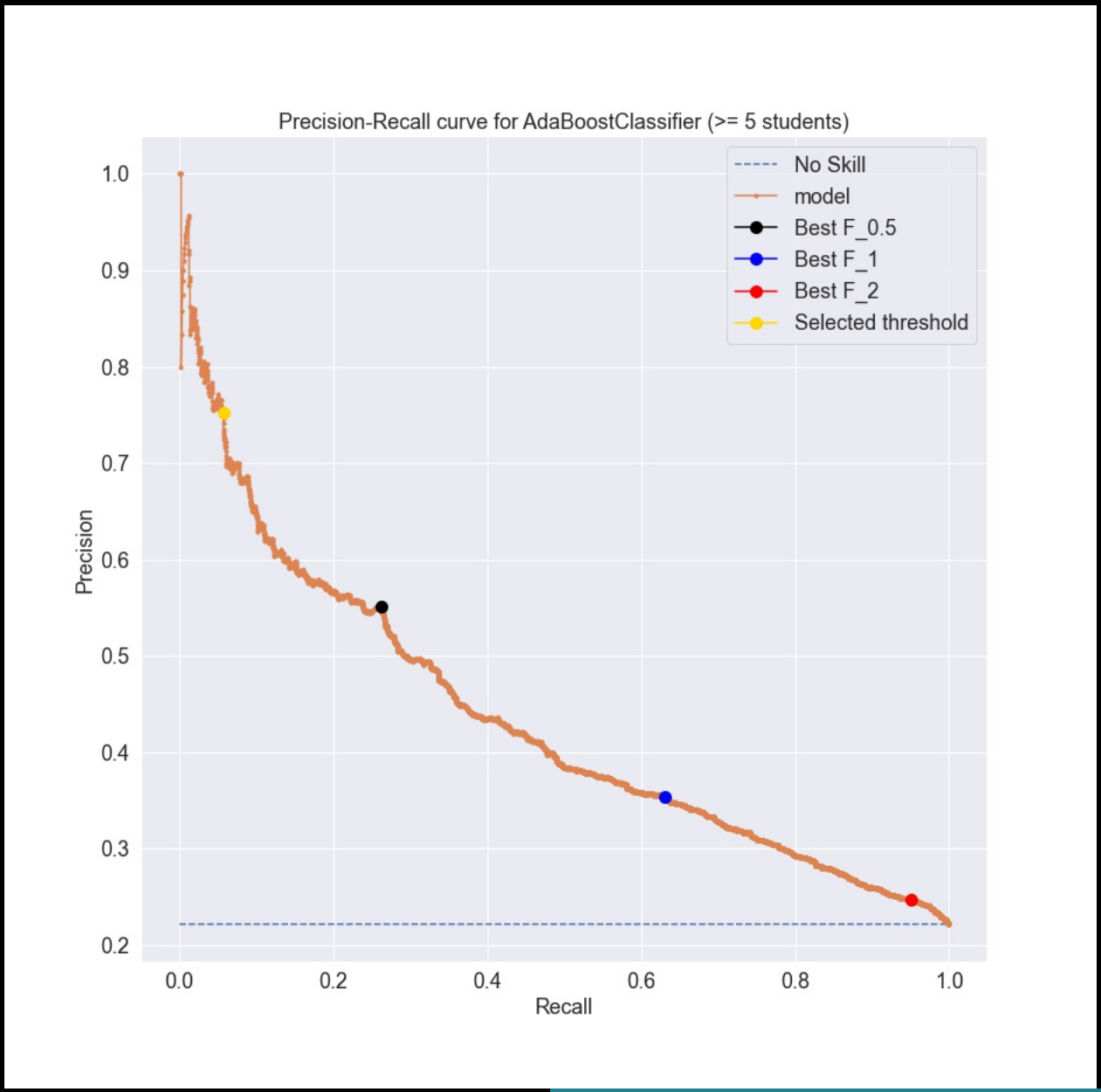
To identify strong candidates: emphasize positive precision at a cost to positive recall.

Selected model: `AdaBoostClassifier` with theshold 0.501.

Positive precision: 75%

Positive recall: 5.7%

Highlighter precision-recall curve



Highlighter confusion matrix

Table: Confusion matrix for AdaBoostClassifier with threshold 0.501 predicting ≥ 5 students

	Predicted False	Predicted True
Actual False	77.4%	0.4%
Actual True	20.9%	1.3%

Note: 99.5% negative recall

Summary and conclusion: Screen

- Use: filter out bad candidates
- Positive precision: 80%
- Positive recall: 27%
- Negative recall: 92%
- **Reduces search committee workload by over 81%**

Summary and conclusions: Highlighter

- Use: identify exceptionally strong candidates
- Positive precision: 75%
- Positive recall: 5.7%
- Negative recall: 99.5%
- Useful for a department looking to expand research scope.

Next steps

- Refine publication data
- Incorporate journal data
- Build a model where the number of students and amount of years are variable