PREDICTING MATH PHD STUDENT ADVISING

SCOTT ATKINSON

1. Introduction

What is a sign of a healthy vibrant mathematics department? For PhD-granting departments, one of the best indicators of the rank and prestige of the department is the quality of their PhD program. Consistently producing PhD graduates is a sign of a healthy department and attracts more PhD students which raises both the revenue and profile of the department. Keeping graduate student offices full of PhD students provides the department with fresh ideas and energy. Advising PhD students motivates faculty members to remain in touch with the frontiers of their research areas. A faculty member who can attract, mentor, and graduate PhD students on a regular basis is extremely valuable to a department. So any PhD-granting department would want to hire faculty members who possess great advising potential. Any department looking to hire a new faculty member would take into serious consideration the advising capabilities of their candidates. Each year mathematics departments in colleges and universities across the world spend significant amounts of time sifting through hundreds of applications for faculty positions. The purpose of this project is to develop classification models to aid departments in narrowing their searches down to only the highest quality candidates.

We will present two models in this report. The first model can be considered as a **highlighter** model. That is, it will identify extremely strong candidates by predicting if a candidate will produce at least 5 students during the first twenty years of the candidate's career. The data show that only 22.2% of all advisors (active for at least twenty years) have graduated at least 5 students in their first twenty years. Thus 5 students in twenty years sets the bar high. The second model is a **screen** model. The screen model will predict if a candidate will produce at least 2 students during the first twenty years of the candidate's career. The data show that 2 students in twenty years is the median amount of students for that time interval. More specifically, 55% of all advisors (active for at least twenty years) have graduated at least 2 students in their first twenty years. So a department may not want to consider any candidates with advising potential predicted to be below the median.

The report is organized as follows. In §2, we discuss the data sources, data preparation and analysis, and we assess feature importance with a logistic regression analysis. In §3 we assess and select our classification models and then choose the proper thresholds for the intended uses of our models. In §4 we draw our conclusions on our findings and discuss potential improvements for the models.

2. Data

2.1. Data sources.

The data for this project are obtained from two sources: The Mathematics Genealogy Project and MathSciNet. The Mathematics Genealogy Project maintains a database of mathematics PhDs, typically recording a mathematician's school, graduation year, thesis title, advisor, and any PhD students they themselves advised. MathSciNet is a database of mathematics publications. MathSciNet maintains author pages for each author in the database. These author pages include relevant data such as total publications, total citations, and collaborators. While full access to MathSciNet requires a subscription, the data obtained from MathSciNet for this project are publicly available.

2.2. Data preparation and analysis.

The data from The Mathematics Genealogy Project (MGP) are assembled into a dataframe with the following attributes: `MGP_ID, Name, MScNetID, Degree, School, Country, Year, Thesis, MSC, num_students, num_descendants, num_advisors, Advisor_k_name, Advisor_k_MGP_ID`. After some cleaning, this dataframe has 259,422 rows.

The data from MathSciNet (MScNet) are assembled into a dataframe with the following attributes: `MScN_ID, Name, Earliest_Pub, Total_Pubs, Total_Rel_Pubs, Total_Citations, CollabIDs, CollabNames, Subjects, Num_Collaborators, Num_Subjects`. After cleaning, this dataframe has 148,190 rows.

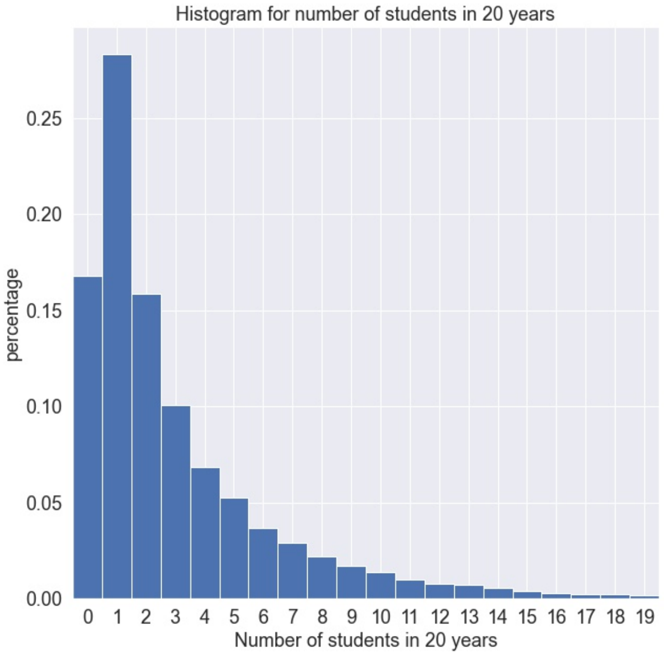

Figure 1. Histogram of the number of students produced in the first 20 years of a career

We then join (inner on MScNet ID) the two dataframes above into a single dataframe and filter down to entries with at least one student and possessing publication, year, and complete student data[1].

To regulate the comparison of progeny among advisors, we consider the number of students each advisor produced in a certain interval of time. We would like the shortest amount of time so to maximize the amount of records (young professors at the beginning of their career not to be left out), but we would also like a long enough time-frame that is both reasonable for the considerations of the department and mimics the overall production over an unlimited time interval. In order to accomplish this, we engineer a `career_length` variable. In the case of missing data we assume a max career length of 40 years and compute the career length according to the following formula:

$$f(x) = \begin{cases} 2020 - x & \text{if} \quad x \geq 1980 \\ \max((L-x), 40) & \text{if} \quad x < 1980 \end{cases}$$

where L denotes the graduation date of the candidate's most recent student. Using career length, we create the `pubs_per_year` variable to put junior and senior mathematicians on more even footing. We also create the `Advisor_students_at_graduation` variable which gives the number of students (counting the candidate) a candidate's advisor had at the time the candidate received their PhD.

We compare the histograms for the number of students produced in the first 5, 10, 15, 20, 25, 30, 35, and 40 years. The interval of 20 years[2] is the shortest amount of time that gives a nicely filled in histogram–see Figure 1. Thus we further filter our dataset to records with career length at least 20–a dataframe with 32568 rows.

From a preliminary correlation comparison, the independent variables we choose for our model are number of publications per year, number of citations per publication, number of collaborators per publication, number of students graduated by the candidate's advisor upon the candidate's graduation, school, and the list of mathematical subject classifications the candidate's publications have addressed.[3] We encode the school

[1] Due to an error in obtaining data, there are ∼800 records obtained with incomplete advisor data. Because of this, a small amount of records (even smaller than the 800) are missing students that should be attributed to them. Due to the relatively small size of this incomplete set, we dropped these records.

[2] This time interval can be easily adjusted based on the department's needs.

[3] Data leakage: the publication data from MathSciNet is only the present-day summary information for each mathematician. So there is some data leakage in obtaining the publications per year, citations per publication, collaborators per publication,

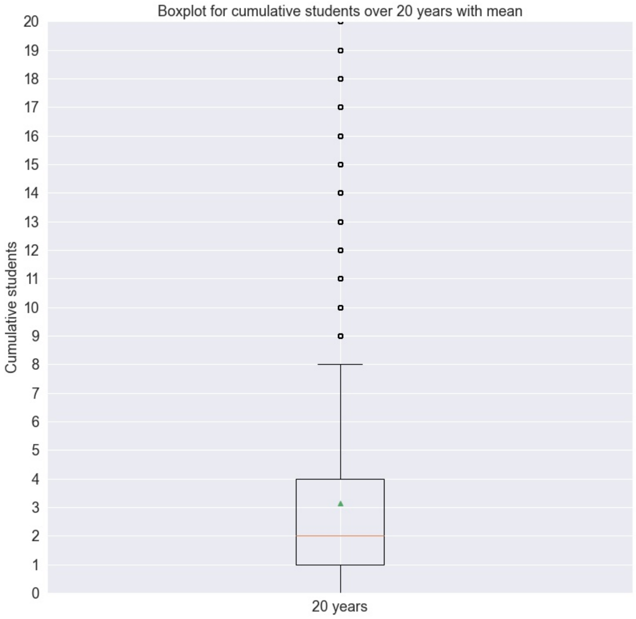

FIGURE 2. Boxplot for the distribution of the number of students produced in the first 20 years. The triangle indicates the mean of the distribution.

variable using dummy variables where the categories are the 125 most frequently appearing schools in our dataset together with an `Other` category that is dropped. We also encode the 68 represented subjects using 68 indicator variables.

Departments are not necessarily concerned with a prediction of an actual amount of students a given candidate may advise in their first 20 years. Departments *would* be interested, however, in knowing that a given candidate will be a productive advisor or not. For this reason the models we build are classification models. In particular, the target variable for each model is a boolean variable indicating whether or not a given candidate will produce at least k students in their first 20 years. The choice of k depends on how such a model is to be used. As mentioned in the introduction, the two use-cases we will consider are as follows:

1. **Highlighter**: recommend extremely strong candidates;
2. **Screen**: keep out weak candidates.

To choose the values of k for our Highlighter model and our Screen model, we consider the boxplot for the distribution of the number of students produced in the first 20 years, shown in Figure 2. We see that the median is 2 and 4 is the third quartile. From this data, a candidate who can produce at least 5 students over twenty years should be considered as exceptionally strong since 5 lies above the interquartile range (78.8^{th} percentile), and a candidate who cannot produce at least 2 students over twenty years may be considered as relatively weak. Thus our Highlighter model will predict if a candidate will produce at least 5 students in their first twenty years, and our Screen model will predict if a candidate will produce at least 2 students in their first twenty years.

There are 25348 records who produced fewer than 5 students over the first 20 years of their careers, and there are 7220 records who produced at least 5 students over the first 20 years of their careers. The histogram in Figure 3 indicates how the candidates with at least 5 students in the first 20 years of their careers are more consistent publishers: The barchart in Figure 4 shows how candidates from the top five schools and working in the top five subjects are more productive than average:

and list of subjects. More refined publication data is provided by MathSciNet but only with a subscription to their database, and a massive scrape of this data would be in violation of their policies. If a model such as the one described in this report is seriously desired, then perhaps an agreement with MathSciNet can be reached to accommodate access to the more refined data to give more accurate values for these variables.

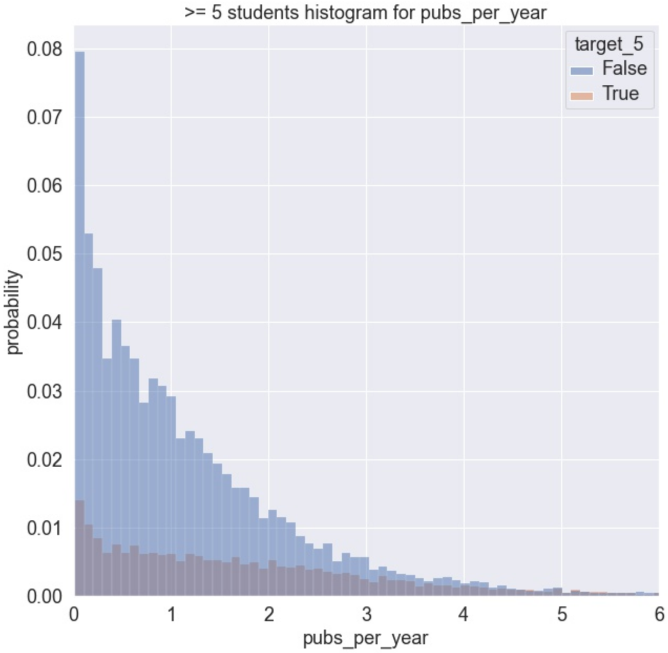

FIGURE 3. Histogram of the number of publication per year separated based on the number of students produced in 20 years (True: ≥ 5 students)

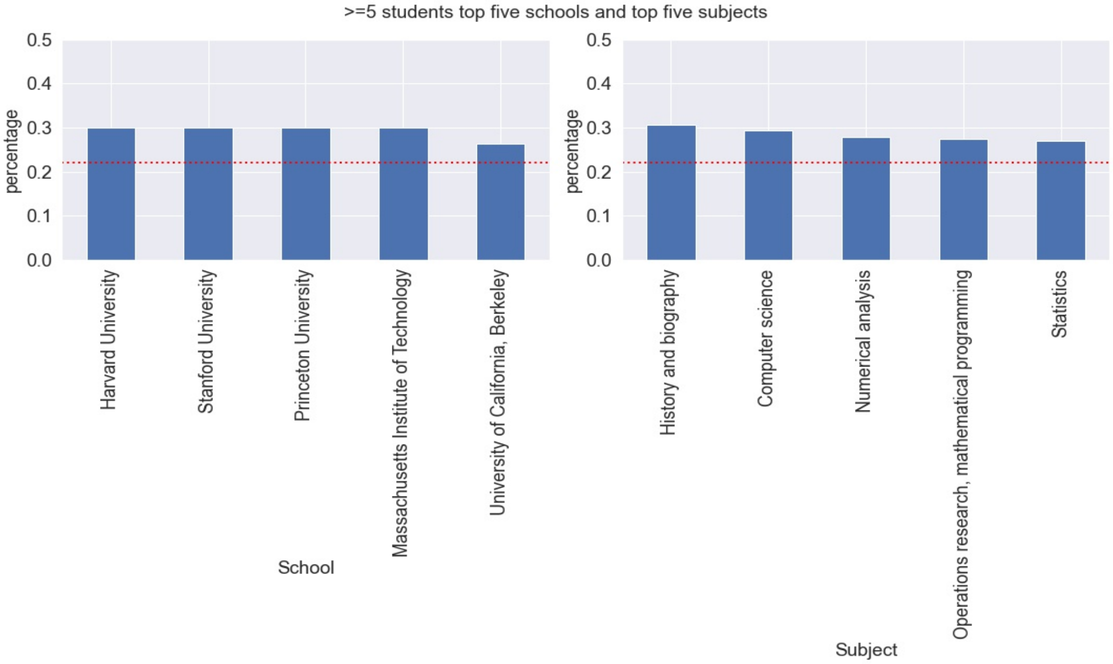

FIGURE 4. Bar charts for the categorical variables giving the percentage of candidates from the top five schools and subjects who advised at least 5 students over twenty years. The red dotted line indicates the overall percentage of advisors with at least 2 students over twenty yeras.

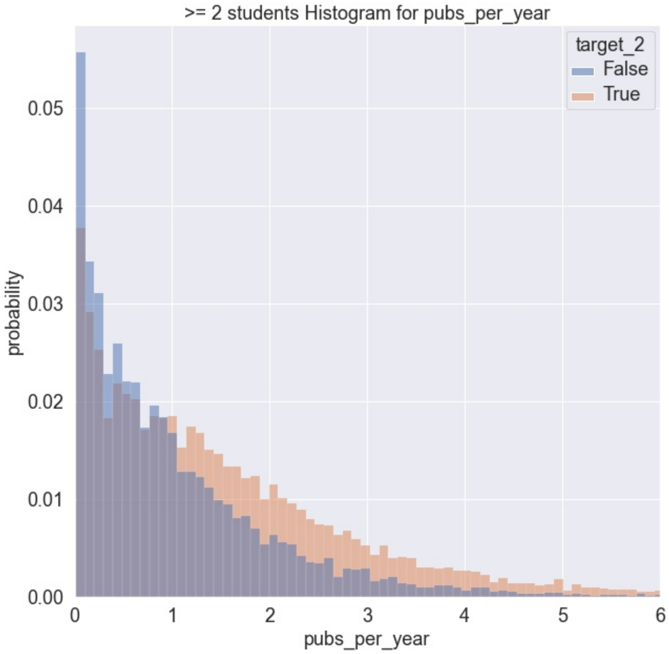

Figure 5. Histogram of the number of publication per year separated based on the number of students produced in 20 years (True: ≥ 2 students)

For the Screen model, there are 14676 records who produced fewer than 2 students over the first 20 years of their careers, and there are 17892 records who produced at least 2 students over the first 20 years of their careers. Figures 5 and 6 are the Screen model's analogs of the above plots. Note how the publication behavior is relatively the same on either side of 2.

2.3. **Regression analysis of feature importance.** In order to study the relative feature importances we perform regression analysis for both the Highlighter and Screen cases. Before proceeding with the regression analysis, we consider the variance inflation factors (VIF) of the variables to see if any variable could potentially negatively impact the model by being too volatile. The VIF analysis returns values indicating that there would be no such problem with volatility.

By performing a rough logistic regression, we are able to obtain a sense for the relative importances of the features. We first scale the features using `RobustScaler` so that the logistic regression model's coefficients are all on the same scale. For the Highlighter model (predicting ≥ 5 students) the most positively impactful school is Johannes Gutenberg Universität Mainz, and the most positively impactful subject is General Applied Mathematics. It is interesting to note that according to this regression analysis on the Highlighter data, the most negatively impactful school is Uniwersytet Warszawski, and the most negatively impactful subject is Difference and Functional Equations. For the Screen model (predicting ≥ 2 students) the most positively impactful school is Westfälische Wilhelms Universität Münster , and the most positively impactful subject is Miscellaneous. The Screen model's most negatively impactful school is Lomonosov Moscow State University, and the most negatively impactful subject is Mechanics. Figures 7 and 8 indicate the relative importances of certain variables for each model.

3. Modeling

3.1. **Model assessment.** To select estimators for each model, we first split the data into 75% training and 25% testing (stratifying on each respective target). On each training set, we perform a grid/randomized search 5-fold cross-validation with four different algorithms: `RandomForestClassifier, AdaBoostClassifier, MLPClassifier, GradientBoostingClassifier`. This is all performed via the scikit-learn API in python. The scoring method is the Area Under the ROC Curve. This score is useful because it takes into account all

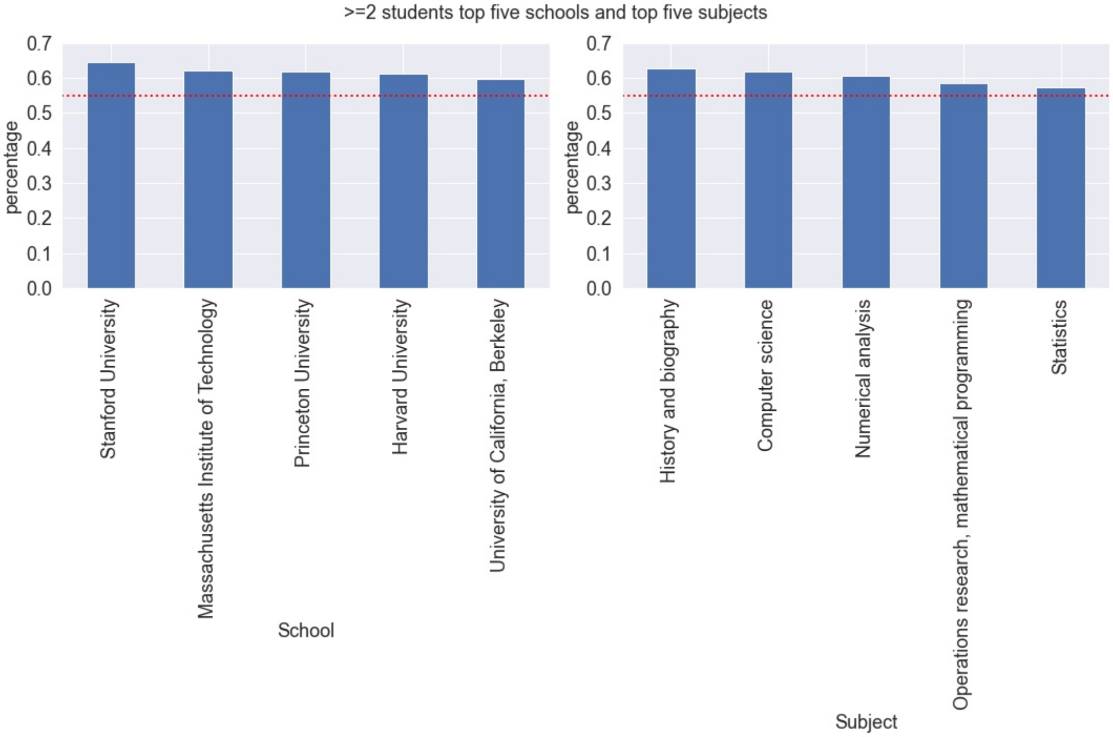

Figure 6. Bar charts for the categorical variables giving the percentage of candidates from the top five schools and subjects who advised at least 5 students over twenty years. The red dotted line indicates the overall percentage of advisors with at least 2 students over twenty yeras.

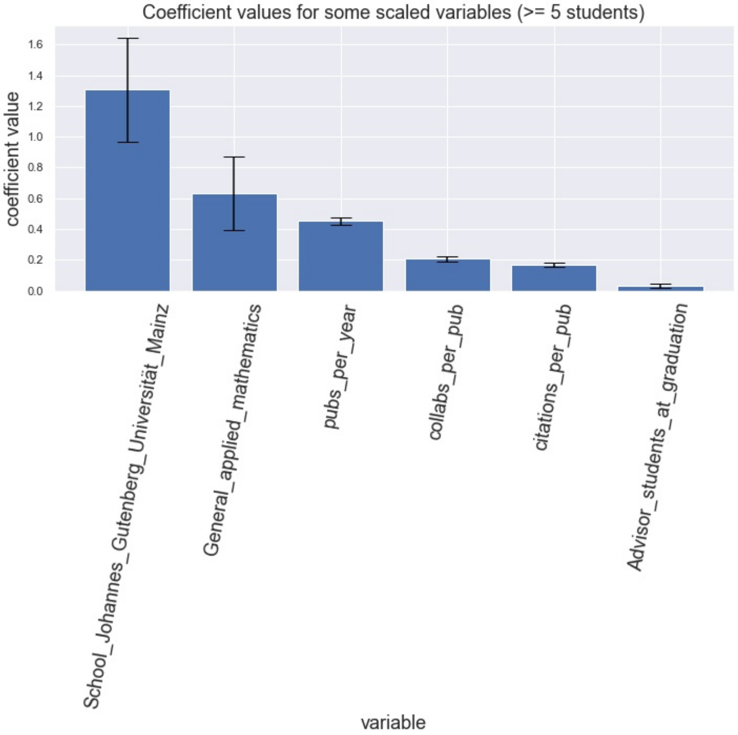

Figure 7. Relative importances of variables with 95% confidence intervals for the Highlighter model.

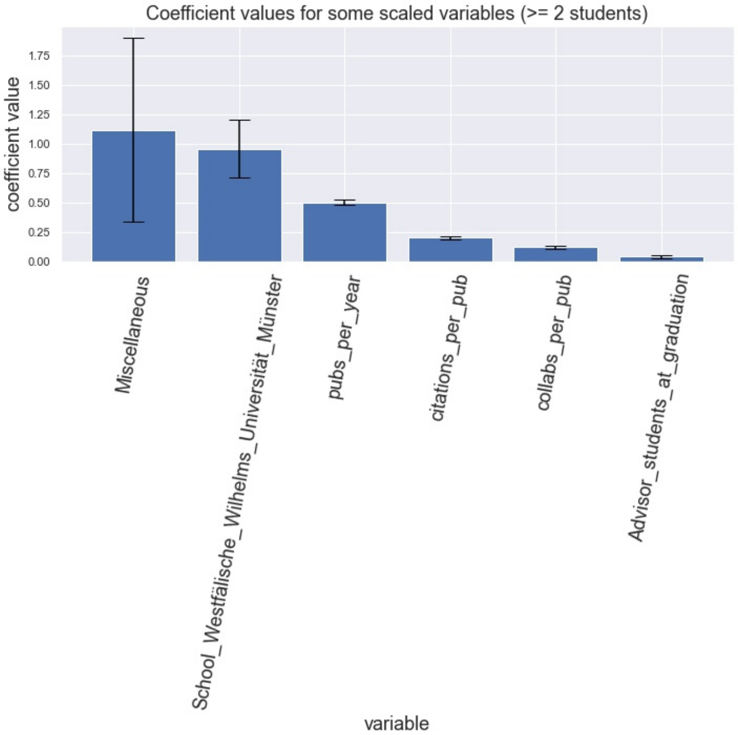

FIGURE 8. Relative importances of variables with 95% confidence intervals for the Screen model.

TABLE 1. Cross-validation results for Highlighter model

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

`GradientBoostingClassifier`	`learning_rate=0.1,`
	`max_depth=3,`
	`n_estimators=316`
		0.7098

possible thresholds for the predicted probabilities given by these estimators. See Tables 1 and 2 indicating the scores and hyperparameters resulting from these searches.

3.2. **Threshold tuning.** The next step is to tune the threshold for each estimator since each estimator in truth returns a probability value for its predictions. We need to select the best threshold for each of our use-cases.

Highlighter: To use the model as a highlighter the emphasis is on positive precision. That is, departments want to reduce the chances that this model falsely classifies a candidate as potentially productive so that they can trust that the model is recommending only worthy candidates. This use of the model justifies prioritizing positive precision with some cost to positive recall. For that reason, we choose the `AdaBoostClassifier` estimator with threshold 0.501 to obtain a positive precision of 75% at the cost of a low recall at 5.7%. Of all the candidates that the model recommends, 3 out of every 4 are strong candidates worth looking into. On the other hand, of all the worthy candidates worth looking into, the model only recommends fewer than 6 out of every 100. Figure 9 shows the precision-recall curve for the fitted AdaBoostClassifier, and Table 3 shows the confusion matrix for this aggressive threshold. This loss of good candidates may be acceptable for a

Table 2. Cross-validation results for Screen model

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
`GradientBoostingClassifier`	`learning_rate=0.01, max_depth=7, n_estimators=379`	0.6716

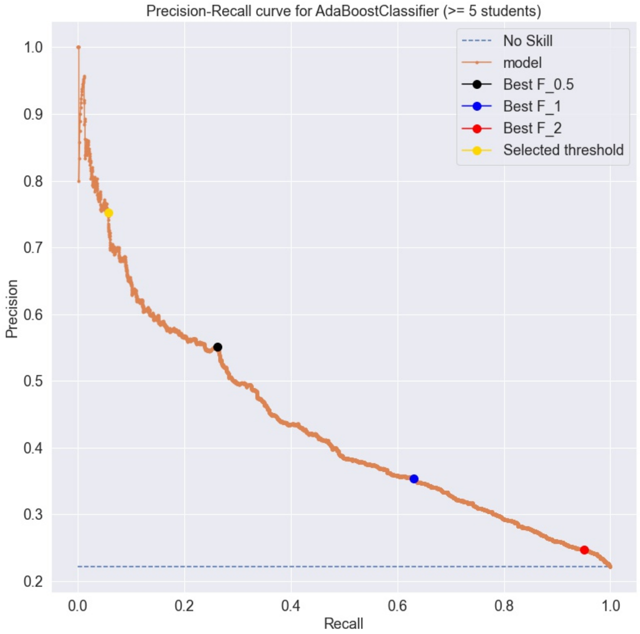

Figure 9. Precision-recall curve for AdaBoostClassifier, predicting ≥ 5 students. The positions giving best $F_{0.5}$-score, F_1-score, F_2-score, and the selected threshold are indicated in black, blue, red, and gold respectively.

department that is looking for some novel recommendations outside of the expertise of the search committee. That is, a department's search committee may have a very good sense for what makes a strong candidate in their areas of mathematics, but they are looking to expand the scope of the department's research and might not have as good a sense for what makes a strong candidate in a foreign area. While this model leaves out a lot of good candidates, the department can be assured that most of the recommended candidates are actually strong.

Screen: We also emphasize positive precision for the Screen model. For the Screen model, the `MLPClassifier` with `RobustScaler` estimator with threshold 0.711 yields positive precision is 0.80 and positive recall is 0.27. See Figure 10 for the precision-recall curve for this estimator. Table 4 shows the confusion matrix for this threshold. Out of all the candidates, 45.0% were "bad"candidates, and 55.0% were "good"candidates. The model recommends that the department consider $3.6\% + 14.7\% = 18.3\%$ of the candidates. The use of this

TABLE 3. 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%

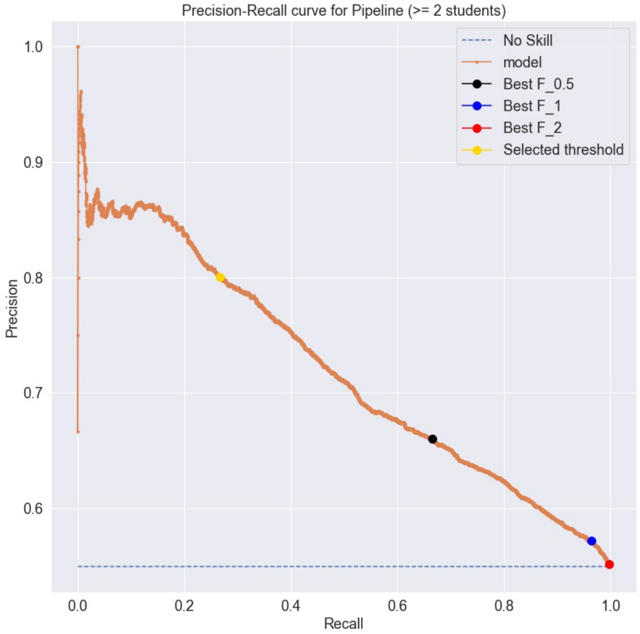

FIGURE 10. Precision-recall curve for `MLPClassifier` with `RobustScaler`, predicting ≥ 2 students. The positions giving best $F_{0.5}$-score, F_1-score, F_2-score, and selected threshold are indicated in black, blue, red, and gold respectively.

TABLE 4. 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%

model removes 81.7% of the candidate pool from the department search committee's workload. Of the recommended candidates, 80.3% are "good."The department using this model would understand that there is a possibility of false positives so further considerations must be taken. In any case, performing further analysis on 18.3% of the original applicant pool is much more appealing than performing such analysis on the entire pool. Also, out of the total collection of "bad"candidates, only 8.1% were recommended by the model. This high negative recall further justifies this model as an effective screen. It should be mentioned that a majority (73.3%) of the "good"candidates were excluded by the model.

4. Conclusion

In this analysis, we obtained two models: the Highlighter model and the Screen model. The Highlighter model is meant to help departments discover extremely strong candidates. This is achieved by its high positive precision of 75% at a cost of low positive recall. Because the positive recall is low (5.7%), departments know that the model is leaving out a larger portion of strong candidates, but the 75% positive precision assures them that most of the recommended candidates are truly strong. The highlighter model can offer value to a department looking to expand the scope of their research. The model can provide novel recommendations that the department may not have the expertise to immediately recognize. This use-case can be further tuned by filtering the data based on the subjects of interest for the department and building a similar model.

The Screen model is meant to reduce the workload of a department's search committee. This model excludes 92.9% of weak candidates, and of the model-recommended candidates, 80.3% meet the $\geq$ 2-students-in-twenty-years bar. The payoff is that this model reduces the search committee workload by more than 81%.

This performance shows that the models serve as a helpful tool for a department's search committee, but of course human discernment should be the primary guide.

As mentioned above, this model can be strengthened and made more accurate with more refined publication data. In particular, dated publication information (rather than summary info) would eliminate most of the data leakage. Furthermore, individual publication data for each publication would provide valuable journal data. The list of journals in which a mathematician publishes is extremely impactful on their professional and scientific reputation. Individual publication data would allow us to use the journals of publication as another categorical variable in further refining our assessment of the potential progeny of a candidate. Journal data will also allow us to perform feature importance analysis to deduce which journals (should) carry more weight in hiring decisions.