



Predicting the future impact of Computer Science researchers: Is there a gender bias?

Matthias Kuppler¹

Received: 15 July 2021 / Accepted: 14 February 2022 / Published online: 7 April 2022
© The Author(s) 2022

Abstract

The advent of large-scale bibliographic databases and powerful prediction algorithms led to calls for data-driven approaches for targeting scarce funds at researchers with high predicted future scientific impact. The potential side-effects and fairness implications of such approaches are unknown, however. Using a large-scale bibliographic data set of $N = 111,156$ Computer Science researchers active from 1993 to 2016, I build and evaluate a realistic scientific impact prediction model. Given the persistent under-representation of women in Computer Science, the model is audited for disparate impact based on gender. Random forests and Gradient Boosting Machines are used to predict researchers' h -index in 2010 from their bibliographic profiles in 2005. Based on model predictions, it is determined whether the researcher will become a high-performer with an h -index in the top-25% of the discipline-specific h -index distribution. The models predict the future h -index with an accuracy of $R^2 = 0.875$ and correctly classify 91.0% of researchers as high-performers and low-performers. Overall accuracy does not vary strongly across researcher gender. Nevertheless, there is indication of disparate impact against women. The models underestimate the true h -index of female researchers more strongly than the h -index of male researchers. Further, women are 8.6% less likely to be predicted to become high-performers than men. In practice, hiring, tenure, and funding decisions that are based on model predictions risk to perpetuate the under-representation of women in Computer Science.

Keywords Impact prediction · h -index · Gender · Discrimination · Machine learning

Introduction

Academia, like all public institutions, faces resource constraints. Funding agencies, hiring committees, and university departments are called to invest scarce public resources into the most promising researchers and research projects. Several authors therefore argue for a data-driven approach to hiring, tenure, and funding decisions (Acuna et al., 2012; Bertsimas et al., 2015). To this end, large-scale bibliographic databases are leveraged to quantify

✉ Matthias Kuppler
matthias.kuppler@uni-siegen.de

¹ Department of Social Sciences, University of Siegen, Adolf-Reichwein-Straße 2, 57076 Siegen, Germany

researchers' past achievements but also to predict their future scientific impact (Ayaz et al., 2018; Dong et al., 2016; Weihs & Etzioni, 2017). Such predictions could be used to target funds at the most promising researchers to increase the efficiency of resource allocations.

Calls for data-driven approaches in academia reflect a general trend towards the integration of prediction-based decision-making into the delivery of public services (Lepri et al., 2018). Data-driven approaches promise to render decision-making processes more accurate and evidence-based and, by limiting decision-maker discretion, less susceptible to human biases and stereotypes. At the same time, concerns are raised that data-driven decision-making may perpetuate unfair discrimination against vulnerable and historically disadvantaged groups (Barocas & Selbst, 2016), especially if past discriminatory decisions are encoded into the data on which the prediction models are trained. Hence, data-driven approaches have the potential to both reduce and reproduce discrimination against historically marginalized groups in academia, e.g., female researchers and researchers from ethnic minorities (Eaton et al., 2020; Moss-Racusin et al., 2012). Before introducing data-driven decision-making into academia, we need to carefully audit our prediction models for discriminatory impact.

In this paper, I empirically evaluate the potential side-effects and fairness implications of scientific impact prediction. I build and evaluate a realistic scientific impact prediction model that could be used to support decisions regarding hiring and funding of post-doctoral researchers. The model draws on data from a large-scale bibliographic data set of Computer Science researchers compiled by Weihs and Etzioni (2017), data that are comparable to those that would be available to actual hiring committees and funding agencies. Given the historic under-representation of women in Computer Science (National Science Board, 2018; NCSES, 2021), model evaluation focuses on detecting gender differences in predictive performance. Scientific impact is defined as a researcher's *h*-index (Hirsch, 2005), a metric adopted by Web of Science, Scopus, and Google Scholar to quantify researcher performance that is already used in hiring decisions (Demetrescu et al., 2020; Reymert, 2021). Two prediction scenarios are investigated: (1) Predicting the *h*-index 5 years into the future based on bibliographic profiles collected in 2005 (*impact prediction*). (2) Predicting whether the *h*-index 5 years into the future is in the top-25% of the *h*-index distribution of the discipline (*high-performer prediction*). The second scenario is especially relevant for stakeholders who are interested in whether a given candidate will outperform their peers (Zuo & Zhao, 2021).

The remainder of the paper is organized as follows: The “**Background**” section provides information on the role of gender in science (“**Gender in science**” section) and scientific impact prediction (“**Measuring and predicting scientific impact**” section). The “**Data and methods**” section introduces the data, prediction setup, and evaluation metrics. The “**Results**” section presents empirical results regarding the predictive performance and fairness implications of the prediction model. The “**Discussion**” section concludes with a discussion of the main findings, limitations, and implications of this paper.

Background

Gender in science

Gender bias in academia is well-documented (Larivière et al., 2013). Female researchers have lower citation impact (Beaudry & Larivière, 2016; Bendels et al., 2018) and lower scientific

productivity (Long, 1992; van Arensbergen et al., 2012) and are less likely to obtain prestigious first and last authorship positions (Holman et al., 2018; West et al., 2013) than male researchers. Female researchers occupy lower academic ranks (Ceci et al., 2014), are less likely to secure funding (van der Lee & Ellemers, 2015; Wennerås & Wold, 1997; Witteman et al., 2019), and have more restricted access to mentorship (Blau et al., 2010; Sheltzer and Smith 2014) and international collaboration (Abramo et al., 2013; Jadidi et al., 2018) networks. Women are more likely than men to leave academia at every career stage (Huang et al., 2020), a tendency called the “leaky pipeline” effect. Science is associated more strongly with stereotypical male than female attributes (Carli et al., 2016; Kessels et al., 2006; Lane et al., 2012; Leslie et al., 2015; Nosek et al., 2002) and women are frequently perceived to lack the competences required for successful scientific careers (Eaton et al., 2020; Moss-Racusin et al., 2012). Research done by women is perceived to be of lower quality (Knobloch-Westerwick et al., 2013) and women receive less credit for their work (Hofstra et al., 2020; Sarsons, 2017; West et al., 2013).

Despite substantial advances towards gender equality (or even female advantage) in PhD graduation rates over the last 30 to 40 years (Miller & Wai, 2015), women remain under-represented among tenure-track and full professorships in the US (Ceci et al., 2014). The gender gap is especially pronounced in STEM (science, technology, engineering, and mathematics) fields (National Science Board, 2018). Computer Science stands out as a particularly male-dominated STEM discipline, with women representing less than 25% of PhD graduates (NCSES, 2021) and full, associated, and assistant professors (NCSES, 2019). Comparable gender disparities are documented for the European academic system (European Commission, 2019).

Establishing equal opportunity in academia is imperative to (1) resolve historical inequalities and (2) foster scientific progress as (gender) diversity increases creativity and innovation in scientific collaborations (AIShebli et al., 2018; Hofstra et al., 2020; Nielsen et al., 2017). Data-driven approaches to the central gate-keeping decisions in academia (hiring, tenure, funding) can help to reach this goal. Research on non-academic labor markets shows that gender bias in evaluations of job candidates is reduced or even eliminated if the evaluation procedure is standardized (Reskin, 2000; Reskin & McBrier, 2000) and based on unambiguous, task-relevant signals of candidate competence (Koch et al., 2015; Heilman, 2012). Predicted future scientific impact is a clear and task-relevant indicator of researcher competence and can help funding agencies and hiring committees to make more informed and equitable decisions. Existing approaches to scientific impact prediction are presented in “[Measuring and predicting scientific impact](#)” section.

On the other hand, however, reliance on impact predictions risks statistical discrimination (Aigner & Cain, 1977; Arrow, 1973). Impact prediction models are trained on historical bibliographic data that encodes past discrimination against female researchers. Learning that female researchers, on average, had lower scientific impact than male researchers in the past, the model systematically predicts lower impact for current female researchers. In this case, women are penalized even if gender is no longer related to scientific impact. Using impact predictions to support hiring, tenure, and funding decisions therefore has the potential to rationalize and perpetuate gender inequality by limiting female researchers’ access to those positions and resources that allow them to conduct impactful research.

Measuring and predicting scientific impact

How can we measure the scientific impact of a researcher? Two metrics of impact are conventionally used in the literature: the cumulative citation count and the h -index (Hirsch, 2005). An h -index of k indicates that the k most-cited papers of a researcher received at least k citations. Both metrics equate scientific impact with peer recognition within the scientific community, reflecting the idea that citation counts represent the “collective wisdom of the scientific community on the paper’s importance” (Wang & Barabási, 2021, p. 182). The metrics come with specific advantages and disadvantages (Hirsch, 2005, 2007; Wang & Barabási, 2021). The cumulative citation count captures a scientist’s total impact but can be skewed by a few outliers (“big hits”) and rewards researchers for co-authoring on many low-impact papers. The h -index is robust to outliers (a single high-impact paper only increases the h -index by 1) and only rewards co-authorship on papers whose citation count exceeds the researcher’s current h -index. A high h -index therefore indicates consistent high-impact work. Several modifications of the h -index have been proposed that give more “credit” to high impact papers (Alonso et al., 2009). Some authors complement citation-based metrics with the overall number of publications (a measure of research productivity) and the number of publications in high-impact journals (defined by the journal impact factor) (Bertsimas et al., 2015).

In an effort to go beyond quantifying the past achievements of researchers, several authors recently attempted to predict future scientific impact (Acuna et al., 2012; Ayaz et al., 2018; Bertsimas et al., 2015; Dong et al., 2016; Weihs & Etzioni, 2017; Zuo & Zhao, 2021). Table 1 gives an overview of existing models for h -index prediction. In the most extensive study to date, Weihs and Etzioni (2017) used machine learning models (random forests and gradient-boosted regression trees) to predict the scientific impact of Computer Scientists up to 10 years into the future. The authors constructed a data set of approximately 800,000 individual researchers active in Computer Science during the period 1975 to 2016. The bibliographic profiles of these researchers in the year 2005 were then used to predict their h -index in the subsequent 10-year period (2006 to 2015). From the bibliographic profiles, Weihs and Etzioni (2017) extracted 44 features that characterize each researcher’s past impact history (e.g., total citation count until 2005), the researcher’s position in the co-authorship network (e.g., centrality), and the researcher’s publication venues (e.g., mean citations per paper of journals in which the researcher published).

Table 1 shows that the models are able to predict the future h -index with relatively high accuracy. In Weihs and Etzioni (2017), for instance, the prediction models were able to explain approximately 83% of the variation in h -indices measured 5 years into the future. The reported accuracy estimates might be overly optimistic, however. First, accuracy usually declines if the time window for the prediction increases (Acuna et al., 2012; Bertsimas et al., 2015; Dong et al., 2016; Weihs & Etzioni, 2017). Second, accuracy is quite low for early career researchers with short publication lists (Penner et al., 2013). Acceptable accuracy with R^2 values around 0.70 and 0.80 are only obtained among researchers with a minimum career age (years since first publication) of 6 to 8 years (Ayaz et al., 2018). Third, the citation count and the h -index are cumulative measures of scientific impact (Zuo & Zhao, 2021). Increases in both metrics over time might be driven by the additional citations accumulated by existing rather than future papers. The accuracy of the prediction models might, therefore, derive from their ability to predict the future impact of past work and not their ability to predict whether a

Table 1 Overview of scientific impact prediction models

Author (year), discipline	Data, inclusion criteria	Outcome	Method	Accuracy
Acuna et al. (2012), Neuroscience	$N = 3,085$, career age 5–12 years, $h(0) > 3$	$h(5)$, $h(10)$	Linear regression with elastic net regularization	$R^2(5) = 0.66$, $R^2(10) = 0.50$
Ayaz et al. (2018), Computer Science	$N = 938,205$, first publication 1936–2007	$h(5)$	Linear regression	$R^2(5) = 0.82$
Bertsimas et al. (2015), Operations Research	$N = 43,047$, first publication before 1995	$h(9)$, $h(16)$	Random forest	$RMSE(9) = 0.40$, $RMSE(16) = 0.90$
Dong et al. (2016), Computer Science	$N = 1,712,433$, $h(0) > 9$	$h(5)$, $h(10)$	Logistic regression	$R^2(5) = 0.90$, $R^2(10) = 0.75$
Mazloubmian (2012), diverse	$N = 150,000$, career age > 5 years	$c(10)$	Multi-level linear regression	$R^2(10) = 0.74$
Weilhs and Etzioni (2017), Computer Science	$N = 16,578$, career age: 5–12 years, $h(0) > 3$	$h(5)$, $h(10)$	Random forest, gradient boosted regression trees	$R^2(5) = 0.83$, $R^2(10) = 0.72$

$h(t)$ is the h -index t years in the future and $c(t)$ is the cumulative citation count t years in the future. $R^2(t)$ is the coefficient of determination and $RMSE(t)$ is the root mean squared error for predictions of the h -index or citation count t years in the future. Career age is the number of years since the first publication of the researcher

researcher will produce high-impact papers in the future (Mazlounian, 2012). Indeed, lower accuracy is obtained when predicting change in *h*-index (e.g., $R^2 = 0.60$ for predicting change over 5-year period) (Weihs & Etzioni, 2017) and citations received by future work (Mazlounian, 2012).

A related literature developed around the problem of rising star prediction. Rising stars are (early-career) researchers with an initially low research profile who subsequently become influential researchers with high scientific impact (Li et al., 2009). In a cumulative effort, researchers developed a series of indices for ranking researchers according to their potential to become highly impactful stars (Daud et al., 2017, 2020; Li et al., 2009; Nie et al., 2019; Zhang et al., 2016b). The indices combine information about the publication history (e.g., publication count, cumulative citation count), the temporal dynamic of the publication history (e.g., one-year change in publication or citation count), the position in co-authorship and citation networks (e.g., number of co-authors, scientific impact of co-authors, centrality in the network), and the prestige of publication venues (e.g., average citation count of papers published in the venue) of researchers. In empirical validations, researchers who are predicted to be rising stars by the indices are indeed found to outperform their lower-ranked peers in terms of their growth in scientific impact (Li et al., 2009; Panagopoulos et al., 2017; Zhang et al., 2016a).

However, even relatively high aggregate accuracy does not guarantee that the prediction models perform equally well for all sub-groups of researchers. We have already seen that existing impact prediction models perform worse for early career researchers than for more seasoned scholars. To the extent that model predictions influence hiring, tenure, and funding decisions, such differences in prediction error rates could translate into systematically biased and discriminatory decisions. The following sections investigate whether such differences in error rates exist between male and female researchers.

Data and methods

Data for this paper are drawn from a large-scale bibliographic data set of Computer Science researchers compiled by Weihs and Etzioni (2017) via the Semantic Scholar API. The data are publicly available online.¹ The data set contains bibliographic information on approximately 800,000 Computer Science researchers, active between 1975 and 2016, who published approximately four million papers. For each researcher, the data set includes their bibliographic profile in 2005 and measures of scientific impact (cumulative citation count, *h*-index) for the years 2006 to 2016. The full set of features that describe researchers' bibliographic profiles is listed in Table 5 (Appendix 1). The features encompass researchers' past impact histories (e.g., total citation count until 2005), positions in co-authorship networks (e.g., centrality), and publication venues (e.g., mean citations per paper of journals in which the researcher published).

I build and evaluate a realistic scientific impact prediction model that could be used to support decisions regarding hiring and funding of post-doctoral researchers. The analysis focuses on post-doctoral researchers for two reasons: (1) Hiring, tenure, and funding decisions among post-doctoral researchers are critical for the long-term retention of (female) researchers in academia. These decisions are also associated with substantial uncertainty

¹ The data were accessed via the GitHub page of Luca Weihs <https://github.com/Lucaweihhs/impact-prediction>. If the GitHub page is no longer available, please contact the author of this article to get access to the data.

for university departments and funding agencies as post-doctoral researchers usually have only a limited publication portfolio to demonstrate their scientific potential (Bertsimas et al., 2015). Such uncertainty is less pronounced among more senior researchers with longer track-records to prove scientific excellence. (2) As suggested by existing research (Ayaz et al., 2018), impact prediction is extremely difficult for researchers with short publication histories. Meaningful predictions are possible, however, for post-doctoral researchers with at least 5 years of publication experience. Hence, impact prediction among post-doctoral researchers is a realistic use case because it is of high practical relevance and technically feasible.

To approximate a sample of post-doctoral researchers, I restrict the analysis sample to researchers with career ages (years since first publication) of 5 to 12 years ($N = 531,502$ excluded). The span of career ages used to select post-doctoral researchers is comparable to the span used by Zuo and Zhao (2021) who derive it from the duration of typical peer review processes. The performance of post-doctoral researchers who apply for tenure or funding should be compared to the performance of their active peers, not the performance of those who left academia. Therefore, researchers who published only one paper until 2005 are excluded to get rid of inactive researchers ($N = 155,234$ excluded). It is reasonable to assume that researchers with a career age of at least 5 years who published only one paper left academia. Finally, I can only include bibliographic profiles that contain full first names that allow me to infer the researcher's gender. Profiles that include only initials or otherwise invalid names are excluded ($N = 13,580$ excluded). It was not possible to determine the gender of $N = 24,552$ researchers (18.1% of all 135,708 valid names) with sufficient reliability. These observations are excluded. The final size of the analysis sample is $N = 111,156$.

Choosing the span of career ages that is used to select post-doctoral researchers is somewhat arbitrary, despite the guidance provided by Zuo and Zhao (2021). As a robustness check, all analyses are run on a more restrictive sample of mid-career researchers with career ages between 4 and 6 years and on a less restrictive sample that includes all researchers with a career age of at least 5 years (see Appendix 3). The final size of the samples is $N = 55,591$ and $N = 185,449$, respectively. The substantive results regarding gender disparities are stable across the different samples.

Tables 6 and 7 (Appendix 2) provide descriptive statistics for the main analysis sample. The sample is composed of 94,038 (84.6%) male and 17,118 (15.4%) female researchers. The share of women in the sample is slightly lower than the overall representation of women in Computer Science (approx. 22.0% in 2019) (NCSES, 2019). Compared to the full data set, the analysis sample contains researchers that are more impactful (higher h -index and citation count), productive (higher number of papers), and central in their co-authorship networks (higher PageRank) and publish in higher-impact journals. Such differences are expected based on the sample inclusion criteria.

Prediction setup

The bibliographic profiles collected in 2005 are used to predict medium-term scientific impact, defined as the researcher's h -index in the year 2010 (*impact prediction*). The same data are used to predict whether the researcher will be in the top-25% of the 2010 discipline-specific h -index distribution (*high-performer prediction*).² Researchers whose

² Note that this prediction task is related to the rising star prediction reviewed in the “[Measuring and predicting scientific impact](#)” section. Rising star prediction, however, aims at identifying researchers with an above-average *growth* in their scientific impact (e.g., the change in the h -index between 2005 and 2010). In

h -index is in the top-25% of the 2010 h -index distribution are referred to as high-performers, all remaining researchers are referred to as low-performers (even though this group also contains researchers with average performance). The continuous outcome makes impact prediction a *regression problem*. The high-performer prediction is a *classification problem* due to its binary outcome. Based on the evidence on gender bias reviewed in the “[Gender in science](#)” section, the gender of researchers is treated as a protected attribute. Women are considered a historically disadvantaged group in academia and the impact prediction models are therefore audited for disparate impact against women.

The analysis sample is split into a training set and a test set using a 70:30 splitting rule. The prediction models are fitted in the training set ($N_{tr} = 77,809$ observations). The predictive performance of the models is evaluated in the test set ($N_{te} = 33,347$ observations). The prediction task includes the following components.

- *Predictors* X : The features, presented in Table 5 (Appendix 1), that are derived from bibliographic profiles in 2005.
- *Gender* G : Gender is considered a protected attribute, with $G = g^*$ for women and $G = g$ for men. The gender of researchers is derived from full names using the commercial service *Gender API*, which was shown to outperform competing services—especially in recognizing Asian names that are quite frequent in the sample (Santamaría & Mihaljević, 2018). The Gender API reports the *accuracy* of the gender prediction (percentage of times that a specific name is associated with the predicted gender in the underlying database) and the number of *samples* of the name in the database. Following advice from the evaluation study of Santamaría and Mihaljević (2018), the gender prediction is only used if accuracy ≥ 60 and samples ≥ 65 to obtain reliable gender classifications with acceptable reduction in the number of classified cases.
- *Observed outcome* $Y \in \mathbb{Z}_+$: The true h -index of the researcher in 2010. For the *high-performer prediction*, the binary label $C \in \{0, 1\}$ indicates whether the 2010 h -index of a researcher is in the top-25% of the observed h -index distribution. The corresponding cutoff value is the sample 75%-percentile $q_{75} = 3$, such that $C = 1$ iff $Y > q_{75}$ and $C = 0$ otherwise.
- *Predicted outcome* $\hat{Y} \in \mathbb{Z}_+$: The predicted h -index of the researcher in 2010. The binary label $\hat{C} \in \{0, 1\}$ indicates whether the h -index of the researcher is predicted to be in the top-25% of the observed h -index distribution ($\hat{C} = 1$) or not ($\hat{C} = 0$).
- *Predicted score* $\hat{S} = \hat{P}(C = 1)$: The predicted probability that the researcher is in the top-25% of the observed h -index distribution. The score is the raw output of the prediction models for the *high-performer prediction*. It is translated into the binary classification \hat{C} by applying a cutoff t such that $\hat{C} = 1$ iff $\hat{S} \geq t$ and $\hat{C} = 0$ otherwise. The value of the cutoff t is chosen by the model builder.

Note that gender is not used as a predictor in the main analyses. Anti-discrimination laws (e.g., Title VII in the Civil Rights Act of 1964 in the US context) recognize gender as a protected attribute. Including gender as a predictor in a model that is ultimately used

Footnote 2 (continued)

contrast, I focus on identifying researchers with an above-average *level* of their h -index in 2010, irrespective of whether the researchers started with a high or low h -index in 2005. Both prediction tasks are relevant for hiring committees and funding agencies that wish to select the best candidates. The results presented in this paper, therefore, complement existing research on rising star prediction.

to grant or deny employment opportunities constitutes an instance of disparate treatment (Barocas & Selbst, 2016). University departments and funding agencies risk litigation if they use models that include gender as a predictor. Realistic impact prediction models, therefore, are unlikely to include gender. For comparison, Appendix 3 reports additional analyses for impact prediction models that explicitly include gender as a predictor.

Models that predict the cumulative citation count are estimated as additional robustness checks. Different prediction targets might be affected by different amounts of gender bias. It has been suggested, for instance, that women publish less than men but receive a similar number of citations for their publications (Symonds et al., 2006). The *h*-index might be a more gender-neutral prediction target because, at a certain point, the higher productivity of men might no longer translate into a higher *h*-index if the additional papers do not receive more citations than the current *h*-index. The results of the robustness checks are reported in Appendix 3.

Prediction models

Models

The performance and fairness implications of four prediction models are compared: standard regression models, Random Forests (Breiman, 2001), Gradient Boosting Machines (Friedman, 2001), and Extreme Gradient Boosting Machines (Chen & Guestrin, 2016). The last three models are prominent ensemble methods that are able to handle prediction tasks with many predictors and non-linear relationships. All models have been used in previous studies on scientific impact prediction or rising star prediction.

- *Linear regression (LR)*: An unpenalized linear OLS regression, including only main effects for all predictors, is used to predict the future *h*-index (*impact prediction*).
- *Logistic regression (LOG)*: An unpenalized logistic regression, including only main effects for all predictors, is used to predict whether the future *h*-index is among the top-25% of the discipline (*high-performer prediction*).
- *Random forest (RF)*: Ensemble method that combines de-correlated regression (or classification) trees to reduce the risk of over-fitting. Each tree in the ensemble is fitted within a separate bootstrap sample. Regression trees are de-correlated by selecting, at each splitting point, a random subset of features that the fitting algorithm is allowed to use. Random Forests are applicable to regression problems (*impact prediction*) and classification problems (*high-performer prediction*).
- *Gradient Boosting Machines (GBM)*: Ensemble method that combines sequentially grown regression (or classification) trees to reduce the risk of over-fitting. Trees are grown sequentially by repeatedly (1) fitting a new tree to the pseudo-residuals from the current model and, then, (2) adding the new decision tree to the current model to update the residuals. Each tree in the sequence is fitted within a separate bootstrap sample. A learning rate controls how much impact each new tree has on the current model. Gradient Boosting Machines are applicable to regression problems (*impact prediction*) and classification problems (*high-performer prediction*).
- *Extreme gradient boosting (XGB)*: An optimized version of the GBM algorithm with finer control of the tree-building process. It adds a pruning parameter that controls the depth and complexity of the single trees in the sequence. Higher pruning parameter

values reduce the risk of over-fitting by preventing splits that do not improve the fit of the tree sufficiently. Each tree is fitted within a separate bootstrap sample to reduce the computational burden and the risk of over-fitting.

Model tuning and cross-validation

All three ensemble methods are associated with hyper-parameters that need to be tuned in order to build a well-performing model. For the Random Forest, the hyper-parameters are the number of trees in the ensemble (*num.trees*) and the number of randomly selected features (*m.try*). For the Gradient Boosting Machine, the hyper-parameters are the number of sequentially grown trees (*num.trees*), the learning rate (*learning.rate*), the number of splits allowed in a single tree (*max.depth*), and the fraction of observations selected into the bootstrap samples (*bag.fraction*). Extreme Gradient Boosting adds the following hyper-parameters: the pruning parameter (*gamma*) and a parameter that controls the minimum number of observations in each node of the tree (*min.child.size*).

The values of the hyper-parameters are selected via 5-fold cross-validation (James et al., 2013). The training data is partitioned into five segments. Each segment is used once as hold-out data. The hold-out data are ignored during model training. The model is trained on the four remaining segments. The hold-out data are used to assess the out-of-sample predictive performance of the trained models for observations not seen during model training. For the *impact prediction* (regression problem), performance is measured via the root-mean-square error (RMSE) averaged across the five hold-out samples.

The $RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$, where \hat{y}_i and y_i are the predicted and observed outcome for the $i = 1, \dots, n$ hold-out observations. For the *high-performer prediction* (classification problem), performance is measured via the prediction accuracy averaged across the five hold-out samples. Accuracy is defined as $ACC = P(\hat{C} = C)$, i.e., the share of correct classifications. The full grid of hyper-parameter values tested during cross-validation is reported in Table 12 (Appendix 4).

Model selection

The hyper-parameters selected for the Random Forest model are *m.try* = 15 (*impact prediction*) and *m.try* = 10 (*high-performer prediction*). In both cases, setting the number of trees to *num.trees* = 500 is sufficient to achieve stable estimates of the out-of-sample performance. The out-of-sample performance of the models is $RMSE = 1.042$ and $ACC = 0.925$, respectively. The hyper-parameters chosen for the Gradient Boosting Machine and the *impact prediction* task are *num.trees* = 400, *learning.rate* = 0.025, *max.depth* = 7, and *bag.fraction* = 0.6. The corresponding model yields $RMSE = 1.032$. For the *high-performer prediction*, the best-performing Gradient Boosting Machine is obtained with *num.trees* = 800, *learning.rate* = 0.01, *max.depth* = 7, and *bag.fraction* = 0.8. The accuracy of the corresponding model is $ACC = 0.927$. The best Extreme Gradient Boosting model for the *impact prediction* is obtained with *num.trees* = 500, *learning.rate* = 0.025, *max.depth* = 7, *bag.fraction* = 0.6, *gamma* = 5, and *min.child.weight* = 4. For *high-performer prediction*, the respective hyper-parameter values are *num.trees* = 500, *learning.rate* = 0.01, *max.depth* = 9, *bag.fraction* = 0.6, *gamma* = 1.5, and *min.child.weight* = 6. The corresponding models yield $RMSE = 1.029$ and $ACC = 0.926$.

Table 2 Confusion matrix (based on the Gradient Boosting Machine)

	Observed	
Predicted	$C = 0$	$C = 1$
$\hat{C} = 0$	23,523	922
$\hat{C} = 1$	2088	6814

Model evaluation

The prediction models are fitted in the training set with the selected hyper-parameters. The predictive performance of the models is evaluated in the test set via the following metrics.

- The R^2 metric compares the performance of the prediction model against a null model that assigns the mean outcome to all observations. $R^2 \in [0, 1]$ can be interpreted as the proportion of total variance in the outcome explained by the prediction model. It is defined as $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$, where n is the number of researchers in the test set, y_i is the observed h -index of the i -th researcher in 2010, \hat{y}_i is the predicted h -index of the i -th researcher in 2010, and \bar{y}_i is the average h -index in 2010 over all researchers in the test set.
- The *past-adjusted* R^2 metric proposed by Weihs and Etzioni (2017). The h -index is non-decreasing over time and highly auto-correlated. The R^2 metric is therefore inflated and overestimates the ability of the model to predict the impact of a researcher’s future work. The past-adjusted $pa-R^2$ removes some of the auto-correlation by subtracting the baseline h -index observed in 2005 from the h -index observed in 2010. It is defined as $pa-R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$, where $z_i = y_{i,2010} - y_{i,2005}$ is the past-adjusted h -index in 2010 and \bar{z}_i is the average past-adjusted h -index in the test set.

Additional performance metrics are used for the *high-performer prediction*. The metrics are based on confusion matrices, such as the one displayed in Table 2, and compare the predicted classification \hat{C} and the observed classification C in the test set. The classification performance of the tuned models in the test set is evaluated via the following metrics.

- *Accuracy (ACC)*. The proportion of correctly classified researchers, defined as $ACC = P(\hat{C} = C)$.
- *Sensitivity/True Positive Rate (TPR)*. The proportion of researchers correctly classified as high-performers among all researchers who actually are high-performers, defined as $TPR = P(\hat{C} = 1|C = 1)$. The False Negative Rate $FNR = 1 - TPR = P(\hat{C} = 0|C = 1)$ is the proportion of true high-performers who are not classified as being high-performers.
- *Specificity/True Negative Rate (TNR)*. The proportion of researchers correctly classified as low-performers among all true low-performers, defined as $TNR = P(\hat{C} = 0|C = 0)$. The False Positive Rate $FPR = 1 - TNR = P(\hat{C} = 1|C = 0)$ is the proportion of true low-performers who are wrongly classified as being high-performers.
- *Positive Predictive Value (PPV)*. The proportion of true high-performers among all researchers classified as high-performers, defined as $PPV = P(C = 1|\hat{C} = 1)$.
- *Negative Predictive Value (NPV)*. The proportion of true low-performers among all researchers classified as low-performers, defined as $NPV = P(C = 0|\hat{C} = 0)$.

Fairness evaluation

Fairness in data-driven decision-making is treated within the legal framework of disparate treatment and disparate impact (Barocas & Selbst, 2016). Disparate treatment is present if the decision procedure directly uses protected attributes (e.g., gender) to grant or deny opportunities. The prediction models studied here do not use gender (or any other protected attribute) as a predictor, such that disparate treatment is not applicable. However, other predictors might work as proxies for gender, leading to indirect disparate treatment. Disparate impact refers to facially neutral decision procedures that, nevertheless, disproportionately disadvantage members of protected groups. Disparate treatment is present if the prediction model systematically misclassifies members of one protected group (e.g., women) more frequently than members of another group (e.g., men). Misclassification implies that members of one protected group are treated more or less favorably than equally qualified members of the other group.

To evaluate the fairness of the impact prediction models, I compare the performance metrics defined above across the protected groups $G = g^*$ (female researchers) and $G = g$ (male researchers) (Mitchell et al., 2021). Let $PM(G)$ be one of the performance metrics defined above. The prediction model is considered fair (according to the metric), if $\frac{PM(G=g^*)}{PM(G=g)} = 1$. That is, the model is considered fair if it performs equally well for male and female researchers. This is a *group-level* definition of fairness in the sense that the metrics require equal error rates across *groups*. The investigation of *individual* fairness (Dwork et al., 2012), the notion that similar *individuals* should receive similar predictions, is left for future investigation.

Results

Impact prediction task

Model evaluation

Performance metrics for the prediction models are presented in column 3 of Table 3. All four models have good predictive performance in the test set, reaching an R^2 of 0.850 or higher. The overall performance of the models is comparable to performance values obtained in prior investigations (see Table 1). As expected, the past-adjusted R^2 is markedly lower, ranging from 0.539 for the linear regression to 0.616 for the XGB. Overall, the XGB performs slightly better than the GBM and the RF. All three ensemble methods surpass the linear regression model.

Fairness evaluation

Figure 1 plots the joint distribution of observed and predicted h -index values, separately for male (blue dots) and female (red triangles) researchers. The predictions are based on the XGB. The two distributions look very similar and there are few outliers. It is apparent, however, that there are gender differences at the top of the h -index distribution. There are a few male researchers with an observed h -index > 30 whereas no female researcher is observed in this area.

Table 3 Performance and fairness metrics for impact prediction

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio women/men (6) = (5)/(4)
XGB	R^2	0.875	0.876	0.868	0.991
	pa- R^2	0.616	0.615	0.621	1.010
GBM	R^2	0.871	0.871	0.866	0.994
	pa- R^2	0.603	0.601	0.614	1.023
RF	R^2	0.871	0.872	0.864	0.991
	pa- R^2	0.602	0.601	0.610	1.014
LR	R^2	0.850	0.852	0.837	0.982
	pa- R^2	0.539	0.541	0.531	0.981

The best value on each metric is highlighted in boldface

XGB Extreme Gradient Boost, GBM Gradient Boosting Machine, RF random forest, LR linear regression

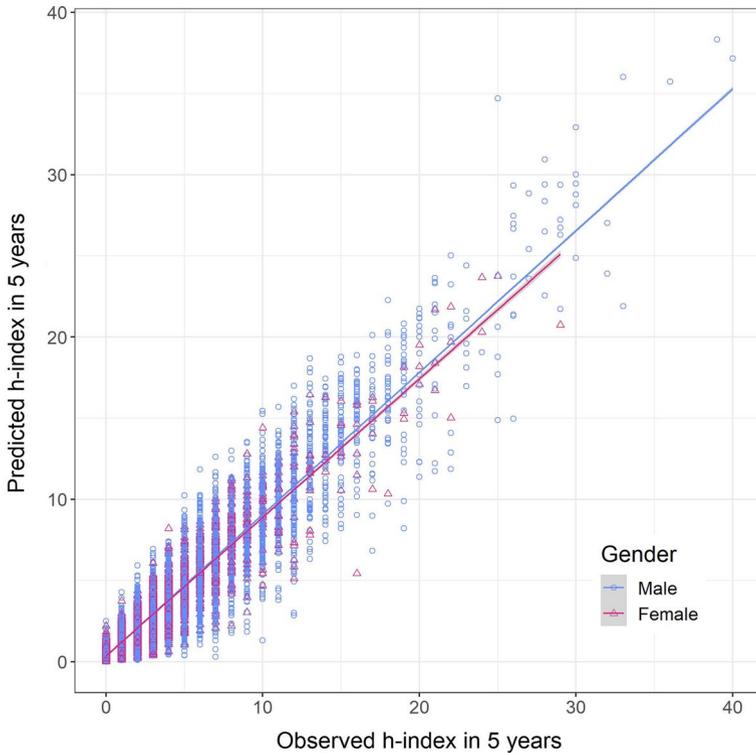


Fig. 1 Observed vs predicted h -index in 5 years (based on XGB model). (Color figure online)

Equation 1 shows a linear OLS regression fitted to the predicted h -index values, where \hat{h} is the predicted h -index, h_{obs} is the observed h -index, and fem is a binary indicator equal to 1 for female and to 0 for male researchers. All coefficients except the main effect of gender are statistically significant at the 5% level. On average, each increase in the observed h -index by 1 is associated with an increase in the predicted h -index by 0.874 for male researchers and by $0.874 - 0.021 = 0.853$ for female researchers. That is, on average, the model under-predicts the h -index of female researchers more strongly than the h -index of male researchers.

$$E[\hat{h}|h_{\text{obs}}, fem] = 0.340 + 0.874 \times h_{\text{obs}} + 0.029 \times fem - 0.021 \times h_{\text{obs}} \times fem \quad (1)$$

Table 3 presents performance metrics separately for male (column 4) and female researchers (column 5). Column 6 presents the fairness ratio, calculated by dividing the value of the metrics among female researchers by the value of the metrics among male researchers. A fairness ratio of 1 indicates perfect fairness, whereas a ratio < 1 indicates unfairness against female researchers and a ratio > 1 indicates unfairness against male researchers. There are only small gender differences in terms of overall model performances. The biggest gender difference is a 2.3% (ratio: 1.023) higher past-adjusted R^2 for female compared to male researchers in the GBM.

Robustness checks

Several checks are performed to test whether the results are robust to changes in the outcome, the predictors, and the sample. The results of these checks are reported in Appendix 3. The predictive performance slightly decreases when predicting the h -index 8 years (instead of 5 years) into the future. The XGB, for instance, achieves an $R^2 = 0.818$ and a past-adjusted $R^2 = 0.591$ in this scenario. The fairness ratios remain virtually unchanged. The models perform slightly better when predicting the citation count 5 years into the future, achieving an R^2 of at least 0.901 and an adjusted R^2 of at least 0.813. Again, there is no strong indication of gender differences in model performance. The inclusion of researcher gender as a predictor does not change the overall performance and fairness of the prediction models. This result does not indicate that gender is irrelevant for predicting scientific impact. Rather, the other predictors in the model might act as proxies for gender. Training the models on a larger sample of all researchers with a career age of at least 5 years (not excluding those with more than 12 years) improves model performance. This result is expected as more seasoned scholars have longer publication histories that make prediction easier (Ayaz et al., 2018). The improved model performance also translates into even smaller gender differences. In contrast, restricting the sample to researchers with a career age between 4 and 6 years reduces model performance. The XGB, for instance, achieves an $R^2 = 0.764$. The fairness of the models also slightly deteriorates. In the XGB, the R^2 is 1.9% (ratio: 0.981) lower for female compared to male researchers. The biggest gender difference is an 8% (ratio: 0.920) lower past-adjusted R^2 for female compared to male researchers in the linear regression.

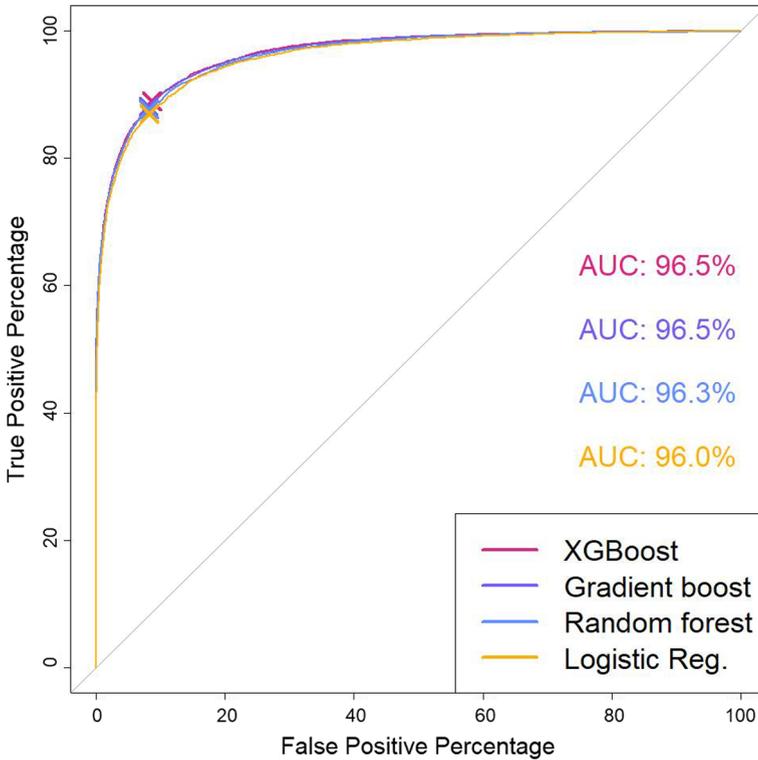


Fig. 2 Receiver operating characteristic (ROC) curve. (Colour figure online)

High-performer prediction task

Model evaluation

Figure 2 presents the receiver operating characteristic (ROC) curves of the four prediction models. The X-axis represents the False Positive Rate (FPR), the Y-axis the True Positive Rate (TPR). A perfect prediction model that correctly classifies all researchers has an ROC curve that reaches the upper left corner of the figure where FPR = 0 (such that True Negative Rate = 1 – FPR = 1) and TPR = 1. The area under the curve (AUC) measures the area between the diagonal and the ROC curve of a model. The AUC varies between 0 and 100. The higher the AUC, the closer the model to the upper left corner and the better the model performance. All four prediction models have an AUC of 96.0 or higher, indicating good predictive performance.

As explained in the “Prediction setup” section, the prediction models produce a score \hat{S} that corresponds to the predicted probability of a researcher to become a high-performer $\hat{P}(C = 1)$. The score is translated into a binary classification \hat{C} by applying a cutoff t such that $\hat{C} = 1$ iff $\hat{S} \geq t$ and $\hat{C} = 0$ otherwise. The cutoff t is chosen such that the resulting classifications maximize the sum of the TPR and the TNR: $\max_{t \in [0,1]}(TPR + TNR)$. The cross-symbols in Fig. 2 indicate the points on the ROC curves that correspond to the chosen cutoff values. For each model, this is the point on the ROC curve that is closest to the

Table 4 Performance and fairness metrics for high-performer prediction

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio Women/Men (6) = (5)/(4)
XGB	ACC	0.908	0.908	0.907	0.999
	TPR	0.890	0.890	0.886	0.995
	TNR	0.913	0.913	0.912	0.999
	PPV	0.755	0.760	0.727	0.957
	NPV	0.965	0.964	0.968	1.004
	$P(\hat{Y} = 1)$	0.273	0.277	0.254	0.918
GBM	ACC	0.910	0.910	0.911	1.001
	TPR	0.881	0.881	0.879	0.997
	TNR	0.918	0.918	0.919	1.001
	PPV	0.765	0.770	0.740	0.962
	NPV	0.962	0.962	0.966	1.005
	$P(\hat{Y} = 1)$	0.267	0.270	0.247	0.914
RF	ACC	0.909	0.908	0.909	1.001
	TPR	0.876	0.878	0.866	0.987
	TNR	0.918	0.918	0.921	1.003
	PPV	0.764	0.768	0.742	0.967
	NPV	0.961	0.961	0.963	1.003
	$P(\hat{Y} = 1)$	0.266	0.270	0.243	0.901
LOG	ACC	0.906	0.906	0.908	1.002
	TPR	0.870	0.870	0.865	0.994
	TNR	0.917	0.917	0.919	1.002
	PPV	0.761	0.765	0.737	0.964
	NPV	0.959	0.958	0.963	1.005
	$P(\hat{Y} = 1)$	0.265	0.269	0.245	0.910

The best value on each metric is highlighted in boldface. XGB for Extreme Gradient Boost, GBM for Gradient Boosting Machine, RF for Random Forest, and LOG for Logistic Regression

upper left corner. The chosen cutoff values are $t_{\text{XGB}} = 0.237$, $t_{\text{GBM}} = 0.228$, $t_{\text{RF}} = 0.288$, and $t_{\text{LOG}} = 0.232$.

Performance metrics for the prediction models are presented in column 3 of Table 4. The metrics are calculated for the classifications obtained by applying the chosen cutoff values. The GBM performs best, although the differences between models are small. The GBM predicts whether a researcher will be a high-performer with high accuracy (ACC) and correctly classifies 91.0% of researchers in the test set. The accuracy is significantly higher than the no information rate (NIR) of 0.768 that is obtained if all researchers are simply classified into the largest class (here: low-performers). The model correctly identifies 88.1% of the high-performers (TPR) and 91.8% of the low-performers (TNR). Among researchers classified as high-performers, 76.5% actually are high-performers (PPV) and among those classified as low-performers, 96.2% actually are low-performers (NPV).

Figure 3 (Appendix 2) displays feature importance scores that measure how much each predictor contributes to the model's predictive performance. The five features with the highest predictive power are (in decreasing order): *h*-index in 2005, citations obtained in 2005, number of papers published in 2004 and 2005, cumulative count

of published papers, and mean citations per year. This ordering is not changed when gender is included as an additional predictor. In fact, researcher gender is one of the predictors with the lowest importance score. Again, this does not indicate that gender is irrelevant but that the other predictors act as proxies for gender.

Fairness evaluation

Table 4 presents performance metrics separately for male (column 4) and female researchers (column 5). Column 6 presents the fairness ratio. Since the substantive results are very similar across all four models, only the GBM is discussed in detail. There are only small gender differences in terms of overall model performance (ACC-ratio = 1.001) and the detection of low-performers (TNR-ratio = 1.001 and NPV-ratio = 1.005). However, female researchers who are predicted to become high-performers have a 3.8% lower probability of actually becoming high-performers than male researchers (PPV-ratio = 0.962).

Interestingly, the probability that a researcher who actually becomes a high-performer is classified as a high-performer does not differ across gender (TPR-ratio = 0.997). This finding is in line with well-known impossibility theorems in the Fair Machine Learning literature (Chouldechova, 2016). Prediction models cannot simultaneously satisfy equality of PPV and TPR whenever base rates differ across groups and the classification is not perfect. Both conditions hold in the present case. The base rate of being in the top-quarter of the *h*-index distribution is 23.6% for male and 20.8% for female researchers in the test set. And overall accuracy is smaller than 1.00.

From the perspective of funding agencies and hiring committees, gender disparities in the TPR might be less concerning because they are more interested in the PPV. Agencies and committees care about how accurate the model is among researchers who are predicted to be in the top-25 and, thus, recommended by the prediction model. However, from the perspective of the individual researchers, gender disparities in the TPR are important. Being denied an opportunity (because of a wrong prediction) despite being qualified can seriously damage the career of a researcher. The present models are more likely to wrongly recommend a woman (PPV is lower among women than among men) than they are to deny qualified women opportunities (almost equal TPR among women and men).

The last row for each model in Table 4 compares the prevalence of positive predictions (i.e., $P(\hat{C} = 1)$) across male and female researchers. Female researchers have a 8.6% lower probability of being classified as a high-performer than male researchers (Prevalence-ratio = 0.914). Table 7 (Appendix 2) compares the predictors across male and female researchers in the test set. The comparison indicates that female researchers have less favorable past impact histories and lower publication counts, on average. It is therefore not surprising that female researchers are less likely to become high-performers. Clarifying whether these gender disparities are the result of discrimination against female researchers or other structural features of the academic system that disadvantage women is beyond the scope of this paper.

Robustness checks

Checks are performed to test whether the results are robust to changes in the outcome, the predictors, and the sample. The results of these checks are reported in Appendix 3. Overall model performance decreases when predicting whether researchers will be high-performers 8 years (instead of 5 years) into the future. The fairness ratios are stable as model performance declines to the same extent for male and female researchers. Model performance is higher when predicting whether researchers are in the top-25% of the citation count (instead of *h*-index) distribution 5 years in the future. Contrary to expectation, gender disparities are not more pronounced than for the *h*-index prediction. The results are robust to including gender as an additional predictor. Performance improves on all metrics (except PPV) when training the models on the larger sample including all researchers with career ages of 5 years or more. In the larger sample, the gender disparity in the prevalence of positive predictions grows, however. In the GBM, for instance, female researchers have a 12.8% lower probability of a positive prediction than male researchers (compared to 8.6% in the original sample). Performance decreases when the models are trained in the more restrictive sample of mid-career researchers with career ages between 4 and 6 years. In addition, the fairness of the models deteriorates. The PPV is 7.2% and the TPR is 4.1% lower among female than male researchers (compared to 3.8% and 0.3% in the original sample, respectively) in the GBM. Female researchers have a 14.5% lower probability of a positive prediction than male researchers (compared to 8.6% in the original sample). The decrease in model performance among mid-career researchers seems to work to the detriment of female rather than male researchers.

Discussion

Data-driven approaches promise to render hiring, tenure, and funding decisions in academia more evidence-based, efficient, and objective. At the same time, reliance on data-driven decision aids that are trained on historical data carries the risk of perpetuating discrimination against historically disadvantaged groups. In line with prior research, I find that the medium-term scientific impact of researchers with a publication experience of at least 5 years can be predicted with high accuracy. Identifying high-performing researchers was the most challenging task for the prediction models—only (PPV =) 76.5% of the researchers predicted to be high-performers actually became high-performers.

Overall, there is no strong evidence of disparate impact based on gender. Nevertheless, the models have a weak tendency to under-estimate the *h*-index of female researchers more strongly than the *h*-index of male researchers (see Eq. 1). In addition, the probability that the models wrongly recommend a researcher is higher among female than among male researchers (PPV-ratio = 0.962).

Implications for practice

Women who actually become high-performers are correctly classified at almost the same rate as men who become high-performers (TPR-ratio = 0.997). However, women are less likely to become high-performers in the first place and, accordingly, have a 8.6% lower probability of being classified as a high-performer than male researchers. Hiring, tenure,

and funding decisions that are based on model predictions therefore risk to perpetuate the under-representation of women in Computer Science. Gains in decision-accuracy (in terms of allocating funds to future high-performers) come at the cost of limiting the chance of female researchers to obtain academic positions and research funding.

To the extent that gender differences in the chance of becoming a high-performing researcher are rooted in structural discrimination against female researchers, there is a need for redress. Hiring committees and funding agencies could, for instance, apply gender-specific cutoff values to select researchers with high potential for future scientific impact (Abramo et al., 2015). Other techniques to mitigate unfairness in machine learning models are reviewed by Caton and Haas (2020). These techniques aim at eliminating unfairness in the training data (pre-processing), making fairness adjustments during model training (in-processing), and adjusting the model output after training (post-processing). Additionally, structural features of the academic system that hinder women's career advancement should be eliminated. Such factors include gender disparities in access to mentorship (Blau et al., 2010; Sheltzer & Smith, 2014) and collaboration (Jadidi et al., 2018) networks, family obligations (Anders, 2004), and allocation of credit (West et al., 2013; Sarsons, 2017), as well as stereotype-based discrimination against women (Eaton et al., 2020; Moss-Racusin et al., 2012).

More generally, targeting funds at researchers who are already successful might reinforce the rich-get-richer dynamics (Matthew effect) that are already present in academia (Merton, 1968). In effect, researchers who represent less prominent positions, ideas, and research agendas may find it increasingly difficult to secure funding and academic positions. In the long run, the reduced diversity of ideas and approaches may hamper scientific progress.

Limitations of impact prediction

There are good reasons why hiring, tenure, and funding decisions should not rely solely on (predicted) performance metrics (Hicks et al., 2015). First, soft skills (e.g., personality) and other qualitative indicators of performance (e.g., teaching and other academic services) are not captured by standard metrics. Second, the practical impact of a researcher's findings in real-world applications is not recognized by standard performance metrics. Existing metrics equate impact with high citation success, i.e., peer recognition within the scientific field. Third, scientific (sub)fields differ in publication and citation practices, making it hard to compare researchers who come from different fields. Fourth, public agreement on a set of performance metrics invites manipulation and scientific misconduct. There is already evidence of attempts to boost citation counts via self-citations, honorary authorship, and forced citations (Flanagin, 1998; Seeber et al., 2019; Wilhite and Fong, 2012). Fifth, the future scientific success of early career researchers with short publication lists is difficult to predict, leading to a high risk of prediction errors (Ayaz et al., 2018; Penner et al., 2013). In face of these challenges, combining qualitative expert judgement with quantitative performance metrics constitutes a promising strategy.

Apart from that, impact prediction and other machine learning models can help to detect and quantify more subtle forms of gender bias – for instance, bias hidden in textual data (Leavy et al., 2020).³ Making such disparities visible is a necessary first step towards mitigating unfair discrimination.

³ I am indebted to an anonymous reviewer for highlighting the potential of machine learning models to detect subtle biases.

Limitations and future research

The results presented in this paper pertain first and foremost to the discipline of Computer Science. Disciplines are characterized by different traditions and practices regarding publication, citation, and credit allocation (Wang & Barabási, 2021) and differ in the historic representation of women (Ceci et al., 2014). While impact prediction models have been tested in multiple disciplines (see Table 1 in “[Measuring and predicting scientific impact](#)” section), more research on their possible disparate impact is needed before they can be integrated into hiring, tenure, and funding decisions across all disciplines. The impact prediction model was only audited for disparate impact by gender. The focus on gender was chosen due to the historic under-representation of women in Computer Science (NCSES, 2019). Other well-known bases of discrimination (ethnicity, disability status, social class background) should be investigated in future work. Such work will need to combine bibliographic information with researcher self-reports on their ethnicity, disability status, and class background as such features cannot be predicted from the information contained in bibliographic databases. A major challenge of working with bibliographic databases is name disambiguation (Sanyal et al., 2021; Tekles & Bornmann, 2019). One researcher can appear under different names or multiple researchers can share the same name, such that it becomes difficult to correctly attribute publications and citations to researchers. In the first case, the scientific impact of the researcher is under-estimated whereas it is over-estimated in the second case. Unfortunately, the data set compiled by Weihs and Etzioni (2017) contains no information that allows me to assess the extent of these two biases. It is unlikely, however, that women and men are affected differently by these two processes, such that the main results on gender differences should not be biased. Finally, this paper focuses on whether the prediction models produce disparate impact. It is, however, unclear how model predictions are used by human decision-makers. Humans are often sceptical towards data-driven decision-support systems, overestimate their own accuracy, and, therefore, resist the recommendations of these systems (Burton et al., 2020). There is also evidence that quantitative metrics of research impact are used as screening tools rather than as ultimate decision criteria (Reymert, 2021). Future research should directly investigate how impact predictions influence hiring, tenure, and funding decisions in academia.

Conclusion

Prior work demonstrated that data-driven approaches have the potential to improve the efficiency of resource allocation in academia (Bertsimas et al., 2015). This paper contributes a more cautionary perspective on the use of data-driven approaches in academia. My results suggest that reliance on impact prediction models can have the unintended consequence of perpetuating gender inequality in access to research positions and funding. A one-sided focus on efficiency may also undermine intellectual diversity and hamper scientific progress in the long run. Given the broad practical and ethical implications of scientific impact prediction, the research community is called to discuss and clarify the criteria and procedures that should guide the allocation of research resources.

Appendix 1: Description of predictors

Table 5 Features for *h*-index prediction (taken from Weihs & Etzioni, 2017)

Description (feature name)
<i>h</i> -index (author_hindex)
Change in <i>h</i> -index over the last 2 years (author_hindex_delta)
Cumulative citation count (author_citation_count)
Cumulative key citation count (author_key_citation_count)
Citations this year (0) and 1 year ago (1) (author_citations_delta_{0,1})
Key citations this year (0) and 1 year ago (1) (author_key_citations_delta_{0,1})
Mean number of citations per paper (author_mean_citations_per_paper)
Change in citations per paper over last 2 years (author_mean_citations_per_paper_delta)
Mean number of citations per year (author_mean_citations_per_year)
Cumulative count of published papers (author_papers)
Number of papers published in last 2 years (author_papers_delta)
Rank of author (between 0 and 1) among all other authors in terms of mean citations per year (author_mean_citation_rank)
PageRank in unweighted co-authorship network (author_unweighted_pagerank)
PageRank in weighted co-authorship network (author_weighted_pagerank)
Career length: years since first paper published (author_age)
Total number of co-authors in last 2 years (author_recent_num_coauthors)
Maximum number of citations for any of author's papers (author_max_single_paper_citations)
<i>h</i> -index (mean, minimum, and maximum) of venues author has published in (venue_hindex_{mean, min, max})
<i>h</i> -index change (mean, minimum, and maximum) of venues author has published in over last 2 years (venue_hindex_delta_{mean, min, max})
Citations per paper (mean, minimum, and maximum) of venues author has published in (venue_citations_{mean, min, max})
Change in citations per paper (mean, minimum, and maximum) for venues author has published in over last 2 years (venue_citations_delta_{mean, min, max})
Number of papers (mean, minimum, and maximum) of venues author has published in (venue_papers_{mean, min, max})
Change in number of papers (mean, minimum, and maximum) for venues author has published in over last 2 years (venue_papers_delta_{mean, min, max})
Rank of venues (between 0 and 1) in which author has published, rank based on mean number of citations per paper published in the venue (venues_rank_{mean, min, max})
Maximum number of citations received by any paper in venue in which author published (venue_max_single_paper_citations_{mean, min, max})
Total number of venues published in (total_num_venues)

The key citation count is the number of times that papers of the researcher are identified as the central reference in another publication. Key citations are identified via the algorithm proposed by Zhu et al. (2015). The PageRank is a measure of network centrality (Brin and Page 1998). A high PageRank indicates that the researcher has a high (in)degree and is connected to others with a high (in)degree. The weighted co-authorship network is weighted by the frequency of past collaborations

Appendix 2: Descriptive statistics

Table 6 Descriptive statistics by inclusion in analysis sample

Feature	Not included	Included	<i>p</i> value
	<i>N</i> = 724,868	<i>N</i> = 111,156	
	Mean (SE)	Mean (SE)	
total_citations_in_5	29.9 (206)	66.9 (216)	< 0.001
total_citations_in_8	41.9 (275)	97.8 (320)	< 0.001
hindex_in_5	1.28 (2.24)	2.68 (2.85)	< 0.001
hindex_in_8	1.51 (2.60)	3.15 (3.44)	< 0.001
author_hindex	0.70 (1.52)	1.58 (1.65)	< 0.001
author_hindex_delta	0.21 (0.49)	0.43 (0.72)	< 0.001
author_citation_count	11.6 (96.2)	20.6 (67.1)	< 0.001
author_key_citation_count	0.25 (2.35)	0.49 (2.21)	< 0.001
author_citations_delta_1	1.72 (14.3)	4.40 (15.6)	< 0.001
author_citations_delta_0	2.12 (16.6)	5.45 (19.2)	< 0.001
author_key_citations_delta_1	0.04 (0.46)	0.11 (0.62)	< 0.001
author_key_citations_delta_0	0.06 (0.59)	0.15 (0.80)	< 0.001
author_mean_citations_per_paper	1.73 (10.0)	2.76 (6.85)	< 0.001
author_mean_citation_per_paper_delta	0.60 (3.64)	0.86 (2.85)	< 0.001
author_mean_citations_per_year	0.90 (4.94)	2.46 (7.31)	< 0.001
author_papers	3.21 (9.27)	6.33 (7.94)	< 0.001
author_papers_delta	0.92 (2.44)	1.75 (3.45)	< 0.001
author_mean_citation_rank	0.09 (0.18)	0.16 (0.21)	< 0.001
author_unweighted_pagerank	0.26 (0.32)	0.35 (0.37)	< 0.001
author_weighted_pagerank	0.26 (0.32)	0.35 (0.37)	< 0.001
author_age	7.67 (7.24)	8.02 (2.26)	< 0.001
author_recent_num_coauthors	3.92 (11.4)	5.10 (11.1)	< 0.001
author_max_single_paper_citations	4.43 (28.3)	9.19 (26.8)	< 0.001
venue_hindex_max	58.4 (54.3)	65.2 (50.8)	< 0.001
venue_hindex_min	45.4 (53.3)	23.4 (39.1)	< 0.001
venue_hindex_mean	50.2 (51.4)	38.8 (37.8)	< 0.001
venue_hindex_delta_max	12.8 (12.3)	14.2 (11.6)	< 0.001
venue_hindex_delta_min	10.1 (12.1)	5.24 (8.81)	< 0.001
venue_hindex_delta_mean	11.1 (11.7)	8.50 (8.56)	< 0.001
venue_citations_max	3.64 (7.72)	5.55 (6.61)	< 0.001
venue_citations_min	2.08 (4.78)	1.30 (1.72)	< 0.001
venue_citations_mean	2.62 (4.96)	2.88 (2.71)	< 0.001
venue_citations_delta_max	0.72 (1.92)	1.25 (1.82)	< 0.001
venue_citations_delta_min	0.30 (1.81)	0.09 (2.13)	< 0.001
venue_citations_delta_mean	0.48 (1.32)	0.59 (0.77)	< 0.001

Table 6 (continued)

Feature	Not included	Included	<i>p</i> value
	<i>N</i> = 724,868	<i>N</i> = 111,156	
	Mean (SE)	Mean (SE)	
venue_papers_max	41,875 (57,955)	43,702 (57,776)	< 0.001
venue_papers_min	33,571 (54,755)	13,964 (38,508)	< 0.001
venue_papers_mean	35,869 (54,244)	22,478 (39,433)	< 0.001
venue_papers_delta_max	14,094 (19,910)	14,605 (19,923)	< 0.001
venue_papers_delta_min	11,326 (18,763)	4656 (13,176)	< 0.001
venue_papers_delta_mean	12,077 (18,609)	7461 (13,534)	< 0.001
venue_rank_max	0.56 (0.25)	0.69 (0.21)	< 0.001
venue_rank_min	0.45 (0.25)	0.36 (0.23)	< 0.001
venue_rank_mean	0.51 (0.23)	0.53 (0.18)	< 0.001
venue_max_single_paper_citations_max	672 (734)	764 (728)	< 0.001
venue_max_single_paper_citations_min	503 (688)	225 (497)	< 0.001
venue_max_single_paper_citations_mean	559 (669)	406 (496)	< 0.001
totalNumVenues	2.15 (4.05)	3.99 (4.16)	< 0.001

Table 7 Descriptive statistics by gender

Feature	Men	Women	<i>p</i> value
	<i>N</i> = 94,038	<i>N</i> = 17,118	
	Mean (SE)	Mean (SE)	
total_citations_in_5	68.8 (221)	56.8 (181)	< 0.001
total_citations_in_8	101 (329)	82.7 (268)	< 0.001
hindex_in_5	2.71 (2.88)	2.49 (2.61)	< 0.001
hindex_in_8	3.19 (3.49)	2.92 (3.17)	< 0.001
author_hindex	1.59 (1.67)	1.48 (1.51)	< 0.001
author_hindex_delta	0.44 (0.72)	0.41 (0.67)	< 0.001
author_citation_count	21.2 (69.2)	17.3 (53.9)	< 0.001
author_key_citation_count	0.50 (2.30)	0.40 (1.63)	< 0.001
author_citations_delta_1	4.52 (15.9)	3.77 (13.4)	< 0.001
author_citations_delta_0	5.59 (19.6)	4.67 (16.9)	< 0.001
author_key_citations_delta_1	0.12 (0.64)	0.10 (0.50)	< 0.001
author_key_citations_delta_0	0.15 (0.82)	0.12 (0.67)	< 0.001
author_mean_citations_per_paper	2.79 (7.00)	2.62 (5.93)	0.001
author_mean_citation_per_paper_delta	0.86 (2.88)	0.87 (2.64)	0.888
author_mean_citations_per_year	2.52 (7.45)	2.13 (6.42)	< 0.001
author_papers	6.43 (8.09)	5.76 (7.02)	< 0.001
author_papers_delta	1.79 (3.52)	1.54 (3.07)	< 0.001
author_mean_citation_rank	0.16 (0.21)	0.16 (0.21)	0.001
author_unweighted_pagerank	0.36 (0.37)	0.33 (0.36)	< 0.001
author_weighted_pagerank	0.36 (0.37)	0.33 (0.36)	< 0.001
author_age	8.05 (2.27)	7.88 (2.23)	< 0.001
author_recent_num_coauthors	5.19 (11.2)	4.63 (10.0)	< 0.001
author_max_single_paper_citations	9.37 (27.0)	8.22 (25.4)	< 0.001
venue_hindex_max	65.4 (50.7)	64.6 (51.6)	0.064
venue_hindex_min	22.9 (38.4)	26.2 (42.2)	< 0.001
venue_hindex_mean	38.5 (37.3)	40.3 (40.5)	< 0.001
venue_hindex_delta_max	14.2 (11.6)	14.1 (11.8)	0.396
venue_hindex_delta_min	5.11 (8.66)	5.91 (9.56)	< 0.001
venue_hindex_delta_mean	8.42 (8.44)	8.91 (9.19)	< 0.001
venue_citations_max	5.62 (6.59)	5.19 (6.71)	< 0.001
venue_citations_min	1.29 (1.57)	1.36 (2.35)	< 0.001
venue_citations_mean	2.90 (2.61)	2.79 (3.19)	< 0.001
venue_citations_delta_max	1.27 (1.85)	1.14 (1.63)	< 0.001
venue_citations_delta_min	0.09 (2.12)	0.10 (2.14)	0.497
venue_citations_delta_mean	0.60 (0.76)	0.56 (0.81)	< 0.001
venue_papers_max	43,628 (57,706)	44,107 (58,161)	0.321
venue_papers_min	13,439 (37,819)	16,847 (41,975)	< 0.001
venue_papers_mean	22,060 (38,867)	24,770 (42,335)	< 0.001
venue_papers_delta_max	14,577 (19,900)	14,758 (20,048)	0.276
venue_papers_delta_min	4477 (12,940)	5636 (14,365)	< 0.001
venue_papers_delta_mean	7318 (13,340)	8249 (14,529)	< 0.001
venue_rank_max	0.69 (0.21)	0.67 (0.21)	< 0.001

Table 7 (continued)

Feature	Men	Women	<i>p</i> value
	<i>N</i> = 94,038	<i>N</i> = 17,118	
	Mean (SE)	Mean (SE)	
venue_rank_min	0.36 (0.23)	0.37 (0.23)	< 0.001
venue_rank_mean	0.53 (0.18)	0.52 (0.18)	< 0.001
venue_max_single_paper_citations_max	766 (727)	755 (736)	0.072
venue_max_single_paper_citations_min	219 (490)	261 (538)	< 0.001
venue_max_single_paper_citations_mean	402 (491)	426 (528)	< 0.001
totalNumVenues	4.04 (4.20)	3.69 (3.90)	< 0.001

Appendix 3: Robustness checks

See Tables 8, 9, 10, 11 and Fig. 3.

Table 8 Performance and fairness metrics for impact prediction

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio women/men (6) = (5)/(4)
Robustness check 1: <i>h</i> -index in 8 years					
XGB	R^2	0.818	0.818	0.815	0.996
	pa- R^2	0.591	0.589	0.605	1.028
GBM	R^2	0.813	0.814	0.811	0.997
	pa- R^2	0.581	0.579	0.596	1.030
RF	R^2	0.814	0.814	0.809	0.994
	pa- R^2	0.582	0.580	0.594	1.023
LR	R^2	0.790	0.792	0.776	0.981
	pa- R^2	0.529	0.529	0.523	0.988
Robustness check 2: citation count in 5 years					
XGB	R^2	0.901	0.902	0.897	0.994
	pa- R^2	0.813	0.813	0.821	1.010
GBM	R^2	0.905	0.906	0.895	0.988
	pa- R^2	0.820	0.821	0.818	0.997
RF	R^2	0.909	0.910	0.903	0.992
	pa- R^2	0.828	0.828	0.831	1.004
LR	R^2	0.908	0.910	0.884	0.972
	pa- R^2	0.825	0.828	0.799	0.965
Robustness check 3: gender as predictor					
XGB	R^2	0.876	0.876	0.868	0.991
	pa- R^2	0.617	0.616	0.621	1.009
GBM	R^2	0.873	0.873	0.867	0.992
	pa- R^2	0.608	0.606	0.617	1.018
RF	R^2	0.871	0.872	0.864	0.991
	pa- R^2	0.602	0.601	0.609	1.013
LR	R^2	0.850	0.852	0.837	0.982
	pa- R^2	0.539	0.541	0.530	0.981
Robustness check 4: include established researchers (career > 12 years)					
XGB	R^2	0.933	0.933	0.932	0.999
	pa- R^2	0.670	0.669	0.678	1.013
GBM	R^2	0.931	0.931	0.929	0.998
	pa- R^2	0.660	0.659	0.665	1.009
RF	R^2	0.930	0.930	0.930	1.000
	pa- R^2	0.654	0.652	0.666	1.021
LR	R^2	0.916	0.917	0.910	0.992
	pa- R^2	0.587	0.589	0.571	0.969

Table 8 (continued)

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio women/men (6) = (5)/(4)
Robustness check 5: focus on mid-career researchers (career 4 to 6 years)					
XGB	R^2	0.764	0.765	0.751	0.981
	pa- R^2	0.496	0.500	0.465	0.930
GBM	R^2	0.763	0.765	0.749	0.980
	pa- R^2	0.494	0.498	0.462	0.927
RF	R^2	0.765	0.767	0.749	0.976
	pa- R^2	0.499	0.504	0.460	0.913
LR	R^2	0.739	0.741	0.726	0.980
	pa- R^2	0.443	0.447	0.411	0.920

XGB Extreme Gradient Boost, *GBM* Gradient Boosting Machine, *RF* random forest, *LR* linear regression

Table 9 Performance and fairness metrics for high-performer prediction

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio Women/Men (6) = (5)/(4)
Robustness check 1: <i>h</i> -index in 8 years					
XGB	ACC	0.895	0.894	0.900	1.006
	TPR	0.861	0.861	0.860	0.999
	TNR	0.904	0.903	0.909	1.006
	PPV	0.698	0.701	0.680	0.971
	NPV	0.962	0.961	0.966	1.006
	$P(\hat{Y} = 1)$	0.253	0.257	0.233	0.909
GBM	ACC	0.885	0.884	0.888	1.004
	TPR	0.877	0.879	0.867	0.986
	TNR	0.887	0.886	0.892	1.008
	PPV	0.666	0.669	0.645	0.964
	NPV	0.965	0.965	0.967	1.002
	$P(\hat{Y} = 1)$	0.270	0.274	0.248	0.904
RF	ACC	0.892	0.891	0.897	1.006
	TPR	0.857	0.858	0.852	0.993
	TNR	0.901	0.900	0.907	1.007
	PPV	0.691	0.693	0.673	0.971
	NPV	0.961	0.960	0.964	1.005
	$P(\hat{Y} = 1)$	0.254	0.258	0.233	0.903
LOG	ACC	0.890	0.889	0.894	1.006
	TPR	0.852	0.852	0.851	0.999
	TNR	0.900	0.899	0.904	1.006
	PPV	0.686	0.690	0.667	0.967
	NPV	0.959	0.958	0.964	1.006
	$P(\hat{Y} = 1)$	0.254	0.258	0.235	0.912
Robustness check 2: citation count in 5 years					
XGB	ACC	0.919	0.919	0.923	1.005
	TPR	0.902	0.902	0.899	0.996
	TNR	0.926	0.925	0.931	1.007
	PPV	0.810	0.811	0.805	0.993
	NPV	0.964	0.964	0.967	1.003
	$P(\hat{Y} = 1)$	0.290	0.293	0.269	0.918
GBM	ACC	0.922	0.922	0.923	1.001
	TPR	0.894	0.895	0.889	0.994
	TNR	0.932	0.932	0.934	1.002
	PPV	0.822	0.824	0.809	0.982
	NPV	0.962	0.961	0.964	1.003
	$P(\hat{Y} = 1)$	0.283	0.286	0.265	0.925
RF	ACC	0.924	0.924	0.925	1.002
	TPR	0.889	0.890	0.882	0.991
	TNR	0.936	0.936	0.939	1.004
	PPV	0.830	0.832	0.821	0.987
	NPV	0.960	0.960	0.962	1.002
	$P(\hat{Y} = 1)$	0.278	0.282	0.259	0.918

Table 9 (continued)

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio Women/Men (6) = (5)/(4)
LOG	ACC	0.922	0.922	0.923	1.001
	TPR	0.885	0.886	0.879	0.992
	TNR	0.935	0.935	0.937	1.002
	PPV	0.828	0.830	0.815	0.982
	NPV	0.959	0.958	0.961	1.003
	$P(\hat{Y} = 1)$	0.278	0.281	0.260	0.925

XGB Extreme Gradient Boost, *GBM* Gradient Boosting Machine, *RF* random forest, and *LOG* logistic regression

Table 10 Performance and fairness metrics for high-performer prediction

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio Women/Men (6) = (5)/(4)
Robustness check 3: gender as predictor					
XGB	ACC	0.908	0.909	0.908	0.999
	TPR	0.886	0.886	0.880	0.993
	TNR	0.915	0.915	0.915	1.000
	PPV	0.760	0.764	0.732	0.957
	NPV	0.964	0.963	0.967	1.004
	$P(\hat{Y} = 1)$	0.270	0.274	0.251	0.915
GBM	ACC	0.901	0.901	0.900	1.000
	TPR	0.900	0.901	0.896	0.995
	TNR	0.901	0.901	0.901	1.001
	PPV	0.733	0.737	0.705	0.957
	NPV	0.968	0.967	0.971	1.004
	$P(\hat{Y} = 1)$	0.285	0.289	0.265	0.917
RF	ACC	0.915	0.915	0.912	0.997
	TPR	0.866	0.867	0.856	0.987
	TNR	0.929	0.930	0.927	0.997
	PPV	0.787	0.793	0.755	0.952
	NPV	0.958	0.958	0.961	1.003
	$P(\hat{Y} = 1)$	0.255	0.258	0.236	0.914
LOG	ACC	0.901	0.901	0.903	1.003
	TPR	0.880	0.882	0.869	0.986
	TNR	0.907	0.906	0.912	1.006
	PPV	0.741	0.744	0.723	0.971
	NPV	0.962	0.961	0.964	1.003
	$P(\hat{Y} = 1)$	0.275	0.280	0.251	0.896
Robustness check 4: include established researchers (career > 12 years)					
XGB	ACC	0.932	0.932	0.933	1.001
	TPR	0.929	0.930	0.920	0.989
	TNR	0.933	0.933	0.936	1.003
	PPV	0.769	0.773	0.742	0.960
	NPV	0.982	0.982	0.983	1.001
	$P(\hat{Y} = 1)$	0.234	0.238	0.208	0.874
GBM	ACC	0.929	0.929	0.928	0.999
	TPR	0.932	0.934	0.918	0.982
	TNR	0.928	0.927	0.930	1.003
	PPV	0.756	0.760	0.726	0.955
	NPV	0.983	0.983	0.983	1.000
	$P(\hat{Y} = 1)$	0.238	0.243	0.212	0.872

Table 10 (continued)

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio Women/Men (6) = (5)/(4)
RF	ACC	0.926	0.925	0.929	1.004
	TPR	0.929	0.930	0.917	0.986
	TNR	0.925	0.924	0.932	1.009
	PPV	0.748	0.750	0.730	0.973
	NPV	0.982	0.982	0.982	1.001
	$P(\hat{Y} = 1)$	0.240	0.245	0.210	0.860
LOG	ACC	0.924	0.923	0.926	1.003
	TPR	0.929	0.930	0.924	0.994
	TNR	0.923	0.922	0.927	1.005
	PPV	0.742	0.745	0.717	0.962
	NPV	0.982	0.982	0.984	1.002
	$P(\hat{Y} = 1)$	0.242	0.246	0.216	0.877

XGB Extreme Gradient Boost, *GBM* Gradient Boosting Machine, *RF* random forest, and *LOG* logistic regression

Table 11 Performance and fairness metrics for high-performer prediction

Model (1)	Metric (2)	Total (3)	Male (4)	Female (5)	Ratio Women/Men (6) = (5)/(4)
Robustness check 5: focus on mid-career researchers (career 4 to 6 years)					
XGB	ACC	0.867	0.866	0.868	1.002
	TPR	0.871	0.877	0.836	0.953
	TNR	0.865	0.863	0.875	1.013
	PPV	0.618	0.625	0.575	0.919
	NPV	0.964	0.964	0.963	0.999
	$P(\hat{Y} = 1)$	0.282	0.289	0.245	0.847
GBM	ACC	0.874	0.873	0.877	1.004
	TPR	0.857	0.862	0.827	0.959
	TNR	0.878	0.876	0.887	1.012
	PPV	0.637	0.644	0.597	0.928
	NPV	0.961	0.961	0.962	1.001
	$P(\hat{Y} = 1)$	0.269	0.276	0.233	0.845
RF	ACC	0.880	0.879	0.883	1.004
	TPR	0.841	0.845	0.817	0.967
	TNR	0.889	0.888	0.896	1.009
	PPV	0.655	0.662	0.614	0.928
	NPV	0.957	0.957	0.960	1.004
	$P(\hat{Y} = 1)$	0.257	0.263	0.224	0.851
LOG	ACC	0.864	0.863	0.870	1.008
	TPR	0.862	0.868	0.825	0.950
	TNR	0.865	0.862	0.879	1.020
	PPV	0.614	0.620	0.580	0.935
	NPV	0.962	0.962	0.961	0.999
	$P(\hat{Y} = 1)$	0.280	0.289	0.240	0.830

XGB Extreme Gradient Boost, *GBM* Gradient Boosting Machine, *RF* random forest, and *LOG* logistic regression



Fig. 3 Feature importance

Appendix 4: Tuning grid

Table 12 Tuning grid for 5-fold cross-validation

Model	Hyper-parameter	Values
XGB (round 1)	num.trees	100
	learning.rate	0.05
	max.depth	1, 3, 5, 7, 9, 11
	min.child.weight	0, 1, 2, 3, 4, 5
	bag.fraction	0.6
	gamma	Regression: 0, 5, 10, 15 Classification: 0, .5, 1
XGB (round 2)	num.trees	250, 500, 750
	learning.rate	.01, .025
	max.depth	best(round 1)
	min.child.weight	best(round 1), best(round 1) \pm 1
	bag.fraction	0.6
	gamma	Regression: best(round 1), best(round 1) \pm 5 Classification: best(round 1), best(round 1) \pm 0.5
GBM	num.trees	200, 300, 400, 500, 600, 700
	learning.rate	.01, .025, .05
	max.depth	3, 5, 7
	bag.fraction	.60, .80
RF	num.trees	500
	m.try	3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Default values of the respective R-packages (ranger, gbm, xgboost) are used for all other parameters. When the best-performing model used the highest or lowest of the tested hyper-parameter values, more extreme values were tested in the next round of cross-validation

Acknowledgements This work benefited heavily from the Impact Prediction data set compiled and made publicly available by Weihs and Etzioni (2017), Allen Institute for Artificial Intelligence. This work was supported by the University of Mannheim’s Graduate School of Economic and Social Sciences. I want to personally thank my PhD supervisor Christoph Kern for his support.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abramo, G., Cicero, T., & D'Angelo, C. A. (2015). Should the research performance of scientists be distinguished by gender? *Journal of Informetrics*, *9*(1), 25–38.
- Abramo, G., D'Angelo, C. A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Informetrics*, *7*(4), 811–822.
- Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Predicting scientific success. *Nature*, *489*(7415), 201–202.
- Aigner, D. J., & Cain, G. G. (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review*, *30*(2), 175.
- Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289.
- AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications*, *9*(1), 5163.
- Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton University Press.
- Ayaz, S., Masood, N., & Islam, M. A. (2018). Predicting scientific impact based on h-index. *Scientometrics*, *114*(3), 993–1010.
- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.
- Beaudry, C., & Larivière, V. (2016). Which gender gap? Factors affecting researchers' scientific impact in science and medicine. *Research Policy*, *45*(9), 1790–1817.
- Bendels, M. H. K., Müller, R., Brueggmann, D., & Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PLOS ONE*, *13*(1), e0189136.
- Bertsimas, D., Brynjolfsson, E., Reichman, S., & Silberholz, J. (2015). OR forum-tenure analytics: Models for predicting research impact. *Operations Research*, *63*(6), 1246–1261.
- Blau, F. D., Currie, J. M., Croson, R. T. A., & Ginther, D. K. (2010). Can mentoring help female assistant professors? Interim results from a randomized trial. *American Economic Review*, *100*(2), 348–352.
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.
- Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.
- Carli, L. L., Alawa, L., Lee, Y., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women scientists. *Psychology of Women Quarterly*, *40*(2), 244–260.
- Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. [arXiv: 2010.04053](https://arxiv.org/abs/2010.04053).
- Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest*, *15*(3), 75–141.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM.
- Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. [arXiv:1610.07524](https://arxiv.org/abs/1610.07524) [cs, stat].
- Daud, A., Aljohani, N. R., Abbasi, R. A., Rafique, Z., Amjad, T., Dawood, H., & Alyoubi, K. H. (2017). Finding rising stars in co-author networks via weighted mutual influence. In *Proceedings of the 26th international conference on World Wide Web Companion—WWW '17 Companion* (pp. 33–41). ACM.
- Daud, A., Song, M., Hayat, M. K., Amjad, T., Abbasi, R. A., Dawood, H., & Ghani, A. (2020). Finding rising stars in bibliometric networks. *Scientometrics*, *124*(1), 633–661.
- Demetrescu, C., Finocchi, I., Ribichini, A., & Schaerf, M. (2020). On bibliometrics in academic promotions: A case study in computer science and engineering in Italy. *Scientometrics*, *124*(3), 2207–2228.
- Dong, Y., Johnson, R. A., & Chawla, N. V. (2016). Can scientific impact be predicted? *IEEE Transactions on Big Data*, *2*(1), 18–30.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference—ITCS '12* (pp. 214–226). ACM.
- Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, *82*(3–4), 127–141.

- European Commission. (2019). *She figures 2018*. Publications Office.
- Flanagin, A. (1998). Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. *JAMA*, *280*(3), 222.
- Friedman, J. H. (2001). Greedy function approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.
- Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113–135.
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.
- Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19193–19198.
- Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity-innovation paradox in science. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9284–9291.
- Holman, L., Stuart-Fox, D., & Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented? *PLOS Biology*, *16*(4), e2004956.
- Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 4609–4616.
- Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018). Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems*, *21*(03n04), 1750011.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: With applications in R*. Springer texts in statistics, vol. 103. Springer.
- Kessels, U., Rau, M., & Hannover, B. (2006). What goes well with physics? Measuring and altering the image of science. *British Journal of Educational Psychology*, *76*(4), 761–780.
- Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: An experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*(5), 603–625.
- Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128–161.
- Lane, K. A., Goh, J. X., & Driver-Linn, E. (2012). Implicit science stereotypes mediate the relationship between gender and academic participation. *Sex Roles*, *66*(3–4), 220–234.
- Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213.
- Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating gender bias in machine learning data sets. In L. Boratto, S. Faralli, M. Marras, & G. Stilo (Eds.), *Bias and social aspects in search and recommendation* (Vol. 1245, pp. 12–26). Communications in Computer and Information Science: Springer.
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*(4), 611–627.
- Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.
- Li, X.-L., Foo, C. S., Tew, K. L., & Ng, S.-K. (2009). Searching for rising stars in bibliography networks. In X. Zhou, H. Yokota, K. Deng, & Q. Liu (Eds.), *Database systems for advanced applications* (Vol. 5463, pp. 288–292). Lecture notes in computer science. Springer.
- Long, J. S. (1992). Measures of sex differences in scientific productivity. *Social Forces*, *71*(1), 159.
- Mazloumian, A. (2012). Predicting scholars' scientific impact. *PLoS ONE*, *7*(11), e49246.
- Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.
- Miller, D. I., & Wai, J. (2015). The bachelor's to Ph.D. STEM pipeline no longer leaks more women than men: A 30-year analysis. *Frontiers in Psychology*, *6*, 37.
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163.

- Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), 16474–16479.
- National Science Board. (2018). Science and Engineering Indicators 2018. Technical Report NSB-2018-1. National Science Foundation.
- NCSES. (2019). Survey of Doctorate Recipients 2019. Technical Report NSF 21-320. National Center for Science and Engineering Statistics. National Science Foundation.
- NCSES. (2021). Women, minorities, and persons with disabilities in science and engineering: 2021. Technical Report Special Report NSF 21-321. National Center for Science and Engineering Statistics. National Science Foundation.
- Nie, Y., Zhu, Y., Lin, Q., Zhang, S., Shi, P., & Niu, Z. (2019). Academic rising star prediction via scholar's evaluation model and machine learning techniques. *Scientometrics*, *120*(2), 461–476.
- Nielsen, M. W., Alegria, S., Börjeson, L., Etkowitz, H., Falk-Krzesinski, H. J., Joshi, A., Leahey, E., Smith-Doerr, L., Woolley, A. W., & Schiebinger, L. (2017). Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(8), 1740–1742.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math me. *Journal of Personality and Social Psychology*, *83*(1), 44–59.
- Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics*, *11*(1), 198–222.
- Penner, O., Petersen, A. M., Pan, R. K., & Fortunato, S. (2013). Commentary: The case for caution in predicting scientists' future impact. *Physics Today*, *66*(4), 8–9.
- Reskin, B. F. (2000). The proximate causes of employment discrimination. *Contemporary Sociology*, *29*(2), 319.
- Reskin, B. F., & McBrier, D. B. (2000). Why not ascription? Organizations' employment of male and female managers. *American Sociological Review*, *65*(2), 210.
- Reymert, I. (2021). Bibliometrics in academic recruitment: A screening tool rather than a game changer. *Minerva*, *59*(1), 53–78.
- Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156.
- Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, *47*(2), 227–254.
- Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review*, *107*(5), 141–145.
- Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, *48*(2), 478–491.
- Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10107–10112.
- Symonds, M. R., Gemmill, N. J., Braisher, T. L., Gorringer, K. L., & Elgar, M. A. (2006). Gender differences in publication output: Towards an unbiased metric of research performance. *PLoS ONE*, *1*(1), e127.
- Tekles, A., & Bornmann, L. (2019). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. [arXiv:1904.12746](https://arxiv.org/abs/1904.12746) [cs].
- van Anders, S. M. (2004). Why the academic pipeline leaks: Fewer men than women perceive barriers to becoming professors. *Sex Roles*, *51*(9–10), 511–521.
- van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2012). Gender differences in scientific productivity: A persisting phenomenon? *Scientometrics*, *93*(3), 857–868.
- van der Lee, R., & Ellemers, N. (2015). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(40), 12349–12353.
- Wang, D., & Barabási, A.-L. (2021). *The science of science* (1st ed.). Cambridge University Press.
- Weiss, L., & Etzioni, O. (2017). Learning to predict citation-based impact measures. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 1–10). IEEE.
- Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature*, *387*(6631), 341–343.
- West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS ONE*, *8*(7), e66212.
- Wilhite, A. W., & Fong, E. A. (2012). Coercive citation in academic publishing. *Science*, *335*(6068), 542–543.

- Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, 393(10171), 531–540.
- Zhang, C., Liu, C., Yu, L., Zhang, Z.-K., & Zhou, T. (2016a). Identifying the academic rising stars. [arXiv:1606.05752](https://arxiv.org/abs/1606.05752).
- Zhang, J., Ning, Z., Bai, X., Wang, W., Yu, S., & Xia, F. (2016b). Who are the Rising Stars in Academia? In *Proceedings of the 16th ACM/IEEE-CS on joint conference on digital libraries* (pp. 211–212). ACM.
- Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2), 408–427.
- Zuo, Z., & Zhao, K. (2021). Understanding and predicting future research impact at different career stages—A social network perspective. *Journal of the Association for Information Science and Technology*, 72(4), 454–472.