Keep It Simple: Picking the Right Data Science Method to Improve Workforce Training Programs

Authors: Camille Preel-Dumas, Richard Hendra, Dakota Denison
MDRC Center for Data Insights

OPRE Report 2023-058 February 2023

In 2019, the Office of Planning, Research, and Evaluation (OPRE) in the Administration for Children and Families (ACF), U.S. Department of Health and Human Services (HHS) awarded **Career Pathways Secondary Data Analysis Grants** to support secondary analysis of data collected to rigorously evaluate a collection of career pathways programs.

Our Study

This brief explores data science methods that workforce programs can use to predict participant success. With access to vast amounts of data on their programs, workforce training providers can leverage their management information systems (MIS) to understand and improve their programs' outcomes. By predicting which participants are at greater risk of dropping out of their program and why, providers can segment their caseloads so that participants receive services better tailored to their needs. Within the data science field, machine learning (ML) has gained popularity for its ability to extract hidden patterns without being explicitly guided by a data analyst. **While these data science methods hold promise, are the added costs and complexity worth it?** We explore the tradeoffs by answering the following questions:

Machine Learning (ML)

The use of computer algorithms and statistical models to find patterns and make predictions from data.

1 What factors are important in predicting a participant's outcome in a program?

2 Are participant outcomes predictable using simple methods, like creating basic risk indicators in Management Information Systems (MIS)? For example, how well does an indicator for prior education predict participant outcomes?

3 What is the added value and cost of incorporating regression and more complex machine learning methods?

Our Data and Sample

Our analysis focuses on participant intake data from an evaluation of the Health Profession Opportunity Grants (HPOG) Program, which provided education and training in occupations in the health care field that pay well and are expected to either experience labor shortages or be in high demand to Temporary Assistance for Needy Families (TANF) participants and other individuals with low incomes.[1] Our sample is restricted to treatment group members who participated in the HPOG 1.0 Impact Study of 36 HPOG programs (N= 5,566).[2]

Our Methods

We define the participant "success" outcome to predict as *a participant completing training, in ongoing training, or currently employed in a healthcare position 15 months after enrollment*.

We then test the added value of using incrementally more complex prediction models:

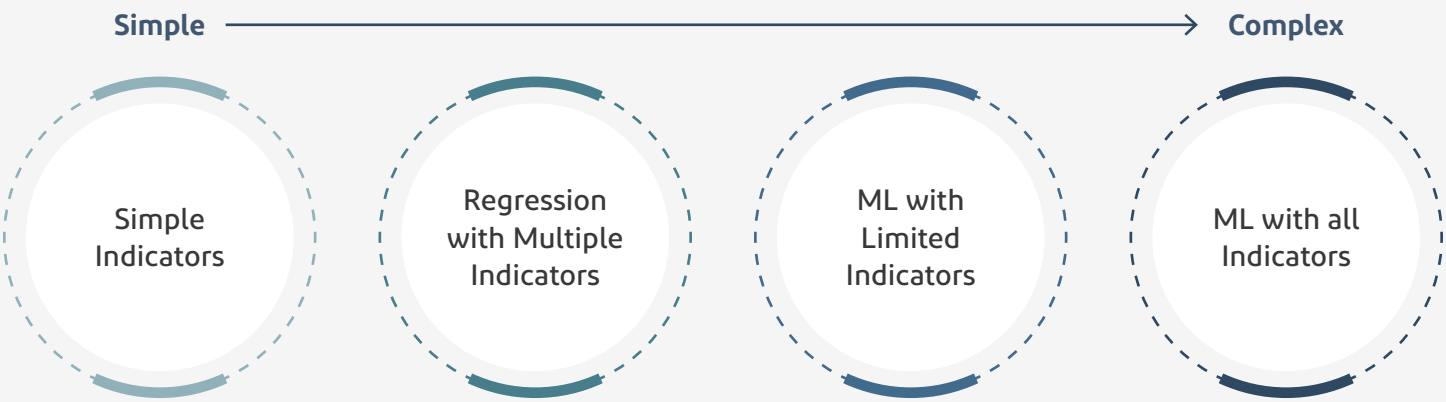
Comparing the success rate for participants with and without a specific indicator (for example, for those experiencing barriers vs. those *not* experiencing barriers)

Conducting regression analysis estimating the relationship between specific indicators and the desired outcome

Running machine learning algorithms

We test the "value" of each method through the F0.5 score, a measure used in machine learning to define how accurately a model predicts a desired outcome. The F0.5 score ranges from 0 to 1, with a value of 1 indicating perfect prediction and a value of 0 indicating completely inaccurate prediction.

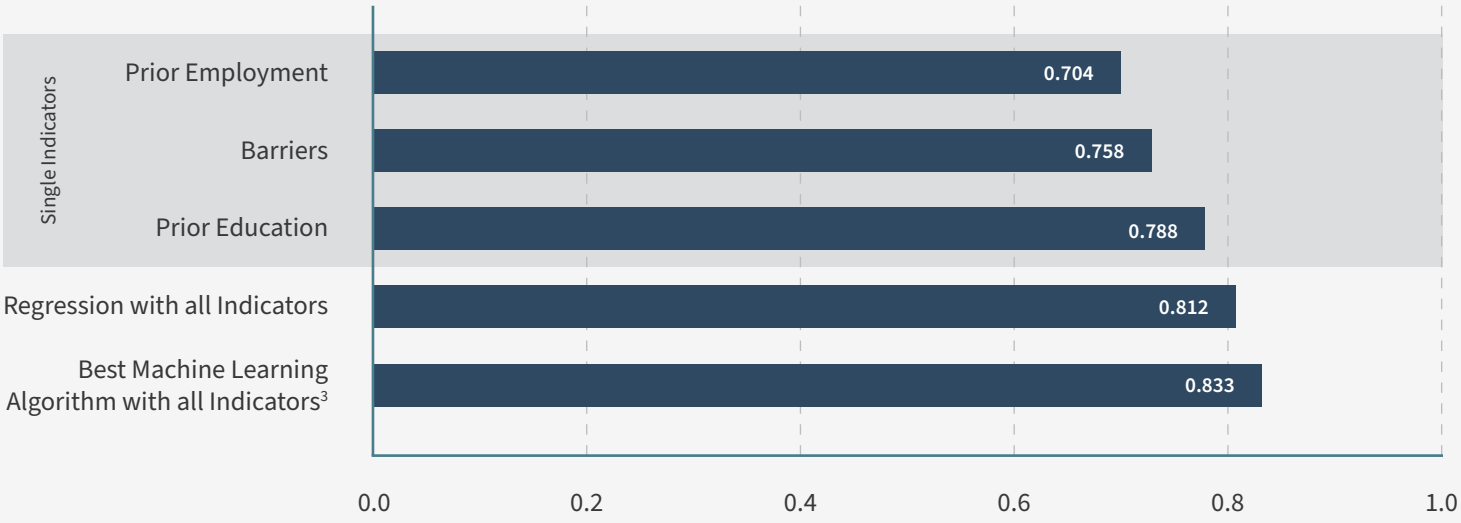
Incremental Model Complexity



Our Findings

The most important factors in predicting participants' success are prior employment levels, prior education levels, and the presence of barriers. Among these, prior education is the most important factor; it is a common pattern for prior education levels to define a participant's journey through workforce training. Prediction can serve as an early warning system to help programs identify participants who need more support, such as childcare and transportation services.

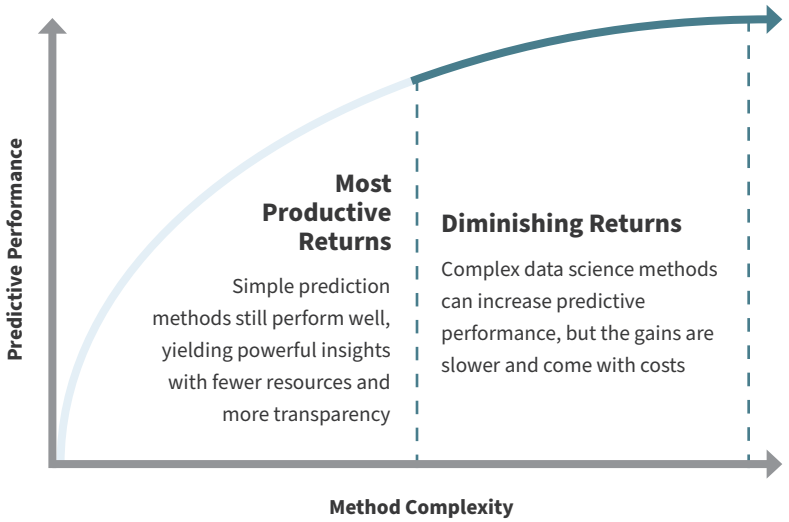
Incremental Improvement in F0.5 Score for each Model



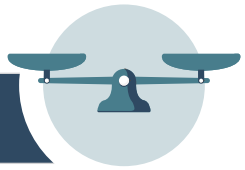
We find that outcomes are predictable, even when simple, cost-effective methods are used. For example, one indicator can be used to predict outcomes. The figure shows that using the prior education indicator to predict outcomes has an F0.5 score of 0.788. This is only marginally lower than the highest performing, most complex machine learning model with an F0.5 score of 0.833. **However, a word of caution: when conducting a simple indicator analysis, one needs to choose carefully which indicator to use**. Among our single indicators, prior education is the most powerful factor for predicting success, but this might not be generalizable to other programs, populations, and data. Larger and more complex datasets might also respond to these methods differently and experience more value from adding regression analysis and machine learning.

Implications for Practitioners

More complex machine learning methods provide small gains in predictive performance, but these gains need to be weighed against several costs, including staff resources, decreased transparency in the models, and bias in the algorithm that can reinforce discrimination and inequity. Machine learning produces estimates that can be hard to interpret, even for those with advanced knowledge. Because these models are made by humans, they will reflect the perspectives and knowledge of those who develop them.[4] If these perspectives are not representative of the target population, or if they perpetuate already existing biases, the models will



produce inaccurate or discriminatory predictions. In addition, the underlying data can also reflect and perpetuate structures of discrimination. For example, wage studies that rely on employment data from state Unemployment Insurance (UI) systems must examine the data's coverage. UI wage data does not capture informal employment and may also miss relatively more employment for populations receiving lower incomes than for populations receiving higher incomes.[5] This means that the employment experiences of communities disproportionately living in poverty—particularly Black and Brown communities—may not be accurately captured in UI data. Machine learning models using such data will produce biased predictions that can perpetuate or exacerbate structures of inequity for these communities.



To weigh the costs and benefits of using ML, workforce providers should examine:

1. **The need for improvements in predictive performance**. How will the results be used? How crucial is it to see improvements in predictive performance, even if marginal? (Ex: in a clinical study of a new medicine's efficacy and potential side effects, any improvement in predictive performance is important)
2. **The size of the data set**. Are simple methods able to capture patterns in the data, or does the size of the data suggest that more complex algorithms might be worth considering?
3. **The budget and timeline for the study**. How many resources are available to test and learn new methods? Will data scientists have the capacity and ability to work with participants and frontline staff to develop a reliable model?
4. **The potential sources of bias in the study**. How might bias be inherent in the dataset? How might it be introduced through the predictive model or implementation of its findings? How can this bias be mitigated to produce a transparent and equitable study?

Whichever method is used, prediction yields powerful insights that help providers tailor their services and ensure their participants receive the supports they need. As a bonus, it doesn't need to be complicated.

[1] HPOG was authorized by the Affordable Care Act (ACA), Public Law 111-148, 124 Stat. 119, March 23, 2010, sect. 5507(a), "Demonstration Projects to Provide Low-Income Individuals with Opportunities for Education, Training, and Career Advancement to Address Health Professions Workforce Needs," adding sect. 2008(a) to the Social Security Act, 42 U.S.C. 1397g(a). There have been two rounds of 32 HPOG grantees: HPOG 1.0, awarded in 2010, and HPOG 2.0, awarded in 2015. More information about the HPOG 1.0 Impact Study is available at **https://www.acf.hhs.gov/opre/project/health-profession-opportunity-grants-hpog-impact-study-2011-2018**.

[2] Abt Associates, and Peck, Laura. Health Profession Opportunity Grants Evaluation, United States, 2010-2018. Inter-university Consortium for Political and Social Research [distributor], 2021-11-29. https://doi.org/10.3886/ICPSR37290.v5

[3] A neural network algorithm performed the best among the different machine learning models we tested.

[4] There are many resources on bias in machine learning, such as **Mitigating Bias in Artificial Intelligence** and **Bias in Pretrial Risk Assessment**

[5] Abraham, Katherine, John C. Haltiwanger, Kristin Sandusky, and James Spletzer. 2013. "Exploring Differences in Employment Between Household and Establishment Data." Chicago, IL: Journal of Labor Economics Vol 31, Number S1 Part 2.

This publication was made possible by Contract Number GS00F010CA/140D0421F0706 and Grant Number 90PE0038 from the Office of Planning, Research, and Evaluation (OPRE), Administration for Children and Families (ACF), U.S. Department of Health and Human Services (HHS). The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services. It was developed in partnership with ICF, the contractor for OPRE's Career Pathways Secondary Data Analysis project. ICF delivers evidence-based solutions, training, technical assistance, and tools for developing and implementing programs that strengthen families and communities.

ICF