Addressing Questions About Bias in Predictive Modeling

By Civitas Learning
Fall 2020

With the increasing emphasis that institutions of higher education are placing on issues of bias and equity in decision making, it's crucial to understand the basis on which those decisions are made. Given that Civitas Learning® uses predictive modeling to help institutions create and execute student success initiatives, and given the potentially significant impact of decisions based on those models, it's important to understand predictive modeling in general and the factors that go into the models that Civitas Learning creates.

What is Predictive Modeling?

First, some terminology. An ***algorithm*** is a set of steps designed to help an entity accomplish a task. Computer systems follow algorithms to perform all of their functions; when computers use algorithms that can improve their own performance without explicit intervention by human programmers, they are said to be engaging in ***machine learning*** (ML).

ML algorithms use statistics to search for patterns in large sets of data. Those searches can be ***supervised*** or ***unsupervised***; that is, programmers can label the data to tell the algorithms what kinds of patterns to look for, or they can leave the algorithms alone to find whatever patterns emerge. Once the algorithms have found patterns in existing data sets, they can use those patterns to predict, or model, future behavior.

Predictive modeling is, in essence, supervised machine learning, in which the algorithms learn relationships between ***training data*** (specifically curated and provided sets of variables or features) and specified outcomes. The trained models can then be applied to real-world operational data to predict future outcomes. Given the expected changes in data characteristics over time and the resulting degradation in model performance, predictive models are retrained as needed.

Is Predictive Modeling Inherently Biased?

The short answer is no: predictive modeling is not inherently biased—the task of its algorithms is to explain data by finding the most significant relationships between inputs and outputs. However, predictive models can fail in three major ways: **a mismatch between training data and real-world operational data, laziness in training the models, and maliciousness in intent.**

Mismatched Data

When training data doesn't accurately represent its corresponding real-world operational data, this mismatch is called ***data nonstationarity***. The most frequent reason for ML algorithm mistakes is historically restrictive or biased data. For example, the mortgage-backed security pricing algorithms that blew up in 2008 failed mainly because they were trained with the most recent three years of data, during which home prices had risen abnormally. Clinical trials for pharmaceuticals have come under fire for a lack of gender, racial, and ethnic diversity, prompting the Food and Drug Administration to institute pharmacovigilance and real-world evidence programs through the use of retrospective ML-based causal inference to augment potentially biased clinical efficacy knowledge from randomized controlled trials (RCTs).

Poor Models

In supervised ML, where programmers train algorithms to seek specific patterns, it matters what patterns the programmers choose. Say, for example, students with a specific demographic characteristic are found to have a 10% lower success rate than the average student. A poor model will—instead of probing for underlying behaviors—simply highlight this group indicator variable and subsequently predict a 10% lower success rate for *all* students with this demographic characteristic. If the predicted success rate is used to make high-stakes decisions such as awarding scholarships, then all students in the group will be penalized. Finding the underlying behavioral factors that go beyond phenotypes requires due diligence, exploratory analysis, and intentionality in terms of how student variables are created and used for helping students in an ethical manner.

Malicious Intent

As with any tool, ML algorithms can be deployed with malicious intent. As with statistics, bad actors can withhold certain pieces of evidence when building models to advance their biased point of view under the name of ML, but most real-world malicious use cases are restricted to an AI-based attack infrastructure. Civitas Learning's models, being social mission–driven, are focused on student success in the context of *reducing* inequity in higher education by relying on influenceable factors that put a student at a major disadvantage and can be mitigated through intervention.

How Does Civitas Learning Avoid Bias in Its Predictive Models?

Civitas Learning uses predictive modeling to help institutions take action to prevent adverse outcomes and help students thrive. Because of known cross-industry experiences in, and understanding of, potential biases in and misuses of ML, Civitas Learning has taken numerous steps from the very beginning to ensure the best possible models and causal inference algorithms for our customers.

Use of Influenceable Derived Variables

Civitas uses Learning Management System (LMS) and Student Information System (SIS) time-series data to compute robust measures of student engagement, various types of LMS activities, short- vs. long-term activity changes, enrollment behavior, and program alignment fit. These variables, which are ***derived*** from actual data (rather than externally imposed) and ***influenceable*** through intervention, power both predictive models and ***impact analysis***, which is the process of quantifying the causal impact of an intervention on student success. Our experience shows that LMS-derived variables are relatively invariant to varying LMS use rates across sections. Further, these derived variables explain far better than group membership why it is that some students are doing worse: it's because, for example, they are less engaged, have less important LMS activities, or experience a sudden drop in engagement factors, not because they belong to the specific group.

Using derived variables rather than imposed variables provides many benefits: a much deeper understanding of prediction scores, higher model accuracy, and the ability to design intentional interventions that are proven to result in student success. The greater the number of a model's influenceable variables, the more meaningful its prediction scores, and the higher its accuracies, the more useful it is in presenting actionable intervention opportunities to help reduce equity or achievement gaps.

A Focus on Malleability and Intervention Impact

Civitas Learning has identified five main categories of influenceable derived student variables—engagement, enrollment behavior, academic performance, pathway progress, and financial aid—which we call ***impact levers with elasticity*** because they are malleable and can be influenced to a greater or lesser degree by means of institutional interventions and policies. Civitas Learning provides institutions with standard and custom impact analysis tools to create evidence-based student success knowledge bases that they can use to personalize interventions to each student based on which impact levers are most elastic in the given case.

Diligent Models

The fewer input variables a predictive model uses, the weaker it is, and the more susceptible it is to inadvertent bias. Civitas uses ***time-series feature engineering*** (which links time-ordered event data to detect and learn more meaningful dynamic patterns associated with student success) to build models on hundreds of input variables, all of which are known to have an effect on student success (typically defined in terms of flexible persistence, successful course completion, term GPA, completion, and job success).

Transparency with Variable Ranking

Civitas uses a number of complementary variable-ranking algorithms to quantify the value of specific types of information. Conceptually, we consider both the marginal predictive power of each input variable and the amount of orthogonal or "new" information it provides relative to the rest of the variables in the predictive model, i.e., its incremental value. For example, actionable and inferred behavioral variables derived from LMS, time-series records in SIS, card swipe, and multiple survey data are ranked much higher than ***point*** or ***raw*** variables, which are directly extracted from raw database tables without further processing. The behavioral variables are far more important in designing intentional interventions that leverage our evidence-based impact knowledge base. This is why Civitas Learning's Illume® application always shows variable ranking for each filtered group.

Social Mission Use Case

Among the hundreds of input variables in our predictive models are demographic variables that categorize students by age, gender, race, ethnicity, socioeconomic status (inferred from Pell and census data), status as first-generation college student, and high school characteristics (free lunch fraction, for instance)—all of which have traditionally been associated with factors that contribute to equity gaps. These variables can't be influenced through institutional interventions or policies and thus are ranked lower in our predictive models: after all, what good will it do to use variables to make predictions if the future can't be improved by acting on those predictions? Demographic variables *can* have specific uses, however, in the goal of improving student success, as institutions can select any group, identify the top influenceable factors within the group, run intentional interventions, and then measure impact for continuous process improvement.

In summary, all of these features help the predictive models that Civitas Learning creates avoid bias and help institutions reduce equity gaps by employing influenceable root-cause variables with high elasticity in interventions.

How Can Institutions Protect Their Predictive Models from Biased Misuse?

Appropriate Use Cases

Because Civitas Learning impact analyses are designed to help institutions build and execute the most effective intervention programs to improve student success, institutions can take care to rely on prediction scores that are based on highly ranked derived variables.

Controlled Access

Institutions that use Civitas Learning applications can limit general access to specific types of data or disable access entirely. Because staff in adviser roles, for example, rarely use non-malleable demographic data in their work, institutions can proactively limit their access to variables that directly influence student performance.

Professional Development

Institutions that plan to use student demographic information for outreach or other targeted programs can take steps to train staff to consider carefully the potential uses and misuses of such data when abstracted from the context of other predictors.

Specific Use Cases

What if data from demographic variables are important for implementing a targeted program?

While it may, in certain specific contexts, be useful to emphasize demographic data as a foundation for targeted interventions, it is unwise and unproductive to do so if *any* of the following three conditions exists:

- **Insufficient exploration of student variables.** If non-malleable demographic data—say, for example, race or gender—is found to be one of the most predictive variables in a given situation, this fact should be seen as a warning sign: it implies that unexplored variables may exist that can be far more predictive and/or have mediating or moderating effects on student success than characteristics expressed through demographics. For example, members of a

small minority group might have a strong affinity towards each other, resulting in cliques outside school speaking in their native tongue, which can slow cultural assimilation and language-skill development. In such a case, it's not the group *membership* that is interfering with student success, but the group *behaviors*.

In the context of the COVID-19 pandemic, LMS use has grown substantially among institutions of higher education and variables derived from it have become particularly significant. Civitas Learning has always relied heavily on robust derived LMS variables to predict student success in course completion and persistence since student engagement is one of the most important non-academic factors to influence through well-designed interventions.

- **Poor model accuracy.** It is virtually impossible to build an accurate predictive model of student success using predominantly demographic variables as specified, for example, in Title VII of the Civil Rights Act of 1964. That is, any model where demographic variables play a crucial role in predictions is likely to be a poor model with low predictive accuracy and little actionable insights to drive student success efforts.
- **Punitive use case.** It is never appropriate to use demographically based predictions to deny services, awards, employment, etc. Equally bad is to use predictions from a defective and poor model to target and harrass good students. Even affirmative action has been controversial because of the zero-sum reality in college admissions.

If the majority of our students identify as a single demographic group, how will this imbalance influence our prediction and impact models?

In this case, models will likely ignore majority/minority indicators and focus on mostly derived student variables. If something truly unique and different exists in the derived variables for minority students, models may not learn these unique characteristics if *N* is very small. However, the more likely scenario is that the characteristics of minority students will resemble those of *some* majority students. When, for example, Civitas performed a model performance analysis across different student groups for an institution with mostly a single demographic group, we found performances to be comparable.

When we include demographic information in our models, how can we test to make sure the model itself isn't biased?

As part of a QA on a predictive model, Civitas checks the ranking of one or more demographic variables. In our experience, we do not generally see these variables in the top model feature subset because of the presence of the numerous derived variables that belong to the five impact-lever categories.

The only exception to this finding was the variable of age for a few institutions where significant differences occurred in persistence rates among traditional students, non-traditional students, and high school dual-enrolled students. In this case, we worked with the customers by highlighting for each student group top malleable predictors that can be influenced through interventions, investigating which impact levers were elastic, identifying programs catered to each impact lever, and designing interventions to help students with low prediction scores in each group.

Some customers ask for model performance statistics across various groups of students to see if material discrepancies exist in performance. Our finding is that performance discrepancies, if they exist, are due to data footprint differences, such as new incoming vs. experienced students. In predictive models designed to assist students through appropriate interventions, prediction 19accuracy is of high priority.

Would running an impact analysis that excludes majority students be a good way to remove bias?

The best way to remove bias without sacrificing statistical power in impact analysis is to use highly predictive and influenceable student variables in building predictive models and conducting ***pilot-control matching***, in which pilot and control students are matched based on their likelihoods of success and receipt of treatment to ensure apples-to-apples comparisons. Matching students in a diverse class of derived student variables is important in predicting student success and provides a healthy balance, without inherent bias, in using drill-down impact analysis to understand how an intervention program affects various groups of students.

Should we just remove all demographic data from our models?

Civitas Learning can easily remove demographic variables from our models upon request. However, because demographic variables are not generally among the most predictive and actionable variables in our models, it is generally unnecessary to eliminate them as a precautionary measure to guard against bias. Further, institutions may want to retain the option of using demographic data as filters to compare population segments or highlight different rankings of powerful predictors and impact levers by group.