A rubric for assessing the legitimacy of predictive optimization

Website: https://predictive-optimization.cs.princeton.edu/

Predictive optimization is a type of automated decision-making in which machine learning is used to predict future outcomes of interest about individuals.

Each of the following items represents a fundamental consideration that can be used to determine the legitimacy of a decision-making algorithm before it is deployed. In our paper, we find that these issues are pervasive across automated decision-making tools, especially in predictive optimization.

Intervention vs. prediction (Sec 3.1)

Automated decision-making algorithms are used to make decisions based on the data they are trained on. The type of decision can affect how well the algorithms work. Even if a tool makes the correct prediction, if the decision taken based on this decision is flawed, the tool cannot work as claimed. A decision taken using automated tools is also called an intervention.

1. Will all individuals who receive the same prediction will respond to the intervention similarly?

Predictive optimization assumes that individuals who receive the same prediction will benefit equally from the same intervention. This is often flawed. For example, if a student is predicted to be likely to drop out of school, an intervention based on increased tutoring may not benefit a student who is dropping out for financial reasons as much as it is benefiting one who is struggling academically.

2. Does the intervention affect the outcome that is being predicted?

If the intervention based on an automated tool affects the outcome directly, it can lead to a feedback loop. For example, like a self-fulfilling prophecy, higher bail amounts set due to a prediction of recidivism can increase the likelihood of recidivism.

3. Do individually optimal predictions lead to a globally optimal intervention?

Predicting each individual correctly does not always correspond to the best overall decisions. For example, individually hiring salespeople may not account for how well they work together, and can lead to overall worse sales.

4. Does it make sense for this intervention to be decided by a prediction about an individual?

Certain kinds of interventions are easier to formulate as a predictive problem than others. For example, in criminal justice, incapacitation is more amenable to a predictive formulation compared to rehabilitation.

Target-construct mismatch (Sec 3.2)

In constructing an application of predictive optimization, some existing data must be chosen for the model to predict. For example, to predict who will do well in college, the application could try to predict the GPA at the end of the 1st year of college. The outcome being predicted is called the target variable. The target variable is typically chosen to roughly correspond to the decision maker's goal—also called the construct. For example, in predicting whether a student will do well at college, success in college is the construct and GPA at the end of 1st year is the target.

5. Do the goals of the decision maker align with the goals of other relevant stakeholders, such as society at large?

There can be a number of different goals behind a particular decision. For example, police may want to maximize the number of speeding tickets while society may want safer roads.

6. Has the decision maker precisely articulated their goal?

The goal of a particular decision can be vague or concrete. For example, an employer may purport to want employees with higher “performance” without clearly saying what aspect of performance they prioritize, as opposed to providing clear metrics that they are focusing on.

7. Does the target variable capture the many goals of the decision-maker?

Ultimately, predictive optimization formulations have one target variable that the algorithm has been trained to predict. For example, even though college admissions officers may have many goals, such as student diversity, student success, etc., an automated admissions system may predict only one thing, such as projected GPA.

8. Does the decision maker’s goal correspond to something measurable?

Given that predictive optimization formulations require a target variable, something measurable will have to be chosen. For example, we cannot measure crime, only arrests.

9. Is the mismatch between the construct and target made transparent?

In most real-world scenarios, there is some mismatch between the decision maker’s goal and the target variable used. For example, healthcare costs may serve as a target for healthcare needs. Making such choices transparent enables informed normative debate.

Distribution shift (Sec 3.3)

10. Is the data the model was trained on representative of the data the model will be deployed on?

In predictive optimization, a model is trained on one set of individuals and deployed on another. Any mismatch between these two distributions can lead to incorrect

predictions. For example, loan data from a recession likely would not generalize to loan data in other economic periods.

11. Is the data obtained under an already existing intervention?

When training data is collected, an intervention is often already in place. For example, an algorithm that predicted whether a patient would be readmitted to a hospital learned that patients with Asthma were at a lower risk because, in the data, patients with Asthma were more likely to receive better care that reduced their risk of readmission.

12. Does the intervention create a feedback loop?

The results of an intervention may directly impact what is being measured. For example, being rejected for a credit card because of an applicant's credit score can lower someone's credit score even more, making them less likely to be accepted in the future.

Limits to prediction (Sec 3.4)

One of the characteristics of predictive optimization is that the prediction target is a future event in an individual's life. Thus, there are many inherent limits to prediction that limit how accurate the system could be.

13. Is the accuracy of the model better than random chance?

Accuracy can be a misleading metric for evaluating predictions. For example, in binary predictions such as recidivism, 99% of individuals in the dataset have the same label. A model which simply outputs the same label every time could have an accuracy of 99%.

14. Is the accuracy of the model better than a simple baseline?

Being better than simply random chance is a low bar, so it's also useful to understand whether a predictor is better than a simple baseline heuristic such as "admit everyone with an income above a given threshold."

15. Is the model's accuracy better than a baseline that only uses historical correlations between demographic factors and the target variable?

Another baseline suggested by Hardt & Kim is a model which uses only historical correlations between demographic factors and the target variable rather than features specific to the individual's circumstances. For example, a model which uses only an individual's gender and race to predict the outcome, rather than aspects of this individual's life that were up to them.

Disparate performance between groups (Sec 3.5)

Disparate performance refers to differences in performance for different demographic groups.

16. Is the contextually relevant fairness definition clearly articulated and justified?

Fairness impossibility theorems state that many reasonable mathematical definitions of fairness are impossible to satisfy simultaneously under most real-world conditions. It's important to justify the fairness criteria used.

17. Is there clear evidence that the relevant notion of fairness is satisfied?

Developers must provide clear evidence that their model satisfies the relevant definition of fairness. For instance, if demographic parity is the chosen fairness definition, it's important to ensure that acceptance rates do not vary between different demographic groups without clear and valid justification.

Contestability (Sec 3.6)

When decision-making algorithms are deployed in consequential settings, they must include mechanisms for contesting such decisions.

18. Is there an explanation accompanying a decision about an individual?

Explanations of why a decision was made are prerequisites for contesting a decision. Additionally, it is important that decision subjects can understand the explanation to contest the model's decisions.

19. Can individuals access or contest the data a model uses about them?

To better contest a model, it is useful for individuals to have access to the data and decisions made about them. For example, in HireVue, job candidates have no insights into the criteria used for evaluation.

20. Can individuals contest a decision made about them?

There also needs to be a system in place for individuals to seek recourse and contest a decision that may be incorrect.

Goodhart’s law (Sec 3.7)

A canonical example of Goodhart’s law is the cobra effect: when the colonial British government offered bounties for dead cobras to reduce the cobra population, the response instead was people breeding more cobras to kill. Similarly, predictive optimization can create unintended incentives for decision subjects to game the system.

21. Is the decision-making system susceptible to strategic behavior?

Like our example, an easier way to accumulate dead cobras is to breed cobras rather than hunting for them in the wild.

22. Do privileged people who understand how the decision-making system works benefit?

Different levels of understanding of a decision-making system can lead to more effective gaming of the system amongst certain groups of people. For example, students at schools with more resources, such as career centers, may receive special training on how to stand out to automated resume screening tools.

23. Do subjects bear the cost of spending lots of effort trying to change the decision-making system's outcome without knowing whether it will affect their chances?

Trying and achieving a desirable outcome from an automated decision-making system can take a lot of effort. For example, job applicants may spend a long time trying to learn a more advanced vocabulary for a particular job interview without knowing if potential employers will consider it during the hiring process. The vocabulary improvement might not lead to better job performance.

Additional Factors that Affect Legitimacy (Sec 4.3)

In addition to the critiques we propose, application-specific considerations are relevant in determining if a decision-making system is legitimate. Consider a travel agency that uses ML to predict whether someone may decide to vacation soon and advertises travel deals to selected individuals. While this example falls under the scope of predictive optimization, it is less morally reprehensible than some other applications since the decision's impact is negligible.

24. Is the decision maker in the public or private sector?

Different sectors have different responsibilities to the general public and are subject to varying forms of governmental regulation.

25. What degree of choice does the subject have in choosing the decision maker?

Individuals have different levels of choice in whether they will be subject to a particular decision-making system. For example, while there are many alternatives for job options, subjects cannot choose which standardized tests they take (SAT for college admissions).

26. How severe is the consequence of a misclassification?

The severity of misclassifications is different for different applications. Setting a high cash bail has severe consequences compared to not getting selected for one of many jobs you apply for.

27. Does the application assign *opportunities* or *burdens*?

Being subjected to the decision-making system voluntarily (e.g., a job application—an opportunity) has lower stakes than being subjected to it mandatorily (e.g., recidivism prediction—a burden).