

# A Causal Approach for Selective Labels

Anonymous Author(s)

## ABSTRACT

We show how a causality-based approach can be used to estimate the performance of prediction algorithms in ‘selective labels’ settings – with particular application to ‘bail-or-jail’ judicial decisions.

### ACM Reference Format:

Anonymous Author(s). 2019. A Causal Approach for Selective Labels. In *Proceedings of ACM Conference (Conference’17)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

## 1 INTRODUCTION

‘Selective labels’ settings arise in situations where data are the product of a decision mechanism that prevents us from observing certain variables for part of the data. A typical example is that of bail-or-jail decisions in judicial settings: a judge decides whether to grant bail to a defendant based on whether the defendant is considered likely to violate bail conditions while awaiting trial – and therefore a violation might occur only in case bail is granted. Such settings give rise to questions about the effect of alternative decision mechanisms – e.g., ‘how many defendants would violate bail conditions if more bail decisions were granted?’. In other words, one faces the challenge to estimate the performance of an alternative, potentially automated, decision policy that might make different decisions than the one found in the judicial data.

The challenge was addressed by Lakkaraju et.al. in [? ], in a setting that involved multiple judges of varying leniency, and under the assumption that defendants are assigned to judges randomly. Lakkaraju et.al. estimate the performance of an automated decision-making algorithm (‘algorithm’, for short) via a technique they call ‘contraction’ – it proceeds as follows:

- It considers a set of judges with same number  $N$  of judged defendants each.
- Judges are ordered from most lenient (most bail decisions) to least lenient. Let  $n_i$  be the number of bail decisions for judge  $\#i$ . We have  $n_{i+1} \leq n_i$ .
- The algorithm considers the  $n_i$  defendants that were granted bail by the  $i$ -th judge.
- It keeps the  $n_{i+1} \leq n_i$  defendants that it finds most likely to violate the bail.
- It makes its own bail-or-jail decision for each of those  $n_{i+1}$  defendants.
- Its performance is measured as the number of defendants that it decides to bail but who, according to the data, eventually violated the bail.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference’17, July 2017, Washington, DC, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

- Its performance is compared to the performance of judge  $\#(i + 1)$ , based on the cases they bailed.

The above procedure gives us a comparison between the performance of the algorithm to that of judges at the  $n_{i+1}/N$  leniency level (leniency measured as the rate of bail decisions). A major drawback of the *contraction* technique is that it requires data to include judges at a given leniency level.

In this document, we describe a different approach based on causal analysis, that allows us to estimate the performance of a decision-making system at any leniency level.

## 2 SETTING

Consider a judge who decides whether to grant bail to a defendant based on whether the defendant is considered likely to violate bail conditions while awaiting trial. We use variable  $T$  to store the outcome of the bail-or-jail decision, with  $T = 1$  denoting a bail decision and  $T = 0$  a jail decision. Whether the defendant violates the bail conditions depends on the bail-or-jail decision  $T$  and the features  $X$  of the defendant.

The decision is based on the following variables. First, the features  $X$  of the defendant, which we assume to be observed. Secondly, the leniency of the judge, expressed as a variable  $R$ . Specifically, we assume that every judge evaluates a given candidate according to the probability

$$P(Y = 0|X = x, \text{do}(T = 1))$$

that the candidate will violate bail conditions ( $Y = 0$ ) if they were granted bail. We write  $Y = 0$  to refer to the case when the defendant does not violate bail, whether bail is granted or not. The  $\text{do}(\text{condition})$  expression signifies that, in evaluating the probability, we consider the event where the condition (here, it is the condition  $T = 1$ ) is imposed to the data-generation process (and therefore alters the generative model). In addition, we assume that every judge would assign the same value to the above probability, given by a function  $f(x)$ .

$$f(x) = P(Y = 0|X = x, \text{do}(T = 1))$$

The assumption that, essentially, all judges have the same model for the probability that a defendant would violate bail is not far-fetched for the purposes of our analysis, particularly taking into account that  $f(x)$  can be learned from the observed data

$$P(Y = 0|X = x, \text{do}(T = 1)) = P(Y = 0|X = x, T = 1)$$

and that data are publicly accessible, allowing us to assume that all judges have access to the same information. Where judges *do differ* is at the level of their leniency  $R$ . Following the above assumptions, a judge with leniency  $R = r$  grants bail to the defendants for which  $F(x) < r$ , where  $F$  is the cumulative distribution.

$$F(x_0) = \int \delta(P(Y = 0|T = 1, X = x) > P(Y = 0|T = 1, X = x_0))dP(x) \quad (1)$$

which should be equal to

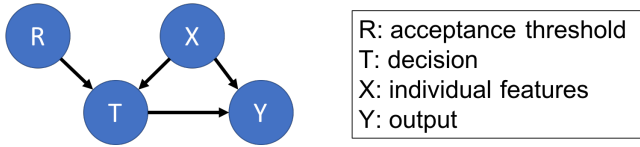


Figure 1: Causal model.

$$F(x_0) = \int \mathbf{P}(x) \delta(\mathbf{P}(Y = 0|T = 1, X = x) > \mathbf{P}(Y = 0|T = 1, X = x_0)) dx \quad (2)$$

The bail-or-jail scenario is just one example of settings that involve a decision  $T \in \{0, 1\}$  that is based on individual features  $X$  and leniency (acceptance rate)  $R$ — and where a behavior of interest  $Y$  is observed only for the cases where  $T = 1$ . The diagram of the causal model is shown in Figure 1. Our results are applicable to other scenarios with same causal model.

## 2.1 Analysis Task

We will use existing machine-learning techniques from the literature to learn function  $\mathbf{f}(x)$ , with the goal to build a decision system that outperforms judges. The challenge we face is to estimate accurately the performance of the decision system – given that we are in a ‘selective labels’ setting. Performance is measured *for a given leniency level* as the rate at which bail is granted *and* the defendant violates it. In other words, performance is measured as the probability that a decision lead to undesired outcome.

## 3 ANALYSIS

We wish to calculate the probability of undesired outcome ( $Y = 0$ ) at a fixed leniency level.

$$\begin{aligned}
 \mathbf{P}(Y = 0 | \text{do}(R = r)) &= \\
 &= \sum_t \mathbf{P}(Y = 0, T = t | \text{do}(R = r)) \\
 &= \mathbf{P}(Y = 0, T = 0 | \text{do}(R = r)) + \mathbf{P}(Y = 0, T = 1 | \text{do}(R = r)) \\
 &= 0 + \mathbf{P}(Y = 0, T = 1 | \text{do}(R = r)) \\
 &= \mathbf{P}(Y = 0, T = 1 | \text{do}(R = r)) \\
 &= \sum_x \mathbf{P}(Y = 0, T = 1, X = x | \text{do}(R = r)) \\
 &= \sum_x \mathbf{P}(Y = 0, T = 1 | \text{do}(R = r), X = x) \mathbf{P}(X = x | \text{do}(R = r)) \\
 &= \sum_x \mathbf{P}(Y = 0, T = 1 | \text{do}(R = r), X = x) \mathbf{P}(X = x) \\
 &= \sum_x \mathbf{P}(Y = 0 | T = 1, \text{do}(R = r), X = x) \mathbf{P}(T = 1 | \text{do}(R = r), X = x) \mathbf{P}(X = x) \\
 &= \sum_x \mathbf{P}(Y = 0 | T = 1, X = x) \mathbf{P}(T = 1 | R = r, X = x) \mathbf{P}(X = x)
 \end{aligned}$$

Expanding the above derivation for model  $\mathbf{f}(x)$  learned from the data

$$\mathbf{f}(x) = \mathbf{P}(Y = 0 | X = x, T = 1),$$

the *generalized performance*  $\mathbf{gp}$  of that model is given by the following formula.

$$\mathbf{gp} = \sum_x \mathbf{f}(x) \delta(F(x) < r) \mathbf{P}(X = x) \quad (3)$$

Equation 4 can be calculated for a given model  $\mathbf{g}(x) = \mathbf{P}(X = x)$  of individual features. Alternatively, we can have an empirical measure  $\mathbf{ep}$  of performance over the  $n$  data points in dataset  $\mathbf{D}$ , given by the following equation.

$$\mathbf{ep} = \frac{1}{n} \sum_{(x, y) \in \mathbf{D}} \delta(y = 0) \delta(F(x) < r) \quad (4)$$

## 3.1 Comments

Roughly speaking, the above formulas should work well if ‘bail’ cases ( $T = 1$ ) cover well the area spanned by the observed features of defendants – i.e., we do not have large areas of  $X$  with no or too few bail cases.

If there are such areas, then we cannot do much about the lack of data. One reasonable modeling choice, however, is to impose the following priors on  $\mathbf{f}(x)$ :

- (1)  $\mathbf{f}(x) \approx 1$  for areas near values of  $X$  for which we have observed data but few bail decisions (i.e., we assume a-priori that a defendant is more likely to violate bail – a belief that will change if the data tell us otherwise);
- (2)  $\mathbf{f}(x) \approx 0$  for areas near unobserved values of  $X$  (i.e., we assume that people who are unlikely to ever be taken to court would probably ‘play nice’ and not violate bail).

Lack of data for large areas of  $X$  is a potential problem for the *contraction* technique of Lakkaraju et.al., as well. Unlike contraction, though, our approach does not require to have data at all leniency levels. Moreover, it is easy to see based on the derivations of Eq.4 that our approach would work identically in the case where defendants are not assigned to judges at random (i.e., if there was a causal relation  $X \rightarrow R$ ).