

Discretionary Trees: Understanding Street-Level Bureaucracy via Machine Learning

Gaurab Pokharel¹, Sanmay Das¹, Patrick J. Fowler²

¹George Mason University

²Washington University at St. Louis

gpokhare@gmu.edu, sanmay@gmu.edu, pjfowler@wustl.edu

Abstract

Street-level bureaucrats interact directly with people on behalf of government agencies to perform a wide range of functions, including, for example, administering social services and policing. A key feature of street-level bureaucracy is that the civil servants, while tasked with implementing agency policy, are also granted significant discretion in how they choose to apply that policy in individual cases. Using that discretion could be beneficial, as it allows for exceptions to policies based on human interactions and evaluations, but it could also allow biases and inequities to seep into important domains of societal resource allocation. In this paper, we use machine learning techniques to understand street-level bureaucrats' behavior. We leverage a rich dataset that combines demographic and other information on households with information on which homelessness interventions they were assigned during a period when assignments were not formulaic. We find that caseworker decisions in this time are highly predictable overall, and some, but not all of this predictivity can be captured by simple decision rules. We theorize that the decisions not captured by the simple decision rules can be considered applications of caseworker discretion. These discretionary decisions are far from random in both the characteristics of such households and in terms of the outcomes of the decisions. Caseworkers typically only apply discretion to households that would be considered less vulnerable. When they do apply discretion to assign households to more intensive interventions, the marginal benefits to those households are significantly higher than would be expected if the households were chosen at random; there is no similar reduction in marginal benefit to households that are discretionarily allocated less intensive interventions, suggesting that caseworkers are using their knowledge and experience to improve outcomes for households experiencing homelessness.

Introduction

Homelessness is an acute and pervasive issue affecting societies worldwide. The lack of stable housing and access to resources poses immense challenges for individuals and families, jeopardizing their well-being and livelihoods. As federal guidelines increasingly emphasize ending homelessness entirely (United States Interagency Council on Homelessness 2022), a need exists for a comprehensive under-

standing of its causes, dynamics, and the strategic allocation of limited resources to provide timely interventions. In the United States, decisions on homeless service allocation are determined through a complex mix of federal guidelines, community and local agency policies, and the interpretation and implementation of these policies by on-the-ground caseworkers.

Homelessness caseworkers are an example of what Lipsky, in a seminal work, calls *street-level bureaucrats* (Lipsky 2010). These are public-facing employees of government agencies who are granted discretion and may (or may not) follow the agencies' policies in each decision they make. Among the reasons Lipsky details for why street-level bureaucrats implement different policies include resource limitations (e.g., the time and information constraints faced by tremendously overworked caseworkers and street-level bureaucrats), as well as the misalignment of individual and agency objectives that may be ambiguous.

This paper aims to use machine learning tools to better understand street-level bureaucrats' behavior. In particular, we are interested in three main questions grounded in the street-level bureaucracy theory: (1) Can bureaucratic decision-making be boiled down to applying a relatively simple set of heuristics/rules? The theory says street-level bureaucrats often adopt such methods to reduce cognitive load and decision-making time. (2) When decisions are not driven by these rules (examples of the application of *caseworker discretion*), are caseworkers nonetheless *consistent* in their decision-making? Inconsistent application of discretionary powers is a major problem because it leads to individuals suffering allocative harms without a procedural basis and could cause considerable bias and inequity. (3) When caseworker discretion is applied, what can we say about the cost/benefit trade-off to the subjects of such discretion?

We leverage a rich dataset of allocations of households to homelessness services during a period in which the allocation process was less formulaic than imposed by the subsequent adoption of decision-making tools (Eubanks 2018; HUD 2017). In addition to the interventions assigned, the dataset contains demographic and other features of the households. We use machine learning methods to ask the question *how predictable are the interventions assigned to households* and to understand the nature of caseworker decisions. We present evidence to support several impor-

tant claims related to the questions above. First, simple rules can capture some, but not all, of the decision-making about which allocations households receive. State-of-the-art “short” decision trees achieve areas under the ROC curve (AUCs) of 0.76 and 0.87 when predicting whether households will receive Emergency Shelter and Transitional Housing, respectively (the two main interventions we study). Second, caseworker decision-making is generally highly consistent beyond what can be explained with simple decision trees. The AUCs for gradient-boosted trees on the same two tasks are 0.95 and 0.94, respectively. Thus, caseworkers are procedurally consistent in their decision-making. Third, caseworkers apply discretion (operationalized as decisions that short decision trees would not have correctly predicted) in an interesting way. Keeping in mind that Transitional Housing (TH) is a more intensive intervention that generally prevents reentry into homeless services (the outcome of interest) better than receiving less intensive Emergency Shelter (ES) that has poorer outcomes, we find that caseworkers typically only apply discretion (in either direction: moving households predicted to do best in TH to ES or predicted ES to TH) on households that would be considered less vulnerable. However, the group that moves from predicted ES to TH (thus, to the more intensive intervention) systematically receives higher expected marginal benefit from this change; conversely, the group that moves from TH to ES does not receive significantly lower expected marginal benefit. These estimates of marginal benefit are based on counterfactual machine learning predictions and were not available in any form to the caseworkers; thus, this third result implies caseworkers can assess and apply discretion in a highly sophisticated manner to the overall benefit of vulnerable populations.

More broadly, the work represents the first to our knowledge that applies techniques from artificial intelligence and machine learning to understand decision-making in the context of street-level bureaucracies, which are specially tasked with many of the core public-facing tasks of local and state governments. Shedding light on how human decision-making works may help us understand how and when automated decision-support may be most valuable to those charged with interfacing with the public in high-stakes allocation and justice scenarios. By gaining a deeper understanding of how organizational policies interact with the street-level bureaucracy typically employed by caseworkers, we can foster greater clarity and transparency in resource distribution processes and start to untangle some of the intricate dynamics between organizational policies and the caseworkers responsible for implementing them.

Background and Related Work

Homelessness Federal guidelines define homelessness as residence in unstable and non-permanent accommodations. This includes emergency shelters, temporary housing, and places not meant for habitation, such as cars, parks, and abandoned buildings (P.L. 112-141). According to de Sousa et al. (2022), more than 580,000 people experienced homelessness on a single night in January 2022 across the United States, while approximately 1.5 million people use homeless

services at some point during each year (Henry et al. 2018). Furthermore, three in ten people experiencing homelessness did so as a part of a family with children under 18 years of age, and one-third of individuals experiencing homelessness exhibited chronic patterns of homelessness (de Sousa et al. 2022). The enduring consequences of homelessness and the resulting upheaval have lifelong implications with substantial societal burdens in terms of reduced productivity, compromised health, and increased expenditures on compensatory social services (Khadduri et al. 2010; Culhane, Park, and Metraux 2011; Fowler et al. 2019a).

There has been some recent work on AI approaches towards homelessness. Much of this work has focused on techniques designed to improve resource allocation to homeless households (Azizi et al. 2018; Kube, Das, and Fowler 2023b, 2019). Other related work has examined what values and efficiency/fairness tradeoffs have been (Mashiat et al. 2022) or should be (Vayanos et al. 2020) encoded into algorithmic approaches. Despite the pivotal role played by allocation decisions in shaping outcomes for homeless individuals and families, substantial gaps remain in understanding the deployment of homeless services (Brown et al. 2018; Fowler et al. 2019b; Shinn et al. 2013). The knowledge gap hinders the development of evidence-based strategies and the establishment of equitable mechanisms for resource allocation within homeless services. As far as we know, this is the first paper to use machine learning approaches to examine caseworker decision-making.

Street-level bureaucracy The seminal work of Lipsky (2010) on street-level bureaucracy is the theoretical underpinning for the questions we ask in this paper. As mentioned in the Introduction, Lipsky notes the tensions between different caseworker objectives and the importance of their ability to use judgment and discretion in their sense of self. It is precisely these tradeoffs we explore in this paper.

Some recent qualitative work has used the lens of street-level bureaucracy in thinking about human-AI interaction. Kawakami et al. (2022) conduct a detailed qualitative investigation into how child welfare workers utilize an AI Decision Support (ADS) system called the Allegheny Family Screening Tool (AFST) in their day-to-day decision-making processes. The study finds that workers often perceive value misalignments between the ADS’s predictive targets and their own objectives. Workers compensated for the limitations of the ADS by relying on their contextual knowledge of individual cases. Organizational pressures and incentives influenced workers’ reliance on the ADS, sometimes leading them to comply with the AFST recommendations even against their best judgment. Additionally, workers felt limited agency in shaping the ADS’s use or improving its accuracy. The paper highlights the need to address human-AI value misalignments.

Kuo et al. (2023) conducted an in-depth study using AI lifecycle comicboarding, a novel method that allowed frontline workers and unhoused individuals to provide specific feedback on an AI system’s design and deployment. Most pertinently to our work, participants expressed concerns about potential biases and harm caused by the algorithm,

particularly in its optimization towards the county’s interests rather than addressing the needs and safety of unhoused individuals. Our work complements this literature thread by analyzing human decision-making (rather than human-AI complementarity) using tools from AI (rather than qualitative methods). It thus sheds a “revealed preference” style light on what caseworkers do in practice.

Interpretability Our main connection to the interpretability literature is how we operationalize our notion of identifying short heuristic rules that caseworkers could use to make quick decisions, reducing cognitive load in “easy” cases. We do so using a relatively new short decision-tree learning algorithm, which comes from interpretable machine learning. Doshi-Velez and Kim (2017) highlights the importance of interpretable machine learning and advocates for a scientific approach. The field of interpretability often focuses on efforts to explain model characteristics, like the LIME framework (Ribeiro, Singh, and Guestrin 2016), SHAP values (Lundberg and Lee 2017), and risk scores (Ustun and Rudin 2019). But using such black-box, model-independent metrics – while offering valuable insights into opaque models – suffers from limitations in the specific context of what we are trying to accomplish. The main issue is that their model-agnostic nature may lead to explanations that fail to capture the idiosyncrasies of specific models, potentially compromising the accuracy of the interpretations. In this study, we embrace Decision Trees (DT) as the model of choice due to their inherent interpretable nature, the importance of which has been stressed by Ustun and Rudin (2019) when making high-stakes decisions. DTs’ model-specificity ensures accurate representations of individual models, providing insights into feature contributions and decision rules. We use a recent algorithmic technique proposed by Souza et al. (2022) to learn short decision trees to try and enhance our understanding of caseworker behavior.

The Data

Through a data sharing agreement, this work uses administrative records from the homeless management information system (HMIS) in St. Louis, MO, collected from 2007 through 2014. Local service providers collect data in real-time using a web-based platform as individuals and families seek federally funded homelessness assistance. The platform is managed by a local non-profit organization contracted with the homeless system, which offers support, user training, technical assistance, and quality control.

The HMIS maintains detailed records, encompassing information on household-level characteristics, such as demographics, housing risk, and eligibility determinations. Additionally, the data captures entry and exit dates from five federally defined types of homelessness assistance in increasing order of their intensity: homelessness prevention (prev), emergency shelter (ES), rapid rehousing (RRH), transitional housing (TH), and permanent supportive housing. The metropolitan area operated a telephone hotline to coordinate service delivery. Records exist on every call, including dates and service referrals. Through the use of household identifiers, information can be linked across different time pe-

riods. Data-sharing agreements with regional homeless systems are in place to ensure privacy and security, allowing access to deidentified records in accordance with Institutional Review Board guidelines. Best practices in data security, including ethics training, are followed for data transfer, storage, and analysis.

Data Cleaning

Service records come from 75 different housing agencies serving households assigned a unique, anonymous household identification number Kube, Das, and Fowler (2023b). Data contains household characteristics available upon entry into the system, as well as information on all entries and exits from different homeless services. We exclude permanent supportive housing for the present study because the service was rarely used as an initial response for first-time entries into the homeless system between 2007 and 2014. In contrast to Kube, Das, and Fowler (2023b) that assessed whether households reentered the homeless system within a two-year follow-up, we aim to predict the interventions assigned to the households. Hence, our target variables are one of the four remaining intervention types, namely: TH, RRH, ES, and Prev. We use a total of 34 features to predict these targets, some of which are listed in Table 1 with a complete list provided in the supplementary material.

Feature Type	Number of Features	Examples
Binary Features	3	Gender, Spouse Present, HUD Chronic Homeless
Non-Binary Categorical Features	17	Veteran Status, Disabling Condition, Substance Abuse
Continuous Features	14	Age, Monthly Income, Wait Time, Calls to Hot-line
Total Features	34	

Table 1: Summary of features in the dataset

As Table 1 illustrates, we used a collection of binary, categorical, and continuous features. We one-hot-encoded the categorical features. We refer the reader to the supplementary material for details regarding the characteristics of the data, including the makeup and distribution of the interventions assigned.

Time context: As mentioned, the data captures first-time entries before 2014. At this time, there was no explicit formula for prioritization for different interventions used by the homeless service system. Thus, caseworkers relied more on their own judgment in making decisions. They balanced potential trade-offs between prioritizing based on household vulnerability (e.g., chronically homeless households with comorbid conditions) and outcome – or allocations expected to be the most beneficial for households. Federal policy shifts following the study period required the adoption of coordinated entry into services based on vulnerability scores, such as measured by the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) (Consulting 2015).

Intervention focus While we report results on the accuracy of simple decision rules and the consistency for all four interventions we study, we focus on two groups – Transitional Housing (TH) and Emergency Shelter (EH) – in analyses. The other two interventions (RRH and Prev) were only available for a part of the study period, as they were funded

through the American Recovery and Reinvestment Act in the wake of the 2008 Great Recession. Other work has shown that Prevention, in particular, is a different type of intervention that is not directly comparable to the others in terms of eligibility and outcomes (Kube, Das, and Fowler 2023b).

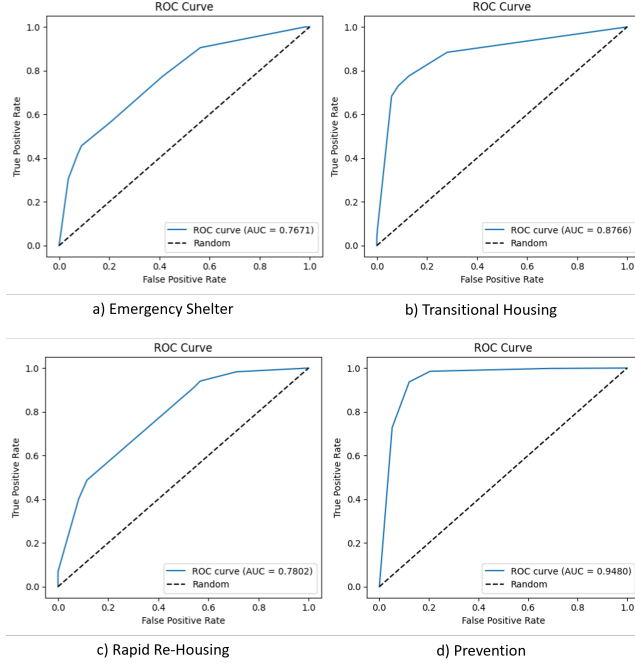


Figure 1: ROC curves for prediction of intervention assignment in a one-vs-all setting for all four interventions using SER-DT. Simple rules are quite effective at predicting interventions overall, implying the use of short rules/heuristics by caseworkers in many cases is plausible and consistent, as predicted by theory.

Results: Simplicity and Consistency in Intervention Assignment

Our main investigatory tool is the predictability of caseworker behavior in assigning households to particular interventions. We employ ML methods to learn the associated decision rules. We start by investigating what fraction of caseworker behavior can be captured by simple decision rules before considering the consistency in intervention assignments.

Simple Decision Rules

Street-level bureaucracy theory suggests that caseworkers develop simple heuristics that enable them to quickly make decisions on most cases, reserving their expertise and time for more complex cases. To test this, we attempt to learn intervention assignments using *short trees*. These are functionally equivalent to simple rules, or the common understanding of heuristics in the behavioral sciences, which are typically contrasted with more cognitively demanding linear models (Hogarth and Karelaia 2007). We use a relatively

Intervention	Average AUC Value	AUC (Best Model)	95% CI (Best Model)
Emergency Shelter	0.7608	0.7671	[0.7452, 0.7890]
Transitional Housing	0.8703	0.8766	[0.8547, 0.8984]
Rapid Re-Housing	0.6508	0.7802	[0.7463, 0.8170]
Prevention	0.9384	0.9480	[0.9371, 0.9589]

Table 2: Average AUC values over 10 different runs, AUC values for the best model, and the 95% confidence interval for the best model predicting different interventions using SER-DT.

new algorithm from the literature, the SER-DT algorithm (Souza et al. 2022). This algorithm is specifically designed to optimize “explanation size” – a custom metric aimed at enhancing explainability using as simple a decision tree as possible.

We applied the method to one-vs-all prediction for the four interventions described above. A partition ratio of 70% for training data and 30% for testing data was employed. Analogous to established methodologies, such as the Classification and Regression Trees (CART) algorithm, the SER-DT method identifies splits that yield sub-trees characterized by diminished ‘impurity’. Unlike other algorithms, SER-DT’s inherent non-greedy nature causes minor variations across runs. This occurs because SER-DT weights the Gini impurity at each node using a hyper-parameter called ‘FactorExpl’ (which was set to 0.97, mirroring the practice from the original paper for a minimum trade-off in accuracy when seeking explainability), and thus, impacting which split gets used on which feature. To ameliorate this slight variation in trees, 10 trees were trained with distinct seeds per intervention. We report the average AUC values of the trees for each intervention in Table 2 along with the AUC values for the best model (taken as the model with the highest test accuracy) and its 95% confidence interval (CI) calculated using DeLong’s method (DeLong, DeLong, and Clarke-Pearson 1988) with a sample size of 1907. Moreover, the structural depth of the trees is constricted to a maximum limit of four tiers. This constraint is formulated in congruence with the guiding principle of preserving the “simplicity” of the decision rules. Greater depth gains higher accuracy at the expense of potentially making the tree more complicated to understand.

Figure 1 shows the AUC curves for the best performing model in each case, while the corresponding trees generated specifically for TH and ES appear in Figure 2. As can be seen, the short rules are good at explaining a chunk of human decision-making regarding intervention assignments based on household characteristics. We see that the short decision tree achieves an AUC of 0.87 in predicting whether a household will receive TH. Examining the tree suggests the basic rules for achieving the high AUC: Assign TH to households with male heads who make \$666.5 or less and either have prior calls to the hotline or are young (less than 23yo); assign TH to households with young (less than 19 years) female heads of household. The assignment of ES is somewhat less predictable by simple rules (AUC of 0.76). The rules, again, are interesting when operationalizing the ratio of the number of calls to wait time as a notion of “urgency” for services. One fundamental rule is that households with female

Trees Generated by SER-DT for Transitional Housing and Emergency Shelter

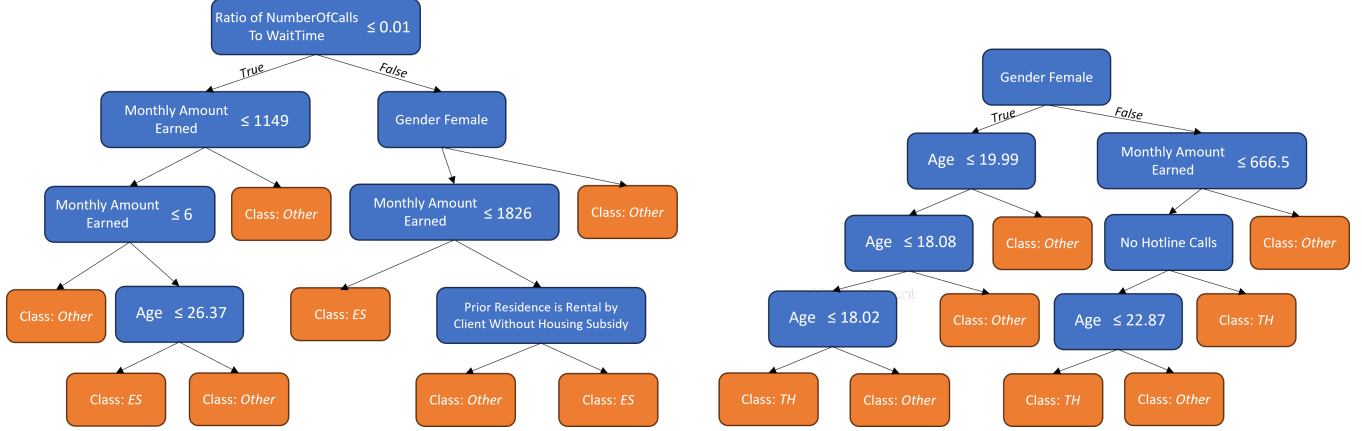


Figure 2: The decision trees generated by SER-DT for the assignment of Emergency Shelter (left) and Transitional Housing (right). The left branches of each node indicate that the parent condition was satisfied, whereas the right branch indicates the condition was false. The generated trees are small but can encompass a significant portion of the correctly made decisions.

heads who have at least some urgency are sent to ES unless their monthly earnings are relatively high (greater than \$1826) **and** their immediate prior residence was an unsubsidized rental. Together with the TH rules, this implies that males are generally prioritized in the simple heuristic rules for transitional housing, possibly because of comorbidities (like alcohol or drug abuse, disability, mental health issues – see Kube, Das, and Fowler (2023a) for a discussion).

For our analysis in the following sections, we take our predicted intervention by SER-DT to be the labels generated by the best performing models i.e. the ones with the highest test accuracy. This could raise a concern about the identifiability of human decisions – for example, how different are predictions made by the same model (SER-DT) due to training set variation? It is important to note that the key question here is about the implications of any model differences on the population they predict would receive different interventions, rather than about differences in models themselves. To test this, we compute rank correlations (Spearman 1904) on ten different SER-DT models, demonstrating a high degree of similarity in the outputs of different models for both TH and ES, affirming consistency in model predictions and alleviating concerns about differences in tree outcomes (see Figure 3).

Complex Decision Making

To understand whether caseworkers consistently assign interventions even when not using simple decision rules, we investigate the predictivity achievable when using a more complex machine learning algorithm. The algorithm of choice here was XGBoost – an ensemble learning algorithm that has gained prominence for its exceptional predictive performance and versatility in a wide range of ML tasks.¹

¹We also do a similar analysis using the CART algorithm. CART has comparable performance to XGBoost in terms of pro-

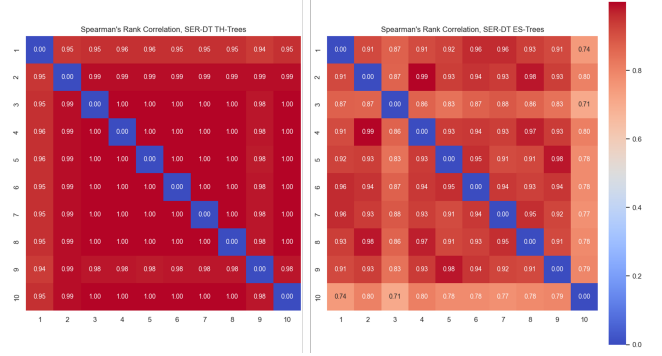


Figure 3: Spearman's rank correlation for ten different models predicting Transitional Housing (left) and Emergency Shelter (right). We observe a high degree of correlation in model output, alleviating any concern regarding differences in tree outcomes.

Again, the dataset was partitioned into training and testing sets, with a ratio of 70% for training and 30% for testing. A systematic grid search approach was employed to optimize hyper-parameters. The explored parameters included the number of estimators (ranging from 50 to 20, increasing in steps of 50) and the maximum depth of the trees (ranging from 2 to 8, increasing in steps of 2).

The optimal hyper-parameters varied across interventions. Specifically, for TH, the optimal configuration was a max depth of 2 with the number of estimators equaling 150. For ES, a max depth of 4 with the number of estimators set to 100 yielded the best performance. RRH achieved optimal results with a max depth of 6 and with the number of estimators at 50, while Prev benefited from a max depth of 2

ductivity. See supplementary material for further details.

Intervention	AUC value	95% AUC CI
Emergency Shelter	0.9452	[0.9392, 0.9513]
Transitional Housing	0.9434	[0.9360, 0.9507]
Rapid Re-Housing	0.9299	[0.9136, 0.9462]
Prevention	0.9875	[0.9846, 0.9904]

Table 3: AUC values and their 95% confidence intervals for the trees used to learn intervention assignments using the XGBoost algorithm.

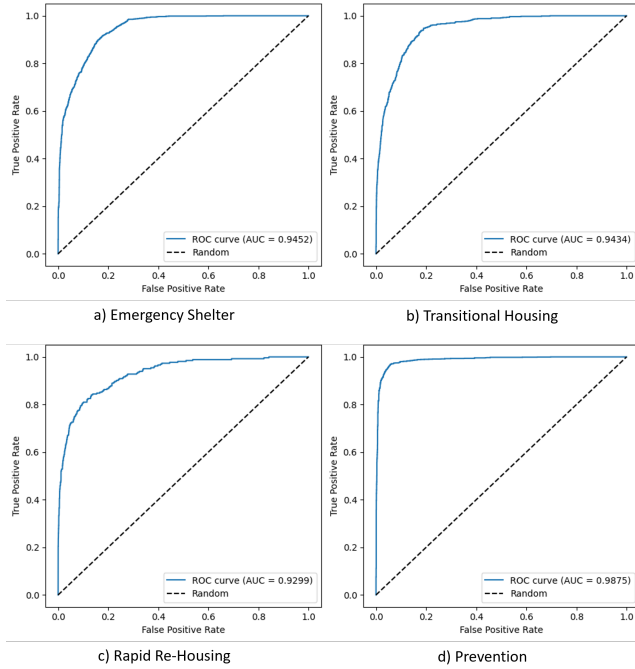


Figure 4: ROC curves for each tree used to learn intervention assignments using the XGBoost algorithm. Predictive discrimination, measured using the area under the ROC curve is very high, implying significant consistency in caseworker decision-making.

and the number of estimators set to 200. Figure 4 displays the ROC curves for the individual decision trees and Table 3 provides details on AUC values and confidence intervals (with a sample size of 4239).

Figure 4 illustrates that the decision rules implemented by caseworkers can be effectively learned with high predictive accuracy. The findings demonstrate that caseworkers are *consistent* in their decision-making. This is important for concerns about procedural justice since there is little arbitrariness in the application of human judgment in these cases. This then raises the question of why caseworkers choose to apply their discretion in these cases, and what it likely means for the outcomes of intervention assignment.

Results: Discretionary Assignments

What explains the application of discretion in the significant fraction of cases where the assignment mismatches that which would have been predicted by the simple decision

rule? In order to better understand this, we consider and operationalize two possibilities that have been hypothesized as major factors in human decision-making on the allocation of scarce homelessness resources: *vulnerability* and *outcome* (Das 2022).

Measuring Vulnerability and Potential Outcomes

To begin, let’s delve into the concept of vulnerability. The Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) is as a widely employed metric to assess vulnerability within homelessness resource allocation (Consulting 2015). Despite its subsequent recognition for introducing biases against specific sub-groups (Brown et al. 2018; Cronley 2022; Shinn and Richard 2022), and its decreasing prevalence, it remained the conventional yardstick for prioritization during the period following our study. Given this context, if vulnerability indeed played a pivotal role in the application of discretion, we would anticipate its correlation with decision-making. While our dataset lacks direct VI-SPDAT scores, we endeavored to closely replicate it using available data to establish a vulnerability score (VS). The full point assignment system we developed is detailed in supplementary material.

Next, we turn to outcomes. While each household has a range of different possible outcomes, local agencies are often evaluated on the basis of success at helping families out of homelessness, that is, whether households were stably housed some time (typically two years) after the homelessness intervention. Therefore we directly use the probability of return to homelessness conditional on different interventions using Bayesian Additive Regression Trees, as estimated in prior work (Kube, Das, and Fowler 2023b,a). Note that the counterfactual estimates provide a sense of how much better one intervention is than another in terms of reducing the probability of returning to homelessness.

We note here again that there is an inherent tension between vulnerability and outcome. Both of these are classic metrics for prioritization in the local justice literature (Das 2022), but it is recognized that they may often produce opposing motivations. Caseworkers frequently are asked to prioritize vulnerability by federal guidelines, while at the same time, the agency is evaluated and funded based on success in getting households out of homelessness (Fowler et al. 2019a). A very vulnerable household may not benefit much from an intensive intervention in some cases, but it may still be appropriate to prioritize them according to our values. These types of trade-offs are what caseworkers have to deal with on a daily basis (Kube et al. 2022).

Investigating Discretion

Can we understand applications of discretion through the lens of vulnerability and/or outcome? We conducted a targeted analysis focusing on a subset of the data where the actual assignments differed from the SER-DT model predictions. Given the employment of a one-versus-all model for each intervention, there exists the potential for an individual household to be predicted to receive multiple interventions. In such instances, the final forecasted intervention for the household is taken from the prediction deemed

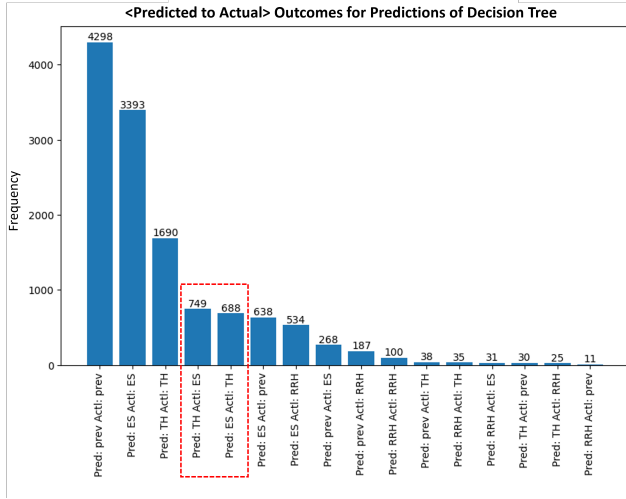


Figure 5: Distribution of the number of records for predicted versus actual assignment of intervention. We focus our analysis of discretion quantification in the subgroups marked within the red box.

most “confident.” Among the 12,546 data points, 9,481 instances aligned with the documented assignments in our dataset, while 3,065 data points (nearly one-quarter) showed disparities. Within the subset of mismatches, we further investigated two distinct groups: 749 data points where our model predicted TH, but ES was assigned; and 688 data points where ES was predicted, but TH was assigned. These subsets represent half of the data points within the discretionary subgroup and are characterized by a clear hierarchy of the intervention’s intensity, with transitional housing being more intensive than emergency shelters. Analyzing the specific subgroups allows us to assess the presence and extent of caseworker discretion in intervention assignments. The full distribution of the predicted versus actual intervention assignments is displayed in Figure 5.

For simplicity, we call the group predicted to receive ES but actually receiving TH “EStoTH” and the group predicted to receive TH but actually receiving ES “THtoES”. We compute vulnerability scores (VS) as described above for each household in the entire population. Additionally, we estimate the marginal benefit of being assigned TH rather than ES as $MB = \Pr(\text{return to homelessness} \mid \text{ES}) - \Pr(\text{return to homelessness} \mid \text{TH})$. A nonparametric test was used to identify differences between the group actually receiving the discretionary assignment and a random group that would have received the assignment. For example, when considering VS, the test compares the average VS score of the EStoTH group with the distribution of average VS scores of 1000 different groups of the same size (688) sampled at random from the population that was predicted to receive ES in the first place. Similarly, the test compares the average MB score of, say, the THtoES group with the distribution of average MB scores of 1000 different groups of the same size (749) sampled at random from the population that was predicted to receive TH in the first place.

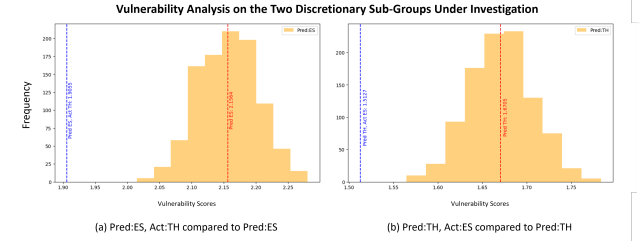


Figure 6: Non-parametric test comparing vulnerability scores between the discretionary subgroups and the population predicted to receive the intervention.

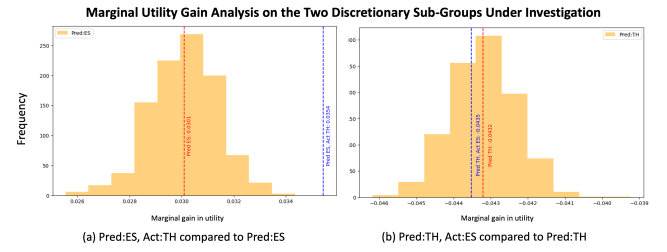


Figure 7: Non-parametric test comparing the marginal gain in utility resulting from the switch in interventions between the discretionary subgroups and the population predicted to receive the intervention.

Vulnerability Figure 6 shows the results of the nonparametric tests for VS. For both the group predicted to receive TH but ultimately assigned ES, and the group predicted to receive ES but ultimately assigned TH, the vulnerability scores of the discretionary subgroups were significantly lower than one would expect if they were randomly sampled from the population of those predicted to receive the (respective) original intervention. The graphs themselves show a histogram (in orange) of the distribution of the 1000 means obtained from the control group. Within the histogram, the vertically dotted line in red represents the “mean of the means” for the control group, while the vertically dotted blue line represents the mean for the actual discretionary group, lying well outside the sampled distribution in both cases. This demonstrates quite clearly that caseworkers are only applying discretion in the context of especially low-vulnerability households.

Marginal Benefit Turning our attention to Figure 7, we are presented with analogous insights derived from the non-parametric tests applied to gauge marginal benefits. A noteworthy distinction surfaces when contrasting the two subgroup analyses. For the EStoTH discretionary group, there is a significant benefit (again, outside the distribution of random samples from those predicted to receive ES), while there is no corresponding significant loss for the THtoES discretionary subgroup. Specifically, the mean value within this subgroup of households nestles comfortably within the 34th percentile spectrum when juxtaposed with the averages computed from the control group. This intriguing discovery suggests that caseworkers, even in scenarios where the im-

perative is to assign households a less intensive intervention, accomplish this task in a manner that circumvents substantial perturbations and maintains a certain level of stability. Impressively, this is achieved while also effectively harnessing discretionary actions to secure tangible gains, steering the transition towards a *more* intensive intervention.

Synthesizing our discoveries within this segment, a coherent narrative emerges. Collectively, our findings shed light on the overarching trend: caseworker discretion, by and large, gravitates towards individuals of comparatively lower vulnerability. Delving deeper, within this subgroup, discretionary measures exhibit a distinct outcome-centric orientation. Caseworkers adeptly wield discretion to enhance outcomes, particularly in the context of the EStoTH transition, all while delicately managing the potential repercussions for the THtoES counterpart—those conceivably requiring re-assignment to respect capacity constraints. Also, note that caseworkers themselves do not have access to any kind of outcome predictions, but still proficiently align their actions with external quantitative machine learning metrics. This alignment, predicated solely on their cumulative experience and judgment, underscores the congruence between human expertise and algorithmic evaluation.

Discussion

The integration of AI tools into various public administration domains, such as child welfare protection, social service delivery, and policing, has given rise to a critical inquiry into the roles that street-level bureaucrats play, both historically and prospectively, in shaping decision-making processes. Amidst the fervor of technological advancements, it is important to understand the intricate dynamics at play between automated systems and the human element inherent to public service interactions. Beyond the mere mechanization of tasks, a pivotal consideration is the form of judgment and discretion exercised by human agents, the “bureaucratic counterfactual” (Johnson and Zhang 2022). In this paper, we show that homelessness service providers were very consistent in their decision-making, alleviating possible concerns about procedural justice. At the same time, their discretionary behavior displays interesting nuances. They typically only target less vulnerable households to make discretionary decisions (choosing the “standard” decision for more vulnerable households). Still, within that, their decisions appear sophisticated and beneficial in maximizing marginal benefit. These results demonstrate consistency and positive applications of judgment. They can also inform how best to target areas of AI tool development in this domain to best assist these human experts in the future.

Acknowledgments

This work was partially supported by the NSF (IIS-1939677, IIS-2127752) and by Amazon through an NSF FAI award. We thank local community partners in helping to conceptualize the challenges facing homeless service delivery.

References

- Azizi, M. J.; Vayanos, P.; Wilder, B.; Rice, E.; and Tambe, M. 2018. Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research: Proceedings of the 15th International Conference*, 35–51. Springer.
- Brown, M.; Cummings, C.; Lyons, J.; Carrión, A.; and Watson, D. P. 2018. Reliability and validity of the Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) in real-world implementation. *Journal of Social Distress and the Homeless*, 27(2): 110–117.
- Consulting, O. 2015. Vulnerability Index-Service Prioritization Decision Assistance Tool (VI-SPDAT) Prescreen Triage Tool for Single Adults.
- Cronley, C. 2022. Invisible intersectionality in measuring vulnerability among individuals experiencing homelessness—critically appraising the VI-SPDAT. *Journal of Social Distress and Homelessness*, 31(1): 23–33.
- Culhane, D. P.; Park, J. M.; and Metraux, S. 2011. The patterns and costs of services use among homeless families. *Journal of Community Psychology*, 39(7): 815–825.
- Das, S. 2022. Local justice and the algorithmic allocation of scarce societal resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12250–12255.
- de Sousa, T.; Andrichik, A.; Cuellar, M.; Marson, J.; Prester, E.; Rush, K.; and Associates, A. 2022. The 2022 Annual Homelessness Assessment Report (AHAR) to Congress Part 1: Point In-Time Estimates of Homelessness. Report, United States Department of Housing and Urban Development, Washington D.C.
- DeLong, E. R.; DeLong, D. M.; and Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3): 837–845.
- Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*.
- Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- Fowler, P. J.; Hovmand, P. S.; Marcal, K. E.; and Das, S. 2019a. Solving Homelessness from a Complex Systems Perspective: Insights for Prevention Responses. *Annual Review of Public Health*, 40(1): 465–486.
- Fowler, P. J.; Wright, K.; Marcal, K. E.; Ballard, E.; and Hovmand, P. S. 2019b. Capability Traps Impeding Homeless Services: A Community-Based System Dynamics Evaluation. *Journal of Social Service Research*, 45(3): 348–359.
- Henry, M.; Mahathay, A.; Morrill, T.; Robinson, A.; Shivji, A.; and Watt, R. 2018. The 2018 Annual Homeless Assessment Report (AHAR) to Congress: Part 1: Point-in-Time Estimates of Homelessness. Technical report, United States Department of Housing and Urban Development, Washington D.C.
- Hogarth, R. M.; and Karelaia, N. 2007. Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, 114(3): 733–758.

- HUD. 2017. Notice CPD-17-01: Notice Establishing Additional Requirements for a Continuum of Care Centralized or Coordinated Assessment System.
- Johnson, R. A.; and Zhang, S. 2022. What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in US Social Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1671–1682.
- Kawakami, A.; Sivaraman, V.; Cheng, H.-F.; Stapleton, L.; Cheng, Y.; Qing, D.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
- Khadduri, J.; Leopold, J.; Sokol, B.; and Spellman, B. 2010. *Costs Associated with First-Time Homelessness for Families and Individuals*. Dept of Housing and Urban Development.
- Kube, A.; Das, S.; and Fowler, P. J. 2019. Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 622–629.
- Kube, A.; Das, S.; Fowler, P. J.; and Vorobeychik, Y. 2022. Just resource allocation? How algorithmic predictions and human notions of justice interact. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 1184–1242.
- Kube, A. R.; Das, S.; and Fowler, P. J. 2023a. Community- and data-driven homelessness prevention and service delivery: Optimizing for equity. *Journal of the American Medical Informatics Association*, 30(6): 1032–1041.
- Kube, A. R.; Das, S.; and Fowler, P. J. 2023b. Fair and Efficient Allocation of Scarce Resources Based on Predicted Outcomes: Implications for Homeless Service Delivery. *Journal of Artificial Intelligence Research*, 76: 1219–1245.
- Kuo, T.-S.; Shen, H.; Geum, J.; Jones, N.; Hong, J. I.; Zhu, H.; and Holstein, K. 2023. Understanding Frontline Workers’ and Unhoused Individuals’ Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23. New York, NY, USA: Association for Computing Machinery.
- Lipsky, M. 2010. *Street-level bureaucracy: dilemmas of the individual in public services*. New York: Russell Sage Foundation, 30th anniversary expanded edition.
- Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 4765–4774. Curran Associates, Inc.
- Mashiat, T.; Gitiaux, X.; Rangwala, H.; Fowler, P.; and Das, S. 2022. Trade-Offs between Group Fairness Metrics in Societal Resource Allocation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1095–1105.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.
- Shinn, M.; Greer, A. L.; Bainbridge, J.; Kwon, J.; and Zuiderveen, S. 2013. Efficient Targeting of Homelessness Prevention Services for Families. *American Journal of Public Health*, 103(S2): S324–S330.
- Shinn, M.; and Richard, M. K. 2022. Allocating homeless services after the withdrawal of the vulnerability index—service prioritization decision assistance tool.
- Souza, V. F.; Cicalese, F.; Laber, E.; and Molinaro, M. 2022. Decision Trees with Short Explainable Rules. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 12365–12379. Curran Associates, Inc.
- Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72.
- United States Interagency Council on Homelessness. 2022. ALL IN: The Federal Strategic Plan to Prevent and End Homelessness. Report, United States Interagency Council on Homelessness, Washington D.C.
- Ustun, B.; and Rudin, C. 2019. Learning Optimized Risk Scores. *Journal of Machine Learning Research*, 20(150): 1–75.
- Vayanos, P.; Ye, Y.; McElfresh, D.; Dickerson, J.; and Rice, E. 2020. Robust active preference elicitation. *arXiv preprint arXiv:2003.01899*.

Supplementary Materials

Appendix A: Data Summary

The dataset includes records on 12,715 households. The target variable here is a one-vs-all prediction label for each intervention, or in general terms what intervention a household receives. We do not include permanent supportive housing in our analysis because it was not assigned as a first-time intervention. A break-down of the number of intervention assignments in the data we analyze is shown below:

Service Type	Number Assigned
Emergency Shelter	4,441
Transitional Housing	2,451
Rapid Re-Housing	846
Prevention	4,977
Total	12,715

Appendix B: The features in the data

A comprehensive description of all the features that we used in the creation of the decision trees is shown in the table below. All categorical variables were one-hot-encoded.

Binary Features	
Feature Name	Description
HUDChronicHomeless	Whether or not the client is chronically homeless.
Gender	In original documentation can take on one of 8 values. But in data is only either Male or Female.
SpousePresent	Variable indicating whether or not client has a spouse.
Categorical Features	
Feature Name	Description
PrimaryRace	The primary race of client. Can be one of 7 categories.
Ethnicity	The ethnicity of client. Can be one of 5 categories.
PriorResidence	The type of residence client had before entering system. Can be one of 25 categories.
VeteranStatus	Variable indicating whether or not client is a veteran. Can take one of 5 values.
DisablingCondition	Variable indicating whether or not client has a disabling condition. Can be one of 3 categories.
ReceivePhysicalDisabilityServices	Variable indicating whether or not client received disability services. Can be one of 5 categories.
HasDevelopmentalDisability	Variable indicating whether or not client has developmental disability. Can be one of 5 categories.
ReceiveDevelopmentalDisabilityServices	Variable indicating whether or not client received services for developmental disability. Can be one of 5 categories.
HasChronicHealthCondition	Variable indicating whether or not client has a chronic health condition. Can be one of 5 categories.
ReceiveChronicHealthServices	Variable indicating whether or not client received services for a chronic health condition. Can be one of 5 categories.
HasHIV/AIDS	Variable indicating whether or not client has HIV/AIDS. Can be one of 5 categories.
ReceiveHIV/AIDSServices	Variable indicating whether or not client received services for HIV/AIDS. Can be one of 5 categories.
HasMentalHealthProblem	Variable indicating whether or not client has a mental health condition. Can be one of 5 categories.
ReceiveMentalHealthServices	Variable indicating whether or not client received services related to mental health. Can be one of 5 categories.
HasSubstanceAbuseProblem	Variable indicating whether or not client has a substance abuse problem. Can be one of 5 categories.
ReceiveSubstanceAbuseServices	Variable indicating whether or not client received services related to substance abuse. Can be one of 5 categories.
DomesticViolenceSurvivor	Variable indicating whether or not client is a survivor of domestic violence. Can be one of 5 categories.
Continuous Features	
Feature Name	Description
Age	The age of the client
Calls	The number of calls made to hotline before entry
Wait	The time (in days) elapsed since first call to hotline to intervention assignment
RatioOfNumCallstoWaitTime	Ratio of number of calls to wait time. Calculated variable to measure 'persistence'
MonthlyAmount	The total income (in dollars) of client per month.
numMembers	The number of members in the client's households.
Children	The number of children in client's household.
Children 0-2	The number of children aged 0 to 2 in client's household.
Children 3-5	The number of children aged 3 to 5 in client's household.
Children 6-10	The number of children aged 6 to 10 in client's household.
Children 11-14	The number of children aged 11 to 14 in client's household.
Children 15-17	The number of children aged 15 to 17 in client's household.
UnrelatedChildren	The number of children in client's household not related to them.
UnrelatedAdults	The number of adults in client's household not related to them.

Appendix C: Creating Vulnerability Scores

The vulnerability scores were calculated using the following table. It was created to replicate the VI-SPDAT as closely as possible given the features that we had in our data.

CRITERIAS	SCORES
Single Person Households	
If single parent with 2 or more children AND/OR a child aged 11 and younger	+1
If NOT single parent with 3 or more children AND/OR a child aged 6 and younger	+1
History of Housing and Homelessness	
If prior residence is anything other than 'emergency shelter' or 'transitional housing for homeless persons', or 'safe haven'	+1
If chronically homeless	+1
Risks	
Used a crisis service i.e., if any one of the following is true: - Received physical disability services - Received developmental disability services - Received chronic health services - Received HIV/AIDS services - Received mental health services - Received substance abuse services	+1
If prior residence is prison	+1
Wellness	
Physical health of client i.e., if any one of the following is true: - Has a chronic health condition - Has HIV/AIDS and Receives HIV/AIDS services - Has physical disability - Has developmental Disability - Has substance abuse problem	+1
Mental health of client i.e., if any one of the following is true: - Has mental health problem - Is domestic violence survivor	+1

Appendix D: Intervention Prediction Using CART

As mentioned in the text, we constructed trees using the CART algorithm in a manner similar to XGBoost. When using CART we performed a grid search based on the following hyper-parameters:

- Measure of impurity: Gini, Entropy
- Max depth : $d \in [1, 10]$
- Minimum samples per split : $s \in [2, 10]$
- Minimum samples in a leaf node: $n \in [1, 5]$

The combination of hyper-parameters that yielded the best results for CART is reported in the table below:

Intervention Type	Criterion	Max Depth	Min Samples in Leaf	Min Samples in Split
Emergency Shelter	Gini	9	4	8
Transitional Housing	Gini	7	4	8
Rapid Re-Housing	Gini	4	3	2
Prevention	Gini	8	1	9

The ROC curves for the trees generated using CART is shown in the figure below:

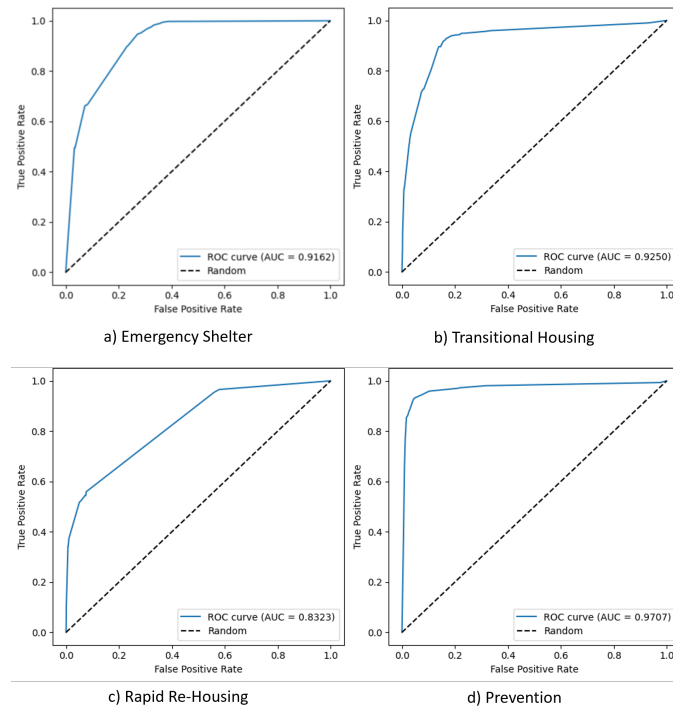


Figure 8: ROC curves for each tree used to learn intervention assignments using the CART algorithm. Predictive discrimination, measured using the area under the ROC curve is very high (similar to XGBoost), further implying significant consistency in caseworker decision-making.

NOTE: Because we use a one-versus-all model for each intervention, it is possible that a single household might be predicted to receive more than one intervention. In the event that this happens, the most 'confident' prediction is taken as the final predicted intervention for the household.