Evaluating Algorithmic Homeless Service Allocation

Wenting Qi[1] and Dr Charalampos Chelmis[1,2*†]

[1]Department of Computer Science, University at Albany, SUNY, Albany, 12222, New York, U.S.A.
[2]Google Scholar profile: https://scholar.google.com/citations?user=-NTZR-kAAAAJ.

*Corresponding author(s). E-mail(s): cchelmis@albany.edu;
Contributing authors: wqi@albany.edu;
†Both authors contributed equally to this work.

Abstract

Improving the homelessness system and evaluating the effectiveness of delivered services are critical to achieve optimal usage of limited social resources as well as to improve the outcomes of the homelessness system. In this context, an increasing number of data science and machine learning methods has been recently applied to the domain of homeless service provision. Given the societal impact of this domain, it is critical to understand the limitations of such methods. However, the performance of algorithmic intervention methods is typically evaluated using abstract metrics that have little meaning for the homeless service allocation domain. We show that domain–agnostic measures are insufficient, and propose a set of new, domain–specific evaluation metrics based on hypothetical, yet realistic "what–if" scenarios. Our empirical analysis demonstrates the value of the proposed measures in understanding the outputs of predictive models and the effect of algorithmic interventions for homeless service provision.

Keywords: complex systems, counterfactual evaluation, fairness, socially important data science

1 Introduction

The general definition of homelessness refers to lack of stable and permanent accommodations, living in shelters or on the street [1–3]. According to a recent global survey of population estimation, more than 1.5% of the global population lacks basic stable and secure accommodations [4]. In the United States alone, homelessness has increased for four consecutive years since 2016, with 580,000 people experiencing homelessness on a single night in 2020 [3].

Factors leading to homelessness are numerous and complex, including but not limited to, poverty, eroding work opportunities, mental illness [5], as well as lack of affordable housing. Communities across the U.S. offer a plethora of homelessness services, many of which are funded by the U.S. Department of Housing and Urban Development. Such services include but are not limited to Emergency shelter, Transitional Housing, Permanent Supporting Housing, Day Shelter, and Street Outreach[1]. Given the scarcity of housing resources and the variety of housing assistance services, it is critical to allocate housing services appropriately and efficiently [6, 7]. Despite risk assessment assistance provided by federal government to help local service assess the eligibility of individuals in need of homelessness service, homeless rates remain high in the United States [7]. Possible reasons include: (i) less available evidence towards homeless characteristics to assist the service provider allocate services, and (ii) inability to assess service matching efficiency based on reducing reentries [7]. With respect to the first concern, Artificial Intelligence (AI) solutions for optimal homeless service allocation have been proposed recently. For example, [8] explored AI's potential to improve the housing system for homeless youth, whereas [7] proposed an optimal service allocation method. [9, 10] explored the feasibility of using Machine Learning (ML) methods to allocate services.

Evaluating the efficiency and fairness of assigned services from algorithmic homelessness service allocation is a critical step to minimize the number of homeless individuals, which is the second concern mentioned above. Current service evaluation methods focus on individual–level or household–level data, including but not limited to, site visits, focus groups, and self–sufficiency assessments [11]. However, such methods require long–term follow up and human resources (e.g., interviewers) [12]. Reentry, a metric widely used in quantitatively evaluating service allocation without the hustle of long–term follow up, refers to individuals experiencing repeated episodes of being homeless [7, 13, 14]. We argue that reentry alone is insufficient to evaluate algorithmic models for homelessness services. Specifically, reentry cannot evaluate systemic fairness (i.e., group–level fairness) as it focuses on individuals. For example, suppose that the reentry rate of certain homeless service (e.g., permanent supporting housing) is 0.01, which means only 1% of homeless people assigned to this service eventually return to homelessness. From the perspective of reentry, the allocation is quite successful (i.e., probability of reentry is low). However, if 90% of the returning individuals (i.e., 1% of homeless people) are female

[1]The abbreviations used in this article are summarized in Table 1.

and assuming an approximately equal ratio of females and males entering the homeless system initially, then the allocation system is biased towards gender despite its low reentry rate. Similarly, accuracy alone is inadequate as an evaluation criterion of the optimality of algorithmic derived policies for homeless service provision. This is because high accuracy largely quantifies the ability of an algorithmic model to learn to replicate the existing allocation system.

In this article, we address the problem of evaluating (and comparing) algorithmic homeless service provision methods along dimensions that go beyond accuracy, and explore in/dependence, accuracy, fairness, and cost. In summary, the main contributions of this article are:

- Proposing novel domain–specific measures to facilitate a fair and meaningful comparison of existing and future algorithmic homeless service provision methods. For completeness, existing measures from the literature are also included.
- The utility of the proposed measures, both for specific features of interest, as well as for arbitrary feature combinations is discussed.
- The usefulness of the proposed measures is demonstrated by evaluating several data science solutions for homeless services allocation in a unique dataset of homeless services administrative records.

The remainder of the article is outlined as follows. Section 2 discusses recent works related to assistance services, allocation systems, and services evaluation methods. Section 3 describes the notation and problem statement. Section 4 introduces the proposed domain–specfic evaluation metrics. Section 5 describes recent algorithmic methods for services provision. Section 6 summaries the experimental setup, whereas, Section 7 presents the experimental results. Finally, Section 8 concludes this article with key takeaways, limitations and future works ideas.

2 Related Work

Recently, machine learning methods have been employed in human–related decision making domains, raising algorithmic fairness concerns. For instance, Asplund et al. [15] showed that sock–puppet browsing is biased to a specific group of users in online housing markets. Public employment services (PES) leverage AI–based methods to assign limited resources to "vulnerable" job seekers [16]. However, in contrast to traditional policy–based manual assignments, AI–based methods can be discriminatory because of correlations between features, even if sensitive features are themselves excluded from the training process. An algorithmic model used in predicting recidivism has been shown to be biased against Afro–American [17]. Reasons for algorithmic decision–making models to introduce biases in the decision making process include biases inside the training set [18] (e.g., missing data, data imbalance, erroneous data), as well as naive use of application–agnostic evaluation metrics (e.g., accuracy), which although emphasize on the predictive power of machine

Table 1 Abbreviations and their corresponding description.

Abbreviation	Description
CoC	Continuum of Care
ES	Emergency Shelter
HMIS	Homeless Management Information System
HUD	U.S. Department of Housing and Urban Development
PSH	Permanent Supportive Housing
RRH	Rapid Rehousing
HP	Homeless Prevention
SNAP	Food Stamp Program
SFaMP	Staying Family Member's Place
SFrMP	Staying Friend Member's Place
PNH	Place Not for Habitation
RCS	Rental by Client with Subsidy
OCS	Owned by Client with Subsidy
SSI	Supplemental Security Income
MI	Monthly Income
SSDI	Social Security Disability Insurance
TH	Transitional Housing
LS	Living Situation
THPTY	Times Homeless Past Three Years
DC	Disabling Condition
PD	Physical Disability
AI	Artificial Intelligence
ML	Machine Learning

learning models cannot be used to evaluate their real–life prediction outcomes. The main focus of this article is how to better evaluate algorithmic decision making models.

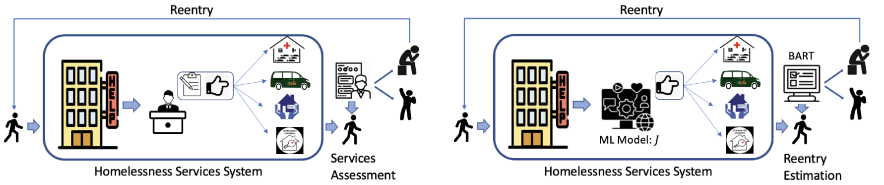
Specific to homeless services provision, several works have applied machine learning related methods, including but not limited to, allocating homelessness service [19], assessing the impact of homelessness service allocation with respect to reentries [14] and prioritizing services allocation based on risk assessment or optimizing allocation based on algorithmic matching outcomes [20]. Specifically, Gurobi optimization has been used to predict bed occupation in a shelter for any given night by tracing the individual trajectories of getting into and out of the shelter [7]. Random Forest, Decision Tree, and Logistic Regression models have been used recently for reentry prediction, with Random Forest achieving the best performance [9]. [8] used the Next Step Tool (NST)[2] to train logistic regression and decision trees for predicting youth's homelessness status after receiving housing assistance based on the youth background and current living states. In practise, homeless services rely primarily on manual evaluation [7]. We consider state–of–art methods for services allocation (i.e., Random Forest [9] and Gurobi [7]), as well as other popular algorithmic models (e.g., Adaboost and K Nearest Neighbors).

[2]NST is a set of multiple–choice and frequency–type questions, which are designed to measure the vulnerability of youth based on their previous history (e.g., socialization, daily function, homelessness experience).

Performance evaluation of automated service allocation system often involves a single dimension (i.e., similarity between automatic allocation and the actual manual allocation results), which although may be used to compare against baselines, may not necessarily probe all aspects of a method's performance in this socially important domain. [21] used domain–agnostic evaluation metrics, including precision, recall, F1–score, and accuracy to evaluate the learning process itself rather than the application of the model. However, high evaluation scores of such domain–agnostic evaluation metrics (e.g., accuracy, precision) only prove that the automatic allocation system can efficiently mimic the manual allocation process, rather than justify the effectiveness of the automatic allocation system itself. To evaluate the potential impact of service allocation, [7] used reentry to measure the effectiveness of the delivered services, and quantify the scale changing degree of three–groups, namely, those that were harmed or benefited by automated service allocation and those that remained unaffected by it. However, [7] only presented statistical results such as the total number of benefited individuals. Deeper relationships between the group compared to the overall population is unclear. In addition, [7] ignores the fact that we wish to maximize the number of individuals non–reentering the homelessness system, while minimizing the number of individuals that do reenter. We propose domain–specific evaluation metrics to quantify the allocation results and separate unaffected groups into two subgroups based on reentry and non–reentry, then evaluate them separately.

3 Algorithmic Homeless Service Allocation

Consider a collection of N records $O_i = (x_i, y_i, r_i), 1 \leq i \leq N$, where x_i is represented by feature vector $x_i = [x_{i1}, ..., x_{iM}]^T$, and M denotes the total number of features. y_i represents the allocated service in reality and $r_i \in \{0, 1\}$ with $r_i = 1$ denoting that the individual has entered the homeless service system multiple times, otherwise entered only once. For consistency with prior art and fair comparison between candidate algorithmic methods, each record is additionally associated with a label of four homeless assistance service, namely, ES, RRH, TH and HP. The service proposed by algorithmic model J is denoted as y'_i, and its corresponding reentry outcome is denoted as r'_i. The anticipated reentry outcome r'_i is estimated by a counterfactual model $\mathcal{B}$. Existing counterfactual models (e.g., DICE [22], AR[23], and CEM[24]) can be used to provide counterfactual feature vectors with respect to certain algorithmic model output, but are unsuitable for the task we explore in this study. Because in such counterfactual methods, the counterfactual feature vector involves all the features (i.e., all M features can be perturbed). However, in our problem setting, the only allowed to perturb "feature" is service (i.e., y_i), and x_i remains unchanged in the counterfactual model. For this reason, we use BART (Bayesian Additive Regression Tree) [25] to predict reentry using counterfactual allocations, as it has been shown to provide coherent probabilistic estimation of heterogeneous treatment effects [7].



(a) (b)

Fig. 1 Manual service allocation process (a), and (b) algorithmic service allocation. BART is used to evaluate the anticipated effectiveness of allocated service.

The overall tasks are: (i) learning an automated allocation model, J, to recommend the most reasonable service for an individual with a previously unseen record O_{N+1}, represented by the feature vector x_{N+1} (i.e., $y'_{N+1} = J(x_{N+1})$); (ii) building and fitting a BART model, $\mathcal{B}$, with actual service; (iii) using $\mathcal{B}$ to estimate r'_i for each individual i assigned to service y'_i by J (i.e., $r'_i = \mathcal{B}(x_i, y'_i)$). Figure 1(a) shows the typical process an individual undergoes in their journey to fulfil their need. A case worker either assigns the individual to a service or refers her to service provider(s) as needed. When an algorithm is used to automate the process (Fig 1(b)), a prediction is made to guide the individual to a service provider. The algorithmic model service allocation process is further evaluated by $\mathcal{B}$.

4 Evaluating Algorithmic Models for Homeless Service Allocation

4.1 Performance Dimensions

Accuracy. In supervised learning classification tasks, the ultimate goal is outputting a correct prediction. In contrast, learning effective algorithmic models to allocate homeless assistance services is more complex. Specifically, there is no ground–truth to quantify whether a prediction result is correct, as with different services, the outcome for a given individual may vary. That is the reason why domain–agnostic evaluation metrics are unsuitable (details in Section 4.2), and a counterfactual model must be used in the evaluation process.

Fairness. Even if an automated allocation system is "perfect" in all aspects, being biased against a certain group is unacceptable [26]. Multiple metrics have been proposed already to quantify the fairness of algorithmic decision–making systems [16], including but not limited to, *Equalized Odds* [27], *Equal Opportunity* [27], and *Demographic Parity* [28, 29]. Equalized Odds evaluates whether the protected and unprotected groups have equal rates for true positives and false positives [27]. Equal Opportunity evaluates whether the protected and unprotected groups should have equal true positive rates [27]. Demographic Parity measures the likelihood of a positive outcome which should be the same regardless of whether the person is in the protected (e.g., female) group [29]. These evaluation metrics focus on a particular protected group (e.g., female,

elder) and mainly evaluate whether the algorithmic decision–making system is biased or harmful to that protected group.

Efficiency. Beyond fairness, the efficiency of an algorithmic decision–making system for homeless service allocation is often measured in its ability to reduce their chance of re–entering the homelessness system in the future, after receiving services. Unfortunately, application independent fairness evaluation metrics cannot quantify allocation efficiency. The evaluation metrics proposed in this work are able to measure both fairness and allocation efficiency, as discussed in Section 4.3.1.

Cost. Developing an algorithmic decision–making system for homeless service allocation has an associated cost, measurable in terms of effort and time for data collection, as well as computational power consumed for model training. Improving the accuracy of an algorithmic system may be impractical for example if more training data cannot be collected[3], or if the quality of (often self–reported) data is questionable[4]. Similarly, compute costs may be considerable if third–party compute resources (e.g., Google Cloud) are used, even if a trained model can be reused over time. Modern deep neural network architectures, for instance, are notorious for their large carbon footprint [31, 32]. At the same time, maintaining and updating trained algorithmic models requires a dedicated trained computer or data scientist, a resource which is often unavailable to homeless serving organizations.

4.2 Application Independent Metrics

Widely used evaluation metrics include but are limited to accuracy, precision, recall, and F1–score [33]. Specifically, the predicted labels of the learning model fall into one of four categories, namely, true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Accuracy is defined as the percentage of correctly classified data over the total number of data, and is calculated by Equation 1. Precision (see Equation 2) refers to the number of data classified correctly within a specific label over the total number of data classified to that label. Recall points to the number of data classified correctly within a specific label over the total number of data that belongs to that label, and is calculated by Equation 3. The F1–score combines recall and precision into a single metric as shown in Equation 4.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

[3] Collecting necessary data for training an algorithmic model requires human effort.

[4] Even though methods to address the problem of training algorithmic decision–making systems in the presence of untrustworthy training data has recently been explored (e.g., [30]), it remains a challenging open problem.

Table 2 Selected features, grouped into four categories.

Category	Feature	Feature Values	Explanation
Basic Information	Age	Age_0-20	Age between 0 to 20
		Age_20-40	Age between 20 to 40
		Age_40-60	Age between 20 to 40
		Age_60_up	Age elder than 60
	Race	Asian	
		White	
		AI&AN	American Indian or Alaska Native
		B&AA	Black or African American
		NH&PI	Native Hawaiian or Other Pacific Islander
		Race_other	
	Gender	Female	
		Male	
		Gender_other	
Living Situation	Times Homeless Past Three Years	THTPY_1	Once
		THTPY_2	Twice
		THTPY_3	Three times
		THTPY_4	Four times or more
		THTPY_other	
	Living Situation	LS_Friend	Staying or living in a friend's place
		LS_Family	Staying or living in a family's place
		LS_Jail	Jail, prison or juvenile detention facility
		LS_ES	Emergency shelter, including hotel or motel paid for with emergency shelter voucher
		LS_NH	Place not meant for habitation
		LS_Rental	Rental by client, with other ongoing housing subsidy
		LS_Owned	Owned by client, with ongoing housing subsidy
		LS_other	
Financial Situation	Monthly Income	MI_None	Monthly income is $0
		MI_1000	Monthly income less than $1,000
		MI_1000_2000	Monthly income between$1,000 to $2,000
		MI_2000	Monthly income more than $2,000
	Earned	Earned_No	No employment Income
		Earned_Yes	Has employment Income
Health Situation	Disability	Disability_No	
		Disability_Yes	
		Disability_other	
	Physical Disability	PD_No	
		PD_Yes	

$$F1\_score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

Such domain–agnostic evaluation metrics are not suitable for homelessness services allocation evaluation, as they can not quantify effectiveness or fairness. Effectiveness means whether the delivered service can assist people in getting out of homelessness or improving their overall quality of life. Fairness refers to the ability of an allocation scheme to operate without incurring bias towards different groups (e.g., female vs male, elderly vs younger).

4.3 Application Specific Metrics

Fairness is a critical issue, particularly when involving algorithmic models in the decision–making process [34, 35]. Provided with biased data algorithmic model can learn to reproduce systemic biases with potentially detrimental effects [36]. For example, algorithms used for predicting recidivism has a much higher false–positive rate for black people than white people [37]. In another scenario, algorithmic model used to automatically rank job candidates has shown to be biased against female [26]. To evaluate possible algorithmic biases induced by algorithmic models [38] as well as biases in the data used to train them, domain–specific metrics are necessary.

As mentioned in Section 1, reentry is widely adopted as a criterion to reflect the effectiveness of the delivered homeless assistance services [7, 14, 39]. Specifically, if the homeless people experience repeated episodes of homelessness (i.e., reentry), the initial or previous delivered assistance services are not considered optimal or effective. In our study, we follow the same idea to quantify the effectiveness of allocated services. Besides, even with a desirable outcome, how do we know the current delivered service is optimal? In other words, what would the outcome be if an individual was pretended with an alternate outcome of assistance service? To compare the current allocated service with other services, we rely on counterfactuals [7]. Specifically, we leverage BART as a counterfactual model to predict reentry. BART is a variant of the Bayesian regression algorithm, which is based on a "sum of trees" model. Each tree is restrained by regularization prior, and BART draws samples from the posterior distribution by the Bayesian back–fitting markov chain monte carlo (MCMC) algorithm [40]. In our study, BART generates posteriors for each individual in the dataset, allowing precise inference for both population–wide level and individual–specific levels. In summary, we use a counterfactual model to predict reentry with allocated service by algorithmic models and compare with the reentry outcome of the actual delivered service to see whether the delivered service is an optimal choice.

4.3.1 Fairness Consideration and Evaluation Metrics

To ensure that allocation outcomes do not disproportionately harm people with certain sensitive characteristics (e.g., age, gender), and inspired by [7], we compare the algorithmic allocated reentry outcome r' with the actual reentry r, across three groups, defined as follows:

- **Group of individuals that benefited (G_b) from algorithmic service allocation:** the set of people predicted not to return to homelessness after being assigned to a service by model J, even though in reality they reenter (i.e., $r = 1, r' = 0$).
- **Group of individuals that were harmed (G_h) from algorithmic service allocation:** the set of people predicted to return to homelessness after being assigned to a service by model J, even though in reality they not reenter (i.e., $r = 0, r' = 1$).

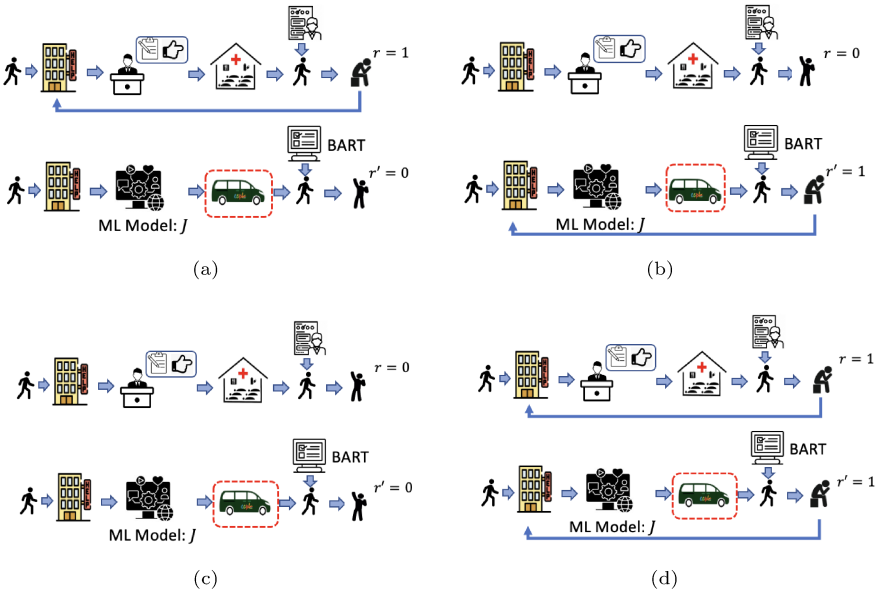


Fig. 2 (a) Group of individuals that benefited G_b; (b) Group of individuals that were harmed G_h; (c) Group of individuals that remained unaffected without reentry G_{u_0}; (d) Group of individuals that remained unaffected with reentry G_{u_1}. In all illustrative scenarios a different service allocation (i.e., Street Outreach as opposed to Emergency Shelter) is recommended by the algorithmic model. Accuracy metrics would consider this a mismatch and count it as error. In contrast, counterfactual evaluation (i.e., BART) can provide insights into the effectiveness (or lack thereof) of the recommended service allocation (even if considered erroneous by ground-truth data).

- **Group of individuals that remained unaffected (G_u) from algorithmic service allocation:** people having the same outcome with algorithmic allocated assistance service compared with the actual outcome (i.e., $r = r'$). Different with [7], we further divide G_u into two subgroups, namely G_{u_1} (i.e., $r = r' = 1$) and G_{u_0} (i.e., $r = r' = 0$) for those individual that reenter and don't reenter the homelessness system, as shown in Figs 2(d) and 2(c), accordingly.

Intuitively, effective service allocation system should maximize G_b and minimize G_h simultaneously. However, the ability of an algorithmic service allocation system to remain unbiased towards different groups of people is more critical. For example, Figure 3 shows two evaluation results of algorithmic model J_A and J_B with respect to G_b and G_h, respectively. Despite the group size of G_b in Fig 3(a) being larger than Fig 3(b), the proportion of female and male is uneven in G_b and G_h. Therefore, service allocation system J_A is biased towards gender, and therefore J_B is better than J_A. The question then is how to define subgroup to quantify fairness. Besides the commonly acknowledged sensitive groups such as age, gender, and race [7], we include important features identified by performing feature importance, as shown in Fig 4 and

(a) (b)

Fig. 3 Evaluation results of model J_A (a) and J_B (b). Different colors denote different groups of gender (i.e., red denotes female, and green denotes male).

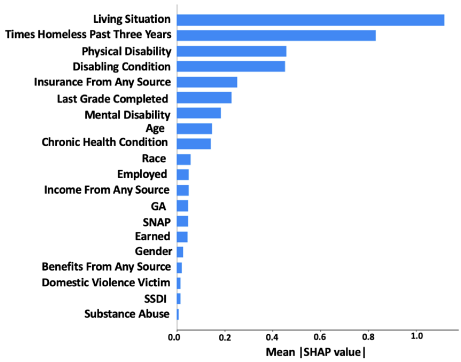


Fig. 4 Feature importance computed by SHAP for XGBoost.

Table 2. Specifically, G_b, G_h, and $G_u(G_{u_0}$,$G_{u_1})$ are computed for each of the features shown in Table 2 as follows.

$$G_b(f_k^m) = \frac{\sum_{i=1}^{N} \mathbb{1}(r_i = 1)\mathbb{1}(r_i' = 0)\mathbb{1}(x_{im} = f_k^m)}{\sum_{i=1}^{N} \mathbb{1}(x_{im} = f_k^m)}, \tag{5}$$

$$G_h(f_k^m) = \frac{\sum_{i=1}^{N} \mathbb{1}(r_i = 0)\mathbb{1}(r_i' = 1)\mathbb{1}(x_{im} = f_k^m)}{\sum_{i=1}^{N} \mathbb{1}(x_{im} = f_k^m)}, \tag{6}$$

$$\begin{aligned} G_u(f_k^m) &= G_{u_0}(f_k^m) + G_{u_1}(f_k^m) \\ &= \frac{\sum_{i=1}^{N} \mathbb{1}(x_{im} = f_k^m)(\mathbb{1}(r_i = r_i' = 1) + \mathbb{1}(r_i = r_i' = 0))}{\sum_{i=1}^{N} \mathbb{1}(x_{im} = f_k^m)}, \end{aligned} \tag{7}$$

where $\mathbb{1}$ is the indicator function. The feature is denoted as f^m, where $m \in \{1, 2, 3...M\}$, and feature value is denoted as $f_k^m \in \{f_1^m, f_2^m, f_3^m ... f_K^m\}$, where K is the total number of distinct values for a certain feature f^m, and may vary across features. In Equation 5, denominator $\sum_{i=1}^{N} \mathbb{1}(x_{im} = f_k^m)$ counts the total number of individuals whose feature value of f^m is k. The numerator in Equation 5 counts the individuals in G_b with certain feature value k of f^m. Equation 6 follows the same idea for G_h. Equation 7 consists of two parts, $G_{u_0}(f_k^m)$ and $G_{u_1}(f_k^m)$ that capture the number of individuals, who reenter

and do not reenter the homeless system respectively with certain feature value of f^m.

Intuitively, we wish algorithmic models to avoid favoring certain populations (i.e., G_b) or hurting others (i.e., G_h). Therefore, we define the following domain–specific metrics:

$$\Delta N(f_k^m) = G_b(f_k^m) - G_h(f_k^m), \tag{8}$$

$$\Delta T(f_k^m) = \Delta N(f_k^m) \frac{\sum_{i=1}^{N} \mathbb{1}(x_{im} = f_k^m)}{N}, \tag{9}$$

Specifically, Equation 8 measures the difference in group size between G_b and G_h for a certain feature value f_k^m. Thus, $\Delta N(f_k^m)$ measures the relative beneficial degree on group–population ($G_{f_k^m}$) (i.e., $\forall x_i \in G_{f_k^m}, x_{im} = f_k^m$). Equation 9 normalizes $\Delta N(f_k^m)$ with the percent of the individuals whose feature value f^m is k. Therefore, $\Delta T(f_k^m)$ quantifies the relative beneficial degree (i.e., the difference between G_b and G_h) on overall population. Note that $\Delta N(f_k^m)$ and $\Delta T(f_k^m)$ always have the same sign. There are several combinations which require further interpretation. Figure 5 provides illustrative examples for combinations of $\Delta N(f_k^m)$ and $\Delta T(f_k^m)$. Note that for illustration purposes, f_k^m denotes the feature of gender with certain feature value of female (assuming N equals to total number of females).

- When both $\Delta N(f_k^m) > 0$ and $\Delta T(f_k^m) > 0$, $G_b(f_k^m)$ is larger than $G_h(f_k^m)$. This is desirable as relatively more people are benefited from decisions made by model J. However, two scenarios need further consideration:
  - $N(f_k^m)$ is large but $T(f_k^m)$ is small. This means that even with high relative benefit within a certain feature value, the benefited people comprise only a small fraction of the overall population. Figure 5(a) shows this case, where the total number of females is 18, and the total number of G_b and G_h are 6 and 3, respectively. Therefore, $\Delta N(f_k^m) = 6 - 1 = 5$, $\Delta T(f_k^m) = 5/18 = 0.27$.
  - $N(f_k^m)$ is small but $T(f_k^m)$ is large. In this case, even with small relative benefit within a certain feature value, the portion of benefited people take a large proportion of the overall population. Figure 5(b) shows this case, where the total number of females is 9, and the total number of G_b and G_h are 6 and 3, respectively. Therefore, $\Delta N(f_k^m) = 6 - 3 = 3$, $\Delta T(f_k^m) = 3/9 = 0.33$.
- When $\Delta N(f_k^m) < 0$ and $\Delta T(f_k^m) < 0$, $G_b(f_k^m)$ is smaller than $G_h(f_k^m)$. This is undesirable, as relatively more people are harmed by service allocations made by model J. Figure 5(c) shows this case, where the total number of females is 9, and the total number of G_b and G_h are 3 and 6, respectively. Therefore, $\Delta N(f_k^m) = 3 - 6 = -3$, $\Delta T(f_k^m) = -3/9 = -0.33$.
- When $\Delta N(f_k^m) = 0$ and $\Delta T(f_k^m) = 0$, $G_b(f_k^m)$ is equal to $G_h(f_k^m)$. In this case, the effectiveness of the service allocation system is neither improved nor impaired compared with the reality status. Figure 5(d) shows this case,

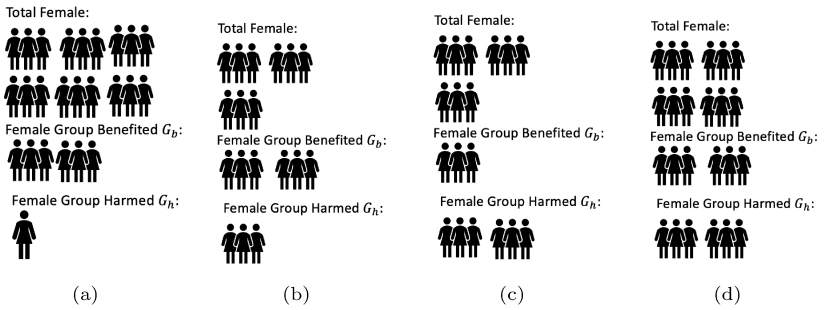


Fig. 5 (a) $\Delta N(f_k^m) > 0$ (large) and $\Delta T(f_k^m) > 0$ (small); (b) $\Delta N(f_k^m) > 0$ (small) and $\Delta T(f_k^m) > 0$ (large);(c) $\Delta N(f_k^m) < 0$ and $\Delta T(f_k^m) < 0$; (d) $\Delta N(f_k^m) = 0$ and $\Delta T(f_k^m) = 0$.

where the total number of females is 12, and the total number of G_b and G_h are 6 and 6, respectively. Therefore, $\Delta N(f_k^m) = 6 - 6 = 0$, $\Delta T(f_k^m) = 0/12 = 0$.

Finally, $\Delta U(f_k^m)$ (defined in Equation 10 below) measures the transition rate from reentry to non–reentry by calculating the difference between $\Delta U_0(f_k^m)$ and $\Delta U_1(f_k^m)$.

$$\Delta U(f_k^m) = \Delta U_0(f_k^m) - \Delta U_1(f_k^m), \tag{10}$$

where

$$\Delta U_0(f_k^m) = \frac{G_{u_0}(f_k^m)}{G_{u_0}(f_k^m) + G_h(f_k^m)} - \frac{G_h(f_k^m)}{G_{u_0}(f_k^m) + G_h(f_k^m)}, \tag{11}$$

and

$$\Delta U_1(f_k^m) = \frac{G_{u_1}(f_k^m)}{G_{u_1}(f_k^m) + G_b(f_k^m)} - \frac{G_b(f_k^m)}{G_{u_1}(f_k^m) + G_b(f_k^m)}. \tag{12}$$

Considering the actual reentry outcomes, the total number of actual non–reentry people (i.e., $r = 0$) should be the sum of the group of unchanged (i.e., G_{u_0}) with non–reentry and the group of harmed (i.e., G_h) whose actual outcome is non–reentry (i.e., $G_{u_0}(f_k^m) + G_h(f_k^m)$). For group of individuals with $r = 0$, $\Delta U_0(f_k^m)$ captures the difference of relative proportion for unchanged non–reentry individuals compared with harmed individuals with allocated service by algorithmic models. Ideally, we want less people whose actual outcome is non–reentry to change their outcome by algorithmic allocated service. This means the larger $\Delta U_0(f_k^m)$ is, the better the result is. $\Delta U_1(f_k^m)$ (shown in Equation 12) follows the same interpretation, which is the smaller $\Delta U_1(f_k^m)$ is, the better the result is because we want more benefited people. Therefore, $\Delta U(f_k^m)$ which calculates the difference between $\Delta U_0(f_k^m)$ and $\Delta U_1(f_k^m)$ reflects the transition ability. In other words, $\Delta U(f_k^m)$ evaluates the beneficial ability of the allocation system.

Table 3 Summary of the proposed evaluation metrics. G_b denotes group of people that is benefited by an algorithmic intervention method. G_h denotes group of people that is harmed. G_u denotes the group of people that were neither negatively nor positively impacted. f_k^m denotes feature m with value k.

Evaluation metrics	Definition	Brief Explanation	Ranges	Advantages	Disadvantages
ΔN	$\Delta N(f_k^m) = G_b(f_k^m) - G_h(f_k^m)$	Measures the difference in group size between G_b and G_h for f_k^m	$(-\infty, +\infty)$	Showing the relative improvement	Cannot show the detailed value of $G_b(f_k^m)$ and $G_h(f_k^m)$
ΔT	$\Delta T(f_k^m) = \Delta N(f_k^m) \frac{\sum_{i=1}^{N} 1(x_{im}=f_k^m)}{N}$	Quantifies the relative beneficial degree on overall population	$(-\infty, +\infty)$	Showing the relative difference on overall population	Cannot show the detailed value of $G_b(f_k^m)$ and $G_h(f_k^m)$
ΔU	$\Delta U(f_k^m) = \frac{G_{u_0}(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} - \frac{G_h(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} - \frac{G_{u_1}(f_k^m)}{G_{u_1}(f_k^m)+G_b(f_k^m)} + \frac{G_b(f_k^m)}{G_{u_1}(f_k^m)+G_b(f_k^m)}$	Measures the transition rate from reentry to non–reentry	$[-2, 2]$	Showing relative difference to (non)reentry in G_u	Cannot show the detailed value of reentry and non–reentry in G_u

Note that $\Delta U(f_k^m) \in [-2, 2]$. To explain why, we consider the range of $\Delta U_0(f_k^m)$ and $\Delta U_1(f_k^m)$ (each one separately), as they are the only two factors that contribute to the calculation of $\Delta U(f_k^m)$ according to Equation 10. We know that $\Delta U_0(f_k^m) = \frac{G_{u_0}(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} - \frac{G_h(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)}$. The lower bound for $\Delta U_0(f_k^m)$ is -1, becasue $\frac{G_{u_0}(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} = 0$ and $\frac{G_h(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} = 1$. Thus, $\Delta U_0(f_k^m) = 0 - 1 = -1$. This extreme case means $G_{u_0}(f_k^m) = 0$, and $G_{u_0}(f_k^m) + G_h(f_k^m) = G_h(f_k^m)$. The upper bound for $\Delta U_0(f_k^m)$ is 1, becasue $\frac{G_{u_0}(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} = 1$ and $\frac{G_h(f_k^m)}{G_{u_0}(f_k^m)+G_h(f_k^m)} = 0$. Thus, $\Delta U_0(f_k^m) = 1 - 0 = 1$. This extreme case means $G_h(f_k^m) = 0$, and $G_{u_0}(f_k^m) + G_h(f_k^m) = G_{u_0}(f_k^m)$. Therefore, the range for $\Delta U_0(f_k^m)$ is $[-1, 1]$. $\Delta U_1(f_k^m)$ follows the similar idea. The lower bound for ΔU is the lower bound of $\Delta U_0(f_k^m)$ subtracting the upper bound of $\Delta U_1(f_k^m)$, which is $-1 - 1 = -2$. The upper bound for ΔU is the upper bound of $\Delta U_0(f_k^m)$ subtracting the lower bound of $\Delta U_1(f_k^m)$, which is $1 - (-1) = 2$. Therefore, the range for ΔU is $[-2, 2]$.

Table 3 summarizes the proposed evaluation metrics. Apart from being applicable to specific features independently, the proposed evaluation metrics can be used to evaluate fairness when considering aggregate features. For instance, one may be interested in evaluating the performance of an algorithmic intervention model with respect to fairness specifically to "back females whose age elder than 60". Naively, one could perform such analysis by examining the model's fairness with respect to race, sex, and age independently. The benefit of aggregating features, however, is that one can get a single measure of fairness even if multiple dimensions (i.e., features) are under investigation. In the particular scenario of back females that are older than 60, we begin by extracting the features (i.e., for illustration propose, we list a small example here and assume the total number of individual is 15). Let m^1 (i.e., Female: 5; Male: 10), m^2 (i.e., Asian: 4; White: 6; Black: 5), and m^3 (i.e., Age $0-60$: 9; Age > 60: 6) denote gender, race, and age, accordingly. Then, the corresponding features values are k^{m^1} (i.e., Female), k^{m^2} (i.e., Black), and k^{m^2} (i.e., Age

> 60). To get the final evaluation results, we organize the selected features and corresponding feature values as $f^{(m^1,m^2,m^3)}_{(k^{m^1},k^{m^3},k^{m^3})}$ (i.e., select the individual whose being female, black, and elder than 60, and the total number of such individuals is 2) which is used to substitute f^m_k (i.e., select the individual based on single feature such as female, the the total number of female is 5) in all proposed evaluation metrics.

4.3.2 Cost

The cost for training and predicting using a computationally expensive algorithmic model can be significant, particularly when a third party computational resource is used to host the model. The *computation cost* of an algorithmic model can be quantified as the CPU/GPU time required to train a model, and the time required to make predictions using the model. Perhaps more important than the computation cost may be the effort required to acquire and assemble training data. To quantify data collection effort, we propose a measure of *data cost (DC)* defined for a particular training and prediction duration in terms of the number, n_s of static features (e.g., gender) that require a onetime collection effort, and the number, n_d, of dynamic features (e.g., income) that need periodic acquisition. Specifically, $DC = \sum_{i=1}^{n_s} s_i + \sum_{i=1}^{n_d} d_i$, where s_i and d_i are the counts of the unique values for feature s_i and d_i respectively.

5 Candidate solutions

Three well–known algorithmic models (K Nearest Neighbors, Random Forest, and AdaBoost) and the Gurobi optimation method proposed in [7] are used in this section to illustrate the usefulness of the proposed metrics.

- **K Nearest Neighbors (KNN):** KNN is a simplistic algorithmic model commonly used for classification without any knowledge of the underlying domain [41]. In our study, we use KNN to output an assistance service by examining the allocation of individuals that are most similar to the one at hand. We use 5–fold cross–validation on the training set to select the reasonable value of the number of neighbors.
- **Random Forest (RF):** RF is constructed by a set of decision trees with random subsets of features. The label is decided by the most votes [42]. We use 5–fold cross–validation to chose the number of trees.
- **AdaBoost (AB):** Boosting has been a popular technique for two–class classification, with multiple proposed variants [43]. We use the Stagewise Additive variant with exponential loss function (SAMME) [44], which experimentally proved its superiority compared with other boosting variants [44]. We use decision tree as a weak classifier and perform 5–fold cross–validation to choose maximum tree depth of weak classifiers.

- **Gurobi:** [7] used Gurobi optimization to find optimal biweekly allocations for homeless individuals given aggregate capacity constriants (e.g., unavailability of beds in a shelter).

6 Experimental Setup

6.1 Dataset

For evaluation purposes, we use a dataset of 50,469 records, corresponding to 38,954 individuals (i.e., each individual may have multiple records) who seek homelessness assistance in the New York Capital Region. Each record comprises individual–level characteristics and allocated services (e.g., Emergency Shelter, Homelessness Prevention, Rapid Rehousing, and Transitional Housing). A complete description of the data elements is available at [45]. The characteristics of each individual are collected and extracted from household relations, education background, living situation, health, and employment situation. The name and Social Security Number of enrolled people in the dataset have been double hashed to protect their privacy. Note that reentry is defined as returning to the homeless service system after previously exiting the system.

6.1.1 Feature Selection

To ensure that only informative features are used to train model J, we perform feature selection before feeding the dataset into training models. Initially, a total of 174 features are available in the dataset. We first removed features such as "Date Created", "Date Deletect", "ClientID", "ExportID", "FirstName", and "NameSuffix", which are irrelevant from a machine learning model perspective. In the next step, we extract *implicit* but important features (e.g., we compute age from date of birth). We subsequently remove uninformative features by performing feature selection. Specifically, we remove features whose fraction of missing values is larger than 60% of selected records. Such features include "WorldWarII" (i.e., whether attend WorldWarII), "VietnamWar" (i.e., whether attend VietnamWar), "DesertStorm" (i.e., whether experience DesertStorm). We additionally remove features with zero variance (i.e., the value is 1 for all selected data instances), such as "NameDataQuality". At this point, the number of features is 34.

To quantify feature importance, we leverage the SHAP [46] package, which uses Shapley values to estimate how each feature contributes to the prediction of a machine learning model [47]. Specifically, a feature importance score is estimated based on the boosted trees that are constructed from the features in the dataset [48]. Therefore, features' importance scores depend on the dataset itself. The higher the SHAP value is, the more important the feature is. In our analysis, we drop features whose importance score is lower than 0.01; such features have very little effect in the model's outcome. The final list of selected features is shown in Figure 4, and summarized in Table 4.

Table 4 Summary of features' type and corresponding number.

Type	Number	Examples
Binary Features	5	Physical Disability, Disabling Condition, Mental Disability
Categorical Features	14	Race, Living Situation, Gender
Continuous Features	1	Age

Table 5 Reentry statistics with respect to train and test sets.

Data	Number of Records	Number of Reentry	Number of Non–Reentry
Train	3499	1049 (29.97%)	2450 (70.03%)
Test	1,167	369 (31.67%)	798 (68.33%)

6.1.2 Data Preparation for BART Experiment

In our experiments, we focus on the subset of those individuals in our dataset that received services after exiting the system (i.e., *reentering* more than once) [7, 13, 14]. This definition includes individuals that for example used an "Emergency Shelter" more than once, or an emergency shelter before being assigned to a "Permanent Supportive Housing" program. The overall number of records referring to such individuals is $38,954$. We additionally focus on records corresponding to "head of household" individuals, since records belonging to dependents and/or spouses often lack socioeconomic, employment and education data. This down selection results in a subset of $24,117$ records.

At the same time, to ensure fair comparison with methods listed in Section 5, we focus on records for four homeless programs, namely Emergency Shelter, Day Shelter, Homelessness Prevention, Rapid Re-Housing. This further reduces the number of relevant records to $18,952$ (i.e., $\sim$ 19k records). We denote this dataset as D_{ori}, and use it to train and test the counterfactual model BART.

6.1.3 Data Preparation for Algorithmic Models Experiment

For a fair experimental evaluation of the Gurobi baseline [7], we tried to replicate, to the extend possible, the setting used by that work. Specifically, the HMIS data reported in [7] was limited in scale (only $7,474$ records in total), and had a reported reentry rate of 43.03% (as opposed to 21.92% in our $18,952$ records dataset). To match that rate, we selected all records between 2013 to 2015 ($4,666$ in total), among which, $1,418$ have entered the homeless system more than once, leading to a reentry rate of 30.3%. We denote this dataset as as D_{ml}, and randomly pre–split it into a training and testing set with a ratio of $3:1$. Table 5 presents the statistics.

6.1.4 Data Preparation for Data Imbalance Experiment

The impact of data imbalance on ML and AI models is well documented, however addressing it remains an open research question [49]. One of the most

widely used solutions to handle this problem in practice, is selecting a subset of available data (e.g., by random sampling) that ensures all classes are equally represented, or excluding the records of severely underrepresented classes.

To study the impact (if any) of data imbalance (i.e., the ratio of records between reentry and non–reentry) on (i) BART, (ii) algorithmic service allocation methods, and (iii) the proposed evaluation metrics, we created subsets sampled from D_{ori} with different reentry rates as follows. We randomly split data instances in D_{ori} (i.e., $\sim$ 19k records) into training and test sets with a ratio of 4 : 1. The number of data instances for training and test sets are 15,161 and 3,791 accordingly, resulting in reentry rates 21.64% and 23.08%, respectively. We fixed the test set, and derived (from the original training set) five separate training sets, with different reentry rates, as shown in Table 6. Specifically, in the original training set (i.e., 15,161 records), the total number of reentries is 3,275. Thus, to create a balanced training set (i.e., reentry rate of 0.5), we randomly sampled 3,275 non–reentry records. To create training sets with varying reentry rates, we fixed the total number of the training set and randomly selected reentry data records according to the corresponding reentry rate. We subsequently randomly selected the non–reentry data instances.

Table 6 Statistical number of training set with different reentry rate.

Reentry Rate	Data Imbalance Ratio (Reentry/ Non–reentry)	Number of Reentry Records	Number of Non–reentry Records	Total Number of Training Data Instances
0.1	1:9	655	5,895	6,550
0.2	2:8	1,310	5,220	6,550
0.3	3:7	1,965	4,585	6,550
0.4	4:6	2,620	3,930	6,550
0.5	5:5	3,275	3,275	6,550

6.2 Model Configuration

6.2.1 BART Model

The BART model has been verified to accurately predict the actual reentry [7]. To build and train model $\mathcal{B}$, we follow the same experimental setting suggested by [7] with the only difference that a limitation on the number of years passed before someone reentering the homelessness system is not imposed. Therefore, a total number of 18,952 records are split into training and testing sets with a ratio of 3 : 1. Specifically, the number of training and testing data instances are 14,214 and 4,783, respectively. We use R package bartMachine [50] to build and fit $\mathcal{B}$.

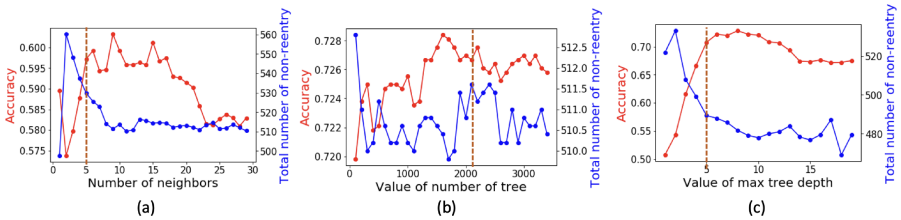


Fig. 6 Parameter tuning results for (a) KNN, (b) RF, and (c) AB accordingly. In each case, the vertical dotted line indicates the best tradeoff between accuracy and non–reentry.

6.2.2 Algorithmic Models Parameter Tuning

We use 5–fold cross–validation to tune and select the parameters (p_θ) for three algorithmic models. For each algorithmic model, the tuning parameter and parameter set ($\mathbf{P}$) are different. In order to select the parameters with the best performance while considering both Accuracy (ACC) and Non–Reentry ($NR = \sum_{i=1}^{N} \mathbb{1}(r'_i = 0)$), we use the following equation:

$$p^*_\theta = argmax_{p_\theta \in \mathbf{P}}(\alpha ACC'(p_\theta) + (1-\alpha)NR'(p_\theta)), \tag{13}$$

where $ACC'(p_\theta) = \frac{ACC(p_\theta) - ACC^{min}(p^{min}_\theta)}{ACC^{max}(p^{max}_\theta) - ACC^{min}(p^{min}_\theta)}$ and $NR'(p_\theta) = \frac{NR(p_\theta) - NR^{min}(p^{min}_\theta)}{NR^{max}(p^{max}_\theta) - NR^{min}(p^{min}_\theta)}$. Both $NR'(p_\theta)$ and $ACC'(p_\theta)$ are normalized. The main idea of the above equation is to achieve the best tradeoff between ACC and NR. α is an adjustable coefficient determining the importance of each metric. For example, $\alpha > 0.5$ gives more importance on the accuracy, which means more preference on allocation output that could get closer to the actual assignment result. $\alpha < 0.5$ implies more importance on reentry, which means more preference on allocation output that could bring practical benefits (i.e., lower the reentry). In our study, we choose $\alpha = 0.5$ so that ACC and NR are equally important.

We select p^*_θ for the three algorithmic models. In the KNN, we tune the parameter p^{KNN}_θ that is the number of neighbors and is denoted as k for better distinction. We choose $p_\theta* = 5$ according to Fig 6 (a). In RF, the tuned parameter p^{RF}_θ is the number of trees. We choose $p_\theta* = 2100$ according to Fig 6 (b). In AB, we use the decision tree as the weak classifier, and then the tuned parameter is the maximum depth of the tree (p^{AB}_θ). We choose $p_\theta* = 5$ according to Fig 6 (c).

7 Analysis

7.1 BART Model Performance

$\mathcal{B}$ predicts that 1,046 individuals (21.56% of the total number of individuals in the testing set) will re–enter the homeless service system, when according to the data 1,093 (23.06%) actually did. Therefore, BART is capable of simulating the

Table 7 Statistics of reentry result based on different algorithmic models and Gurobi optimization for assistance service allocation.

Allocation Methods	Number of Reentry	Percentage of Reentry
AB	**294**	**25.19**
RF	377	32.30
KNN	407	34.87
Gurobi	476	40.78
Actual Reentry	369	31.61

Table 8 Statistics of three groups based on different algorithmic models and Gurobi optimization for assistance service allocation.

Allocation Methods	G_b	G_h	G_{u_1}	G_{u_0}
AB	**233**	**167**	**127**	**640**
RF	218	230	147	572
KNN	188	235	172	572
Gurobi	176	296	184	515

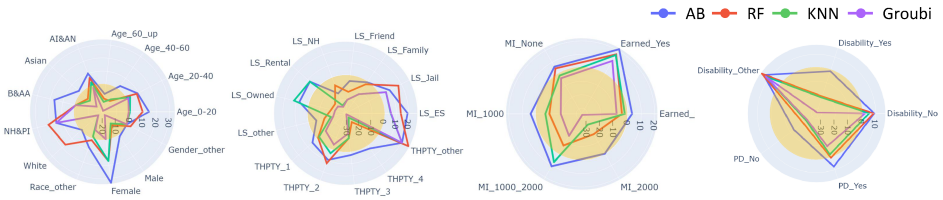
real reentry/non–reentry situation with given allocation services. We use the trained $\mathcal{B}$ to evaluate the results of assistance service allocated by algorithmic models.

7.2 Algorithmic Models Evaluation

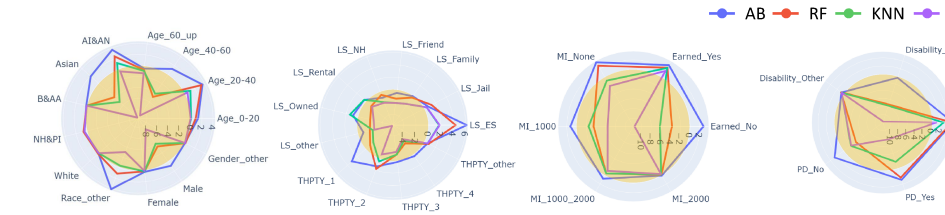
Table 7 summarizes the reentry results based on the delivered services by different algorithmic models. AB achieves the best performance by lowering the reentry from 369 to 294. The performance of RF and KNN are better than the comparative method Gurobi. However, it is worse than the actual reentry outcome.

Based on the discussion on Section 4.3, we compute the overall G_b, G_h and G_u for each algorithmic model and Gurobi method. Detailed statistical results are shown in Tables 9 and 12 in the Appendix. Ideally, we want G_b and G_{u_0} to be as high as possible, whereas G_h and G_{u_1} to be as low as possible, as we wish more people to be benefited, and fewer people to be harmed. In this context, AB still achieves the best performance, whereas RF and KNN both outperform Gurobi.

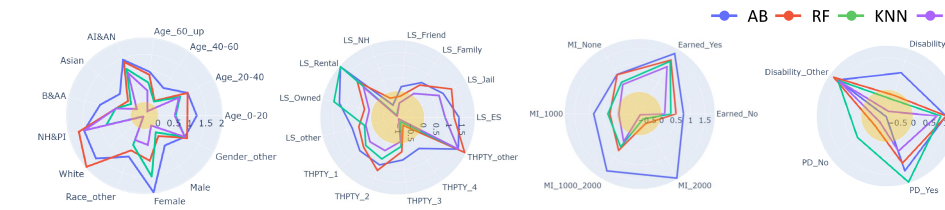
As mentioned in Section 4.3, we tend to avoid the algorithmic model like J_A shown in Fig 3(a). Therefore, we visualize the three groups on feature value level for that selected importance and sensitive features shown in Table 2. Note that features that belong to the category of basic information (i.e., Age, Gender, Race) are considered sensitive features. We use radar plot to visualize the fairness evaluation metrics ΔN, ΔT, and ΔU. Table 9 shows the boundary line of acceptable model performance with respect to ΔN, ΔT, and ΔU. Note



(a) ΔN Comparison



(b) ΔT Comparison



(c) ΔU Comparison

Fig. 7 Fairness evaluation result. Refer to Table 2 for explanation of variables.

that the unacceptable criterion ΔT is defined as smaller than -5%, with 5% being the tolerance value suggested by [7].

Figure 7(a) shows the results of ΔN based on four categories features. The highlighted middle circle is served as the boundary (i.e., $\Delta N(f_k^m) = 0$). Each radius refers to a certain feature value. Ideally, we want the radius point located outside of the boundary circle. For each feature value, the points located inside the boundary means $\Delta N(f_k^m) < 0$, which implies that the number of benefited people is less than the hurt people. Evaluation results for ΔT and ΔU shown in Fig 7(b) and Fig 7(c) follow the same idea. Results of ΔN and ΔT should be interpreted together. For example, in the feature of Age, the feature value of Age_60_up means people who are older than 60. The ΔN result for AB is -7.22%, which might be misinterpreted as the allocation result of AB hurts elderly people. However, the ΔT result for AB is -0.51%, which refers to a small number of people (e.g., six people). It is acceptable when the percentage of total hurt people is less than 5% (i.e.,$\Delta T > -5.0\%$) as suggested in [7]. The interpretation of ΔU follows the discussion in Section 4.3. Therefore, the desired value of $\Delta U(f_k^m)$ is a positive number, and the larger, the better. If the value of $\Delta U(f_k^m)$ is negative, then we conclude that the transitional

Table 9 Criterion for three evaluation metrics.

Evaluation metrics	Acceptable	Unacceptable
ΔN	$\Delta N \geq -5\%$	$\Delta N < -5\%$ and $\Delta T < -5\%$
ΔT	$\Delta T \geq -5\%$	$\Delta T < -5\%$
ΔU	$\Delta U \geq 0$	$\Delta U < 0$

performance with allocated service by the algorithmic model is poor. To sum up, to quantitatively explore the service allocation results by three algorithmic models and the baseline work, we classify the possible combination of ΔN, ΔT, and ΔU into two categories which are acceptable and unacceptable shown in Table 9.

According to experimental results presented in Fig 7, we further analyze the service allocation results of each method as follows.

- AdaBoost (AB): The total number of Unacceptable cases (i.e., f_k^m) defined in Table 9 is 2, including the value of Place Not for Habitation in feature Living Situation and the feature value of No Physical Disability. The rest results of the feature values belong to the acceptable cases. There are no unacceptable cases shown in the sensitive features (i.e., Basic information). AB model achieves a relatively high score (i.e., the highest score of all three metrics for certain features) for several feature values, including female, having employment earned, previous living in the emergency shelter, and having physical disability.
- Random Forest(RF): 12 unacceptable cases exist in the allocated results of the RF model. 2 among them belong to sensitive features that are age between 40 to 60 and older than 60. The rest 25 feature cases belong to acceptable, and the RF model achieves a relatively high score in the feature values, including female, having employment earned, monthly income is zero, having employment earned, and having a physical disability. The results of the RF model are less satisfactory than AB because of hurting sensitive features and more unacceptable cases.
- K Nearest Neighbors (KNN): 11 unacceptable cases are shown in the allocated result of the KNN model, and one of them belongs to the sensitive feature, which is the age is elder than 60 years old. The rest of the cases belong to acceptable. The feature values with a relatively high score are female, American Indian or Alaska Native, having employment earned, monthly income is zero, and having a physical disability. The results of the KNN model are also worse than the AB model.
- Gurobi: The total number of unacceptable cases in Gurobi is 19, and 5 of them belong to the sensitive features. The feature values with a relatively high score in Groubi are age between 20 and 40 years old, having employing earned, previous living in the emergency shelter, the information of times being homeless past three years is unknown or missed (i.e., THTPY_other). The performance of Gurobi is the worst compared with the three algorithmic models.

In summary, AB achieves the best performance with fewer unacceptable cases, and while not hurting any groups of individuals on sensitive features. Although the homeless services provision method employed in [7] (i.e., Gurobi) was reported to perform well (i.e., algorithmically assigned homeless services reduce reentry), our experiments using data collected from the New York Capital Region show that Gurobi is not a favorable method since it was found to result in increased overall reentry rate. This result illustrates the need to select a method among alternatives using domain–specific evaluation criteria, such as those proposed in this article, as opposed to naively selecting the best performing method with respect to domain agnostic metrics, such as predictive accuracy.

7.3 Impact of Data Imbalance

To explore the impact of the data imbalance on BART model, we use *difference rate*, which we define as the discrepancy between the number of predicted reentry and true reentry records. Figure 8 shows the results. Evidently, BART can more accurately learn when the training set is balanced (i.e., reentry rate is close to 0.5). Note that in our experiment, the most accurate model was obtained for a reentry rate of 0.4. However, upon close inspection, the difference in performance is negligible.

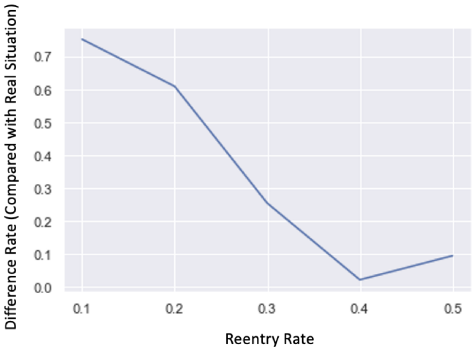


Fig. 8 Experiment results of difference rates on test set of BART model with training by training sets with different reentry rate.

To explore the impact of data imbalance on algorithmic models, as well as the proposed evaluation metrics, we use five separate training sets (see Section 6.1.4) to train services allocation models, then use the corresponding trained BART model to predict reentry based on the predicted service allocations. Table 10 shows the results. When the reentry rate is low, the perceived benefit rate is high. This is because, with a fewer training reentry data for both BART and services allocation models, the output is biased to non–reentry. Therefore, the number of benefited individuals increased with changing to non–reentry state. However, the high benefit rate is not equal to the real benefits

of those services reallocation models. Specifically, with increasing reentry rate, the benefit rate keeps diminishing.

The key takeaway from this analysis is twofold. First, data imbalance can adversely impact model output, so a balanced training dataset is required for both BART and the algorithmic models to be accurate. Second, the proposed evaluation metrics, which are designed to evaluate models' output, can be potentially used to indicate problems with the training dataset, such as biases as the result of data imbalance (e.g., when the perceived benefit is deemed to be too "good to be true").

Table 10 Tables for experiments statistical results for allocation models and Gurobi methods of domain–specific evaluation metrics

	AB			RF			KNN			Gurobi		
Reentry Rate	G_b	G_h	G_{u_0}/G_{u_1}	G_b	G_h	G_{u_0}/G_{u_1}	G_b	G_h	G_{u_0}/G_{u_1}	G_b	G_h	G_{u_0}/G_{u_1}
0.1	0.26	0.05	0.67/0.02	0.28	0.07	0.65/0.00	0.21	0.04	0.70/0.05	0.22	0.04	0.69/0.04
0.2	0.21	0.07	0.70/0.02	0.21	0.07	0.70/0.01	0.12	0.09	0.78/0.01	0.10	0.11	0.77/0.01
0.3	0.13	0.05	0.81/0.01	0.12	0.09	0.71/0.08	0.11	0.10	0.70/0.09	0.09	0.14	0.69/0.08
0.4	0.09	0.13	0.74/0.04	0.08	0.09	0.70/0.14	0.04	0.23	0.60/0.13	0.02	0.31	0.54/0.13
0.5	0.05	0.28	0.59/0.07	0.06	0.21	0.64/0.09	0.02	0.37	0.50/0.11	0.02	0.46	0.43/0.09

8 Conclusion and Future Work

In this study, we evaluated the performance of algorithmic models for homeless service allocation. To include fairness in the evaluation process, we proposed three application specific evaluation metrics. Using the proposed metrics, we compared several data science solutions for homeless services allocation in a unique dataset of homeless services administrative records.

Avenues for future study include exploring allocation fairness at the individual level, as well as different criteria for service effectiveness assessment. For instance, while reentry can be a good indicator of those still in need of assistance after receiving homelessness services, it does not capture cases where people request assistance but are placed on a waiting list due to resource constraints (e.g., unavailability of beds in a shelter). Finally, the proposed application specific evaluation metrics for fairness are computed for each individual feature. As a result, it can be time consuming to manually inspect a large number of important features as the dimensionality of the data increases in order to best inform model selection. To address this "scalability" issue, either an aggregate metric can be developed or an explainable machine learning metamodel can be developed to communicate to practitioners which features may be most important in their selection of a fair (while at the same time being accurate) model. This is a challenging problem in itself and can therefore be a promising research direction.

9 Statements and Declarations

Conflict of Interest. The authors declare no competing interests regarding the publication of this article.

Data Availability. The dataset that support the findings of this study is not publicly available, as it has been provided to the authors under a data sharing agreement with CARES of NY, Inc. Information on how to obtain the dataset and reproduce the analysis is available from the corresponding author on request.
Ethics Approval. This study was conducted in accordance to all relevant policies and procedures set forth by the University at Albany Institutional Review Board for the protection of human subjects.
Funding. This material is based upon work supported by the National Science Foundation under Grant No. ECCS–1737443.

References

[1] Volker Busch-Geertsema, William Edgar, Eoin O'Sullivan, and Nicholas Pleace. Homelessness and homeless policies in europe: Lessons from research. In *Feansta: A report prepared for the European Consensus Conference on Homelessness; 9-10 December 2010*, 2010.

[2] NU CEPAL. Report on the 2010 world programme on population and housing censuses. 2007.

[3] M. Henry, R. Watt, A. Mahathey, J. Ouellette, and A. Sitler. The 2020 annual homelessness assessment report (ahar) to congress: Part 1: Point–in–time estimates of homelessness. *The ANNALS of the American Academy of Political and Social Science*, 2020.

[4] Paragraphs Page. Report of the special rapporteur on adequate housing as a component of the right to an adequate standard of living, and on the right to non-discrimination in this context on her mission to indonesia. 2015.

[5] Why are people homeless? *National Coalition*, 2007. http://nationalhomeless.org/publications/facts/Why.pdf.

[6] Mohammad Javad Azizi, Phebe Vayanos, Bryan Wilder, Eric Rice, and Milind Tambe. Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. In *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 35–51. Springer, 2018.

[7] Amanda Kube, Sanmay Das, and Patrick J Fowler. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 622–629, 2019.

[8] Hau Chan, Eric Rice, Phebe Vayanos, Milind Tambe, and Matthew Morton. Evidence from the past: Ai decision aids to improve housing systems

for homeless youth. In *AAAI Fall Symposia*, pages 149–157, 2017.

[9] Yuan Gao, Sanmay Das, and Patrick Fowler. Homelessness service provision: a data science perspective. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] Charalampos Chelmis, Wenting Qi, Wonhyung Lee, and Stephanie Duncan. Smart homelessness service provision with machine learning. *Procedia Computer Science*, 185:9–18, 2021.

[11] Cynthia Nagendra, Jason Satterfield, Laura Gillis, and Vicki Judice. Resource allocation and monitoring strategies. 2010. https://files.hudexchange.info/resources/documents/ResourceAllocationandMonitoringStrategies_Presentation.pdf.

[12] Theresa Dostaler and Geoffrey Nelson. A process and outcome evaluation of a shelter for homeless young women. *Canadian Journal of Community Mental Health*, 22(1):99–112, 2009.

[13] Yin-Ling Irene Wong, Dennis P Culhane, and Randall Kuhn. Predictors of exit and reentry among family shelter users in new york city. *Social Service Review*, 71(3):441–462, 1997.

[14] Molly Brown, Danielle Vaclavik, Dennis P Watson, and Eric Wilka. Predictors of homeless services re-entry within a sample of adults receiving homelessness prevention and rapid re-housing program (hprp) assistance. *Psychological services*, 14(2):129, 2017.

[15] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 24–35, 2020.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[17] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. There's software used across the country to predict future criminals and it's biased against blacks. 2016, 2020.

[18] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[19] Till Von Wachter, Marianne Bertrand, Harold Pollack, Janey Rountree, and Brian Blackwell. Predicting and preventing homelessness in los angeles. *California Policy Lab and University of Chicago Poverty Lab*,

2019.

[20] United States Department of Housing and Urban Development. Hmis guides and tools. *Retrieved December 14, 2020, from https://www.hudexchange.info/programs/hmis/hmis-guides/*, 2019.

[21] Blake VanBerlo, Matthew AS Ross, Jonathan Rivard, and Ryan Booker. Interpretable machine learning approaches to prediction of chronic homelessness. *Engineering Applications of Artificial Intelligence*, 102:104243, 2021.

[22] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[23] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[24] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

[25] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[26] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

[27] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[28] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American journal of epidemiology*, 186(9):1026–1034, 2017.

[29] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[30] Wenting Qi and Charalampos Chelmis. Improving algorithmic decision–making in the presence of untrustworthy training data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1102–1108. IEEE, 2021.

[31] Aslı Bozdağ. Local-based mapping of carbon footprint variation in turkey using artificial neural networks. *Arabian Journal of Geosciences*, 14(6):1–15, 2021.

[32] Titouan Parcollet and Mirco Ravanelli. The energy and carbon footprint of training end-to-end speech recognizers. 2021.

[33] Nigel Williams, Sebastian Zander, and Grenville Armitage. A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 36(5):5–16, 2006.

[34] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[35] Sriram Vasudevan and Krishnaram Kenthapadi. Lift: A scalable framework for measuring fairness in ml applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2773–2780, 2020.

[36] Danielle Leah Kehl and Samuel Ari Kessler. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. 2017.

[37] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[38] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[39] Boyeong Hong, Awais Malik, Jack Lundquist, Ira Bellach, and Constantine E Kontokosta. Applications of machine learning methods to predict readmission and length-of-stay for homeless families: The case of win shelters in new york city. *Journal of Technology in Human Services*, 36(1):89–104, 2018.

[40] Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.

[41] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. Efficient knn classification algorithm for big data. *Neurocomputing*, 195:143–148, 2016.

[42] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[43] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[44] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[45] United States Department of Housing and Urban Development. HMIS Data Standards Manual. *Retrieved May 26, 2021, from https://www.hudexchange.info/resource/3824/hmis-data-dictionary/*, 2020.

[46] Yue Bi, Dongxu Xiang, Zongyuan Ge, Fuyi Li, Cangzhi Jia, and Jiangning Song. An interpretable prediction model for identifying n7-methylguanosine sites based on xgboost and shap. *Molecular Therapy-Nucleic Acids*, 22:362–372, 2020.

[47] Amir Bahador Parsa, Ali Movahedi, Homa Taghipour, Sybil Derrible, and Abolfazl Kouros Mohammadian. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136:105405, 2020.

[48] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[49] Swati V Narwane and Sudhir D Sawarkar. Machine learning and class imbalance: A literature survey. *Industrial Engineering Journal*, 12(10), 2019.

[50] Adam Kapelner and Justin Bleich. bartmachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*, 2013.

Charalampos Chelmis, Assistant Professor in Computer Science at the University at Albany, State University of New York, and Director of the Intelligent Big Data Analytics, Applications, and Systems (IDIAS) Lab, conducts research on data–intensive computing involving interrelated data, and social good applications. He has served and is serving as Co–Chair, TPC member or reviewer in numerous international conferences and journals including TheWebConf, ASONAM, and ICWSM. He is currently Associate Editor of Social Network Analysis and Mining Journal (SNAM). He earned his Ph.D. and M.Sc. degrees in Computer Science in 2013 and 2010, respectively from the University of Southern California, and B.S. in Computer Engineering and Informatics from the University of Patras, Greece in 2007.

Wenting Qi, received the B.S. degree in automation from the Beijing University of Technology, China in 2017, and earned M.S. degree in Electrical Engineering from the University of Southern California in 2019, Los Angeles, CA, USA. She is currently working towards the Ph.D. degree in Computer Science at the University at Albany, Albany, NY, USA. Her research interests include noisy detection, hierarchical classification, and explainable machine learning.

Table 11 Tables for experiments statistical results for algorithmic models and Gurobi methods of three domain–specifc evaluation metrics I

Features Values		AB			RF			KNN			Gurobi		
Features	Values	ΔN	ΔT	ΔU	ΔN	ΔT	ΔU	ΔN	ΔT	ΔU	ΔN	ΔT	ΔU
Age	0-20	14.42	1.28	1.22	9.61	0.85	0.93	0.00	0.00	0.56	-0.96	-0.08	0.53
	20-40	9.19	3.59	1.11	8.53	3.34	1.11	2.84	1.11	0.84	1.31	0.51	0.80
	40-60	2.86	1.28	0.63	-10.89	-4.88	0.14	-9.36	-4.19	0.11	-20.07	-8.99	-0.26
	≥60	-7.22	-0.51	0.96	-10.84	-0.77	0.87	-13.25	-0.94	0.62	-19.27	-1.37	0.37
Gender	Female	10.11	3.68	1.45	6.58	2.39	1.38	3.05	1.11	1.14	-1.41	-0.51	1.01
	Male	2.98	1.88	0.60	-6.11	-3.85	0.26	-8.15	-5.14	0.13	-14.94	-9.42	-0.11
	Other	16.66	0.08	1.0	0.00	0.00	0.5	0.00	0.00	0.5	0.00	0.00	0.5
Race	AI&AN	15.15	0.42	1.52	21.21	0.59	1.69	15.15	0.42	0.8	15.15	0.42	1.52
	Asian	0.00	0.00	1.6	16.66	0.08	2.0	-16.66	-0.08	-0.39	-16.66	-0.08	-0.39
	B&AA	7.95	4.19	0.95	2.11	1.11	0.74	-0.81	-0.42	0.56	-5.84	-3.08	0.39
	NH&PI	33.33	0.17	2.0	16.66	0.08	1.0	16.66	0.08	1.5	0.00	0.00	0.5
	White	2.10	0.85	0.71	-8.42	-3.42	0.35	-9.89	-4.02	0.21	-17.68	-7.19	-0.05
	Other	0.00	0.00	1.01	3.22	0.08	1.09	0.00	0.00	1.01	0.00	0.00	1.01
LS	ES	19.20	6.16	1.35	13.59	4.37	1.16	5.33	1.71	0.8	5.59	1.79	0.81
	Jail	9.19	0.68	0.96	17.24	1.28	1.33	5.74	0.42	0.78	5.74	0.42	0.78
	SFaMP	-0.65	-0.08	0.60	-2.63	-0.34	0.48	-12.50	-1.62	0.08	-12.50	-1.62	0.08
	SFiMP	-4.87	-0.51	0.16	-13.82	-1.45	-0.22	-20.32	-2.14	-0.49	-20.32	-2.14	-0.49
	PHP	-14.11	-1.02	-0.54	-7.05	-0.51	-0.25	-24.70	-1.79	-0.96	-25.88	-1.88	-1.01
	RCS	7.69	0.42	1.93	-20.00	-1.11	1.05	7.69	0.42	1.93	-23.07	-1.28	0.73
	OCS	7.51	0.85	1.29	-17.29	-1.97	0.28	12.03	1.37	1.51	-22.55	-2.57	0.09
	Other	-6.80	-0.85	0.77	-13.60	-1.71	0.51	-19.04	-2.39	0.28	-21.08	-2.65	0.23

Table 12 Tables for experiments statistical results for algorithmic models and Gurobi methods of three domain–specifc evaluation metrics II

Features Values		AB			RF			KNN			Gurobi		
Features	Values	ΔN	ΔT	ΔU	ΔN	ΔT	ΔU	ΔN	ΔT	ΔU	ΔN	ΔT	ΔU
Earned	No	3.17	2.74	0.76	-4.36	-3.77	0.50	-7.24	-6.25	0.32	-13.30	-11.48	0.11
	Yes	21.25	2.91	1.66	16.87	2.31	1.40	16.25	2.22	1.35	11.24	1.54	1.13
MI	0	6.29	3.59	0.82	4.79	2.74	0.81	-1.34	-0.77	0.51	-3.59	-2.05	0.42
	$\leq$ 1,000	3.72	1.28	0.84	-10.17	-3.51	0.28	-7.19	-2.48	0.35	-18.61	-6.42	-0.03
	1,000-2,000	10.58	0.77	1.52	-7.05	-0.51	0.71	7.05	-0.51	0.57	-15.29	-1.11	0.36
	$\geq$ 2,000	0.00	0.00	1.81	-16.66	-0.17	-0.36	-25.00	-0.25	-0.54	-33.33	-0.34	-0.72
THPTY	One	4.81	3.08	0.92	-1.33	-0.85	0.72	-2.40	-1.54	0.61	-8.70	-5.56	0.39
	Two	9.35	1.37	0.94	12.28	1.79	1.17	3.50	0.51	0.69	-4.09	-0.59	0.34
	Three	2.97	0.25	0.63	-10.89	-0.94	0.13	-10.89	-0.94	0.10	-15.84	-1.37	-0.09
	Four	3.50	0.34	0.43	-22.80	-2.22	-0.67	-27.19	-2.65	-0.83	-30.70	-2.99	-0.98
	Other	20.58	0.59	1.63	26.47	0.77	1.91	20.58	0.59	1.63	20.58	0.59	1.63
DC	No	8.88	5.31	1.02	8.45	5.05	1.03	5.01	2.99	0.86	2.00	1.19	0.74
	Yes	-0.96	-0.34	0.60	-20.53	-7.28	-0.14	-21.49	-7.62	-0.28	-33.09	-11.73	-0.67
	Other	14.54	0.68	1.09	16.36	0.77	1.24	12.72	0.59	1.08	12.72	0.59	1.03
PD	No	-13.20	-1.19	-0.79	-21.69	-1.97	-0.34	-24.52	-2.22	0.32	-25.47	-2.31	-0.53
	Yes	7.54	6.85	0.98	0.56	0.5	0.73	-1.97	-1.79	1.35	-8.38	-7.62	0.33