Problem Statement

On a single night in January of 2023, more than half a million people in the Unites States were homeless, and the number of people experiencing homelessness is growing (De Sousa et al., 2023). Homeless housing situations matter: people experiencing homelessness have a 3.5 times all cause mortality rate higher than otherwise observably similar housed individuals (Meyer et al, 2023). In recognition of these stark differential outcomes, governments across the United States have allocation substantial funds to assist households in finding stable housing. For instance, in 2023 the federal grant program to fund local homeless programs, Continuums of Care, allocated $3.16 billion dollars (HUD, 2024), which is on top of the local dollars raised by governments and philanthropy to address the problem (Lee 2021). Despite the significant amount of funds, funding level remain far below the need (Kim, 2023) and there is widespread national and local recognition that homelessness remains a persistent problem. Until these funds materialize, governments will need to do more with less, that is, assist more people in finding stable housing with the same amount of resources.

To that end, recent applications of machine learning have attempted to aid service providers by A) predicting who might be more likely to exit homelessness absent services and B) given a fixed amount of services help prioritize which household experiencing homelessness are most likely to benefit from those services (Kube et al., 2023). Kube et al. accomplish this by leveraging a Bayesian Additive Regression Tree to predict the likelihood that an individual experiencing homelessness will exit homelessness if the received services and if they did not receive services. These probabilities are used as weights an "assignment problem" optimization algorithm with fixed capacity (i.e. number of available services). The authors find that were homeless services allocated based on predicted outcomes and optimized in accordance with which services are available when 5.5% fewer households would request additional homelessness services within the next two years.

An additional consideration the authors note is that reallocation of services is not a pareto-improving, in other words, while the system overall exits more households from homelessness not all individuals are better off after reallocation. For example, allocation of resources differed under historical assignment and the authors optimal assignment by gender[1]: one of seven federally protected classes under the Fair Housing Act.[2] For governments bound by ethics, public opinion, and the legal system an unequitable allocation of resources may limit the adoption of algorithmic solutions that improve the system overall, but have deleterious impacts for specific protected class. Thus, the primary question/problem this project seeks to address is:

1. Is there a reformulation of the algorithm of described by Kube et al., that can both maximize the number of households exiting homelessness ***and*** introduce flexibility for governments to balance optimizing performance of the overall system with equity concerns for sub-populations?

[1] Both males and females are predicted to be better off under optimal assignment, however, relative to females males housing outcomes are expected to improve more.
[2] For more information see:
https://www.hud.gov/program_offices/fair_housing_equal_opp/fair_housing_act_overview

2. Additional extensions of the work done by Kube et al. may include an explication of the statistical rational for using predicted probabilities of exit, investigating the predictive capacity of models that omit all protected class information, and varying different timelines in the optimization algorithm.

Background and Data

In Pierce County, Washington in 2022, more than 9,000 individuals experienced homelessness, the highest number since electronic records were maintained.[3] Of those 9,000 individuals 1,500 exited to a known permanent housing. And of those 1,500 individuals an estimated 225 will return to homelessness within the next two years.[4] Using data from Pierce County, it our intent to address our research question. Data on homelessness from Pierce County is maintained in the Homelessness Management Information System (HMIS). HMIS maintains records on individuals when they enter and exit services for homelessness, the type of service, where they exit services to, demographic information, household information, and employment information.

When a client contacts a provider of homeless services about needing assistance the client enters Coordinated Entry. The Coordinated Entry process works as follows: 1) the client is assessed to see if they are literally homeless or fleeing domestic violence, if so 2) a diversion conversation follows, in which an alternative housing arrangement absent assistance from a provider is sought, if non-can be found 3) a prioritization interview is conducted, 4) the results of the interview generate a prioritization for housing interventions score, and if 5) another household leaves services and a spot opens up for housing interventions, the household with the highest prioritization score is selected for that intervention. That is, periodically throughout the year the Coordinated Entry system makes decisions about which households will get scarce housing interventions based on a household's prioritization score. Prioritization scores are intended to capture "barriers to accessing housing" and "vulnerability factors".[5]

The three primary housing interventions available for households on the prioritization list are Rapid Rehousing (RRH), Transitional Housing (TSH), and Permanent Supportive Housing (PSH). Rapid rehousing provides short and medium-term rental assistance, TSH provides supportive services and housing for up to 24 months, and PSH provides long-term housing with no set end date.[6] Due to the scarcity of slots in these programs, the majority of individuals who enter the Coordinated Entry System and who are literally homeless do not receive any of these three housing interventions.[7]

There are different ways a successful exit from homelessness could be defined. Typically, a successful exit includes a window of time in which if a household is monitored for reentry into

---

[3] For more information see (accessed 3/10/2024): https://open.piercecountywa.gov/stories/s/7wee-rgqc
[4] Between 2013 and 2020 15% of individuals exiting homelessness to permanent housing were observed to be homeless within 2 years. Source: https://open.piercecountywa.gov/stories/s/7wee-rgqc
[5] See chapter 3.3.6 of https://www.piercecountywa.gov/DocumentCenter/View/111563/Policy-Ops-Manual_Ch-3---Coordinated-Entry__Pub-0122
[6] For more information see: https://www.hudexchange.info/homelessness-assistance/coc-esg-virtual-binders/coc-program-components/coc-program-components-overview/
[7] See https://open.piercecountywa.gov/stories/s/7wee-rgqc

homelessness. For example, if a household exits to a stable housing situation, but returns to homelessness within that window, then the exit is not considered successful. HUD defines a "returns to homelessness" metric, which tracks the percent of people who exit homelessness successfully and return to homeliness within two years.

Per policy, providers are required to contact households that do not receive an intervention within 90 days of their prioritization interview to conduct an exit interview or to reaffirm their continued request for a housing intervention. However, a significant proportion of households cannot be re-contacted, and are recorded as having an unknown exit destination in HMIS. Thus, a successful exit could mean that the provider has been able to confirm a household's acquisition of stable housing, or it could mean that the household does not request additional services for some period of time after their initial request. Due to the ambiguity around how to define a successful exit, we construct 6 different measures of a successful outcome that A) vary the window with which to look for new instances of homelessness (6, 12, and 24 months) and B) define a successful exit if no new requests for services occur within that timeframe and if no new requests for services occur within that timeframe plus having a confirmed exit to a stable housing situation. We will evaluate the performance of our models on these different outcomes and determine an outcome to use.

Information about the household is collected during the prioritization interview. During these interviews, demographic, economic, health, and living situation information is collected. All members of a household are interviewed, but the head of the household is identified during the interview. Thus, depending on the variable, information can be collected from the head of the household or, in some cases, aggregated from the household. For example, the head of the household is asked about their veteran status, but so too are each of the household members. A variable is constructed for the head of the household, but also one that considers whether any members of the household were veterans. We construct three feature sets: one with all head-of-household and household variables including potentially colinear variables such as the veteran's status, one where variables are primarily taken from the head of the household,[8] and one where variables are primarily taken from aggregating household information.[9] The majority of these variables are categorical or binary. We consider two different encodings: weight of evidence encodings (De La Bourdonnye & Daniel, 2021) and one hot encoding.

Methodology

Learning an optimal and equitable prioritization algorithm for homelessness interventions has three primary steps:

1. Learn an algorithm to predict the likelihood of exiting homelessness, which can be used to estimate the likelihood of exiting homelessness whether a household received any of the three interventions or none

[8] Not all variables can be taken from the head of the household alone. For example, we include a variable for the number of people in the household and the number of children.

[9] Not all variables can be reasonably aggregated from the household. For example, there isn't a reasonable way to aggregate race and ethnicity. For these we use the head-of-household information.

2. Describe an algorithm for assigning housing interventions based on the results from step 1
3. Learn weights to be used in the algorithm described in step 2 to balance treatment assignment across a specific sub-population

To motivate the methodology for determining an optimal and equitable treatment assignment, we first consider a simpler objective: maximizing the number of households exiting homelessness.

Let be the event that individual *i* exits homelessness, $x_{i,j}$ be the 0/1 indicator for whether or not individual *i* was assigned treatment j, where when j = 1 the individual is assigned no treatments. Then:

$$y_i = \begin{cases} 0 \; with \; probability \; 1 - p_i \\ 1 \; with \; probability \; p_i \end{cases}$$

$$S = \sum_{i=1}^{N} y_i$$

As stated above our goal is to maximize the expected number of people exiting homelessness by manipulating the treatment assignments, $x_{i,j}$.

$$\max_{x_{i,j}} E[S] = \max_{x_{i,j}} E\left[\sum_{i=1}^{N} y_i\right] = \max_{x_{i,j}} \sum_{i=1}^{N} E[y_i] = \max_{x_{i,j}} \sum_{i=1}^{N} p_i * 1 + (1 - p_i) * 0$$

$$= \max_{x_i,j} \sum_{i=1}^{N} p_i$$

By the law of total probability and that treatment assignments are mutually exclusive:

$$p_i = \sum_j P\left(y_i = 1 \middle| x_{i,j}\right) P\left(x_{i,j}\right) = \sum_j p_{i \mid x_{i,j}} \, x_{i,j}$$

However, since we are optimizing over $x_{i,j}$, i.e. determining these assignments, then $P(x_{i,j})$ is either 1 or 0 depending on whether or not we assign individual i to treatment j, thus $P(x_{i,j})$ can be replaced by $x_{i,j}$. Thus our objective is the following:

$$\text{Eq.1)} \quad \max_{x_i,j} \sum_{i=1}^{N} p_i = \max_{x_i,j} \sum_{i=1}^{N} \sum_j p_{i \mid x_{i,j}} \, x_{i,j}$$

For the first step in solving this optimization problem, we need to learn the values of $p_{i \mid x_{i,j}}$. That is, we need to estimate individual *i*'s likelihood of exiting homelessness when they were assigned to treatment *j* (note again that when j=1, we consider this the case most households realize, namely, no housing intervention). Specifically, we need a flexible model that considers non-linear heterogeneous treatment effects so that we can find a mapping from a households attributes and treatment assignments to their likelihood of exit (Hill & Murray, 2020).

If Z_i is vector of household information for household *i,* and each of the $x_{i,j}$ then our goal is to estimate:

$$\text{Eq. 2)} \quad y_i = f(Z_i)$$

Such that we can recover the $p_{i \mid x_{i,j}}$. For example, if we were to use logistic regression then this would become

$$y_i = \begin{cases} 0 \; if \; p(z_i) < 0.5 \\ 1 \; Otherwise \end{cases}$$

Where

$$p(z_i) = \frac{1}{1 + e^{-(z_i^T \beta)}}$$

To learn this mapping we plan on testing different classification algorithms, feature sets, feature encodings, features transformation, and use the combination that performs best. Please the "Evaluation Strategies" section for more details. Specifically, we will separate our dataset using an 80/20 train-test split and perform 5-fold cross-validation on the training data over our search space. Our search space will include:

- Algorithms – Elastic Net, Random Forest, Bayesian Additive Regression Trees
- Hyperparameter tuning
- 6 different outcomes, as mentioned in the data section
- 3 different feature spaces, as mentioned in the data section
- 2 different feature encodings, as mentioned in the data section
- Whether or not to perform feature selection
- Whether or not to perform principle components analysis

For principle component analysis, we keep as many components necessary to retain 95% of the original variation in the data, but no more.

Once a mapping that corresponds to equation 2 is built, we will recover the probability that household *i* exits homelessness if they were assigned each of the *j* housing interventions. Using these estimated probabilities we will develop an optimization model based on equation 1. In practice slots for housing interventions become available at different points in time and because we want to develop and optimization model that incorporates equity considerations for sub-populations, we will need to modify equation 1. Following Kube et al., we optimize across weeks, though in practice different providers may have preferences for how often to assign slots when they become available. That is:

- Define x_{ijt} as an indicator variable that represents whether household *i* in week *t* is assigned intervention *j*.
- Use p_{ijt} to represent the probability that household *i* exits homelessness in week *t* if assigned intervention *j*.
- Define C_{jt} as the number of available slots for intervention *j* in week *t*.

- Define j=1 to be the "housing intervention" of no intervention, of which there are no capacity constraints on.

Without incorporating equity considerations, the optimization model becomes:

$$\text{Eq. 3) } \max_{x_{ijt}} \sum_{i=1}^{N} \sum_j p_{ijt}\ x_{ijt}$$

$$\text{Subject to: } \sum_i x_{ijt} = C_{jt}\ \forall \text{j } \neq 1$$

$$\sum_j x_{ijt} = 1$$

$$x_{ijt} \in \{0, 1\}$$

This optimization problem is formulated as an Integer Program. Kube et al., show that it can be reformulated as a weighted bipartite b-matching problem that can be solved in polynomial time. However, as indicated above, this optimization formulation does not account for considerations of equity. To account for equity considerations between arbitrary sub-populations g and keep the problem within polynomial time we modify equation 3 as follows.

- Define α_g as the proportion of group g in the total homeless population, based on historical data.
- Calculate $\gamma_{gt}*$, the proportion of group g assigned to some treatment up to time t^*, and compare it with α_g to assess disproportionality.
- $E_{g,t^*} = \gamma_{gt^*}/\alpha_g$ is proportionality of assignment to treatment for group g at time t^*.
- $r_{gt^*} = E_{g,t^*}/E^B_{g,t^*}$ is the risk ratio of group g relative to the base group B at time t^*.

When the non-reference group has been assigned a housing intervention more often than the base group relative to their population size then r_{gt^*} is greater than one, when assignment is equal it is one, and when assignment is lower than the base group's it is less than one. Thus, we modify Equation 4 by adding the following term : $+(\sum_{j\neq 1} C * \left(1 - r_{gt^*}\right)x_{it^*j}))$.

$$\text{Eq 4) } \max_{x_{ijt}} \sum_{i=1}^{N} \sum_j p_{ijt}\ x_{ijt} + (\sum_{j\neq 1} C * \left(1 - r_{gt^*}\right)x_{it^*j}))$$

$$\text{Subject to: } \sum_i x_{ijt} = C_{jt}\ \forall \text{j } \neq 1$$

$$\sum_j x_{ijt} = 1$$

$$x_{ijt} \in \{0, 1\}$$

The problem introduces a new hyperparameter C which is a weight designed to increase or decrease the likelihood that a household is assigned treatment based on whether or not the group

they belong to is over or underrepresented in past assignments to housing interventions (Note the exclusion of j = 1, which is the state of no treatment). For example, if the group is overrepresented then $1 - r_{gt^*}$ is negative and has the effect of penalizing an assignment to x_{it^*j}. With this problem set up the problem then becomes finding a value of C that will center the long-term risk ratios around 1 for all groups. We will adjust the optimization objective to include a fairness term that penalizes, or rewards assignments based on the difference between actual and expected proportions. To find C, we will bootstrap our data, perform the optimization for different values of C, and calculate r_{gt^*} at the end of the optimization run.

Evaluation Strategies

The likelihood of successfully exiting homelessness is unfortunately low with fewer than 20% of households exiting homelessness successfully and not reentering services within two years. To evaluate the performance of different learning algorithms along with the parameters that define the search space indicated in the methods section, we will use the area under the Receiver Operating Curve (AUROC) as our performance metric (Hossin, & Sulaiman 2015). The average out-of-sample AUROC will be used within the cross-validation framework described above. Furthermore, because prioritization scores are intended to identify households with challenges obtaining housing, they should be predictive of which households are able to exit homelessness. Thus, we plan to compare our chosen model against a logistic regression model which uses a single predictor, prioritization score, to predict exits from homelessness. Finally, once a final algorithm and parameters are chosen, we will evaluate our model's ability to correctly predict the classes of the testing data, when the model was built on the training sample.

1. **Weekly Optimization Execution:**

   Execute the optimization process weekly, adjusting the weights based on the previous weeks' results to improve fairness over time.

2. **Evaluation and Adjustment:**

   At the end of a predefined period, evaluate the outcomes in terms of both the number of successful exits from homelessness and fairness across sub-populations.

   Adjust the weight (*C*) based on the evaluation to better balance the trade-off between maximizing successful exits and ensuring equitable treatment allocation.

3. **Long-term Implementation and Monitoring:**

   Implement the optimized assignment process as a standard operational procedure.

   Continually monitor and adjust the weighting factor to respond to changes in the population or intervention effectiveness, ensuring ongoing optimization of both exit rates and fairness.

References

De La Bourdonnaye, F., & Daniel, F. (2021). Evaluating categorical encoding methods on a real credit card fraud detection database. arXiv preprint arXiv:2112.12024.

De Sousa, T., Andrichik, A., Prestera, E., & Rush, K., Tano, C., Wheeler, M., & Abt Associates (2023). The 2023 annual homelessness assessment report (AHAR) to Congress. *The US Department of Housing and Urban Development*

Hill, J., Linero, A., & Murray, J. (2020). Bayesian additive regression trees: A review and look forward. Annual Review of Statistics and Its Application, 7, 251-278.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

Kim, G. (2023, January 26). Ending homelessness in king county will cost billions, regional authority says. *The Seattle Times*.

Kube, A. R., Das, S., & Fowler, P. J. (2023). Fair and efficient allocation of scarce resources based on predicted outcomes: implications for homeless service delivery. Journal of Artificial Intelligence Research, 76, 1219-1245.

Lee, D. (2021). Is need enough? The determinants of intergovernmental grants to local homeless programs. *Journal of Urban Affairs*, *43*(7), 995-1009.

Meyer, B. D., Wyse, A., & Logani, I. (2023). *Life and death at the margins of society: the mortality of the US homeless population* (No. w31843). National Bureau of Economic Research.

U.S. Department of Housing and Urban Development (HUD) (2024). Biden-Harris Administration Awards $3.16 Billion in Homelessness Assistance Funding to Communities Nationwide. HUD No. 24-018