Predicting the Slide to Long-Term Homelessness: Model and Validation

Sandeep Purao
Info & Process Management
Bentley University
Waltham, MA, USA
orcid.org/0000-0002-9101-1105

Monica Garfield
Computer Info Systems
Bentley University
Waltham, MA, USA
mgarfield@bentley.edu

Xin Gu
Mathematical Sciences
Bentley University
Waltham, MA, USA
gu_xin@bentley.edu

Prakash Bhetwal
Computer Info Systems
Bentley University
Waltham, MA, USA
bhetwal_prak@bentley.edu

***Abstract*—In spite of numerous programs and interventions, homelessness remains a significant societal concern. Long-term homelessness is particularly problematic because it can be increasingly difficult to escape from, and because it represents a continuous drain on societal resources. This paper develops a model for predicting long-term homelessness in response to a simple question: if an individual becomes homeless, what influences the individual's slide to long-term homelessness? The data we analyze to answer the question comes from the City of Boston. The model points to race, veteran status, disability, and age as key factors that predict this slide. The paper describes and illustrates the model along with problems encountered in data preparation and cleansing, prior scholarly work that helped to shape our decisions, and collaboration with participants in the ecosystem for homeless care that complemented the model-building effort. The results are important because they point to possible policy interventions (programs and funding) and process improvements (at homeless shelters) to mitigate this slide.**

Keywords*—homelessness, long-term homeless, big data, prediction model

I. Introduction

On a single night in 2018, roughly 553,000 people were experiencing homelessness in the United States [1]. The number of individuals with "chronic patterns of homelessness" increased by two percent between 2017 and 2018 [1]. In spite of national attention to the problem and investments of time and effort by scholars, understanding patterns of homelessness remains difficult [2]. The problems can be traced to the fragmented ecosystem responsible for caring for the homeless [3, 4], practices for data gathering dictated by peculiar requirements and carried out by often overburdened volunteers who cannot assess the downstream benefits [5], obstacles to data sharing across agencies and institutions that are within the ecosystem [6, 7], and problems related to terms and definitions (tied to funding sources and policies) which can lead to some dysfunctional decisions. In the presence of these obstacles, the promise of data analytics to better understand specific concerns related to homeless can be difficult to pursue.

Homelessness is a significant concern in terms of its scale [8]. Being homeless is a condition that takes a toll on the individual, and also presents difficult challenges for the society [9]. A number of problems can be identified against this backdrop such as health and crime costs [9], infrastructure and policies needed to integrate data across multiple systems [6], dependency on government funded housing [10] and estimating an individual's vulnerability to homelessness [11]. In response, much scholarship has tried to address different aspects of the problem including practices for data gathering [12], approaches to data analysis [7], developing intervention possibilities [4], and others. The specific focus of our work is to examine the possibilities afforded by the data that is already gathered within the ecosystem of care for the homeless in spite of known problems.

The goal of this work is, therefore, to take first steps towards understanding a specific concern related to homelessness: the slide to long-term homelessness. We define the term 'long-term homelessness' as an experience for an individual who has been homeless (during a given year, defined as a continuous period of 365 days) for a significant fraction of time. For the purpose of this paper (without loss of generality), we operationalize this significant fraction as at least one out of three days, i.e. (365/3 = 122 days of the year), tying this operationalization to initial analysis of data available. The definition allows us to focus on those individuals in our data set that have the most number of days of homelessness (and hence, use the most number of bed nights at the shelters). Our definition is different from the oft-quoted term 'chronically homeless.' The AHAR report [2018] defines that term as: "an individual with a disability who has been continuously homeless for one year or more or has experienced at least four episodes of homelessness in the last three years where the combined length of time homeless in those occasions is at least 12 months." Our definition of long-term homelessness also bears a similarity to this definition in that it uses the same fraction (one out of three) to identify individuals who spend the greatest number of days in a shelter but it is different in that we use a narrower band (one year, a continuous period of 365 days) instead of using a three year window.

Understanding the transition to long-term homelessness (as we define it) is important for several reasons. First, societal resources to address homelessness continue to be directed at prevention, often targeted at specific target populations [1]. However, we do not yet have a nuanced understanding of what might cause individuals to slide into long-term homelessness. This knowledge can help shape policies and programs. More specifically, our characterization can then be used to promote

978-1-7281-0650-2/19/$31.00 ©2019 IEEE
DOI 10.1109/CBI.2019.00011

action (e.g. detecting a slide to long-term homelessness can be useful for anticipatory interventions that prevent the transition to becoming chronically homeless) [13], Second, although the scholarly community seems to have coalesced around a definition of chronically homeless, in practice, it appears to be quite difficult to track and use, given the problems related to data capture and data sharing. The definition we propose and the computations we describe have the potential for application with existing practices. Third, discovering contributors to this slide to homelessness can pinpoint specific target populations as well as opportunities for practices that can improve the effectiveness of care programs for the homeless.

Recent point-in-time counts show the numbers of homeless as NY: 78,676; LA: 49,955; and Seattle: 12,112 [1]. Our focus is the city of Boston (6,146 homeless, City of Boston, 38th Annual Homeless Census) because the City has made strides in getting the scale of the problem under control, which now allows us to focus on understanding the problem in a more nuanced manner. Our work relies on data contributed by the City of Boston / Pine Street Inn and covered the period from Jan 2014 to May 2018. We describe how the work progressed, including efforts to better understand the peculiar characteristics of data and meta-data available (based on visits to the homeless shelter), challenges encountered for cleansing and validating the data available, and building multiple models for understanding factors that contribute to the slide to long-term homelessness. The efforts were influenced by a review of prior work in research related to homelessness.

The key contribution of our work is a model that discovers and points to four factors – race, veteran status, disability, and age – as key influences on the transition to long-term homelessness. The paper describes and illustrates the model along with problems encountered in data preparation and cleansing, prior work that helped to shape our decisions, and collaboration with participants in ecosystem for homeless care that complemented the model-building effort.

The reminder of the paper is organized as follows. Section 2 reviews prior work. In section 3, we describe the research setting and the research approach. Section 4 outlines and illustrates the model along with the validation efforts. In section 5, we discuss implications of our work and next steps.

II. Prior Work

A. Homelessness: Chronic and Long-term

Being homeless is not a condition, it is something you experience. The term 'homeless' describes "a person who lacks a fixed, regular, and an adequate nighttime residence" [1]. The definition from HUD (2012) clarifies the term as unsheltered persons occupying a "place not designed for ... sleeping accommodation for human beings." Regulatory, technological and political obstacles make lasting solutions for combatting homelessness difficult [6]. The importance of the problem can be traced to The United National Universal Declaration of Human Rights (Article 25), which states: "Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care." The same AHAR report (2018), which reported more than half a million people experiencing homelessness in the US also found that homelessness increased for the first time in seven years.

Homelessness is a multifaceted issue that requires a range of agencies (shelters, hospitals, correctional facilities) in order to respond to the many concerns of the homeless community [14]. Much prior scholarship identifies four categories of homeless people. These include: literally homeless, at imminent risk of homelessness, homeless under federal statutes, and fleeing / attempting to flee domestic violence. The categories are important because they establish eligibility for different programs (e.g. individuals in category 2 are not serviced by street outreach or rapid rehousing)[1]. Other categorizations of homeless people also impact their eligibility for various programs. For instance, there are programs that target American veterans, minors, families, the disabled, and the chronically homeless. A chronically homeless individual is defined as someone who is continuously homeless for a year or more or has 4 or more episodes of homelessness adding to a year or more in the previous three years. For some people they are not categorized as chronically homeless because they perhaps stayed at shelters 3 times in 3 years for a total of 12 months, whereas they would be considered chronically homeless if they stayed at shelters 4 or more times in 3 years and their total stays accumulate to 12 months. Furthermore, how one is classified in terms of disabilities also impacts chronic homeless designation (i.e. the individual needs documentation of their disability, which many homeless do not have or wish to have). The chronically homeless typically account for about 10% of the homeless population and consume about 50% of the shelter bed space yearly [15]. Our work focuses on the long-term homeless so that the focus is not merely meeting the HUD definition for the purpose of counting, but rather, for the purpose of identifying the long-term homeless more quickly, so that this can lead to more effective action at different points within the ecosystem of care.

B. Causes of and Contributors to Homelessness

Causes of and contributors to homelessness vary. A prominent and often-cited source, the U.S. Conference of Mayors' Report on Hunger and Homelessness [16] points to several factors that can cause or contribute to homelessness. These include: lack of affordable housing, unemployment, poverty, low wages, mental illness and the lack of needed services, and substance abuse and the lack of needed services. For women, domestic violence continues to be a another leading case of homelessness [17]. Another study [18] estimates that at least 30% of homeless women in Minnesota reported becoming homeless due to domestic violence.

Scholarship related to causes of homelessness can be placed in two categories: one uses macro level data at a city or state level and ones that use individual level data [19]. Regardless of the starting point, scholars agree that the cases of and contributors to homelessness are rarely single variables but rather a range of variables with varying degrees of importance

[1] https://www.hudexchange.info/resources/documents/HomelessDefEligibility%20_SHP_SPC_ESG.pdf

[20]. The emphasis on a single variable can, however, be traced to a trigger or tipping point (on the verge of homelessness or recidivism of homelessness), which may vary between individuals in terms of the impact a particular issue has on them (mental illness, poverty, lack of social supports, job status, education level, etc.) and their resiliency to withstand varying degrees of obstacles they face. These triggers may be unidimensional (e.g. acute physical issues that require immediate health interventions) or multidimensional (e.g. foster care history, joblessness that may need training, substance abuse interventions, education and other support inventions to enable one to obtain and hold a job) [11, 21] Shah et al. 2017).

Once one becomes homeless, different variables impact the intensity of homelessness (the number, frequency or percent of possible nights an individual stays on the streets). One study found that gender, family support, veteran status, combat status and education level had the largest impact on the number of nights an individual stays on the streets [19]. Five demographic factors (age, family size or type, race and ethnicity, pregnancy status, employment status, citizenship status and receiving public assistance) have been cited as the most significant predictors of shelter readmission [22]. Other studies reassert that demographic characteristics and housing conditions were the most significant risk factors affecting shelter reentry, with enduring poverty and disruptive social experiences also important conditions [23]. This selective survey allows us to characterize prior research in two important ways that lends support to our work. First, the intensity of the homelessness experience can be conceptualized in different ways. Our work, therefore, provides an important extension to these efforts the HUD definition as a starting point to derive and compute what we call long-term homelessness. Second, results from prior work demonstrate the importance of searching for contributors to the intensity of homelessness, which our work explores as well.

C. Using Data Analytics to Understand Homelessness

Gathering and analyzing data remains difficult in the context of homelessness, as it does for several other non-profit contexts. A particular area where scholars and government agencies have often focused is understanding the scale and incidence of homelessness. To explore this, much funding and program assessment relies on Point in time counts that generally can take place across locations on a selected single night in January {Henry, 2018 #1}. On this night, volunteers document the number of sheltered and unsheltered homeless persons in a geographic area. The sheltered homeless count is captured via head counts in homeless shelters, safe havens and transitional housing. For counts beyond the homeless shelters, volunteer groups identify visit places where the homeless tend to congregate (such as parks, bridges, semi-sheltered areas). Scholars have, however, shown that large numbers (in the case of New York, 31% and by some estimates 2.5 to 10.2 times the number of homeless may go unaccounted in PIT counts) may choose places classified as not visible, at least partially because of laws that criminalize homelessness [24], be staying in a hotel for a single night, or “doubling” up in a home {Ellickson, 1990 #57}. Prior work also points to several other issues related to using point in time counts [12] [2].

Beyond such counts, data analytics has been used to detect, diagnose and monitor homelessness [14, 25] . These efforts have shown that there remain a number of issues that may hamper our ability to do so. One such issue is the manner in which the data is collected (as described above). Another is the impact of missing data in our ability to analyze data sets related to homelessness [26], Still another issue is that the intake forms used by the various shelters in one continuum of care are not all the same and the data is not collected in similar methods across the various shelters [19]. Much of the past work has also been hampered by common data related issues such as dealing with incomplete data [27] and limited historical data [28].

In spite of such problems, prior work has attempted to apply data analytics in a variety of ways, including the development of predictive models for identifying homeless persons who are likely to become high-cost users of public services [29-32] relying on a database that contains administrative records, which provide information on risk factors such as demographics, clinical variables and service utilization variables for the current and previous years as well as cost of service data. Similar to these efforts, our work relies on administrative records related to entries, exits and demographics – data we obtain from the City of Boston. We acknowledge that our efforts will, therefore, face some of the same challenges as in prior work. However, by freeing ourselves of definitions that force specific types of measurements (i.e. the HUD definition of chronically homeless), we are able to let the data speak for itself as we investigate how various reported characteristics of homeless individuals predict the likelihood that once that individual has been documented by a shelter, they will slide into long- term homelessness.

III. Research Approach

A. Setting

The context for our work is the City of Boston, and Pine Street Inn (PSI), the largest homeless shelter in New England. To appreciate the scale of the problem, we return to the 2018 Annual Homeless Assessment Report [1], which documents that the total homeless population in Massachusetts in 2018 was estimated to be 20,068. With a 14.2% increase from 2017, Massachusetts had the sixth largest population of homeless individuals across all states. More specifically, according to the City of Boston Annual Homeless Census, the number of homeless individuals in the City of Boston numbered 6,146 in 2018. This represented a three percent drop from 6,327 in 2017. The data also showed that in 2017, emergency shelters in the City of Boston provided shelter to 5,331 individuals (of the 6,327 homeless).

Emergency shelters are part of the ecosystem of care for the homeless, described as the continuum of care (CoC), and represent the first line of defense for the homeless {Poole, 2003 #66}. Pine Street Inn (PSI) is the largest homeless shelter in New England, and the largest provider of emergency services to homeless individuals in the City of Boston. PSI provides street

[2] https://www.nlchp.org/HUD-PIT-report2017

outreach, front door triage, and emergency shelter as part of its emergency services. Of these three services, emergency shelter is the largest one with 670 beds across four shelters around the city of Boston that PSI operates. Once an individual arrives at PSI in search of shelter, PSI provides beds, meals, programs, and services to transition the homeless out of homelessness. This is done through programs such as permanent housing, workforce development, legal support and substance recovery services. PSI provides these services as part of the ecosystem of care (the continuum of care, CoC) for the homeless in the city of Boston. As a part of this role, it must routinely collect data about the homeless, and coordinate with the city officials as well as national efforts for understanding as well as helping the homeless.

B. *Data Access and Data Cleansing*

The data we relied on for this research is the data that PSI and the City of Boston collect, primarily for the purpose of reporting to organizations such as HUD. This is important because various programs that provide funding for the homeless require the collection and reporting of such data. This includes data on the use of emergency services in the City of Boston. The dataset we obtained from the City of Boston dataset, therefore, covers the data on emergency services. The complete data set covers about 4.5 years, from the beginning of 2014 till May 2018. To contextualize the data, the research team visited PSI Men's Inn and observed the data collection and service processes. In-depth interviews and ongoing discussions with PSI executives further complemented our understanding of the veracity of the data and surfaced both, questions and concerns of interest, that could be explored with the dataset.

The dataset did, however, require significant cleansing and preparation to make it ready for analysis and modeling. Many issues were encountered during the process. Incomplete data and empty data points were the biggest challenges. The reasons for this problem were several, traced to the data gathering process at the shelter. First, the intake step at PSI did not require the individuals to answer questions related to demographics or those in the HUD assessment. Second, as we learned from the interviews and discussions, sometimes the individuals themselves did not know the answers to some questions. A third reason was human error during data entry because of different people, including volunteers, who entered the data in the database and their ability to understand, appreciate or follow through on the need for accurate data collection varied significantly.

Consider an example: During the intake process, when an individual guest responds that s/he does not know about a certain health condition, some volunteers followed the practice of leaving the question blank while others followed the practice of coding it as 'client does not know.' Both could be problematic but the former would lead to missing values in the dataset. There were other problems as we inspected the data and found examples that appeared to be errors. For example, some individuals had reported that their Social Security Income was above $75,000 per month. Other problems included clear inconsistencies in the data. For example, there two fields in the dataset recorded income/benefit information. One is a binary variable that indicates whether the individual has income/benefit (1) or not [23]. The other is the amount of income (per month in dollars). For some individuals, information in these two fields was in conflict. For example, in our data, an individual with no income (the value of the binary variable recorded as 0) was accompanied by an income amount (dollars per month) in the other field.

To better understand and resolve these problems, we worked with the PSI executives. Our discussions included specific examples and exploring possible approaches we could use for resolving these problems, informed by the insights from the PSI executives who contributed deep understanding of complexities surrounding the data collection processes. Based on these, the research team could make decisions about how to address these problems in the dataset. For example, in some cases, the research team was able to derive rules to detect and fix flawed data. This was done in a systematic way, first for each attribute, and then, based on combinations of attribute (such as the example described above). This sometimes required correcting the data specific records, and in other cases, removing some of the attributes or combinations of attributes from consideration because of the noise. For example, after much discussion with the PSI executives, the research team decided not to use the income attribute for building a predictive model but used it as a reference to make the binary income condition variable more accurate. To do this, simple rules were created to cleanse the data (e.g. if an individual shows no income, i.e. value of binary variable as 0, but reports an income amount per month, i.e. value of income greater than 0, then change income condition value from 0 to 1, and vice versa). Other examples of data cleansing included examining entry and exit records (e.g. exit preceding entry) were investigated for possible errors. These efforts allowed us to ensure possible errors were detected and fixed before moving to analysis and model-building.

C. *Generating Predictive Models*

Based on this data set we began to build our predictive model. Predictive modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining that works by analyzing historical and current data and generating a model to help predict future outcomes. In predictive modeling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available [33].

A predictive model uses factor/s to predict the outcome of a target variable. With the goal of predicting the likelihood of an individual using the emergency services to become a long-term homeless individual, we identified our target variable as the length of homelessness. The length of homelessness of the historical shelter users was converted into a binary variable representing long-term users and non-long-term users. To identify factors that influence the length of homelessness, and use them as input factors in our model, we investigated both, prior research as well as attributes available in our dataset. Exploratory analysis of the available dataset was done, and further input factors were identified. Because of the binary nature of our target variable, we used a logistic regression technique for building the predictive model. We started building the predictive model for the available dataset using the factors identified in prior work and iteratively incorporating other factors that we found significant from our exploratory analysis. To explore possible improvements, we performed feature engineering by converting some of our input factors into binary variables. The predictive model including the analysis on the significance of different factors and the correlation with the target variable was shared with the PSI team. Upon receiving feedback on the findings, the model was finalized and checked for consistency.

IV. Model Development

The model development started with an analysis of the data to understand the patterns and identify constructs of interest. The analyses relied on records of 22,693 individuals for whom entrances and exits into the CoC were recorded by Pine Street Inn between [2014.01.01] and [2018.05.31]. We used the complete set of data records, focusing on specific attributes such as Enrollment Entry Date (the earliest entry date of all enrollments for an individual) and Enrollment Exit Date (the latest exit date of all enrollments for an individual) to determine when each individual entered into or exited from the shelter. A number of other attributes (such as Personal ID, Enrollment ID and Project ID) were used to identify a particular individual so that entrances and exits could be matched. We encountered several problems during this analysis including duplicate records for an individual, with conflicting information. For example, for one Personal ID (one individual), sometimes we encountered several records, with the race, DOB and other demographic and health information different in each record. This made the identification of individual cases challenging. We resolved these by using date for record creation and update to determine the most recent records and retained the newest information to minimize data error. There was also a large number of missing values in the dataset. For example, 24.88% of the individuals had missing disability condition information, and 16.77% of all individuals did not contain veteran information. Imputing these variables is beyond the scope of our research. To ensure that we based our model on data that would be reliable, we removed those records, where critical values were missing. The analysis proceeded with this dataset.

A. Computing Length of Homelessness

Initial results from the analysis showed that there were 22,693 records of homeless individuals with at least one completed homelessness episode. Each episode was defined as one entry date and one exit date. The data showed that there were multiple episodes for some individuals. We first extracted the length of each enrollment (without overlap) for every individual, then we summed the length of different enrollments together as the total length of homeless for each individual during the four years and five months research window. Obviously, some individuals became homeless earlier than others, for example, some individuals came into the system from 2014 while some from 2017. To account for how long the individual was in the system during our window of analysis, we used the total number of days that an individual stayed in the system as the numerator (see Equation 1 below with variations).

$$avg\ (length\ (homeless_k)) = \sum length\ (episode_{ik}) \ / \ \{time\text{-}periods\} \qquad [1A]$$

$$= \sum length\ (episode_{ik}) \ / \ \{episodes\}\{Poole,\ 2003\ \#66\}\{Poole,\ 2003\ \#66\} \qquad [1B]$$

$$|\ i=1..n,\ k=1..K$$

where i refers to episodes, k refers to individuals.

Figure 1 shows the distribution of the average length of homelessness per year for the 22,693 homeless individuals following equation [1A]. The range of average length of homelessness for all individuals was from 0 days to 359.86 days per year with a mean value of 65.61 days per year. For a majority of the individuals (75.1%) the average length of homelessness per year was less than 100 days.

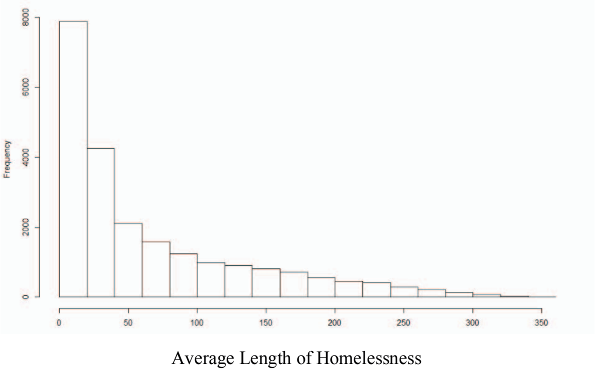


Fig. 1. Average length of homelessness per year (22,693 cases)

These numbers directly suggest that some individuals (with a longer average length of homelessness per year) were the recipients of more bed nights from the emergency shelter. This, initial analysis, pointed out that (as might be expected) there were large disparities in how the homeless shelter services were used by different individuals. One simple interpretation of this result is that greater public services and dollars are spent on individuals with longer length of average homelessness lengths. The question of interest for the research team could, then, be stated as the following. If we assume that the initial entry into experiencing homelessness is traced to misfortune or factors beyond the control of the individual, what influences their slide into long-term homelessness? To explore this question, we returned to the data and applied additional statistical analyses techniques to uncover factors that would influence this slide to long-term homelessness.

B. Exploring Demographic Characteristics

This exploration started by identifying possibilities available in the dataset such as demographic characteristics, socio-economic information and health indicators (including disability types) for every individual. No initial filtering was used to discard any of the attributes. Instead, we cast a wide net to explore all possible factors that may have a correlation with long-term homelessness. Table 1 below lists the factors that we considered as potential independent variables (excluding missing values, and values of client refusing to answer and client doesn't know):

TABLE I. POTENTIAL INDEPENDENT VARIABLES

Variables Description and Type
(AI) American Indian: Is the individual American Indian (native)? (*B*)
(AS) Asian: Is the individual Asian? (*B*)
(AF) African American: Is the individual Black /African American? (*B*)
(PI) Pacific Islander: Is the individual Native HI/ Pacific Islander? (*B*)
(WH) White: Is the individual White? (*B*)
(HI) Hispanic: Is the individual Hispanic? (*B*)
(G) Gender: What is the individual's gender? (Is it Male?) (*B*)
(V) Veteran: Is the individual a veteran? (*B*)
(Age) Date of Birth: What is your DOB? (to compute Age)
(Dom) Domestic Violence Victim: Is the individual a Victim? (*B*)
(Inc) Income: Does the individual have income? (*B*)
(Sub) Substance Abuse: Was there substance abuse? (converted to Ordinal)
(Dis) Disabled: Does the individual have a disability? (*B*)
(DisT)Disability Type: What is the disability type? (converted to *B* for each)

(*B*: Binary variable; *Gender*: Female, Male, Transfemale, Transmale, Does not identify as male, female or transgender; *Substance abuse*: No, Alcohol abuse, Drug abuse, Both alcohol and drug abuse; *Disability type*: Physical, Developmental, Chronic condition, HIV, Mental health)

TABLE II. DIFFERENCES LONG- VS. SHORT-TERM GROUPS – A

Group	***Variables (see Table I)***						
	AI	**AS**	**AF**	**PI**	**WH**	**HI**	**G**
Short-term	0.011	0.016	0.397	0.013	0.536	0.207	0.750
Long-Term	0.013	0.014	0.485	0.009	0.468	0.205	0.738

TABLE III. DIFFERENCES LONG- VS. SHORT-TERM GROUPS - B

	Variables (see Table I)				
Group	**V**	**Age**	**Dom**	**Inc**	**Sub**
Short-term	0.063629	42.32759	0.17763	0.472659	0.61715
Long-term	0.053934	47.96416	0.184598	0.628301	0.773144

TABLE IV. DIFFERENCES LONG- VS. SHORT-TERM GROUPS - C

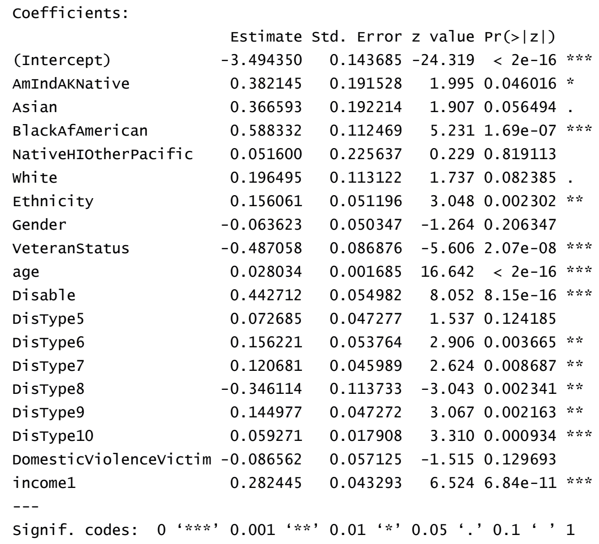
Group	**Variables (see Table I)**						
	Dis	**DisT5**	**DisT6**	**DisT7**	**DisT8**	**DisT9**	**DisT10**
Short-term	0.62	0.306	0.142	0.344	0.029	0.512	1.064
Long-term	0.77	0.443	0.200	0.481	0.032	0.623	1.230

Building on results from the initial data analysis, we reasoned as follows. Because the majority of individuals (75.1%) had an average length of homelessness per year that was under 100 days; we identified an individual as long-term homeless if he/she was homeless longer than that number. We experimented with different values for this definition using 100 days as the anchor. For the purpose of results described in this paper, we used the heuristic of 1 in 3 days, i.e., we classified an individual as long-term homeless if his/her average length of homelessness per year was greater than 1 in 3, so, greater than 121.67 days (365/3) for the year. Individuals with an average length of homelessness less than that were then, considered not long-term homeless. The results we describe including the model for predicting the slide into long-term homelessness are tied to this definition. However, we emphasize that our experiments with varying this value produced similar predictive models, albeit with different parameters for the factors. The model development then progressed by treating the experience of long-term homelessness as a binary variable; coding individuals who were experiencing long-term homelessness as 1, and others as 0.

Following this reasoning, 4,566 individuals were classified as experiencing long-term homelessness (Long-term Group). This accounted for 20.12% of the total number of individuals. The rest, 18,127 individuals (79.88%) were considered as experiencing short-term homelessness (Short-term Group). To understand and assess whether and how the two groups were different, we performed several tests to examine the differences across several characteristics, at the significance level of 5%. Tables II through IV below show the proportions of individuals with different conditions for each group and the average across the two groups. We can see that the Long-term group has more American Indian, Black/African American, individuals with disability, individuals who are domestic violence victims and individuals who report having some income. The average age of individuals in the Long-term group is greater than those in the Short-term group and the Long-term group has a higher proportion of female than the Short-term group.

C. Developing a Predictive Model

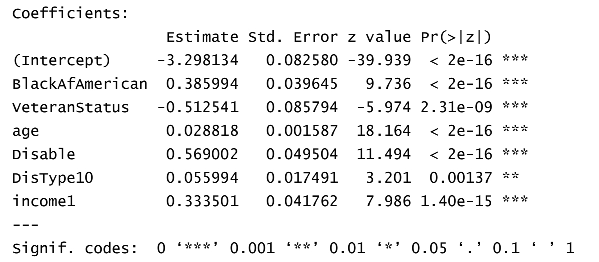
Following the initial analysis (extracting the variables and exploring differences across the short-term and long-term groups), we built a logistic regression model. As we described above, we only used the data of those samples without critical missing values, the sample size for this model is 15,667. The model was intended to identify factors correlated with the dependent variable of interest: the probability of an individual being a long term homeless, p, from 0% to 100% given the status of the individual as a homeless individual. In other words, we modeled the log-odds of the probability of the individual being in the Long-term Group $log\ (p/(1\text{-}p))$ based on the presence or absence of the independent variables listed in the tables above. Figure 2 shows the initial model with all variables. The model showed that variables significantly associated with membership in the Long-term Group (at a significance level below 0.001) included the following: (***AF***) being African American, (***V***) being a veteran, (***Age***) being (above a certain) age, (***Dis***) having a disability condition, (***DisT10***) having substance abuse disability and (***Inc***) having (reported an) income. Several other factors including belonging to other races (***AS***, ***WH***, ***HI***, ***PI***), being of a certain gender (***G***), having a physical disability (***DisT5***) and being a victim of domestic violence (***Vic***) were not significant at the significance level of 5%.



```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -3.494350   0.143685 -24.319  < 2e-16 ***
AmIndAKNative            0.382145   0.191528   1.995 0.046016 *  
Asian                    0.366593   0.192214   1.907 0.056494 .  
BlackAfAmerican          0.588332   0.112469   5.231 1.69e-07 ***
NativeHIOtherPacific     0.051600   0.225637   0.229 0.819113    
White                    0.196495   0.113122   1.737 0.082385 .  
Ethnicity                0.156061   0.051196   3.048 0.002302 ** 
Gender                  -0.063623   0.050347  -1.264 0.206347    
VeteranStatus           -0.487058   0.086876  -5.606 2.07e-08 ***
age                      0.028034   0.001685  16.642  < 2e-16 ***
Disable                  0.442712   0.054982   8.052 8.15e-16 ***
DisType5                 0.072685   0.047277   1.537 0.124185    
DisType6                 0.156221   0.053764   2.906 0.003665 ** 
DisType7                 0.120681   0.045989   2.624 0.008687 ** 
DisType8                -0.346114   0.113733  -3.043 0.002341 ** 
DisType9                 0.144977   0.047272   3.067 0.002163 ** 
DisType10                0.059271   0.017908   3.310 0.000934 ***
DomesticViolenceVictim -0.086562   0.057125  -1.515 0.129693    
income1                  0.282445   0.043293   6.524 6.84e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 2. Initial Logistic Regression Model (with all variables from the dataset)

A revised model was developed by removing variables that were not considered significant, the final model is shown below.



```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -3.298134   0.082580 -39.939  < 2e-16 ***
BlackAfAmerican  0.385994   0.039645   9.736  < 2e-16 ***
VeteranStatus   -0.512541   0.085794  -5.974 2.31e-09 ***
age              0.028818   0.001587  18.164  < 2e-16 ***
Disable          0.569002   0.049504  11.494  < 2e-16 ***
DisType10        0.055994   0.017491   3.201  0.00137 ** 
income1          0.333501   0.041762   7.986 1.40e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 3. Logistic Regression Model (after removing not significant variables)

The final model results showed that (***AF***) being African American, (***V***) being a veteran, (***Age***) being (above a certain) age, (***Dis***) having a disability condition, and (***Inc***) having (reported an) income are still significant at significance level of 0.001, while (***DisT10***) having substance abuse disability is now only significant at level of 0.01. This means that, the confidence of the impact of other significant variables on long-term homeless is higher than that of ***DisT10***. However, no matter the variables are significant at level of 0.01 or 0.001, they are all considered to be valid results to show the correlation between these variables and long-term homeless.

The results can be interpreted in the following manner. Holding other variables constant: if an individual is a veteran, the probability of belonging to the Long-term Group (being long-term homeless) is lower, if an individual is African American or disabled, the probability of belonging to the Long-term Group (being long-term homeless) is higher. Further, among individuals with disability, if the individual has substance abuse problems, the probability of belonging to the Long-term Group (being long-term homeless) is higher. As the individual gets older (as age increases), the probability of belonging to the Long-term Group (being long-term homeless) increases. If the individual has reported income, the probability of belonging to the Long-term Group (being long-term homeless) is also higher. Intuitively, this result seems not to make sense. However, it may be due to the fact that much of this income may be coming from Social Security Disability Insurance. This produces high correlation between the factors ‘has disability’ and ‘reports income,’ and therefore, individuals with disability and reporting income tend to have a higher probability of being long term homeless.

D. Model Validation

To validate the model, we constructed a receiver operating characteristics (ROC) curve. The ROC curve shows the performance of our predictive model, which essentially operates as a classification model that allows us to place individual instances in one of two groups (short-term group vs. long-term group) and tracing this membership to independent variables (see Table I above). The performance of our classification model can, then, be assessed by examining the True Positive (Recall) and False Negative (Precision) at different classification thresholds. When plotted, the ROC curve allows computation of the area under the curve (AUC) AUC as an aggregate measure of performance across all possible classification thresholds, i.e., the probability that the model ranks a random positive example (membership in the Long-term group) more highly than a random negative example. In effect, AUC provides an indicator that shows the efficiency of the model. The AUC for the model we constructed is shown below, which shows AUC as 0.6364. The closer to 1 the AUC is, the better the model. If the AUC is 0.5, that means the model is equivalent to random selection. If the AUC is greater than 0.5, that means the model is better than random selection. The AUC for our model, therefore, indicates that our model is 13.64% more efficient than random selection in general.

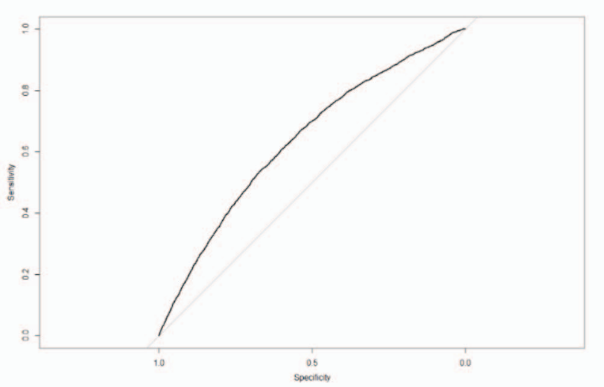


Fig. 4. Computing AUC for Validation

Since our model has better performance than random selection into the two groups, we can use it to identify the individuals with higher probability of belonging to the Long-term Group with all the independent variable information.

E. Use Scenario

For example, if there are two homeless individuals we want to know who is more likely to become a long-term homeless in future. Without any model, we have to use domain knowledge to make an educated guess. However, with our model and the information of independent variables in our model, we can calculate the probability with confidence. The process can be elaborated as follows:

With the model in place, we can envision the following use scenario. Consider Jay, an individual who shows up on Tuesday at 10am at the PSI. Jay happens to be a Black/African American, 60 years old, has a disability, reports having an income, and reports no substance abuse. With this information, the volunteer helping Jay with the intake may use the predictive model to compute the log odds that Jay would belong to the Long-term Group is he is admitted into the shelter. This computation would be -0.280557 (-3.298134 + 0.385994 + 0.028818 *60 + 0.569002 +0.333501). The model would suggest that the probability that Jay would belong to the Long-term Group is 43.03%. Armed with this knowledge, the volunteer would be able to make an informed decision about possible interventions for Jay. Consider Emily who walks in at 1030am at PSI. She is 40 years old, happens to be Asian, i.e., not Black/African American, a veteran, without disability and reports no income. With these attributes, the model computes the log odds of Emily of belonging to the Long-term Group as -2.657955 (-3.298134-0.512541+0.028818*40). The model computation suggests that the probability that Emily would belong to the Long-term Group is 6.55%. The volunteer helping with the intake at PSI is able to better understand and distinguish the possible trajectories of the two guests. Although individual circumstances will dictate the actual paths, the predictive model will suggest propensities that the PSI volunteers can take into account. Based on the ROC (see Figure 4), PSI volunteers can have greater confidence (13.64% higher than a random decision) that the second individual has lower probability of becoming long-term homeless.

V. Discussion and Implications

The model we have developed shows that understanding the likelihood of 'slide into long-term homelessness' may provide useful pointers for (a) prioritizing efforts to ensure that individuals in certain target groups are given more attention to prevent such a slide, or (b) selecting or devising different interventions for individuals in certain target groups may prevent such a slide. Use of the model can, thus, enhance the triage process in a shelter, where the intake of new individuals includes decisions about how to best provide the services needed. The model may also be used to evaluate, retroactively, decisions made over a period of time (which may, in turn, can lead to revisions to the model parameters).

It is important to place the work we have described in the context of other, comparable efforts. One of the works we reviewed earlier [15] reported that individuals who are homeless for long periods (in their case, the chronically homelesss) account for about 10% of the homeless population and consume about 50% of the shelter bed space yearly. Other scholars have found that five demographic factors (age, family size or type, race and ethnicity, pregnancy status, employment status, citizenship status and receiving public assistance) were the most significant predictors of shelter readmission [22]. Other studies reassert that demographic characteristics and housing conditions were the most significant risk factors affecting shelter reentry, with enduring poverty and disruptive social experiences also important conditions [34]. A comparison of our model with these prior studies suggests two possibilities. One may be that the landscape has shifted since these studies were conducted, and therefore, our model may point to more relevant variables today. A second possibility is that the models may be tied to specific contexts (these prior efforts were carried out in different cities), and it may be more appropriate to use results that reflect the context in different cities. This remains an open question.

Prior scholarship has also tried to examine the issue of resource allocation in a more effective manner, both for the purpose of responding to the specific and peculiar needs of each individual; as well as ensuring effectiveness of programs based on resources spent in aggregate. In a recent systematic review on the effectiveness of interventions to reduce homelessness only 43 studies (published in 78 articles) were found that used randomized control trials of interventions and their impact on homelessness at least one year after the intervention [35]. This work found that interventions that perform the best to reduce homelessness included efforts such as: high intensity case management, housing first, critical time intervention, abstinence-contingent housing, housing vouchers, residential treatment, and some combinations of these programs. At the same time, it is important to acknowledge the specific and peculiar needs of each individual who may need these services. This requires identifying effectiveness of programs provided and resources spent in order to make more informed decisions about what interventions are most likely to help a homeless individual get out of homelessness faster and helps the City tackle the homelessness in better way. Individuals who do not qualify for specified programs but can be supported with flexible resources can benefit the most with identification of alternative interventions that are more effective to specific group of population.

We acknowledge that our work has some limitations. First, the data we have drawn represents a slice in time (four and a half years), and tied to a single city. It is possible that the significance of different factors will vary if the data were to be drawn from different population centers and/or for different time periods. The exercise so far, and the predictive model we have shared in this paper, however, point to significant possibilities. The model, along with the potential use scenario we have outlined points to use of data analytics not only for the purpose of maximizing resource efficiency but also for addressing specific problems that an individual may face if they are unfortunate to experience homelessness. The use scenario also illustrates the anticipatory manner in which the model can be used to identify and work with at-risk individuals, those with a greater propensity to slide into homelessness.

Finally, we would like to recommend using predictive model to support decisions on type of intervention and time of intervention rather than to solely depend on historical shelter use or homelessness data. Based on current practice, a homeless person has to qualify to receive interventions like temporary housing or permanent supportive housing. The biggest part of this qualification is historical length of homelessness. When a person becomes homeless and ends up in an emergency shelter, they have to wait, up to years in many cases, to receive permanent form of support. During this time frame, they can cost the city a significant amount of money because of the use of medical services, emergency services, etc. In addition, they are also likely to deteriorate in health and increase substance abuse. Identifying these individuals in advance, therefore, has the potential to both reduce the problem of homelessness and also the cost to the city. What we do not yet know is the different kinds of services that may be available and applicable to each individual based on their characteristics, or even the historical effectiveness of these programs. However, a predictive model such as the one we have outlined here, can be the basis for initial investigations necessary for the choice of appropriate programs for each individual. Exploring these remain on our future research agenda.

Acknowledgment

We acknowledge the ongoing discussions and access to de-identified data from Pine Street Inn. We also acknowledge the time from employees and volunteers at Pine Street Inn who responded to the invitation to participate in the study. The work was supported by a Seed Grant from the Big Data Innovation Network at Bentley University.

References

[1] M. Henry, A. Mahathey, T. Morrill, A. Robinson, A. Shivji, and R. Watt, "The 2018 Annual Homeless Assessment Report (AHAR) to Congress. Part 1: Point-in-time estimates of homelessness," Dec 2018 2018.

[2] R. J. Calsyn and L. A. Roades, "Predictors of Past and Current Homelessness," *Journal of Community Psychology,* Article vol. 22, no. 3, pp. 272-278, 1994.

[3] Z. Glendening and M. Shinn, "Risk Models for Returns to Housing Instability Among Families Experiencing Homelessness.," *Cityscape,* vol. 19, no. 3, pp. 309-330, 2017 2017.

[4] A. Turner and D. Krecsy, "BRINGING IT ALL TOGETHER: INTEGRATING SERVICES TO ADDRESS HOMELESSNESS," *School of Public Policy Publications,* Article vol. 12, no. 1, pp. 1-30, 2019.

[5] M. M. Jones, "Creating a Science of Homelessness During the Reagan Era," *Milbank Quarterly,* Article vol. 93, no. 1, pp. 139-178, 2015.

[6] D. Culhane, "The Potential of Linked Administrative Data for Advancing Homelessness Research and Policy," vol. 10, no. 3, 2016.

[7] D. P. Culhane, "The Cost of Homelessness: A Perspective from the United States," *European Journal of Homelessness,* vol. 2, no. 1, pp. 97-114, 2008.

[8] K. Dittmeier, S. H. Thompson, E. Kroger, and N. Phillips, "PERCEPTIONS OF HOMELESSNESS: DO GENERATIONAL AGE GROUPS AND GENDER MATTER?," *College Student Journal,* Article vol. 52, no. 4, pp. 441-451, 2018.

[9] M. McLaughlin and M. R. Rank, "Estimating the Economic Cost of Childhood Poverty in the United States," *Social Work Research,* Article vol. 42, no. 2, pp. 73-83, 2018.

[10] R. C. Ellickson, "The homelessness muddle," 1990.

[11] H. Toros and D. Flaming, "Silicon Valley Triage Tool," 2016.

[12] C. Smith and E. Castañeda-Tinoco, "Improving Homeless PointIn-Time Counts: Uncovering the Marginally Housed," *Social Currents* pp. 1-14, 2018.

[13] P. Cloke, Paul Milbourne, and Rebekah Widdowfield, "Making the Homeless Count? Enumerating Rough Sleepers and the Distortion of Homelessness," *Policy and Politics,* vol. 29, no. 3, pp. 259-279, 2001.

[14] V. K. Mago *et al.*, "Analyzing the impact of social factors on homelessness: a Fuzzy Cognitive Map approach," 2013.

[15] R. Kuhn and D. P. Culhane, "Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization," *American Journal of Community Psychology,* Article vol. 26, no. 2, p. 207, 1998.

[16] "The U.S. Conference of Mayors' Report on Hunger and Homelessness," 2016, Available: https://endhomelessness.atavist.com/mayorsreport2016.

[17] V. Tischler, A. Rademeyer, and P. Vostanis, "Mothers experiencing homelessness: Mental health, support and social care needs," vol. 15, no. 3, pp. 246-253, 2007.

[18] (2015). *Homelessness in Minnesota*. Available: http://mnhomeless.org/minnesota-homeless-study/reports-and-fact-sheets/2015/2015-homelessness-in-minnesota-11-16.pdf

[19] J. Jarvis, "Individual determinants of homelessness: A descriptive approach," *Journal of Housing Economics,* vol. 30, pp. 23-32, 2015.

[20] A. J. Levitt, D. P. Culhane, J. DeGenova, P. O'quinn, and J. J. P. S. Bainbridge, "Health and social characteristics of homeless adults in Manhattan who were chronically or not chronically unsheltered," vol. 60, no. 7, pp. 978-981, 2009.

[21] D. Flaming and P. Burns, "Crisis indicator: Triage tool for identifying homeless adults in crisis," 2011.

[22] Y.-L. I. Wong, D. P. Culhane, and R. J. S. S. R. Kuhn, "Predictors of exit and reentry among family shelter users in New York City," vol. 71, no. 3, pp. 441-462, 1997.

[23] C. Kontokosta *et al.*, "Predictors of Re-admission for Homeless Families in New York City: The Case of the Win Shelter Network," 2017.

[24] K. Hopper, M. Shinn, E. Laska, M. Meisner, and J. J. A. J. o. P. H. Wanderling, "Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods," vol. 98, no. 8, pp. 1438-1442, 2008.

[25] M. Johnstone, C. Parsell, J. Jetten, G. Dingle, and Z. Walter, "Breaking the cycle of homelessness: Housing stability and social support as predictors of long-term

well-being," *Housing Studies,* vol. 31, no. 4, pp. 410-426, 2015.

[26] A. E. Montgomery, D. Szymkowiak, J. Marcus, P. Howard, and D. P. Culhane, "Homelessness, Unsheltered Status, and Risk Factors for Mortality: Findings From the 100 000 Homes Campaign," *Public Health Rep,* vol. 131, no. 6, pp. 765-772, Nov 2016.

[27] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[28] N. R. Faria *et al.*, "The early spread and epidemic ignition of HIV-1 in human populations," vol. 346, no. 6205, pp. 56-61, 2014.

[29] C. E. Adair *et al.*, "Outcome Trajectories among Homeless Individuals with Mental Disorders in a Multisite Randomised Controlled Trial of Housing First," *Can J Psychiatry,* vol. 62, no. 1, pp. 30-39, Jan 2017.

[30] A. Basu, R. Kee, D. Buchanan, and L. S. J. H. s. r. Sadowski, "Comparative cost analysis of housing and case management program for chronically ill homeless adults compared to usual care," vol. 47, no. 1 Pt 2, p. 523, 2012.

[31] D. R. Holtgrave *et al.*, "Cost-utility analysis of the housing and health intervention for homeless and unstably housed persons living with HIV," vol. 17, no. 5, pp. 1626-1631, 2013.

[32] M. E. Larimer *et al.*, "Health care and public service use and costs before and after provision of housing for chronically homeless persons with severe alcohol problems," vol. 301, no. 13, pp. 1349-1357, 2009.

[33] Z. H. Delen D, "The analytics paradigm in business research," *Journal of Business Research,* vol. 90, pp. 186-195, 2018.

[34] M. B. Shinn, D. R. Rog, and D. P. J. D. P. Culhane, "Family homelessness: Background research findings and policy options," p. 83, 2005.

[35] H. Munthe-Kaas, R. C. Berg, and N. Blaasvær, "Effectiveness of interventions to reduce homelessness: a systematic review," vol. 3, 2018.