mathematics

MDPI

Article

Using K-Means Cluster Analysis and Decision Trees to Highlight Significant Factors Leading to Homelessness

Andrea Yoder Clark [1,*], Nicole Blumenfeld [2,*], Eric Lal [1], Shikar Darbari [1], Shiyang Northwood [1] and Ashkan Wadpey [1]

[1] School of Business, University of San Diego, 5998 Alcala Park, San Diego, CA 92110, USA; elal@sandiego.edu (E.L.); sdarbari@sandiego.edu (S.D.); snorthwood@sandiego.edu (S.N.); awadpey@sandiego.edu (A.W.)

[2] 2-1-1 San Diego, P.O. Box 420039, San Diego, CA 92124, USA

* Correspondence: andreayoderclark@sandiego.edu (A.Y.C.); nblumenfeld@211sandiego.org (N.B.); Tel.: +1-619-370-1907 (A.Y.C.)

Abstract: Homelessness has been a persistent social concern in the United States. A combination of political and economic events since the 1960s has driven increases in poverty that, by 1991, had surpassed 1928 depression era levels in some accounts. This paper explores how the emerging field of behavioral economics can use machine learning and data science methods to explore preventative responses to homelessness. In this study, machine learning data mining strategies, specifically K-means cluster analysis and later, decision trees, were used to understand how environmental factors and resultant behaviors can contribute to the experience of homelessness. Prevention of the first homeless event is especially important as studies show that if a person has experienced homelessness once, they are 2.6 times more likely to have another homeless episode. Study findings demonstrate that when someone is at risk for not being able to pay utility bills at the same time as they experience challenges with two or more of the other social determinants of health, the individual is statistically significantly more likely to have their first homeless event. Additionally, for men over 50 who are not in the workforce, have a health hardship, and experience two or more other social determinants of health hardships at the same time, the individual has a high statistically significant probability of experiencing homelessness for the first time.

Keywords: data science; machine learning; data mining; k-means; cluster analysis; decision trees; homelessness; behavioral economics

Citation: Yoder Clark, A.; Blumenfeld, N.; Lal, E.; Darbari, S.; Northwood, S.; Wadpey, A. Using K-Means Cluster Analysis and Decision Trees to Highlight Significant Factors Leading to Homelessness. *Mathematics* **2021**, *9*, 2045. https://doi.org/10.3390/math9172045

Academic Editors: Anca Andreica and Oliviu Matei

Received: 18 June 2021
Accepted: 12 August 2021
Published: 25 August 2021

1. Introduction

1.1. Homelessness in the United States

Homelessness continues to be a significant social issue in the United States. A combination of political and economic factors has compounded over time to contribute to the changing landscape of poverty in America [1,2]. Those experiencing extreme poverty teeter at the edge of homelessness. In 1963, Macdonald's seminal piece in the New Yorker popularized the term the "Invisible Poor" as popular culture began to come to terms with the fact that mass poverty may not have been eradicated by New Deal and Post War era prosperity [2]. Gaps between the rich and the poor slowly began to rise, and rates of homelessness grew to unprecedented levels in the mid-1970s, with higher rates of poverty and homelessness concentrated in America's cities [1,3]. The 1980s saw the shrinking of the welfare state [1,3]. In the 1990s, globalization resulted in the relocation of industry to other countries in search of lower-wage workers, thereby drastically reducing the number of well-paying jobs for those without access to higher education [1,3]. The resulting increase in the population of the working poor was hit hard in the 2000s as the combined effect of two global recessions increased rates of poverty to levels not seen since the 1928 Depression [1,2,4].

By 2012, 15% (46.5 million) of the US population was living at the poverty line [5], up from 14% of the population (35.7 million) in 1991 [1]. While the relatively insignificant proportional increase of those experiencing poverty in America over the last decade seems hopeful, concerns with the accuracy of these numbers remain due to how the "federal poverty level" is calculated [1,3]. Timmer, Eitzen, and Talley suggest that, if 1991 numbers properly accounted for inflation, the number of people in the US living at the poverty line in 1991 would have been 20 million higher than reported levels [1]. What is more, including an adjustment to account for inflation in poverty level calculations would not solve the problem entirely. As early as 1963, Macdonald noted the impact of regional costs of living on how far one's income can go [2], raising questions about the utility of a universal "federal poverty level" income calculation at all. For example, a family living at the poverty line in one region can potentially have more buying power than a family living at the poverty line in another area, underscoring the importance of place, and local factors like regional cost of living, in an examination of poverty and those living at the edge of homelessness [1,2,6]. Federal poverty level calculations are relevant to the consideration of homelessness in America because access to most social services uses the level of one's income in relation to the "federal poverty level" to determine whether or not, or to what degree, one is eligible to receive social service aid. If our current method of calculating the "federal poverty level" does not account for factors like inflation and regional cost of living, many who could potentially access aid are left without social service support, making the experience of homelessness more likely.

Finally, rising concerns around increasing poverty rates, and subsequent homelessness, in the 1990s led to the first widespread studies of homelessness. Much of this research has led to reductions in the homeless population, despite increasing socio-economic gaps. According to HUD calculations, from 2007 to 2020, we saw a 10.3% reduction in the homeless population in the US [7]. While this reduction is hopeful, the transiency of the homeless population makes accurate accounts of the homeless hard to come by. Where HUD numbers typically place the annual number of U.S. homeless in the hundreds of thousands, advocacy groups like the National Coalition for the Homeless, estimate numbers of homeless in the US to be in the millions [3,7].

1.2. Methodological Approaches to Studying Homelessness in the United States

Many studies of homelessness in the US during the 1990s are grounded in the anthropological ethnographic research tradition and build off of the survey research methodology conducted by economists and sociologists in the 1960s and 1970s [2,3,8]. A large proportion of this research sought to explain the rise of the homeless population in the US and can be broken down into three schools of thought: (1) the individualistic approach (2) the structural approach, and (3) the "politics of compassion" argument [1]. Individualists explain the experience of homelessness as a result of individual choices, behaviors, and experiences related to addiction, illness, mental health, and lack of work ethic [3,9,10]. The structuralist approach focused on economic factors like the impacts of globalism, the rise of high tech highly skilled jobs, and the shortage of low-income housing as some of the primary causes of homelessness [3]. The "politics of compassion" argument is a combination of the prior two theories. Some have used the "politics of compassion" argument to make a distinction between the "old homeless" before the 1970s, whose experience could be explained by the individualist perspective; and the "new homeless", whose experience of homelessness in recent decades has been more impacted by changing economic and structural forces.

This study will adopt the third "politics of compassion" approach that recognizes both individualistic and structural explanations around the experience of homelessness. One distinction in this study is that the current authors do not differentiate between the "old" and "new" homeless, and instead believe that the experience of homelessness can be ascribed to both individual choices and behaviors, as well as the effect of larger economic and political trends. This perspective is consistent with the principles of modern-day behavioral economics [1,6].

In behavioral economics, the impact of place is considered central to an understanding of one's ability to access upward mobility in a community [1,6]. In the experience of homelessness, the impact of place is governed largely by government policy related to regional and national homeless laws, as well as laws that determine access to social services [1,3]. Access to social services is often based on where one's income stands relative to the current federal poverty level calculation. Further, policies on the homeless are generally centered on how to reduce the homeless population, and when that is not possible, how to service, rehabilitate and sometimes relocate the homeless [1,8]. While the body of research on homelessness established in the 1990s has paved the way for impressive reductions in the homeless population in the last decade [7], this study postulates that investigations into the issue of poverty may offer opportunities to help prevent the experience of homelessness altogether.

Other areas of research exploring the homeless experience investigate the relationship between demographic characteristics and homelessness. Early research on the experience of homelessness from the 1980s began pointing to decreasing differences in the experience of homelessness between genders, as higher rates of female-led single-parent homes emerged in the homeless population [6,8,11]. Later, in the 1990s, the term the "feminization of homelessness" was coined [12,13] to explain increasing rates of homeless women acting outside of traditional gender-based behavior norms as the men in their communities began to disappear due to high rates of incarceration and drug-related violence [11,14,15]. The "feminization of homelessness" theory states that as men began to be incarcerated and exposed to violence at higher rates, the impact of gender in the experience of homelessness became less significant.

On the other hand, the impact of age as it relates to homelessness has consistently been demonstrated to be significant [1,3,11,16]. The impact of age in the experience of homelessness makes sense, as the economic implications of living at or below the poverty line at the end of life offer less opportunity to overcome the economic challenges that could lead to homelessness. Studies have shown that those experiencing homelessness later in life are statistically significantly more likely to be facing health challenges, adding further obstacles [16]. Experiencing economic challenges early in life is still very hard to overcome, however, research has shown that those experiencing homelessness at a younger age are less likely to face mental health challenges which, when present, can be some of the largest obstacles to overcoming homelessness [1,3,16,17].

The large body of ethnographic and survey research on small samples of homeless from the 1960s to the 1990s, combined with the advent of low-cost high-powered computers to analyze and store data that became available en masse in the mid-2000s, has paved the way for today's behavioral economists to begin studying the issue of homelessness. Historically, behavioral economics has had little focus on the issue of homelessness [11]. This is largely due to a lack of quantitative data at the level of the individual homeless person [7,16]. Traditionally, gathering data to explore the experience of homelessness has been challenging given that most homeless have little interaction with traditional government systems [3,7]. In the past, data gathered to study the homeless are collected through interviewing the homeless on the streets either via surveys or in longer interviews in the ethnographic tradition [3,7]. This cannot be done on a large scale and has thereby prevented the accumulation of large data sets on this population. In recent years, data at the level of the region has been organized by HUD and is collected once a year in over 3,000 cities, where volunteers walk the city identifying and interviewing the homeless [7]. HUD's data are limited, although it has been the most reliable data set to date for the homeless population.

Since 2011, communities have begun to come together to share data on those they serve across social sector institutions, largely fed by the Collective Impact movement [18,19]. Collective Impact, with its' emphasis on shared measurement across social services in communities, has opened up the potential for more advanced research on homeless individuals as clients interfacing with social sector agencies can be tracked across agency

databases [18,19]. Shared measurement, in the collective impact tradition, depends on the development of an information hub, or a data warehouse, that stores data across social sector agencies in a region [19]. When shared measurement is implemented in communities, the collective impact backbone organization will build and manage the data warehouse that is accessible by all partner agencies in the community addressing social issues [19]. This data warehouse can be used to track trends across communities and allows for coordinated community action to effect social change [19]. This study asserts that the ongoing collection of social sector data by collective impact organizations around the country could offer new data sources for research on the experience of homelessness that, when validated and compared against HUD homeless point in time count databases, could establish a more reliable accounting of the number and experiences of the homeless in a region.

1.3. Recent Advances in Data Science Approaches to Studying Social Issues

Relatively few studies have used machine learning techniques like cluster analysis and decision trees to examine homelessness given the historic lack of quantitative data on those categorized as homeless [16]. The advent of the Collective Impact movement has furthered opportunities for quantitative research in this area through the development of more sophisticated data collection, storage, and management systems in social agencies [18–20]. Applying more advanced analytics strategies to data gathered from these shared data systems will provide opportunities to validate and expand on the insightful early research seeking to understand and alleviate homelessness in America.

One example of a study that used quantitative machine learning strategies to better understand the homeless population ran a cluster analysis model [16] to identify patterns in the length of stay among individuals entering homeless shelters based on demographic characteristics. This research identified that people exhibiting behavior termed as episodically homeless had an average stay in the homeless shelter of five months, tended to be employed, older, and have a higher income [16]. On the other hand, those identified as chronically homeless had an average stay in the homeless shelter of nine months or higher, were usually in their thirties, did not have a stable job, and had the highest average family size [16]. Finally, individuals identified as transitionally homeless had an average shelter stay of eight days and were the youngest in the study cohort [16]. Those identified as transitionally homeless in this study also had the fewest identified mental health concerns [16]. Even though such trends could be found through qualitative analysis and more descriptive quantitative techniques, cluster analysis provided "well-defined and robust divisions between the groups in the shelter population, which might not have been picked up by exploratory or descriptive analysis" [16]. This prior research validated the findings of early homeless research that gender did not present as a significant factor in the experience of homelessness [12–14], however, age did consistently present as a significant factor in the length of stay at homeless shelters [13,16], with younger people staying at the shelters for a period of days vs. the month-long stays seen in their older counterparts. This dramatic discrepancy in the amount of time in a shelter based on age gives credence to the hypothesis that economic experiences of extreme poverty could be overcome more easily for younger populations. Further, this paper informed the choice of cluster analysis as an appropriate model for the current study.

Much of the early work applying data science and machine learning to understand social issues has become informed by data science techniques in the field of behavioral economics [1,6]. Behavioral economics explores the results of one's choices, how they relate to the larger economic policies and environment, then establishes the combined impact of individual choices and environmental conditions on one's future economic outcomes [1,6,21]. A behavioral economics methodological approach to research on homelessness naturally supports the "policy of compassion" argument [3] that acknowledges the influence of both individual and structural impacts on one's exposure to a homeless experience. In the Moving to Opportunity Experiment [6], Chetty and Ackerman studied the impact of parents' choice of neighborhoods on children's development. Chetty's research has been

instrumental in identifying how regional policy decisions have had the peripheral effect of creating unequal levels of poverty in various geographies [6]. This study found that the choice to move to a lower-poverty neighborhood with less segregation significantly improves college attendance rates and future earnings for children who were below age 13 when their families moved [6].

Different from a purely individualistic argument that homelessness is the inevitable outcome of a certain percentage of the population who are unique in some key characteristic or behavior [3,22], behavioral economists suggest that homelessness is the result of a series of life experiences, impacted by regional political and economic forces, that have changed lived environments and impacts one's available choices [1,6,22]. For example, as rates of income inequality have grown in specific regions, like inner cities, families in those areas are more likely to experience precarious financial situations due to local policy and environmental realities, such as the availability of fewer jobs, which could put them in the position of making choices that present a loose: loose scenario.

1.4. Moving beyond Descriptive Studies of Homelessness

This study seeks to expand understandings of homelessness beyond the ethnographic accounts of the 1990s by examining call center data collected and stored by 2-1-1 San Diego as individuals call in to obtain access to social services. In essence, 2-1-1 San Diego is a non-profit information and referral hub, accessed through an easy-to-remember three-digit dialing code. Further, 2-1-1 San Diego acts as the community's backbone organization for a larger collective impact movement [18,19]. Realizing the value of shared data and measurement approaches consistent with the collective impact epistemology [18,19], 2-1-1 San Diego launched The Community Information Exchange (CIE) in 2018. CIE is a collective impact data-sharing hub that tracks key socioeconomic, demographic, and social data gathered from those calling the 2-1-1 San Diego call center. Further, 2-1-1 San Diego's mission is to serve as a nexus to bring community organizations together to help people efficiently access appropriate social services and provide vital data and trend information for proactive community planning. Organizations across San Diego County have leveraged CIE's cloud-based data warehouse to share information for individual care coordination, and have used real-time data for community-wide coordination in times of crisis. This collection of regional data around social service clients, their needs, and the available resources will allow an examination of patterns in environmental factors and behavioral choices that may occur before a client becomes homeless.

2. Materials and Methods

2.1. Study Sample

For this study, 2-1-1 San Diego provided access to anonymized client intake data that included demographic and housing information. All records in the sample represent someone who was calling 2-1-1 to access social services. Client records included in the study were identified as having a homeless event, or not, by cross-referencing 2-1-1's client intake database with the HUD Homeless Management Information System (HMIS) database.

The study sample included 3673 original records from 2019, representing one year's worth of calls to 2-1-1 San Diego. The sample was more-or-less evenly split between those who had experienced a homeless event (42%) and those who had not (55%), allowing for sufficiently balanced groups within our independent variable. A small portion of the data set had missing data for the variable that identified whether or not the respondent had experienced a homeless event (3%). There were significantly more females (71%) represented in the sample overall than males (26%). Notably, out of the 42% of the overall sample who had experienced a homeless event, about half of those (23%) were men. In San Diego's actual 2019 *Homeless Point in Time Count Report*, 69% of all recorded homeless were men [23]. This could point to evidence that men are over-counted in traditional studies of the homeless that depend on interviews or surveys of those homeless that are

observed on the street. This also points to the fact that there may be many more female homeless persons that have not been interviewed in the HUD numbers, further validating earlier research that suggests HUD's account of the homeless in America is significantly under-representing actual numbers [3].

The numeric racial and ethnic breakdown of the study sample is outlined in Table 1. above and is represented as a percentage of the total study population in the text. follows: white (27%), African American (26%), Hispanic (21%), Multi-Racial (5%), Asian (2%), Native American (1%), Pacific Islander (1%), and Other (1%). The study sample under-represents whites (45% of County population, but 27% of the sample), Asians (12% of County population, but 2% of the sample), and Hispanics (34% of County population, but 21% of the sample) in proportion to their larger presence in San Diego County's population overall [24]. On the other hand, the study sample significantly over-represents African American's (5.5% of the County population, but 26% of the sample), in comparison to the proportional make-up of San Diego County's population overall [24]. Those included in the sample that identified as Native American (1% of the county population and 1% of the sample), Pacific Islander (1% of the county population and 1% of the sample), multi-racial (5% of County population and 5% of the sample) and other (2% of county population and 1% of the sample) were similarly represented in San Diego County's actual population [24]. The over-representation of the African American population is also seen in poverty line calculations and HUD's 2019 Annual Point in Time Count of San Diego's homeless population as well [23,24].

Table 1. Socio-economic distribution of study sample.

Housing	Gender Identity		Race and Ethnicity							
Homeless	**Male**	**Female**	**White**	**African American**	**Asian**	**Hispanic**	**Native American**	**Other**	**Pac.Islander**	**Multi**
Yes	454	1103	440	393	20	340	21	27	16	91
No	492	1534	566	561	57	539	17	49	31	89

According to the 2019 Homeless Point in Time Count, San Diego had 8102 homeless throughout the county with the majority of homeless residing in Downtown San Diego (63%), and 13% of homeless located in the county's eastern suburbs [23]. The sample size captured in this study represents 19% of all reported homeless in San Diego county in the study year [23]. One difference in the study sample vs. the numbers of homeless represented during HUD's annual point in time count is the under-representation of men and the over-representation of women in the sample relative to actual homeless populations. Men represented 69% of all homeless in San Diego during the study period, however men represented about half of the study sample (23% of the 43% who had experienced homelessness) [23]. Finally, in San Diego in 2019, over 40% of those who experienced a homeless event eventually became housed, but 26% of those who became housed had a subsequent homeless event. This represents the highest rate of recidivism in the homeless experience in all metros that year [23].

2.2. Cluster Analysis

Cluster analysis is a common machine learning approach used to group a set of records into different clusters based on the similarities across different factors that were entered into the model. Different from classification analysis, cluster analysis is an unsupervised learning algorithm, which means there are no assumptions made about the possible relationships among each data point [25,26]. In the present study, cluster analysis was used to create clusters of individuals with similar characteristics and life experiences, who also experienced homelessness, to better understand different pathways that could potentially lead to homelessness.

2.3. K-Means Cluster Analysis

Clustering algorithms outperform traditional descriptive analytics techniques by ensuring that groups are composed of the most similar records by segregating observations based on distance from the centroid across all characteristics. In the present study, the Python sci-kit learn package was used to apply the K-means clustering algorithm. In K-Means clustering, the number of groups (k), is built around midpoints, called "centroids", so that each observation is closest to its' own group's "centroid" [27]. K-means clustering uses the mean of the group as the centroid metric. This process starts by mapping each observation into a plane using the values of the variables associated with it. Once mapped, the midpoint, or centroid, is generated. Distances between each observation and every centroid are then calculated, and each observation is assigned to the closest centroid, the collection of which becomes a group [27]. After the group assignment, the new centroid for each group is recalculated. Distances between each observation and the new centroid are calculated and observations are reassigned to the closes centroid once more. The process continues until each observation cannot be assigned to a new centroid [27]. Elbow analysis was used to determine the number of (k) clusters. Calculated in Python, an elbow chart graphically depicts how much variance each cluster will have based on the number of groups (k) used. The ideal number of groups is determined by identifying the point at which, if more groups are added, there is no statistically significant difference in the reduction of variance within groups [27]. For this study, Elbow analysis identified that four clusters (k = 4) was the ideal number of groups, as segregating records into more than four groups would not significantly reduce the within-cluster variance.

2.4. Factors Introduced to the K-Means Model

In the present study, the K-means clustering algorithm was applied to the 2-1-1 San Diego data sets. The specific factors introduced into the K-means model included key demographic information, as well as respondents' scores on the hardship indicator sub-scales that measured level of security across different social determinants of health categories. The social determinants of health were defined by the World Health Organization [28] and measure one's security in areas of life that have been demonstrated to be instrumental in maintaining a standard level of living above the poverty line [28].

For each of the social determinants of health hardship indicator sub-scales, each respondent could have a hardship indicator score of high, medium or low, that corresponded to their level of security in that social determinant of health category [29]. The social determinants of health hardship indicator sub-scales measure (a) housing instability, (b) food insecurity, (c) medical financial constraints, (d) transportation barriers, (e) utility payment insecurity, (f) criminal justice involvement, and (g) employment instability [28]. For this study, hardship indicators were grouped into High, Medium, and Low levels based on where the individuals' total score for that hardship indicator fell within the relative distribution of total scores across all respondents. The final hardship indicator sub-scale scores of high, medium or low were fed into the model. A Total hardship indicator score was calculated by taking an average of all of the total scores on each hardship indicator sub-scale.

Anonymized client demographic factors included in the model include age, gender, employment status, race, education, and whether or not the client was at or below the Federal Poverty Level (FPL). Additionally, this study feature engineered new variables from the underlying data that were assigned to each client record [30]. Additional feature engineered variables included (1) the total number of calls to 2-1-1 San Diego during the study period, (2) the total number of referrals to external social service agencies and (3) the total number of high scores across all the hardship indicator sub-scales. Researchers in this study also practiced variable reduction techniques like combining variables into one value (such as aggregating all types of unemployment into one value) or binning variables—for example, creating age ranges, rather than using the actual age of each respondent [30].

While cluster analysis is a powerful analytic tool to identify patterns of variables that co-occur together in specific sub-groups [26], the strategy is limited in that it does not provide the statistical significance of variables [27], making it challenging to identify those variables that have the largest effect on the probability of becoming homeless. To address this disadvantage of the cluster analysis model, the study applied a decision tree model to validate the findings of the cluster analysis and select the most significant variables that were most likely to lead to the condition of becoming homeless.

2.5. Decision Tree Analysis

Decision trees are decision-making tools that split the data into consecutively more pure groups, or nodes until the tree cannot split further based on the model constraints [31]. This process allows us to determine which specific variable lead to the splitting of nodes. Those variables that split the nodes have more weight toward the outcome. In this case, if the outcome is becoming homeless, the variable that split the node is significantly related to that client becoming homeless [31]. When a node is split, the decision tree model will identify which variable is splitting the node and the value, or threshold, at which that variable split the node. The value at which the node is split becomes the threshold point for decision-making when the results are applied [31].

Decision trees have been used in similar studies in the past to identify factors that increased the risk of youth becoming homeless in the future [17]. Historical studies have found that length since last stable housing and mental health status can play roles of greater importance in determining the future risk of homelessness [16,17]. Given that some life experiences can impact one's probability of becoming homeless more than others for specific groups, we felt that it was necessary to understand the most significant variables using decision tree analysis to create actionable results.

In the present study, decision tree analysis was applied to the same data used for cluster analysis, with the addition of the cluster that the client had been assigned during cluster analysis. This allowed us to determine which variable was most statistically significant to leading an individual into homelessness for specific groups.

3. Results

3.1. K-Means Cluster Analysis

Our analysis resulted in four final clusters that encompassed all records in the data (see Figure 1). Cluster 3 (n = 514) had the most clients that were previously housed who did eventually become homeless. This group had two commonalities, a utility hardship indicator of any level and employment.

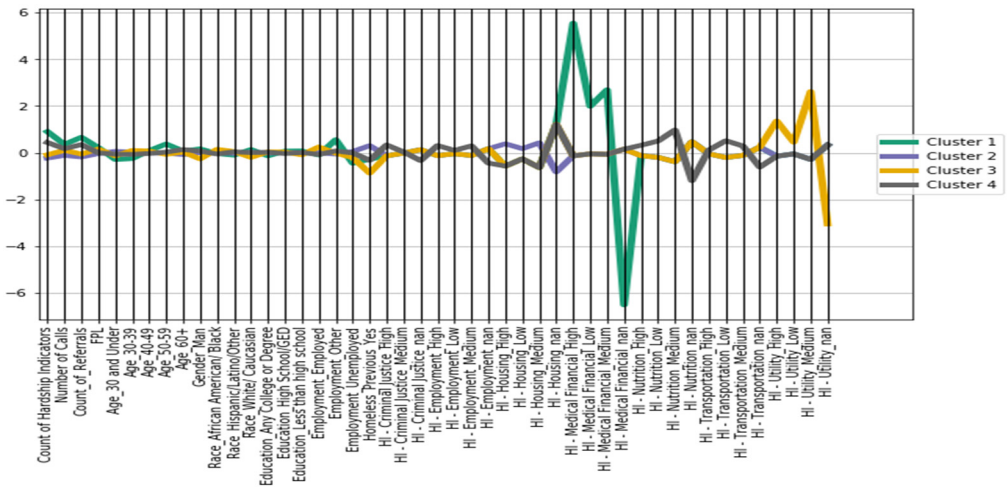

Figure 1. Cluster analysis results.

In Cluster 1 (n = 115), those clients who did eventually become homeless were elderly, mostly male, white, individuals not in the workforce who exhibited medical financial hardship indicators. Having a medical or financial crisis alone was not enough to cause a person to experience homelessness. Those that entered into homelessness in this group also had a high count of total hardship indicators overall indicating other challenges they were facing at the same time as the primary medical or financial crisis, (a high count of calls to 211 San Diego), and a high number of referrals to external social service agencies. This group had a higher risk of becoming homeless given the total number of hardship indicators experienced at one time.

Cluster 2 (n = 3320) and Cluster 4 (n = 1521) are almost exclusively clients that were previously homeless and/or were homeless at the time of the study. Not surprisingly, the main hardship indicator they exhibit is housing. We did not focus on these clusters for our final recommendations to 2-1-1 San Diego, as this seems to represent the chronically homeless population. The purpose of this study is to determine leading indicators for those who have not experienced homelessness before who may be at high risk, therefore the data for these clusters are less relevant.

3.2. Decision Tree Analysis

The results from our decision tree (see Figure 2. below) showed that a prior experience of homelessness was the strongest statistically significant predictor of becoming homeless in the future. This is consistent with previous studies of homelessness where it was found that "persons with a history of homelessness in HUD's HMIS (Homeless Management Information System) were 2.6 times more likely to return to homelessness than others" [5]. This phenomenon was reflected in the San Diego 2019 *Homeless Point in Time Count Report* as well, where 26% of the 40% of homeless who became housed, became homeless again [27]. Given that the present study is most interested in identifying those who are most at risk for experiencing their first homeless event, we will focus on the next most important indicator and the next layer of results in the tree.

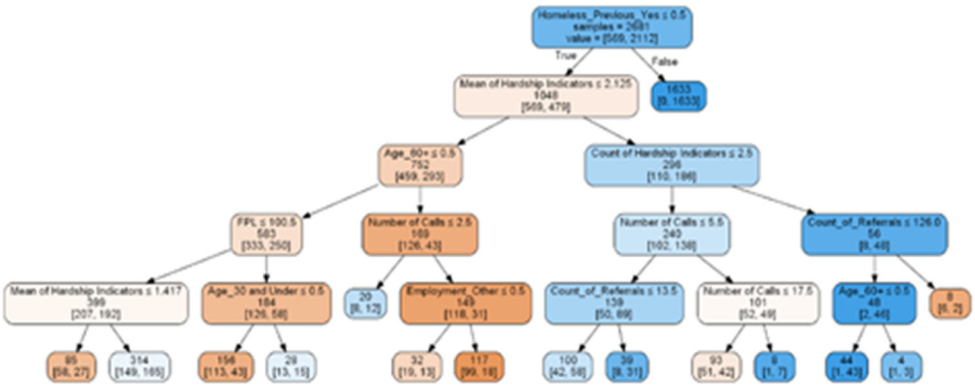

Figure 2. Decision-tree analysis results.

The second most important variable in predicting the experience of homelessness in our study was the mean of hardship indicators score, as can be seen in Figure 3 below. This indicates that the number of hardships one is facing at one time is the next most important factor impacting the probability that one will experience a homeless event.

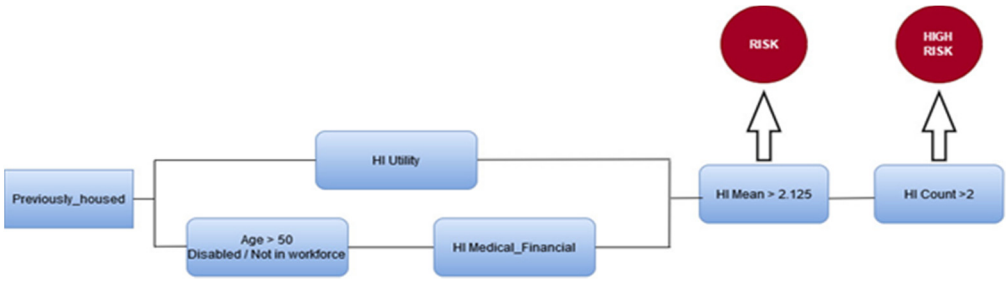

Figure 3. Co-occurring factors leading to increased risk of experiencing homelessness.

For those that were not previously homeless, our data showed that:

1. If a client had a hardship indicator in the area of utilities AND demonstrated a mean hardship indicator score of two, meaning they had two other hardships while also facing an inability to pay their utility bills, the probability of becoming homeless in the future increased to a statistically significant level.
2. If a client had a hardship indicator in the area of utilities AND demonstrated a mean hardship indicator score of more than two, meaning they had more than two other hardships while also facing an inability to pay their utility bills, the probability of becoming homeless in the future was highly statistically significant.
3. The final pathway to homelessness occurred if a housed client was over 50 years old, male, disabled, and/or not in the workforce AND had ANY medical financial hardship and more than two other hardships at the same time. Respondents with these characteristics were highly statistically significantly more at risk of becoming homeless.

4. Discussion

This study points to the importance of the compound impact of two or more hardships in the social determinants of health categories [29] co-occurring at the same time as either a utility hardship indicator, regardless of age; or if over 50 and male, a health financial hardship indicator. The utility hardship indicator looks at the individual's utility bill status (e.g., shut off, past due) to determine the severity and immediacy of a basic need [31]. Individuals whose bill has been shut off had a utility hardship indicator score of high, those with a bill past due had a score of medium, and those with payment concerns and the utility bill is more than 25% of income had a utility hardship indicator score of low [29]. Our results demonstrated that anyone with a utility hardship indicator at any level—high, medium, or low—AND two other hardship indicators were statistically significantly more likely to become homeless.

The health financial hardship indicator looks across the assessments to determine if the individual is experiencing financial strain related to medical costs or medical debt [31]. This indicator is defined by the level of difficulty paying for basic needs (e.g., housing, food) due to a financial hardship related to a disability, accident, or medical condition, barriers related to medical costs (e.g., cost of prescriptions or medical procedures), and percent of average monthly income spent on medical costs [31]. Individuals with the highest difficulty and lowest income are considered high, and those with moderate difficulty alongside moderate income are considered medium, with the lowest difficulty and higher incomes considered low [29]. Our results indicated that the medical financial hardship indicator was statistically significant only for those who were male, over 50, not in the workforce, AND who were also experiencing more than two other hardships from the social determinants of health categories.

Interestingly, gender did present as a significant factor related to becoming homeless in this study, offering counter-evidence to the "feminization of homelessness" hypothesis put forth in the 1990s [8,15–17]. It is relevant to note, that gender only became significant for those who were older in our study, while the "feminization of homelessness" research

base typically was focused on those who were just entering adult life. Replicating the "feminization of homelessness" hypothesis is warranted, where future studies take age into account. Given the gender imbalance in this study where men represented only about a quarter of the homeless in the sample, but 69% of all actual homeless individuals in the population [27], the fact that a statistically significant gender effect is still present in the sample is important. It is also worth noting that women may be under-represented in traditional accounts of the homeless that rely on physical counting or interviewing those homeless visible on the streets. Homeless point in time count methodologies may unintentionally under-represent women, as women may simply be less visible on the streets than men. Given that there were fewer men in the sample, and that a gender effect was still seen, it is likely that a gender effect on homelessness is valid in this study.

The San Diego County 2019 *Homeless Point in Time Count Report* [27] further validates the behavioral economics assertion of the importance of place when considering economic social issues like homelessness [3,4], as San Diego was highlighted as the metro with the highest rate of recidivism for those that had become housed after experiencing homelessness in 2019. The report states that 26% of the original 40% of homeless who became housed experience another homeless event [27]. While this rate is high relative to other geographies, it is consistent with HUD's findings that once a person experiences homelessness, they are 2.6 more times likely to have another homeless event [32]. Historical challenges with the calculation of the federal poverty level are important to reflect on as it relates to recidivism rates around the experience of homelessness [2,3]. The body of literature that asserts the need to account for inflation and local costs of living when calculating the federal poverty level [2,3] could offer potential avenues for local policy to help reduce homeless recidivism in San Diego. Finally, the importance of place established in this study as it relates to levels of homelessness underscores the need for replication of these findings across geographies.

Study results provide 2-1-1 San Diego, and local San Diego policymakers, information that can they can use to proactively identify clients most at risk for becoming homeless. These results provide ways to triage those most at risk and in need of immediate services to prevent clients from experiencing homelessness at critical junctures. This information will be helpful for 2-1-1 San Diego, as they work with groups that address regional homelessness issues, such as the Regional Task Force on Homelessness. Additionally, this information can be utilized with community partners, such as local utility companies, to help them shape their policies around serving customers with the highest needs who are at most risk for utility shut-offs. Further, 2-1-1 could also use this information to provide clients most at risk for homelessness with emergency funds to prevent a utility shut-off and/or address other financial hardships that may be occurring.

5. Conclusions

Preventing the initial experience of homelessness is especially important as prior research demonstrates that once a person has experienced homelessness, the person is 2.6 times more likely to experience homelessness again [32]. The current study is novel in its' application of data science to help understand and prevent the first experience of homelessness by using machine learning strategies like cluster analysis and decision trees to identify statistically significant early indicators of an impending homeless event. The findings of this study demonstrated statistically significant early indicators of a first homeless event using data collected from a Collective Impact Data Sharing Hub, 2-1-1 San Diego's Community Information Exchange, pointing to the following conclusions and recommendations for future research.

5.1. Collective Impact Regional Data Hubs, like 2-1-1 San Diego's CIE, Offer New Sources for High-Quality Quantitative Data on the Homeless Population, That Could Be Used to Replicate Study Findings and Expand Research of the Homeless Population in Other Geographies

Future studies could explore the use of collective impact hub data [18–20] for advanced statistical quantitative research on the experiences of homelessness in other geographies.

Further, by tracking those interacting across social service agencies, and comparing against HUD's annual homeless point in time count numbers, the number of those homeless in a community could be represented across multiple data collection methodologies, both of which are likely not capturing the full extent of homeless in a region. There are limitations to relying on either data set alone as a single source of truth around the numbers of homeless in a region. Collective Impact Hub data only represents those homeless who have made an effort to receive services and point to an under-representation of men when compared to HUD's point in time count for the same year [23]. On the other hand, HUD data rely on the homeless being visible [32]. Additionally, findings from this study point to the potential under-representation of women in HUD's account of homeless populations, as women represented over 3/4 of this study's sample population (77%) vs. only 31% in HUD's 2019 *Homeless Point in Time Count Report*. The potential under-representation of women in HUD's data, and the under-representation of men in the study sample point to limitations in both data sets when used alone. Therefore, recommendations for future studies are to consider the demographic make-up of both data sets to create a potential range represented as a proportion of the population to approach a more thorough representation of homeless demographic characteristics in a region.

5.2. Current Calculations of the Federal Poverty Level Do Not Include Regional Cost of Living and Inflation Adjustments, Potentially Leading to Higher Rates of Homeless Recidivism in Certain Geographies. Replicating Study Findings in Other Geographies Could Help Validate This Hypothesis

Findings of this study found that in 2019, San Diego experienced the highest rates of recidivism around a homeless event than any other geography in HUD's 2019 annual point in time count efforts [23]. These results, and the high cost of living in southern California metros, specifically as it relates to housing costs, could be diminishing one's buying power for those living at the edge of the federal poverty level in these geographies. These findings point out that there may be a case for local policymakers to expand the definition of the "federal policy level" locally to take into account higher costs of living, thereby opening up access to social services to those who may be living just above the current "federal poverty level". There is a need for future studies that investigate the impact of regional cost of living for those living at the federal poverty level in different geographies.

5.3. Measuring Social Determinants of Health Hardship Indicators When One Accesses Social Services Provides Additional Information to Further Assess Overall Risk across Complex Social Experiences, like the Experience of Homelessness

This study found that complex social hardships, like homelessness, are often related to a "perfect storm" of co-occurring needs that present at the same time. The social determinants of health sub-scales [28] have the potential to inform a wide variety of social concerns and would be useful to understand co-occurring risks across many domains. Collecting and storing client responses to the social determinant of health sub-scales would be useful to unify data collected across different collective impact hubs in different geographies and could potentially create a unified lens across the social sector to inform more holistic therapeutic approaches to social service outreach. Finally, if the social determinants of health sub-scales were used to unify social sector data across geographies, clients could be tracked across regions as well, furthering more accurate tracking of services. Future research that includes the social determinant of health sub-scale data across different collective impact data sharing hubs could further validate study findings when measuring early indicators of homelessness in other geographies.

5.4. Having a Utility Hardship Indicator at Any Level and Two or More Other Social Determinants of Health Hardships That Co-Occur at the Same Time Create a Statistically Significant Probability of an Impending Homeless Event

While replicating study findings across geographies would be useful to validate this work, it would also be useful to study the impact of local policy changes made as a result of these findings to determine if the same early indicators of homelessness in

San Diego are found over time and if there are reductions in the number of first-time homeless experiences. Ideally, if recommended policy changes are made, like providing financial support for those with a utility hardship—while also addressing the other co-occurring social determinant of health needs at the same time—we would hope to see the numbers of first-time homeless in San Diego diminish. Additionally, if future studies using 2-1-1 San Diego's data continue to find that a utility hardship indicator remains a statistically significant indicator of an impending homeless event, the study findings can be considered more reliable and will reinforce continued policy support.

5.5. A Health Medical Hardship Indicator at Any Level and More Than Two Other Hardships That Co-Occur Create a Highly Statistically Significant Probability of an Impending Homeless Event for Men Who Are Not Working and 50 or Older

Addressing health care for the elderly living at or near federal poverty levels continues to be an ongoing concern in America. While this issue was not addressed by this paper, study findings point to the re-occurring challenges faced by health care costs of the elderly that can ultimately lead to homelessness. Future studies that summarize current social service approaches to address these issues, and the resulting impact on reducing the elderly who are homeless would be worth examining.

This study, based in the behavioral economic tradition, validates the "politics of compassion" argument to explain homelessness [3], demonstrating that local economic and political systems can impact the lived experience, leading to a more limited set of choices available for individuals experiencing extreme poverty, resulting in a higher likelihood of homelessness [1,6]. The importance of economic and political laws in mediating the probability of the experience of homelessness, is validated by study findings that point to the inability to pay utility bills as an early indicator of a potential homeless event; or, high costs of healthcare for the elderly, especially elderly men living at or below the poverty line, which also have a high statistically significant probability of leading to a homeless event.

This study highlights many areas where local policy can be adjusted to help reduce the experience of homelessness by providing utility bill payment assistance, assistance with health care costs for the elderly, and/or opening up eligibility to social services that were previously inaccessible by expanding the "federal poverty level" threshold access requirements for social service aid. Future studies should replicate these findings, both in San Diego and in other geographies, to determine the persistence of the early indicators outlined in this study.

Author Contributions: Conceptualization, A.Y.C. and N.B.; methodology, A.Y.C., E.L.; validation, N.B., A.Y.C.; formal analysis, E.L.; resources, A.Y.C., N.B., E.L., S.N., S.D., A.W.; data curation, E.L., S.D., A.W.; writing—A.Y.C., E.L., S.N., S.D., A.W.; writing—review and editing, A.Y.C., N.B.; visualization, E.L.; supervision, A.Y.C.; project administration, N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is not publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kneebone, E. The Changing Geography of Disadvantage. In *Shared Prosperity in America's Communities*; Watcher, S.M., Ding, L., Eds.; University of Pennsylvania Press: Philadelphia, PA, USA, 2016; pp. 41–56.
2. Macdonald, D. Our Invisible Poor. New York, NY, USA. 11 January 1963. The January 19, 1963 Issue. Available online: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1078&context=prism (accessed on 24 July 2021).
3. Timmer, D.A.; Eitzen, S.; Talley, K.D. *Paths to Homelessness: Extreme Poverty and the Urban Housing Crisis*; Westview Press: Boulder, CO, USA, 1994.

4. Saez, E. *Striking it Richer: The Evolution of Top Incomes in the United States (Updated with 2012 Estimates)*; University of California: Berkeley, CA, USA, 2013.
5. Denavas-Walt, C.; Procter, B.; Smith, J. *Income, Poverty and Health Insurance Coverage in the United States: 2012*; Current Population Reports; US Census Bureau: Washington, DC, USA, 2013.
6. Chetty, R.; Ackerman, W. The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment: Data- Set. *Am. Econ. Rev.* **2016**, *106*, 855–902. [CrossRef] [PubMed]
7. National Alliance to End Homelessness. *State of Homelessness: 2020 Edition—National Alliance to End Homelessness*. Available online: https://endhomelessness.org/homelessness-in-america/homelessness-statistics/state-of-homelessness-2020/ (accessed on 21 January 2021).
8. Susser, I. Poverty and Homelessness in US Cities. In *Cultural Diversity in the United States*; Susser, I., Patterson, T.C., Eds.; Blackwell Publishers: Maiden, MA, USA, 2001; pp. 229–249.
9. Jones, J. The Widening Gap Between Rich and Poor. *Crit. Anthropol.* **1993**, *13*, 247–267. [CrossRef]
10. Liebow, E. *Tally's Corner*; Little Brown Publishing: Boston, MA, USA, 1967.
11. Jencks, C. *The Homeless*; Harvard University Press: Cambridge, MA, USA, 1994.
12. Sidel, R. *On Her Own: Growing Up in the Shadow of the American Dream*; Viking Press: New York, NY, USA, 1990.
13. Sidel, R. *Women and Children Last*; Basic Books: New York, NY, USA, 1992.
14. Williams, T. *The Cocaine Kids: The Inside Story of a Teenage Drug Ring*; Addison-Wesley Publishers: Boston, MA, USA, 1989.
15. Dehavenon, A. Where did all the men go? An Etic Model for the Cross-Cultural Study of the Causes of Matrifocality. In *Where Did All the Men Go? Female-Headed Households Cross-Culturally*; Mencher, J., Okongwu, A., Eds.; Westview Press: Boulder, CO, USA, 1993; pp. 53–69.
16. Kuhn, R.; Culhane, D.P. Applying Cluster Analysis to Test a Typology of Homelessness by Pattern of Shelter Utilization: Results from the Analysis of Administrative Data. *Am. J. Community Psychol* **1998**, *26*, 207–232. [CrossRef] [PubMed]
17. Chan, H.; Rice, E.; Vayanos, P.; Tambe, M.; Morton, M. Evidence From the Past: AI Decision Aids to Improve Housing Systems for Homeless Youth. In *Proceedings of the National Clearinghouse on Homeless Youth & Families*; University of Southern California Center for Artificial Intelligence in Society and Chapin Hall at the University of Chicago; The Westin Arlington Gateway; Arlington, VA, USA, 2017.
18. Kania, J.; Kramer, M. Collective Impact. *Stanford Social Innovation Review*. 2011. Available online: https://ssir.org/articles/entry/collective_impact (accessed on 28 July 2021).
19. Kolker, A. *Community Data Sharing 101: General Warehouse Design*. January 2021. Available online: https://www.nfocus.com/community-data-sharing-101-general-warehouse-design/ (accessed on 28 July 2021).
20. Porter, N.D.; Verdery, A.M.; Gaddis, S.M. Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PLoS ONE* **2020**, *15*, e0233154. [CrossRef] [PubMed]
21. Congdon, W.J.; Kling, J.R.; Mullainathan, S. Poverty and Inequality. In *Policy and Choice: Public Finance through the Lens of Behavioral Economics*; Brookings Institution Press, 2011; pp. 149–151. Available online: http://www.jstor.org/stable/10.7864/j.ctt127x9c.9 (accessed on 21 January 2021).
22. Roleff, T.L. *The Homeless: Opposing Viewpoints*; Greenhaven Press: San Diego, CA, USA, 1995.
23. San Diego Regional Task Force on Homelessness. 2019 RTFH Homeless Point in time Count. 2019. Available online: https://www.rtfhsd.org/wp-content/uploads/AnnuallayoutRevised9_3_20.pdf (accessed on 28 July 2021).
24. US Census. US Census Quick Facts: San Diego County, California. 2019. Available online: https://www.census.gov/quickfacts/fact/table/sandiegocountycalifornia,CA/PST045219 (accessed on 28 July 2021).
25. Aggarwal, C.C. An Introduction to Cluster Analysis. In *Data Clustering: Algorithms and Applications*; Aggarwal, C.C., Reddy, C.K., Eds.; CRC Press: Baton Rouge, FL, USA, 2014; pp. 2–15.
26. Qualtrics. Cluster Analysis: Definition and Methods. Available online: https://www.qualtrics.com/experience-management/research/cluster-analysis/ (accessed on 21 January 2021).
27. Reddy, C.K.; Vinzamuri, B. A Survey of Partitional and Hierarchical Clustering Algorithms. In *Data Clustering: Algorithms and Applications*; Aggarwal, C.C., Reddy, C.K., Eds.; CRC Press: Baton Rouge, FL, USA, 2014; pp. 88–93.
28. World Health Organization. *The Economics of Social Determinants of Health and Health Inequalities: A Resource Book*. Available online: https://ebookcentral.proquest.com/lib/sandiego/reader.action?docID=1612011 (accessed on 21 January 2021).
29. Health Leads USA. *Learning from the Implementation of CSCA*. Available online: http://healthleadsusa.org/wp-content/uploads/2020/08/Learning-from-the-Implementation-of-CSCA.pdf (accessed on 21 January 2021).
30. Alelyani, S.; Tang, J.; Liu, H. Feature Selection for Cluster Analysis: A Review. In *Data Clustering: Algorithms and Applications*; Aggarwal, C.C., Reddy, C.K., Eds.; CRC Press: Baton Rouge, FL, USA, 2014; pp. 30–35.
31. Ma, X. *Using Classification and Regression Trees: A Practical Primer*; Information Age Publishing: Charlotte, NC, USA, 2018; pp. 1–52.
32. Homeless Hub. *Homelessness Recurrence in Georgia: Descriptive Statistics, Risk Factors, and Contextualized Outcome Measurement*. Available online: https://www.homelesshub.ca/resource/homelessness-recurrence-georgia-descriptive-statistics-risk-factors-and-contextualized (accessed on 21 January 2021).