Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City

Article *in* Journal of Technology in Human Services · January 2018

DOI: 10.1080/15228835.2017.1418703

5 authors, including:

Boyeong Hong
New York University
7 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Awais Malik
New York University
6 PUBLICATIONS **57** CITATIONS
SEE PROFILE

Constantine E. Kontokosta
New York University
83 PUBLICATIONS **888** CITATIONS
SEE PROFILE

Journal of Technology in Human Services

ISSN: 1522-8835 (Print) 1522-8991 (Online) Journal homepage: http://www.tandfonline.com/loi/wths20

Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City

Boyeong Hong, Awais Malik, Jack Lundquist, Ira Bellach & Constantine E. Kontokosta

To cite this article: Boyeong Hong, Awais Malik, Jack Lundquist, Ira Bellach & Constantine E. Kontokosta (2018): Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City, Journal of Technology in Human Services, DOI: 10.1080/15228835.2017.1418703

To link to this article: https://doi.org/10.1080/15228835.2017.1418703

Published online: 31 Jan 2018.

JOURNAL OF TECHNOLOGY IN HUMAN SERVICES
https://doi.org/10.1080/15228835.2017.1418703

Routledge
Taylor & Francis Group

Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City

Boyeong Hong[a], Awais Malik[a], Jack Lundquist[b], Ira Bellach[c], and Constantine E. Kontokosta[a]

[a]Department of Civil and Urban Engineering and Center for Urban Science and Progress, New York University, Brooklyn, New York; [b]Center for Urban Science and Progress, New York University, Brooklyn, New York; [c]Women In Need NYC, New York, New York

ABSTRACT

New York City faces the challenge of an ever-increasing homeless population with almost 60,000 people currently living in city shelters. In 2015, approximately 25% of families stayed longer than nine months in a shelter, and 17% of families with children that exited a homeless shelter returned to the shelter system within 30 days of leaving. This suggests that "long-term" shelter residents and those that re-enter shelters contribute significantly to the rise of the homeless population living in city shelters and indicate systemic challenges to finding adequate permanent housing. This article focuses on our preliminary work with Win (Women-in-Need) shelters to understand the factors that predict readmission and length-of-stay of homeless families. We create a unified, comprehensive database of the homeless population being served by Win shelters, accounting for more than 6,000 homeless families. We apply logistic regression models and an unsupervised clustering algorithm to identify predictors of re-entry and long-term length-of-stay. Citizenship, age, medical conditions, employment, and history of foster care or shelter stays as a child are found to be significant predictors. The results of the K-means clustering identify three primary groups, consistent with previous typologies characterized by transitionally homeless, episodically homeless, and chronically homeless.

ARTICLE HISTORY
Received 14 December 2017
Accepted 14 December 2017

KEYWORDS
Clustering; homelessness; machine learning; predictive modeling; shelter services

Introduction

Homelessness for families and children has become a substantial social, public health, and housing policy challenge for U.S. cities (Biel et al. 2014). New York City has one of the largest homeless populations, with more than 127,000 people sleeping in its shelter system at some point during fiscal year (FY) 2016, and approximately 60,000 living in shelters on any given night. A New York City Department of Investigation (DOI) probe of Department of

CONTACT Constantine E. Kontokosta ✉ ckontokosta@nyu.edu 📇 Department of Civil and Urban Engineering and Center for Urban Science and Progress, New York University, 370 Jay Street, 12th Floor, Brooklyn, NY 11201.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/wths.

Homeless Services' shelters in 2015 found "serious deficiencies" in the shelters for families with children, citing unsafe and unhealthy living conditions, and recommending aggressive and immediate reforms (NYC DOI, 2015). In 2015, approximately 25% of families stayed longer than nine months in a shelter, and 17% of families with children that exited a homeless shelter returned to the shelter system within 30 days of leaving. This suggests that "long-term" shelter residents and those that re-enter shelters contribute significantly to the rise of the homeless population living in city shelters and indicate systemic challenges to finding adequate permanent housing.

Women in Need (Win) is a nonprofit agency that provides shelter to almost 10,000 homeless women and children (10% of all homeless families of New York City), and is the largest homeless shelter provider in the city. Win's goal is to enable homeless women to become self-sufficient as they transition from a Win shelter into permanent housing. The Win shelter network houses close to 4,700 people each night. The purpose of our research is to assist Win, and other homeless shelter providers, to identify at-risk families and individuals as they enter the shelter system. To support this goal, Win has provided anonymized sociodemographic data for residents that have exited their shelter over the past five years, accounting for more than 6,000 households. Using machine learning classification algorithms and unsupervised clustering, this article presents our initial exploratory analysis to identify the factors that are associated with long-term stays and multiple re-entries into the shelter system.

This article begins with a brief review of previous studies focused on predictors and determinants of homeless family outcomes. The next section describes our data and the integration of these datasets to create a unified database of the homeless population served by Win shelters. We then develop a predictive model of homeless family shelter entries and stay duration, examining the probability of readmission back into the shelter system following exit and the likelihood of a homeless family becoming a long-term stayer. We also employ an unsupervised clustering algorithm to identify distinct subgroups based on stay patterns. We conclude with a discussion of the implications of our findings and the use of data analytics in homeless services, as well as future research directions.

Background

While several studies focus on the reasons why an individual or family initially enters a homeless shelter (Early, 2004; Shelton et al., 2009; Shinn et al., 2007), only a few studies have examined the factors associated with homeless families' stay patterns, despite the importance of these factors to outcomes for those in shelters. Wong and colleagues (1997) analyzed readmission patterns of New York City's homeless population, and examined certain demographic characteristics, time variables, and reasons for homelessness with respect to the likelihood of readmission. The authors used eight years of client data from the Homeless Emergency Referral System database, and found that five demographic features (age, family size, race and ethnicity, pregnancy status, and receiving public assistance) were the most significant predictors of shelter readmission (Wong et al., 1997). Additionally, the study found that families that exited into subsidized housing stayed in the shelter system for a longer time, but had a lower rate of readmission.

In a particularly relevant study, Culhane and Kuhn (1998) examine the time-to-exit for homeless individuals in New York City and Philadelphia. Using data from the early 1990s, the authors apply a discrete-time logistic hazards regression model to understand the determinants of longer shelter stays. The study finds that older individuals with mental health or substance abuse problems tend to have greater difficulty exiting the shelter system. Also, in a separate paper, Kuhn and Culhane (1998) use early clustering methods to identify classes of homeless clients based on their stay patterns. Based on identified case profiles for homelessness, they describe three groups: transitionally homeless, episodically homeless, and chronically homeless. Transitionally homeless are characterized by those that enter the shelter system only a single time, and stay for a relatively brief period. Episodically homeless experience a number of stays, each for up to a few months. Chronically homeless are those clients that spend extended periods of time within the shelter system. In many cases, shelters become a de facto substitute to permanent housing options. Kuhn and Culhane (1998) find support for these classifications from the results of their cluster analysis, as does a study using data from Vancouver, Canada (Patterson, Somers, & Moniruzzaman, 2012), although both studies focus on homeless individuals rather than families.

Shinn and colleagues (1998) examined predictors of entry into homeless shelters and subsequent housing stability. The authors surveyed 266 New York City families as they requested shelter, and compared them with a random sample of 298 families from the welfare caseload. The responding families were interviewed five years later to evaluate their housing stability, and families with prior history of shelter use were removed from the follow-up study (Shinn et al., 1998). The authors found that demographic characteristics and housing conditions were the most significant risk factors affecting shelter entry, with “enduring poverty” and “disruptive social experiences” also important conditions. Access to subsidized housing was the most crucial factor in long-term housing stability: the odds of housing stability were 20 times greater for recipients of housing subsidies or tenants in subsidized housing (Shinn et al., 1998).

Recently, the Institute for Children, Poverty and Homelessness (2017) provided an in-depth analysis of demographic patterns and social dynamics of family homelessness in New York City. The Institute suggests that the growing population of the shelter system is primarily caused by long-term staying clients and repeated entries, not by new homeless families (ICPH, 2017). This is supported by the fact that 17% of families with children that exited a homeless shelter returned to the shelter system within 30 days of leaving. This report suggests that in order to reduce homelessness in New York City, the families returning immediately to homeless shelters need to be provided more stable housing solutions upon exiting shelters, such as receiving subsidized housing. This study provides an important foundation for our analysis.

Data and methods

We begin by integrating datasets provided by Win with New York City records from the Department of Homeless Services, the Department of City Planning, and NYC311 to create a unified, comprehensive database of the homeless population being served by Win shelters. To better understand risk factors associated with readmission and long-term stays in shelters, classification models are developed to predict the odds of readmission and length of shelter stay based on demographic and socioeconomic characteristics and family history. We also apply a K-means clustering unsupervised learning algorithm to identify possible subgroups of homeless clients based on stay patterns within the sample. This will provide reinforcement for, or potential modification of, the typology of homelessness previously identified in the literature.

Data cleaning and processing

As part of our initial data collection process, Win provided anonymized social and demographic data on the homeless population that entered and exited its shelters over the past five years (FY 2013–FY 2017). The data are split into four different files: the first dataset included information about client demographics (Demographic Data) such as gender, race and ethnicity, medical history, criminal history, family history, education level, employment status, income, and so forth; the second identified reasons for client exit (Exits Data); the third summarized incidents involving clients that occurred while in the shelter (Incidents Data); and the fourth summarized the living conditions for individual families while in the shelter (Occupancy Data). A significant challenge in merging these datasets is the presence of three separate identifiers representing the individual (Client Assistance and Re-housing Enterprise System or CARES ID), the individual's family (FAMILY ID), and the individual's current case number (CASE ID). We create a new identifier for each unique individual concatenating the three identifiers associated with each record. After merging the various datasets, the total length-of-stay for each individual can be calculated by summing the lengths of stay for each distinct period of residence for the same family (identified by head of household). An individual with more than one distinct period of residence was considered a multientry client.

The integrated dataset contains 6,093 unique individuals (heads of case). Following the merge of the datasets, we further clean the data for our analysis. We exclude any clients whose last exit was after April 2016, which is one year before the end of our study period. We remove these cases to ensure sufficient time for a client to potentially re-enter the shelter system. We also drop observations where there is no length-of-stay data, and remove outliers in the income, age, and monthly income variables. We define an outlier as any values greater than three standard deviations from the mean for that variable. The total cleaned dataset for analysis consists of 3,721 cases.

We combine these client data with information about individual Win shelter facilities, including physical characteristics, inspection reports, and operational data, such as energy use. To develop a more complete picture of facility conditions, we also extract and geolocate NYC311 complaints for the facility and the surrounding area. We do this, and further collect neighborhood-level data on crime and public health records, to test the impact of specific facility placement on outcomes of homeless individuals and families. We are also interested in testing neighborhood effects, to examine whether the neighborhood surrounding the shelter has an influence on likelihood of readmittance, and on the probability of curfew violations or other incidents and infractions that could negatively impact outcomes. While we test the impact of specific shelters on both length-of-stay and readmission, further examination of shelter incident and neighborhood effects will be explored in future work.

Machine learning methods

The challenges of identifying at-risk homeless families are nontrivial. In studies of predicting families that may become homeless in the future, the relatively small proportion of those that actually become homeless creates an unbalanced sample that can confound predictive models (Shinn, Baumohl, & Hopper, 2001). The practical impact of high false negative and false positive rates could mean either that individuals that need special assistance do not receive it, or resources are dedicated to those that would not otherwise require additional services. While additional services rarely will create hardship for the recipient, misallocating limited resources can have serious consequences for constrained shelter providers. More concerning, then, are false negatives, which would mean that those most at-risk are left unsupported. Given the potential impact of this type of oversight, we evaluate and tune our models to maximize recall or sensitivity, thereby reducing the potential false negative rate. The trade-off here, however, is an increase in the false positive rate.

Logistic regression is a widely used linear classification approach for both binary and multinomial problems (Friedman, 2001), and has been applied to numerous problems in the social and human services (Camasso & Jagannathan, 1995; Morrow-Howell & Proctor, 1993). The logistic regression model predicts the logit transformation of the probability of an outcome given a set of independent variables in the form:

$$logit(P) = log\frac{P(X)}{1 - P(X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_v X_v \tag{1}$$

where P(X) is the probability of an outcome given X, X are the model's independent variables, and β are the regression coefficients. To maximize the interpretability of the coefficients, we calculate the odds ratio by taking the exponential of Equation 1 for both sides of the equation. The model specification can then be rewritten as:

$$odds = \frac{P(X)}{1 - P(X)} = e^{\beta_0} + e^{\beta_1 X_1} + e^{\beta_2 X_2} + \ldots + +e^{\beta_v X_v} \tag{2}$$

From Equation 2, a unit change in X_v results in the odds changing by a factor $e^{\beta v}$, which can be defined as the measure of association between dependent and independent variables. The odds ratio explains the change in the likelihood of the outcome variable given a unit change in an independent variable, controlling for other variables in the model (for more on logistic regression, see Hilbe 2011).

Finally, for each coefficient β we compute the Wald statistic as:

$$Wald = \left[\frac{\beta}{s.e.(\beta)}\right]^2 \tag{3}$$

where *s.e.(β)* is the standard error of the coefficient. The Wald statistic is the square of the t-statistic and is used as a measure of variable importance.

As noted, previous research has identified several classes of clients based on their stay and entry patterns (Culhane & Kuhn, 1998; Kuhn & Culhane, 1998). In addition to our logistic regression model, we apply K-means clustering to identify groups of homeless individuals with similar length-of-stays and number of re-entries over time. The K-means algorithm is a partitioning unsupervised learning algorithm that separates the data into K equal variance groups, by minimizing the within-sample sum of squares. To select K, the number of clusters, we execute the algorithm using a range of values for K, and select the value that minimizes the variance within the clusters, while maximizing the variance among different clusters.

Results

Descriptive statistics

Figure 1 shows a Sankey diagram of the entrants and exits into the Win shelter network during the study period. The diagram shows the number of

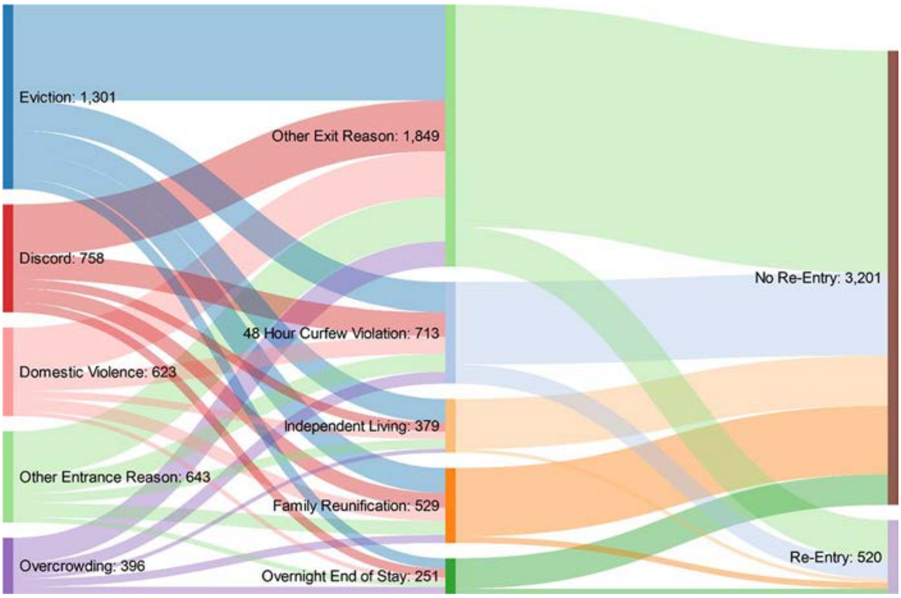

Figure 1. Sankey diagram of entrants and exits during the study period.

homeless families by reason for homelessness and reason for exit, and whether they re-entered the shelter system after exit.

Table 1 presents the descriptive statistics for the variables included in our analysis, grouped by single and multiple entrants and long-term (>9 months) stayers. Overall, 85% of the clients were single entrants, and as many as 1,867 head of households (31%) were identified as long-term stayers.

From a cursory examination of the descriptive statistics in Table 1, we find few distinct differences between groups. We do observe a greater percentage of clients with a history of being in a shelter as a child for multiple entrants, and find long-term stayers to have a higher median monthly income than other groups. Conversely, long-term stayers are less likely to have U.S. citizenship and are found to have relatively high rates of known medical issues. However, the descriptives do not indicate any clear variations between paired groups. We explore these factors in the results from our logistic regression model.

Logistic regression results

We present the output for the logistic regression models predicting the probability of readmission (Table 2) and length-of-stay (Table 3). For the length-of-stay model, we transform the variable to a binary for whether the particular stay was longer than nine months (270 days).

The predictive power of the readmission model is reasonably strong, with a Area Under Curve (AUC) of 0.70 (shown in Figure 2). We control for the year of exit to account for the time period of our sample, and the differential

Table 1. Descriptive statistics by number of entries and length-of-stay.

	Single entrant	Multiple entrant	Short-term entrant (≤9 months)	Long-term entrant (>9 months)
N	5156 (85% of total)	937 (15% of total)	4226 (69% of total)	1867 (31% of total)
DSAT flag (yes)	36 (0.7%)	5 (0.5%)	32 (0.8%)	9 (0.5%)
Foster care as child (yes)	22 (0.4%)	4 (0.4%)	16 (0.4%)	10 (0.5%)
DSAT score (mean)	0.19	0.14	0.19	0.15
Pregnant (at time of entry) (yes)	446 (8.6%)	98 (10.0%)	403 (9.4%)	141 (7.6%)
Veteran (yes)	47 (0.9%)	8 (0.9%)	39 (0.9%)	16 (0.9%)
On parole (yes)	16 (0.3%)	1 (0.1%)	11 (0.3%)	6 (0.3%)
Mental health Flag (yes)	563 (10%)	113 (12%)	448 (11%)	228 (12%)
Prior shelter history as a child (yes)	932 (18%)	184 (20%)	848 (20%)	268 (14%)
Age group (median group)	31–40 years	31–40 years	21–30 years	31–40 years
Family size (median group)	>2	>2	>2	>2
Monthly income (median)	$975	$925	$931	$1037
Employed (yes)	2303 (45%)	393 (42%)	1761 (41%)	935 (50%)
Race/ethnicity	"Black" 3042 (59%)	"Black" 582 (62%)	"Black" 2497 (59%)	"Black" 1145 (61%)
	"White" 145 (2.8%)	"White" 28 (3.0%)	"White" 122 (2.9%)	"White" 51 (2.7%)
	"Hispanic" 1867 (36%)	"Hispanic" 305 (33%)	"Hispanic" 1551 (36%)	"Hispanic" 621 (33%)
	"Other" 102 (2.0%)	"Other" 22 (2.3%)	"Other" 74 (1.7%)	"Other" 50 (2.7%)
U.S. citizen (yes)	3408 (66%)	646 (69%)	2961 (69%)	1093 (58%)
Known medical condition (yes)	1922 (37%)	385 (41%)	1546 (36%)	761 (41%)

Table 2. Results of the logistic regression model for readmission.

No. observations:	3721
Df residuals:	3696
Df model:	24
Pseudo R^2:	0.04902
Log-likelihood:	−1431.4
LL-null:	−1505.2
LLR p-value:	9.40E-20

	coef	std err	z	P>|z|	Odds ratio
History as youth in shelters or foster care	−0.1869	0.123	−1.519	0.129	0.830
Age	−0.0814	0.007	−11.627	0.000	0.922
Family size	0.0110	0.043	0.255	0.799	1.011
Employment binary (1 if employed)	0.1553	0.100	1.558	0.119	0.856
2012 time dummy	0.4443	0.111	4.018	0.000	1.559
2013 time dummy	0.4702	0.125	3.767	0.000	1.600
2014 time dummy	0.6054	0.140	4.312	0.000	1.832
2015 time dummy	0.5288	0.127	4.148	0.000	1.697
2016 time dummy	0.1045	0.108	0.965	0.335	1.110
Reason for homelessness: discord?	0.4350	0.150	3.897	0.000	0.557
Reason for homelessness: Domestic violence?	−0.5856	0.158	−3.713	0.000	0.557
Reason for homelessness: Eviction?	−0.5665	0.132	−4.288	0.000	0.568
Reason for homelessness: Overcrowding?	−0.2850	0.170	−1.674	0.094	0.752
Binary variable: Identify as White?	−2.4054	0.931	−2.584	0.010	0.090
Binary variable: Identify as Hispanic?	−1.7260	0.386	−4.476	0.000	0.178
Citizenship binary: (1 if citizen)	0.1217	0.107	1.133	0.257	1.129
Binary variable: Health condition?	−1.3088	0.362	−3.612	0.000	0.270
Resident of Junius facility?	−0.2710	0.131	−1.921	0.055	0.763
Resident of Liberty facility?	−0.1811	0.128	−1.419	0.156	0.834
Age and binary for Hispanic ethnicity	0.0475	0.012	3.948	0.000	1.049
Age and binary for White racial identity	0.0735	0.026	2.862	0.004	1.076
Age and binary for medical condition	0.0435	0.011	3.944	0.000	1.044
First exit in the summer?	−0.1345	0.136	−0.990	0.322	0.874
First exit in the fall?	−0.4059	0.150	−2.708	0.007	0.666
First exit in the winter?	−0.3614	0.148	−2.450	0.014	0.697

probability of re-entry given the amount of time covered by the dataset following a particular exit. We find that younger clients are less likely to be readmitted, controlling for other factors, although this relationship reverses when age is interacted with race or medical condition. Those with health conditions are generally less likely to re-entry the shelter system, as are those that exit the shelter during the fall or winter seasons. Overall, we do not observe any statistically significant relationships between the particular shelter facility of residence and the likelihood of re-entry.

The model for predicting long-term length-of-stay results in a pseudo-R^2 of 0.069, suggesting relatively low explanatory power. Clients without U.S. citizenship are more likely to be long-term stayers, as are those who are employed while in the shelter system. Those that experienced foster care or were in a shelter as a child are less likely to have a long-term stay in a shelter as an adult, as are larger families. Overall, this results largely reinforce the patterns observed in the descriptive statistics.

Table 3. Result of logistic regression model, length-of-stay (1 = greater than nine months).

No. observations:					3770
Df residuals:					3759
Df model:					10
Pseudo R^2:					0.06926
Log-likelihood:					−1876.3
LL-null:					−2015.9
LLR p-value:					3.800e-54
	coef	std err	z	P>|z|	Odds ratio
Citizenship (1: citizen, 0: noncitizen)	−0.9740	0.080	−12.167	0.000	0.378
Reason for homelessness: Overcrowding	−0.7204	0.151	−4.755	0.000	0.487
Reason for homelessness: Domestic violence	−1.1596	0.138	−8.402	0.000	0.314
Employment (1: employed, 0: unemployed)	0.3548	0.081	4.361	0.000	1.426
Reason for homelessness: Discord	−1.2390	0.134	−9.257	0.000	0.290
Foster care or shelter history as a child	−0.3632	0.112	−3.256	0.001	0.695
Hispanic (1: Hispanic, 0: non-Hispanic)	−0.2488	0.086	−2.897	0.004	0.780
Reason for homelessness: Eviction	−0.4034	0.103	−3.921	0.000	0.668
Season (1: Spring, 0: other seasons)	0.2171	0.103	2.109	0.035	1.242
Age	0.0027	0.003	0.880	0.379	1.002
Family size	−0.0508	0.029	−1.743	0.081	0.950

Clustering results

Our clustering algorithm results in three identified groups as presented in Table 4. The largest group, Group 1, has a moderate number of stays and typical length-of-stay of approximately five months. This group tends to have a relatively higher average monthly income ($1,500 per month), is generally older, and more likely to be employed. The reason for homelessness for this group also tends to be housing-related, with eviction from previous housing

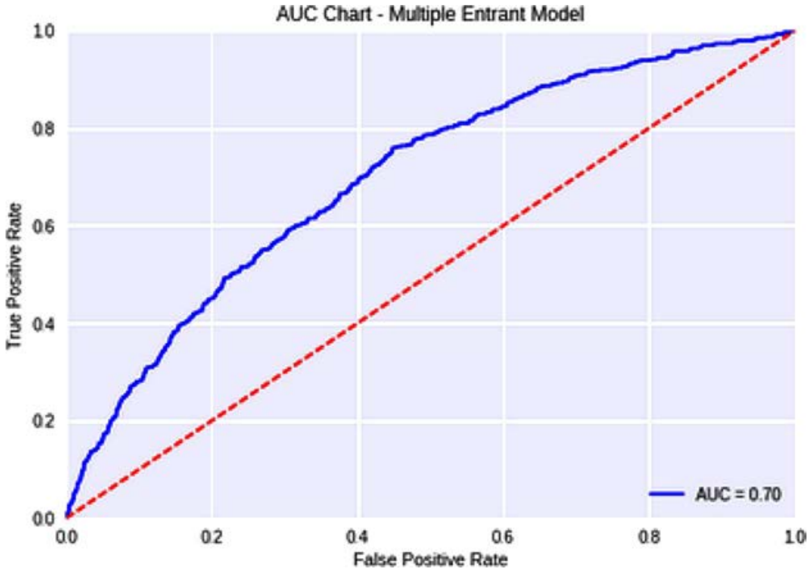

Figure 2. AUC Receiver Operating Characteristic (ROC) for readmission model.

Table 4. Results of the K-means clustering algorithm.

	Group 1 (n = 1978)	Group 2 (n = 959)	Group 3 (n = 784)
Length-of-stay (days)	152.21	263.25	7.98
Number of entries (mean, all entrants)	1.18	1.20	1.11
Monthly income (mean)[a]	$1,500.35	−$9.83	−$15.65
Age (mean)	33	32	31
Family size (mean)	3.3	3.4	3.1
History in foster care or shelter as child	18%	19%	20%
Mental health issue	10%	14%	9%
Employed	51%	34%	24%
Reason for homelessness: Discord	19%	17%	28%
Reason for homelessness: Domestic violence	14%	18%	22%
Reason for homelessness: Eviction	38%	35%	26%
Reason for homelessness: Overcrowding	11%	11%	12%
U.S. citizenship	63%	54%	72%
Medical condition	94%	92%	65%

[a]Groups 2 and 3 have negative average incomes, reflecting payments or debts owed.

or overcrowded conditions identified as the two most likely causes and accounting for nearly 50% of the shelter entries. Group 1 is, therefore, consistent with the characteristics of those considered episodically homeless.

Group 2 is the second largest group in the sample, and is characterized by longer average length-of-stay and more frequent stays. Head of households in Group 2 are typically in their early 30s and have the largest average family size (3.4 persons). This group is the least likely to hold U.S. citizenship status, while as many as 38% have had a long-term (>9 month) stay at some point during the study period. Those in Group 2 face an above average incidence of mental health issues and do not, on average, have any regular income. Group 2 clients are similar to those classified as chronically homeless.

Finally, clients in Group 3 experience the fewest entries into the shelter system, with very short duration stays, averaging just eight days. These clients are, on average, the youngest of the three groups and the least likely to have mental health problems or serious health conditions. Importantly, those experiencing homelessness in this group are typically forced into the shelter system because of domestic violence or discord, reasons that account for more than half of all entries. Group 3 individuals were also the most likely of any of the clusters to have experienced foster care or been in the shelter system as a child. Group 3 can be associated with those described in the literature as transitionally homeless.

Discussion and considerations for data science applications to homeless services

Homelessness is one of the most pressing social problems facing low-income residents and city policy makers. Although the homeless population in the United States had declined since 2007 (National Alliance to End Homelessness, 2017), many cities face significant rates of homelessness and a growing number of those at-risk, fueled by increasing housing costs, limited affordable housing supply, and decreasing availability of stable, lower-skill job opportunities. For city agencies and nonprofit providers of homeless shelter and social services, such as Win, meeting the needs of a greater number of homeless families, while simultaneously facing reductions in funding, is a daily challenge.

The application of data science and data technologies to urban and social challenges has received significant attention in recent years (Bettencourt, 2014; Kitchin, 2014; Kontokosta, 2017). Spurred by the need to more efficiently use resources, while providing increased levels of service, city governments and nonprofit organizations have looked to data analytics as a means to extract actionable insight from existing and new streams of data. While a range of machine learning and artificial intelligence algorithms have successfully been used in other industries and sectors, their application to social and urban policy and operations remains exploratory and ad hoc. This is in large part due to the complexity of these problems, which involve competing values, goals, and priorities and where latent or implicit bias in collected data can exacerbate underlying mechanisms for discrimination (Boyd & Crawford, 2012; Kontokosta, Hong, & Korsberg, 2017).

Shelter providers and homeless service agencies, such as the NYC Department of Homeless Services, collect a substantial amount of data on homeless individuals entering and exiting the shelter system. While in residence at a shelter, case workers conduct regular interviews and consultations with clients to support basic services and improve the likelihood of transition into stable, long-term housing. While substantial insight can be gained from these data, the focus of these organizations as it pertains to data largely remains one of data collection for reporting and record keeping, rather than for analysis and operational decision-making. This mindset has three non-trivial consequences. First, data collection methods are designed for one-time reporting, without consideration of future processing and analysis of these data. This leads to, for instance, hand-written case notes and free-form responses to standardized questions. Data are then stored for archival purposes, making them difficult to extract, process, and visualize.

Second, relevant data on homeless families and their conditions are stored by multiple different agencies and organizations, and there is little sharing and integration of data across these siloes. For example, tracking a homeless family as they move from one shelter provider to another is a difficult, if not impossible, task for any single shelter provider. While city agencies may have access to these data in some form, there are no universally accepted methods for identifying all records generated from different sources and shelters.

Finally, integration of data-driven decision-making into organizations not accustomed to these tools and processes creates significant barriers to using analytics for day-to-day operations. Part of this is the result of human and computational infrastructure limitations: in many cases, employees simply do not have the skills needed to interpret and apply data analytics to their work, and information technology—including database management—is often characterized by legacy systems using outmoded software. The more significant challenge to data-driven decision-making is created by rigid, bureaucratic processes, often present in social service agencies, that dis-incentivize innovation and foster a lack of trust in new methods (Provost & Fawcett, 2013). While there is a rationale for this type of path dependent approach in light of the seriousness of these agencies' responsibilities, the aversion to new ways of organizational management can limit relatively low-cost and low-risk opportunities for the use of data.

Conclusion

Our research attempted to predict readmission of homeless families back into the shelter system following an exit and the likelihood of a client becoming a long-term stayer, using data from Win shelters in New York City. After building a robust database of homeless family characteristics, we analyze the effect of factors such as age, race, family type, employment status, citizenship status, and reasons for homelessness. In general, the logistic models provide modest predictive power for both readmission and length-of-stay. Citizenship, age, medical history, employment, and history of foster care or shelter stays as a child are found to be significant predictors. The logistic models are designed to provide insight into the influence of specific attributes on the probability of the outcome variable. In the future, we will explore other classification models, such as decision trees, to potentially improve model performance, although at the expense of model interpretability.

The clustering results strongly support previous qualitative and empirical research on homeless typologies. Our three identified groups align closely with previous definitions of typical stay profiles. The largest group are those defined by a moderate number of stays and average length-of-stay, consistent with those defined as *episodically homeless*. The second group has the most entries and the longest length-of-stay for each entry, on average 263 days. Individuals in this group have no monthly income and the largest family size, and also face the highest probability of having a mental health issue. This group appears to use the shelter system as a substitute or alternative to permanent housing, similar to homeless clients classified as *chronically homeless*. Finally, the third, and smallest group, is clearly distinguished by very short (approximately eight days on average) stays and the fewest number of entries. The most common reason for homelessness cited by those in Group 3 relates to various types of domestic violence. This group can be considered *transitionally homeless*. Unlike previous studies, however, we find the *episodically homeless* cluster as the largest of the three. This could be the result of the nature of the housing market in New York City during the study period, which exhibited a limited supply of available subsidized housing and rising rents (Bassuk & Rosenberg, 1988; NYU Furman Center, 2016).

We anticipate further improvements to our models to strengthen the results. These include accounting for the interactions between continuous variables (such as age) and the categorical predictors, as well as extracting additional relevant indicators from the Win datasets. Additional data on clients outside of the Win shelter network would also provide for a greater depth of understanding of the likelihood of re-entry.

Although preliminary, our findings can provide resource-constrained shelter service providers with insight that, when combined with their expertise and on-the-ground knowledge, can enhance the likelihood of vulnerable populations getting the support and interventions they need to improve the potential for positive outcomes. We recognize, however, that the proposed methods will not address the overall magnitude of the homelessness problem. While we intend our models to be used to help reduce readmission rates and shorten expected length-of-stay, at best these outcomes will only increase the number of beds available at any given time. Comprehensive policy to reduce homelessness requires a renewed commitment to increasing the supply of subsidized housing and providing meaningful economic opportunities for low-income households (Culhane, 1992; Shinn et al., 2001).

Finally, we believe our work with Win will serve as a benchmark for establishing data-driven operations and risk assessment to improve outcomes for homeless families. Although this type of analysis does not address persistent systemic challenges that exacerbate the severity of the homelessness crisis, it can aid in delivering timely and effective services and interventions for individual families in need. In addition to the predictive models presented here, our ongoing work with Win is intended to form the basis for a network of "smart shelters" that can integrate data analysis and other emerging technologies to support advancements across the entire shelter system. This network will serve to support improved outcomes for homeless families both during and after their time in the shelter system by integrating additional insight from data analytics to provide new tools to help shelter providers meet the needs of their clients.

Acknowledgments

Our thanks to Win for their partnership in this research, especially to Christine Quinn, Ira M. Bellach, and Meghan Linehan for their support and guidance. We would also like to thank CUSP MS students Xueqi Huang, Kristi Korsberg, Dara Perl, and Avikal Somvanshi for their work on a preliminary version of this analysis presented at the 2017 Bloomberg Data for Good Exchange.

Funding

This work is supported, in part, by the John D. and Catherine T. MacArthur Foundation. This research has been conducted in accordance with NYU Institutional Review Board approval IRB-FY2017-1135. Any opinions, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect the views of any supporting institution. All errors remain the responsibility of the authors.

References

Bassuk, E. L., & Rosenberg, L. (1988). Why does family homelessness occur? A case-control study. *American Journal of Public Health*, *78*(7), 783–788. doi:10.2105/ajph.78.7.783

Bettencourt, L. M. (2014). The uses of big data in cities. *Big Data*, *2*(1), 12–22. doi:10.1089/big.2013.0042

Biel, M. G., Gilhuly, D. K., Wilcox, N. A., & Jacobstein, D. (2014). Family homelessness: A deepening crisis in urban communities. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(12), 1247–1250.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118x.2012.678878

Camasso, M. J., & Jagannathan, R. (1995). Prediction accuracy of the Washington and Illinois risk assessment instruments: An application of receiver operating characteristic curve analysis. *Social Work Research*, *19*(3), 174–183.

Culhane, D. P. (1992). The quandaries of shelter reform: An appraisal of efforts to "manage" homelessness. *Social Service Review*, *66*(3), 428–440. doi:10.1086/603931

Culhane, D. P., & Kuhn, R. (1998). Patterns and determinants of public shelter utilization among homeless adults in New York City and Philadelphia. *Journal of Policy Analysis and Management*, *17*, 23–43.

Early, D. W. (2004). The determinants of homelessness and the targeting of housing assistance. *Journal of Urban Economics*, *55*(1), 195–214.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Hilbe, J. M. (2011). Logistic regression. In *International Encyclopedia of Statistical Science* (pp. 755–758). Springer Berlin Heidelberg.

Institute for Children, Poverty & Homelessness. (2017). *On the map: The dynamics of family homelessness in New York City 2017*. Retrieved from http://www.icphusa.org/new_york_city/map-dynamics-family-homelessness-new-york-city-2017/.

Kitchin, R. (2014). The real-time city?. Big data and smart urbanism. *Geo Journal*, *79*(1), 1–14. doi:10.1007/s10708-013-9516-8.

Kontokosta, C. E. (2017). *Urban informatics for social good: Definitions, tensions, and challenges*. In Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering (pp. 52–56). Pittsburgh, PA: ACM.

Kontokosta, C., Hong, B., & Korsberg, K. (2017). *Equity in 311 reporting: Understanding socio-spatial differentials in the propensity to complain*. Data for Good Exchange 2017, arXiv preprint arXiv:1710.02452.

Kuhn, R., & Culhane, D. P. (1998). Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization: Results from the analysis of administrative data. *American Journal of Community Psychology*, *26*(2), 207–232. doi:10.1023/a:1022176402357

Morrow-Howell, N., & Proctor, E. (1993). The use of logistic regression in social work research. *Journal of Social Service Research*, *16*(1–2), 87–104. doi:10.1300/j079v16n01_05

National Alliance to End Homelessness. (2017). *2016 annual report of the State of Homelessness in America*. Retrieved from http://endhomelessness.org/wp-content/uploads/2016/10/2016-soh.pdf

NYC Department of Investigations. (2015). *Probe of department of homeless services' shelters for families with children finds serious deficiencies*. Retrieved from https://www1.nyc.gov/assets/doi/reports/pdf/2015/2015-03-12-Pr08dhs.pdf

NYU Furman Center. (2016). *State of New York City's Housing and Neighborhoods in 2016*. New York, NY: New York University.

Patterson, M. L., Somers, J. M., & Moniruzzaman, A. (2012). Prolonged and persistent homelessness: Multivariable analyses in a cohort experiencing current homelessness and mental illness in Vancouver, British Columbia. *Mental Health and Substance Use*, *5*(2), 85–101. doi:10.1080/17523281.2011.618143

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59. doi:10.1089/big.2013.1508

Shelton, K. H., Taylor, P. J., Bonner, A., & van den Bree, M. (2009). Risk factors for homelessness: evidence from a population-based study. *Psychiatric Services*, *60*(4), 465–472.

Shinn, M., Baumohl, J., & Hopper, K. (2001). The prevention of homelessness revisited. *Analyses of Social Issues and Public Policy*, *1*(1), 95–127. doi:10.1111/1530-2415.00006

Shinn, M., Gottlieb, J., Wett, J. L., Bahl, A., Cohen, A., & Baron Ellis, D. (2007). Predictors of homelessness among older adults in New York City: Disability, economic, human and social capital and stressful events. *Journal of Health Psychology*, *12*(5), 696–708.

Shinn, M., Weitzman, B. C., Stojanovic, D., Knickman, J. R., Jimenez, L., Duchon, L., … Krantz, D. H. (1998). Predictors of homelessness among families in New York City: From shelter request to housing stability. *American Journal of Public Health*, *88*(11), 1651–1657. doi:10.2105/ajph.88.11.1651

Wong, Y. L. I., Culhane, D. P., & Kuhn, R. (1997). Predictors of exit and reentry among family shelter users in New York City. *Social Service Review*, *71*(3), 441–462. doi:10.1086/604265