NYU Furman Center

Center for Innovation through Data Intelligence

Research Brief

Predicting Homelessness For Better Prevention

Background

Homelessness is among the most pressing public policy challenges facing New York City. More than 125,000 individuals passed through homeless shelters in 2016, among whom more than 70 percent were in families. Much public attention has been given to the scale of the homelessness crisis in New York City and the significant challenge of addressing it. While there are some interventions that have proven effective at reducing the likelihood of shelter entry, it is difficult to reach at-risk households to deliver prevention assistance before they become homeless. Devising effective means of directing homelessness prevention services to those at greatest risk is therefore a key policy issue. To help improve existing means, the Center for Innovation through Data Intelligence (CIDI) in the Mayor's Office partnered with New York University's Furman Center to use data on human services, buildings, and neighborhoods to predict families' risk of homelessness. This brief summarizes our research design and the key insights from this work.

Approach

The study uses administrative data on receipt of public benefits, including cash assistance and Medicaid, linked to information on homeless shelter applications and stays, building characteristics, and neighborhood characteristics from the years 2006 to 2015. We use machine learning methods to predict shelter application and entry in the year 2015 as a function of these characteristics in previous years. We evaluate the quality of our predictions on a withheld test sample using common machine learning performance metrics. (or two years in some cases) among adult family members.

To better understand whether algorithm-driven predictions can enhance homeless prevention programs, we explore whether our algorithms can identify families at higher risk of homelessness than families currently seeking out and receiving prevention assistance on their own. We first estimate what share of families currently receiving prevention assistance from Homebase, the city's primary homelessness prevention program, would have become homeless had they not received assistance. We then compare this to the share that becomes homeless in an equivalently sized group of high-risk households identified by our algorithms. We find that the highest risk families identified by our algorithms apply to shelter at considerably higher rates than Homebase recipients would have had they not received Homebase. This suggests that our predictive model, which identifies a combination of factors that result in high likelihood of future homelessness, could be used to direct to Homebase individuals who would benefit from Homebase, but do not currently seek it out on their own.

[1]Our data covers 2006-2014 for Medicaid and Cash assistance from HRA, 2003-2015 for data from DHS, and 2006-2015 for data from the New York City Housing Court.

How well can we predict homelessness?

Predicting applications or stays in homeless shelters is difficult. While far too many families suffer from spells of homelessness in New York City, less than one percent of low-income families in our sample stay in shelters in any given year, making prediction challenging. Many features that might be relevant to measure an individual's risk of homelessness, such as previous housing instability or depth of peer and family network support, either do not exist in most data sets or are difficult to measure.

We evaluate how well we can predict homelessness using two measures. Recall measures the percentage of the total homeless population that we predict to be homeless. Precision measures the percentage of persons we predict to be homeless that actually become homeless. Generally, there is a trade-off in these measures: we can reach a large share of the homeless (high recall) by predicting many households to be homeless, but at the cost of lower precision. Conversely, we can achieve very high precision by predicting few households to be homeless, resulting in low recall. In this brief we focus our results on a middle ground, reporting the precision achieved at recall levels of 10%, 20%, and 50%. To calculate precision and recall, we estimate our predicted risk score in our test data set, sort families from highest to lowest risk, and draw cutoffs. Above each cutof we predict all households to be homeless. We then calculate the associated recall and precision for that given cutoff:

$$\text{Recall} = \frac{\text{Predicted Homeless AND Actually Homeless}}{\text{Predicted Homeless}}$$

$$\text{Precision} = \frac{\text{Predicted Homeless AND Actually Homeless}}{\text{Actually Homeless}}$$

We plot the full precision-recall curve in Figure 1, which represents the precision and recall at all possible cut-offs.[2] The farther the curve is to the top right-most corner, the better overall predictive performance.

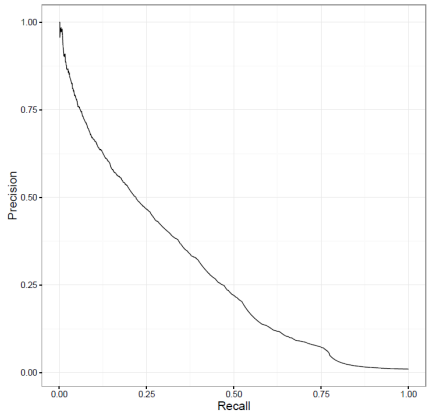

Figure 1: Precision-Recall: Predicting Shelter Applications in 2014

This plot predicts shelter application in 2014 using data from 2013 and prior years.

Table 1 selects three levels of recall on this curve (10%, 20%, and 50%) and reports the associated precision. For example, in our sample there were roughly 41,000 adults applying to family shelter in 2014. If we wanted to reach 10 percent of them (4,161) for a targeted intervention we would need to reach out to 6,275 persons. This improves upon random guessing by 66 times, and reaching 20% and 50% of adults applying to family shelter improves upon random guessing by 52 and 22 times, respectively.

Table 1. Individual-Level Prediction Results: Predicting Shelter Application in 2014

Recall: Share of Total Shelter Applications Correctly Predicted	10%	20%	50%
Shelter Applications Correctly Predicted	4,161	8,323	20,807
Total Shelter Applications Predicted (Total Outreach)	6,275	15,980	94,625
Precision: Shelter Applications Correctly Predicted/Total Shelter Applications Predicted (Total Outreach)	66%	52%	22%
Improvement over Random Guessing	66x	52×	22×

[2]Figure 1 and all subsequent figures use predictions generated from the random forests algorithm. The best performing algorithm we tested was boosted trees, but for computational reasons we present results from random forests, which achieves comparable performance. See Collinson et al. (2016) for details.

What are the risk factors identified in this work?

This project does not attempt to evaluate the causal drivers of homelessness. Instead, we focus on identifying factors that are strongly predictive of homelessness but that do not necessarily cause it. This distinction is important. For example, receipt of Cash Assistance from HRA is strongly predictive of homelessness, but nearly all homeless families are enrolled in some type of assistance upon arrival at shelter. A strong positive relationship between cash assistance and homelessness does not mean that cash assistance causes homelessness. Instead, this relationship could be driven by this reverse causality or correlations between cash assistance eligibility and other factors that are strongly related to risk of homelessness, such as being poor.

We summarize the top risk factors for our individual-level analysis and building-level analysis in Table 2. While the top risk factors are remarkably consistent across specific models and year, the precise ordering can differ, so we present them unordered. At the individual level, the factors associated with elevated risk of homelessness include having previously stayed in a shelter; having ever previously received TANF; being evicted; and living in a building that has previously sent someone to family shelter. At the building level, these factors include the building having recently housed families that applied to shelter; nearby and neighborhood buildings having previously housed people that applied to shelter; various indicators of low building or owner quality (housing code violations, ordered repairs, and litigations); and the number of units in the building. Importantly, we find that building- and neighborhood-level variables contribute to overall predictive performance above and beyond individual characteristics alone

Table 2. Top Ten Predictors, Unordered

Individual-Level Prediction	
Variables	Risk
Ever Stayed in Homeless Shelter	+
Ever Received TANF	+
Eviction in t	+
Building had Previous Shelter Entrant	+
Apply for PA in t	+
Active on PA in t	+
Denied PA in t	+
Sanctioned from PA in t	+
Shelter Code 6: Hotel/Motel	+
Shelter Code 13: DV Program Housing	+

Building-Level Prediction	
Variables	Risk
NYCHA Building	+
Number of Units in Building	+
HPD Housing Code Violation	+
HPD Ordered Repair	+
HPD Litigation against Owner This Year	-
HPD Litigation against Owner Ever	+
Shelter Application This Year	+
Shelter Application Ever	+
Shelter Application Rate, Tract	+
Shelter Application Rate, Block	+

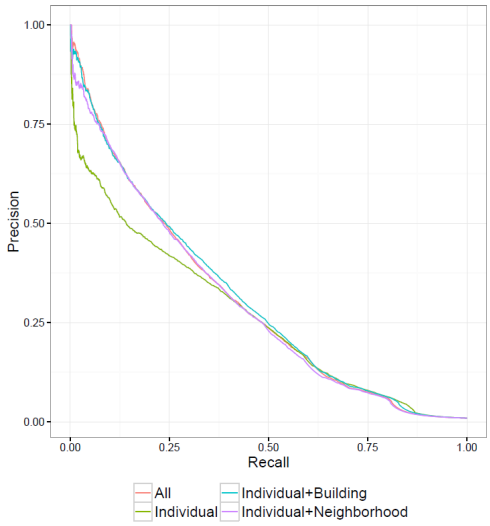

Figure 2: Precision-Recall: Variable Comparison

This figure plots the precision-recall curves from predicting shelter application in 2014 using different sets of variables from 2013 and earlier.

Can algorithm-driven predictions enhance existing homeless prevention programs?

To better understand whether algorithm-driven predictions can enhance homeless prevention programs, we explore whether our algorithms can identify families at higher risk of homelessness than families currently seeking out and receiving prevention assistance on their own. This would suggest potential gains from reaching out to these families. We first estimate what share of families currently receiving prevention assistance from Homebase would have become homeless had they not received assistance. We then compare this to the share that becomes homeless in an equivalently sized group of high-risk households identified by our algorithms.

We find that the highest risk families identified by our algorithms apply to shelter at considerably higher rates than Homebase recipients would have had they not received Homebase. In other words, the baseline risk of families we find is higher than the baseline risk of families currently seeking out Homebase. By combining estimates on the effectiveness of Homebase from the randomized evaluation of the program conducted by Abt Associates (Rolsten et al. 2013) with data on shelter applications by Homebase recipient, we estimate that nearly 20 percent of Homebase recipients would have applied for shelter without assistance within 24 months.[3] Given that only about 1 percent of HRA beneficiaries apply for shelter in a given year, this is an impressively high number – suggesting Homebase already serves relatively risky households. We then track shelter applications over the same period for an equivalently sized group identified by our algorithm as being high risk but not receiving Homebase.[4] In the group identified by the algorithm as high risk, 35 percent applied for shelter within 24 months. This suggests that the algorithm's predictions could be useful in ensuring that Homebase remains targeted to families at greatest risk.

At the building level, we compare our best model to an approximation of current outreach practice by using different sets of variables. Figure 3 plots the precision-recall curves using these. "All variables" predicts shelter applications based on our full set of building and neighborhood characteristics. "Shelter variables" predicts shelter applications based on a building's or neighborhood's history of housing families that would become homeless (derived from DHS data). And "Courts variables" predicts shelter applications using just hosuing court variables. Current targeting of prevention services uses the location of evictions and the addresses of shelter applications and exits to guide outreach. We evaluate how combining these sources with other readily available building- and neighborhood-level data in our predictive models can improve outreach. We find that our model with all building and neighborhood characteristics along with the DHS and Courts variables substantially outperforms the best performing models using either the courts or shelter variables. Given that much of the building and neighborhood data we use is publicly available, these could feasibly be used to guide outreach

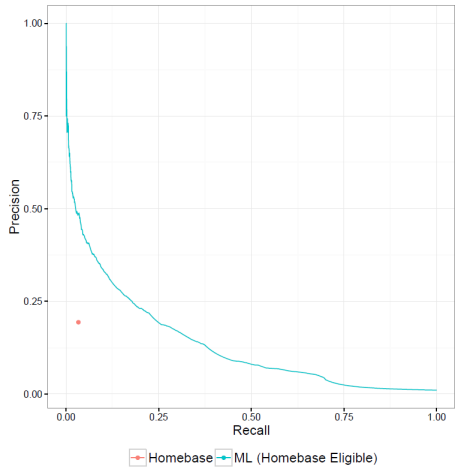

Figure 3: Current Homebase Recipients vs. Algorithm High-Risk Eligibles

This figure plots the precision-recall curves from random forests for Homebase-eligible households (blue) along with our estimate of the precision and recall of Homebase in 2013. In both instances, the outcome is application for shelter within 24 months.

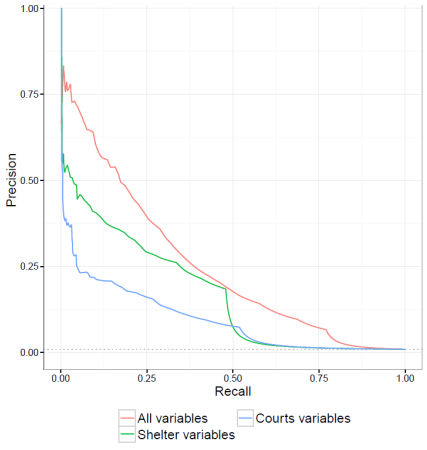

Figure 4: Comparing Different Methods of Building Outreach

This figure plots the precision-recall curves from random forests for buildings using different sets of variables as inputs. Data are from years 2013 and before and the outcome being predicted is shelter application in 2015.

Policy implications

This research suggests several important points for homelessness assistance in New York City.

- The use of predictive targeting in the provision of homelessness prevention services can help ensure that programs find those most likely to benefit from assistance.

  Our machine learning predicted risk scores—which use benefits information, history of shelter interactions, housing court activities, and building and neighborhood attributes—can identify a prevention-eligible population that is roughly 1.5 times more likely to be apply for shelter within 24 months than those currently receiving prevention services through Homebase.

- Prediction-driven outreach to neighborhoods or buildings can enhance existing approaches to efficiently direct services to buildings housing the highest-risk families.

  Our best building-level prediction model is 30% more accurate at identifying building that house families at risk of entering shelter than comparison models built just from the information currently used to direct building-level outreach. These improved predictions could be used to enhance the cost effectiveness of outreach.

- Building and neighborhood characteristics can improve assessment of individual risk of homelessness.

  Building and neighborhood shelter entry histories are important predictors of future homelessness, above and beyond individual characteristics. This is true for those who have previously been homeless and for those who have not. While we cannot discern whether this is a result of certain buildings or neighborhoods attracting vulnerable families (through low rents, few rental requirements, or availability of certain services) or the result of causal factors that might push families into shelter, it is still true that these characteristics, many of which are publicly available, improve predictive accuracy.

[3]This calculation uses data on Homebase recipients from 2013 and shelter applications within 24 month of Homebase receipt. [4]We limit this group to individuals not in shelter or actively applying for shelter to approximate Homebase eligibility.

References

[1] Collinson, Robert, Eileen Johns, Jessica Raithel, Davin Reed, and Maryanne Schretzman (2016). "Predicting Family Homelessness using Machine Learning." Working Paper.

[2] Rolston, Howard, Judy Geyer, and Gretchen Locke (2013). "Evaluation of the Homebase Community Prevention Program: Final Report." Abt Associates.