Predicting homelessness with individual, building, and neighborhood characteristics

NYU Furman Center and NYC Center for Innovation through Data Intelligence (CIDI)
CIDI: Eileen Johns, Jessica Raithel, and Maryanne Schretzman
NYU: Robert Collinson, Ingrid Gould Ellen, and Davin Reed

Context and background on predicting homelessness

New York City has been a pioneer in trying to understand, predict, and prevent homelessness. Our project therefore benefits from and builds on a number of previous projects.

In 2003 the City commissioned a study conducted by the Vera Institute of Justice to understand the pathways that lead families to the shelter system. This included determining: what neighborhoods families come from immediately before they enter shelter and what characteristics distinguish these neighborhoods from others; what factors contribute to families entering the shelter system; and what factors contribute to families returning to the shelter system after placement in permanent housing. The study included a geographic analysis which informed the siting of DHS's first Homebase community prevention programs and interviews of over 300 families which informed Homebase' program design.

After Homebase was implemented for some time, the City sought to understand how effective it was at targeting services. This work led by Marybeth Shinn and coauthors (2013), titled "Efficient Targeting of Homelessness Prevention Services for Families,"[1] entailed analysis of administrative data of over 11,000 households applying for Homebase preventive services to obtain household characteristics thought to be associated with shelter entry. They then merged these survey data with shelter entry data from the Department of Homeless Services (DHS) and ran regression models to identify which characteristics best predicted shelter entry and to obtain risk scores for each household. They found that their models were significantly better at predicting risk than were Homebase workers, suggesting opportunities for data-driven prediction and outreach tools to supplement existing efforts.

A current study by the Human Resources Administration (HRA) is applying these methods to administrative data to see if they can also yield predictive power. They are using household information on HRA Cash Assistance receipt, HRA demographic characteristics, and foster care involvement and regression methods similar to those in Shinn et al. to identify the five biggest risk factors associated with shelter application. They use these five factors to reach out to households deemed at-risk to inform them of existing services that might help prevent shelter application, such as Homebase and HRA emergency assistance/ rental arrears grants.

A third project used information at the building level to predict shelter entry. DHS combined administrative housing courts data on where evictions occur with their administrative data on where shelter applicants come from to highlight the link between eviction and shelter entry. They then provided this information as an interactive map to their Homebase outreach workers, who could use the maps to see in which buildings and neighborhoods evictions had recently occurred and thus better target outreach to those areas.

[1] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3969118/

How this project builds on previous work

Building and neighborhood characteristics

This project builds on previous work in two key ways. First, it adds building and neighborhood characteristics to the more commonly studied household characteristics to explore the importance of these characteristics in predicting homelessness risk. For example, it is known from the DHS project and other work that a number of shelter applicants have a history of eviction, so it is important to understand the role that eviction and housing court involvement more generally might have in predicting homelessness. The results also shed light on potential for building and neighborhood characteristics to help predict other types of risk of interest to policymakers.

We define building characteristics as characteristics of the building from which a family enters shelter. Examples include whether the building was recently foreclosed upon, whether it was recently sold to another landlord, whether it is rent stabilized or recently left rent stabilization, or whether it is public housing. Building characteristics help capture important risk factors not easily observed at the household level. These could be factors of the buildings themselves (like having a "difficult" landlord) or of the families living in those buildings (if families at risk tend live in certain types of buildings). We include all characteristics in three different ways: value in the year of shelter application (current), value in the year previous to shelter application (previous), and change from previous year to current year).

Table 1: Building characteristics and sources

Building characteristic	**Source**
Building class	Department of Finance (DOF) - RPAD master file
Tax class	Department of Finance - RPAD master file
Total units	Department of Finance - RPAD master file
Number of stories	Department of Finance - RPAD master file
Foreclosure (lis pendens)	Public Data Corporation
Housing code violations: non-hazardous	Housing Preservation and Development (HPD)
Housing code violations: hazardous	HPD
Housing code violations: immediately hazardous	HPD
Housing court litigation against landlord	Housing Preservation and Development (HPD)
Emergency repair	HPD
Tax delinquent	DOF
Building sale	Automated City Register Information System (ACRIS)
Multiple dwelling rental registration	HPD
Rent stabilized	Rent Guidelines Board
Public housing (NYCHA)	Department of Finance - RPAD master file

Neighborhood characteristics we define as all of the household and building characteristics described above aggregated to the Census tract-level. For example, from the household -level eviction variable we create counts of evictions in each neighborhood and this new variable in the prediction model. While previous work has shown which neighborhoods send more households to shelters, we hope to shed light on what characteristics about these neighborhoods drive this relationship. As with

the building characteristics, neighborhood characteristics should help pick up risk factors not readily observed at the household level. Moreover, since they are simply aggregations of individual and building characteristics already created, they can be included in prediction models at almost zero additional cost.

Machine learning methods

Second, this project uses predictive algorithms from the field of machine learning in addition to standard regression methods like those used by Shinn et al. and HRA. Regression has many strengths, including availability and ease of interpretation, and thus rightly plays an important role in prediction work. However, other machine learning methods offer a number of benefits as well: they focus on prediction for new samples and avoid overfitting through cross-validation; are very flexible and good at leveraging information from large data sets; and often yield higher predictive accuracy than standard regression. Thus, the combination of regression and other machine learning methods may provide the best possible prediction and insight into difficult problems such as predicting homelessness risk.

Recommendations to policymakers

1) **Adding building and neighborhood characteristics to models can improve predictive accuracy compared to models with just individual characteristics.**

We ran models to predict shelter application among a data set of households receiving Medicaid or Cash Assistance from the Human Resources Administration. We predicted shelter application in 2013 based on household, building, and neighborhood characteristics in 2013 and 2012. Models we tested included regression, random forests, and boosting.

Our results suggest that household, building, and neighborhood characteristics all play an important role in determining homelessness risk and that models that include building and neighborhood characteristics perform better than those that do not. Many of these building and neighborhood characteristics are already collected by various city agencies, so integrating them into prediction models can be a relatively low cost method of improving predictive accuracy.

The table below shows the 10 best predictors from a random forests model that included over 100 different variables. As can be seen, the top 10 variables include household, building, and neighborhood characteristics.

Table 1: Strongest predictors of shelter application

Variable	Variable type	Importance ranking
Number shelter applicants from other buildings in neighborhood	Neighborhood	1
Building classification	Building	2
Gross square feet	Building	3
Number residential units	Building	4
Number HPD emergency repairs	Neighborhood	5
Number immediately hazardous housing code violations	Neighborhood	6
Tax class (building)	Building	7

Number hazardous housing code violations	Neighborhood	8
Year built	Building	9
Number of sales	Neighborhood	10

Most of these building characteristics are publicly available from various agencies through the city's OpenData portal. We have already constructed them for the years 2003 to 2014 and can make them easily available to other agencies through CIDI. Neighborhood characteristics can be easily constructed from individual and building characteristics at any level of geography. Here we have constructed them at the Census tract level, but any level is possible (block, block group, Census tract, community district, borough, etc.) if one has a sample with addresses available for geocoding.

2) **Machine learning methods can provide additional accuracy, flexibility, and insight to complement standard regression methods.**

Prediction methods only work as well as the data and variables available to them. The first step in any project therefore relies on theory and the expertise and experience of staff at different agencies to determine broadly what kinds of variables should be included in prediction models. However, theory doesn't always say exactly how these variables should be included. For example, Cash Assistance receipt may be an important predictor of homelessness risk but how should it be incorporated in a model? For example, should it be included as receipt in the same year as shelter application or the previous year? Should it be included as a change in receipt from the previous year to the present year (either starting or dropping out of the program)? Furthermore, theory is often agnostic on how exactly different variables might interact or what interactions should be allowed. Machine learning methods make it possible to include many different formulations of variables and simply allow the algorithms to determine which are the best predictors and how they interact. Thus, a theoretical approach to which types of variables should be included combined with an atheoretical approach to exactly how they should be included and allowed to interact may yield the best possible predictive accuracy.

3) **These methods can help identify a handful of "red flag" predictors that can be used to create and deploy cost-effective homelessness risk assessment tools across different agencies.**

The inclusiveness and flexibility of machine learning methods make them very adaptable to the particular needs of different agencies. For example, classification trees are an easily interpretable machine learning method that illustrate two of the greatest strengths of learning over standard regression methods: the ability to select just a handful from among many possible variables; and the ability to allow for complex interactions between the variables without specifying them directly. Together, these abilities create the possibility to easily select a small number of predictors from among many predictors to use in risk assessment tools. For example, if an agency plans to design a risk assessment tool based on just a few key predictors, one can still run a classification tree model with hundreds of variables and ask the model to return the few most important predictors and how they interact, thus yielding the best possible predictive accuracy given constraints surrounding development and deployment of the assessment tool.