Predicting Family Homelessness Using Machine Learning

Robert Collinson
New York University

Follow this and additional works at: https://repository.upenn.edu/admindata_conferences_presentations_2017

https://repository.upenn.edu/admindata_conferences_presentations_2017/3

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_conferences_presentations_2017/3
For more information, please contact repository@pobox.upenn.edu.
Predicting Family Homelessness Using Machine Learning
Predicting Family Homelessness Using Machine Learning

Robert Collinson ¹ Eileen Johns ² Jessica Raithel ² Davin Reed ¹ Maryanne Schretzman ²

¹New York University
²NYC Center for Innovation through Data Intelligence

November 14th, 2017
Outline

1. Background
2. Data and Methods
3. Prediction Quality
4. Prediction Content
5. Prediction Comparison
6. Summary
Disclaimer The views expressed here are those of the authors and should not be construed as representing those of the New York City Mayor’s Office, the Center for Innovation through Data Intelligence (CIDI), the Human Resource Adminstration (HRA), the Department of Homeless Services
Background

- More than 125,000 individuals passed through homeless shelters in 2016, among whom more than 70 percent were in families

- One-in-Five homeless families nationally is in New York City

- Many tools to prevent homelessness:
  - Subsidized housing
  - Emergency assistance
  - Legal aid to tenants in housing court
Prevention

- Successful prevention depends on:
  - Effectiveness: Does assistance help families avoid homelessness?
  - Efficiency: Does assistance reach the families most likely to become homeless (high risk)?
Prevention

• Successful prevention depends on:
  • Effectiveness: Does assistance help families avoid homelessness?
  • Efficiency: Does assistance reach the families most likely to become homeless (high risk)?
Our Goals

- Can new data and new methods better identify high risk families?
  - Improve cost effectiveness of outreach
- Guide existing outreach
- Suggest new possibilities for outreach
- Help reduce family homelessness
Sample

- Predict homelessness among all families who have any history of receiving Medicaid or Cash Assistance in NYC
  - Medicaid enrollment, 2006-2014
  - Cash assistance, 2001-2014

- Includes 90+ % of family shelter applicants
Predictors

Variables used to predict homelessness are from a number of sources:

- HRA history of Medicaid enrollment, 2006-2014
- HRA history of cash assistance (application, receipt, sanction, and denials), 2001-2014
- HRA demographic characteristics: education, age, marital status, and family structure
- DHS history of shelter application and entry, 2003-2014
- Housing Court history of court interactions, 2006-2015
- Housing code violations, litigations against owners, mandated repairs, etc.
- Block- and tract-level aggregates of these individual and building characteristics
Machine Learning Methods

• Predict homelessness using machine learning

• Greater flexibility and focus on out of sample prediction

• Better identification of high risk families

• Random forests, boosted trees, boosted logit, neural networks
Our Approach

- Train models in one data set and evaluate predictions in withheld test data set
- Focus on prediction in recent years: 2013, 2014, and 2015
- To predict in 2015, use data from 2014 and prior
- Predict at both individual and building level
Definitions

- Evaluate quality of predictions using precision and recall

\[
\text{Precision} = \frac{\sum_{i=1}^{n} 1(\text{Predicted Homeless} \& \text{Actually Homeless})}{\sum_{i=1}^{n} 1(\text{Predicted Homeless})} \tag{1}
\]

\[
\text{Recall} = \frac{\sum_{i=1}^{n} 1(\text{Predicted Homeless} \& \text{Actually Homeless})}{\sum_{i=1}^{n} 1(\text{Actually Homeless})} \tag{2}
\]
### Predicting Shelter Applications in 2014

<table>
<thead>
<tr>
<th>Share of Applications Predicted</th>
<th>Applications Correctly Predicted</th>
<th>Implied “Outreach”</th>
<th>Apps Predicted / “Outreach”</th>
<th>Beats Random-Guess</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>4,161</td>
<td>6,275</td>
<td>66%</td>
<td>66×</td>
</tr>
<tr>
<td>20%</td>
<td>8,323</td>
<td>15,980</td>
<td>52%</td>
<td>52×</td>
</tr>
<tr>
<td>50%</td>
<td>20,807</td>
<td>94,625</td>
<td>22%</td>
<td>22×</td>
</tr>
</tbody>
</table>
Predicative Performance: Precision-Recall

Precision: 66%  
Recall: 10%  
Total Shelter Applicants: 41,000  
Implied Outreach: 6,200  
Correctly Predicted: 4,100
Predicting Shelter by Variable Type

![Graph showing precision and recall for different variable types: All, Individual + Building, Individual, Individual + Neighborhood.](image-url)
Prediction Quality Summary

- Precision of 60-70% among top 10 percent riskiest families
- Results are similar across years
- Building and neighborhood characteristics improve prediction quality
Prediction Content Summary: What Makes a Good Predictor?

- **Best** Shared by many homeless and few non-homeless: Previously Homeless

- **Broadly Useful** Shared by many homeless but also many non-homeless: Age 20-25, Received Family Assistance, Building Had Previous Shelter Applicant

- **Narrowly Useful** Shared by some homeless but few non-homeless: Eviction, Recent Public Assistance Application.
Prediction Comparison: Homebase

- Can we identify a higher risk set of families than those currently receiving homelessness prevention services (Homebase)?

- This is difficult because we need to know who would be homeless had they not received Homebase.

- Compare risk of families seeking Homebase to risk of equivalently size group of families identified by our models.
Prediction Comparison: Homebase

- 19% of Homebase families enter shelter within 24 months
- 35% of model-identified families enter shelter within 24 months
Actual Homebase Outreach Versus Machine Learning Outreach

Predicting Family Homelessness Using Machine Learning Collinson et al. (2016)
Building-Level Prediction

- Predict which buildings will send families to shelter using building and neighborhood characteristics

- Similar prediction quality and content as in individual models

- Find higher-risk buildings than existing DHS outreach based on locations of evictions and previous shelter entrants
Predicting Shelter Application in 2015

![Graph showing prediction quality for different variable sets.]
Summary

- **Performance**: Predict shelter entry well, particularly among highest risk families and buildings
  - 60-70% precision at 10% recall

- **Comparison**: Find higher risk families and buildings than existing outreach
  - 16 percentage point increase in precision over Homebase (1.5 x more likely to apply to shelter)

- **Content**: Identify best predictors to help guide existing outreach
  - Previous shelter history most important
  - Building and neighborhood characteristics matter

- **Practice**: City is incorporating building-level prediction into dashboard to guide outreach for Homebase
Thank You

- Rob Collinson: rcollinson@nyu.edu
- Davin Reed: davin.reed@nyu.edu