Bridging the Gap Between Big Data and Social Services

Adeen Flinker, PhD

SumAll.org

247 Centre Street, 6th Floor

New York, NY 10013

adeen@sumall.org

adeen.f@gmail.com

Abstract

The goal of health and human service agencies is to benefit the general public as well as protect at-risk populations from worsening social concerns. While there has been a growing focus on prevention, predictive models can be hard to translate into solutions that can be effectively implemented. The recent proliferation of big data sources has created an unprecedented opportunity to leverage data in order to focus work with vulnerable populations and provide predictive-based intervention prior to the worsening of an individual's situation. For example, publically available court records indicating an imminent eviction can be used in order to identify a population at a greater risk of becoming homeless. Prevention services can be provided to these identified individuals prior to their becoming homeless. This intervention, which precedes actual homelessness, not only helps an individual or family, but is also cost effective for the city. Such an approach requires integrating solutions across multiple levels: data integrity, predictive analytics, and implementing an effective intervention process. There are not many organizations that have the necessary tools, ability and knowledge to follow through on all these levels in order to deliver an effective outcome. In this perspective we would like to introduce a predictive-based social intervention approach and examine the associated challenges that must be addressed.

Main Text

The for-profit world is no stranger to leveraging big data predictive analytics to inform marketing decisions. Collecting, analyzing and predicting patterns are an integral part of marketing approaches used by many companies across industries. Critically, such approaches require an infrastructure for: 1. Data collection and augmentation; 2. Linking datasets; 3. Predictive analytics; 4. Incorporating results into an efficient marketing strategy.

While companies such as Amazon, Google and Netflix have exemplified the use of predictive analytics to drive marketing customization, they have both resources and monetary for-profit interests to do so. Social services, in contrast, have limited resources and technical know-how at their disposal in order to make efficient use of big-data analytics. Moreover, social services are designed to both help those in need, as well as protect the privacy of those requesting services, which is not always the case in for-profit industries. There has been growing attention to preventive-focused approaches both federally[1] and in social work[2,3]. Recently, statistical models have been created to assess the risk for homelessness and enhance prevention efficiency in New York City[4]. Such models provide an invaluable assessment tool for gauging the homelessness risk of individuals and families and it can be incorporated into new social service marketing approaches by leveraging big data analytics.

The SumAll.org team (see Acknowledgments) had the opportunity to analyze big data related to homelessness in New York City. Following strict confidentiality and security protocols, several anonymized datasets were shared for data mining including information related to eviction notices as well as families who became homeless and entered city provided shelter. Unlike large businesses that have dedicated infrastructure for data collection, specifically designed for data mining, the datasets available to most social services are legacy-based and designed for various different purposes (such as reporting and accountability). The first challenge in using big-data to improve preventive services involves translating the data into a usable format and linking the various datasets. Such an endeavor requires significant technical knowledge and resources that may not be readily available to social service organizations. We were able to provide assistance in processing the data as well as linking the different datasets (automatic error correction, address geolocation, algorithms reducing matching complexity time from $O(N^2)$ to O(NlogN), etc.). Linking the datasets allowed for a visualization of shelter entries (Figure 1) as well as estimating the time elapsed between eviction notice and shelter entry (mode of 139 days, 4-5 months depending on borough).

We focused our attention on this critical prevention window. The earliest indicator of an eviction was a report social services received including recent eviction notices from the housing court system. Processing the report required social services to manually sort through thousands of records for several weeks and subsequently reach out by mail to the entire processed list. We first automated the process

including geolocated data thereby reducing the work necessary and time delay from weeks to several hours. Next, we augmented the dataset by adding purchased third party demographic information, which is standard practice in most marketing businesses. We assisted in linking the augmented dataset with other databases regarding previous history with social services. Lastly, we applied a set of weights to the most informative features and calculated a risk score. While previous models for assessing the risk for homelessness in New York City were based on features derived from personal interviews[4] (i.e., Report of having been in shelter, Moves in the past year, Discord with landlord, Disruptive experiences in childhood, see reference 4 for full details) we had to restrict our features to those available directly from the data sources and purchased demographic information (e.g., Shelter history match, Age, Children in family, Education, etc.). Based on the risk scores, the automatically identified population (potentially homeless) was now instantly available to social service staff for preventative intervention in order to avoid a shelter stay.

In order to achieve a desired outcome, it was critical to address the bottleneck of the intervention process - establishing an efficient communication channel with potentially homeless individuals. We leveraged the augmented data combined with current marketing approaches. While our goal is not to sell a product, we can offer a social service strategy using a directed marketing campaign. We first updated the outreach letters both in aesthetic design and semantic content in order to convey clear, non-threatening messages. Content was redesigned to remove barriers by

incorporating different languages as well as addressing the stigma associated with asking for help. The letter was redesigned less formally and incorporated visual icons to communicate different services. We also conducted feedback sessions with individuals who met with social service staff in order to optimize the design. We then provided the prevention service with a revamped outreach package that was customized based on the data. Individuals with a higher risk score were assigned additional reminder letters for follow up. A map schematic was provided with geolocated risk scores on both the block and individual level (Figure 2). Finally, the prevention service received each individual's relevant information from the augmented data. This approach allowed the prevention service to conduct focused outreach for high-risk individuals, guided by a data map of different neighborhoods in addition to the customized traditional mailing outreach.

To fully leverage the data and marketing campaign, we expanded the communication channels and created an infrastructure for cost-benefit analysis. The third party demographic information we purchased allowed for a separate e-mail outreach. We provided response infrastructure for both traditional phone as well as text message communication. Based on the data, we purchased Facebook ads using specific demographics and geographical location with high probability of eviction. Each communication channel (letter, email, text message, Facebook ad) was tagged with a separate phone number so we could then follow and analyze the response distribution. These communication channels, together with the risk factor data, serve as a tool for optimizing capacity of the prevention service. The degree of

outreach can be modified depending on the amount of resources available to social service in a given time period. A program can reach out to a greater (or fewer) number of individuals by controlling a threshold of the risk score and employing different communication channels as a function of their average response distribution.

A preliminary analysis based on several months of data collected during the marketing campaign, showed that an overwhelming percentage of responses were in the form of phone calls (97% phone calls, 3% text messages) which originated predominately from the mailed letters (98.4% letters, 1.4% Facebook ads, 0.2% email). Given the distribution of cash resources invested by SumAll.org (approximately 1000$ for letters, 400$ for Facebook ads, 400$ for email addresses) and the overwhelming responsiveness to letters, the most cost effective avenue was a mailing campaign. Nevertheless, the digital outreach may develop to be more cost effective as prices for advertising drop and the use of digital avenues as a communication channel with social services mature.

Our experience with leveraging big data in the social service arena spans multiple domains including data integration, predictive analytics and implementing an effective marketing campaign to ensure those in need are matched with the appropriate level of preventive service. However, in this case our scope was limited to homeless prevention services of the city that partnered with us. Potentially, big data can be integrated across departments and used to customize specific services

for individuals. Indeed, one of the factors that hindered our efforts was that the housing court database and the social service databases did not have a shared unified key for each individual. Furthermore, while some departments had access to geolocation services, limited resources prevented integrating such data into automated processing. The system we implemented identified individuals at a higher risk (based on a selective feature set) and allowed social services to initiate a direct, customized mailing outreach campaign. We believe that going forward it is critical for cities to develop systems that are designed for big data analysis and support seamless integration across service domains. Such a unified framework should support modern analytic tools while also considering the privacy aspects of municipal data and the ethical questions regarding how to best interact with vulnerable populations.

A large portion of our marketing campaign depended on paid services from third party providers. In addition to augmenting our dataset we purchased Facebook location based ads and we will use tailored ads based on emails in the future (just the individual). We strongly urge businesses, and display advertising platforms in particular, to subsidize such efforts for social causes. For example, Google Ad Grants subsidizes $10,000 a month in AdWord advertising for qualified non-profits. While the relevant keyword namespace for evictions was overcrowded, this avenue could be highly cost effective for other projects. We believe that our partnership with New York City represents a proof of concept for the use of big data in the social service

arena and can be developed in the future into platforms with rich data services and large scale automation removing technical barriers for social services.

References

1. G Colbrum. The Federal Commitment to Homelessness Prevention: A Silver Lining of the Economic Crisis. Poverty & Public Policy 2014; 6(1):33-45.
2. EL McCave & C Rishel. Prevention as an Explicit Part of the Social Work Profession: A Systematic Investigation. Advances in Social Work 2011; 12(2):226-240.
3. EL McCave, C Rishel, M Morris. Prevention as an Explicit Part of the Social Work Profession: Part Two of a Systematic Investigation. Advances in Social Work 2013; 14(2):554-555.
4. M Shinn, A Greer, J Bainbridge, et al. Efficient Targeting of Homelessness Prevention Services for Families. American Journal of Public Health 2013; 103(S2):S324-S330.

Acknowledgments

We would like to thank the members of the team who made this project possible: Stefan Heeke, Casson Stallings, Adeen Flinker, Deniz Zorlu, Abhimanyu Ramachandran, Phil Martin, Aash Anand, Sol Eun, Melissa Mowery, Sara Zuiderveen. We would like to thank SumAll.com for their generous support by providing the seed funds that created SumAll.org.

Figure Legend

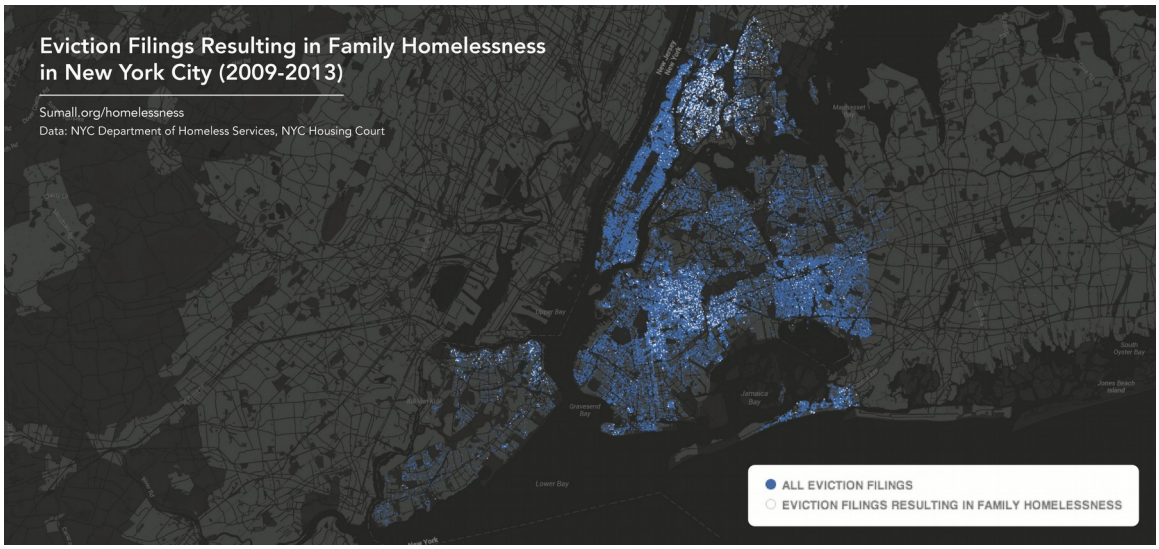

Figure 1.

Eviction filings overlaid with filings that resulted in homelessness.

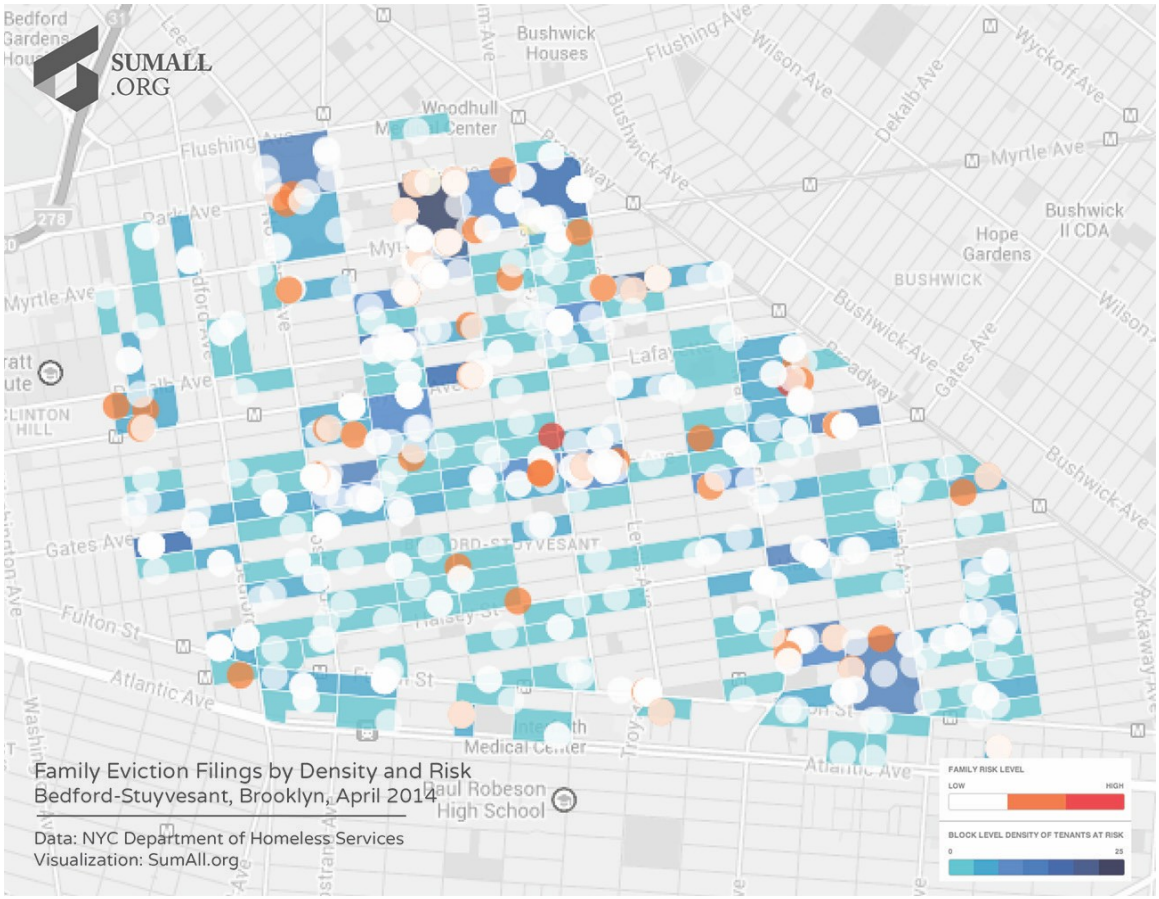

Figure 2.

An example of a data map provided to social services. Geolocated risk scores of both individual families (red color scale) and block level density (blue color scale) are superimposed on regions served by social services.