Nayar Prize II Phase I Quarterly Progress Report  
April 2017

Project: A Data-Driven Crime Prevention Program  
Team: Professors Miles Wernick, Yongyi Yang, and Lori Andrews  
Students: David Haro Alonso, Eric Navarro, and Raven Zeer

Progress Summary of Nayar Prize II Phase I

Background. We are developing a crime-prevention program, focused on social services, for the city of Elgin, Illinois, in partnership with the Elgin Police Department (P.D.). The technological underpinning of the project will be a machine-learning prediction model that identifies individuals who are at high risk for involvement in serious crimes, so that social services can be directed toward those individuals. In a previous study, members of our team (Wernick, Yang, and Alonso) created a prediction model for Chicago, where the focus was on shootings, owing to the particular needs of Chicago. In this study, our aim is to create a new model that is suitable for more general crime problems, and applicable to a broader array of cities, including small towns.

Overview of progress in the second quarter (Q2): At the end of Q2 we are slightly ahead of schedule, having already developed two preliminary versions of the prediction model. During Q2 we developed a totally new approach to the problem that we had not envisaged at the time of our proposal. Preliminary results described below suggest that this new concept is a promising technical solution that is also very reasonable from a legal perspective. As described in our proposal, our model development is proceeding in an iterative fashion, in which each version represents a refinement on the previous one. Our nomenclature is that version numbers beginning with "0" are pre-delivery versions in the initial development phase. Version 1.0 will be the first version delivered to Elgin P.D. and used in the crime-prevention program.

Prediction model Version 0.1: Prediction model v0.1 was an implementation of the starting-point concept that we had described in our initial proposal. Because Elgin is focused on a broad array of serious crimes (not only shootings), we sought to predict an individual’s risk for involvement in at least one serious crime incident in the upcoming year (where the definition of “serious” crimes, listed in our initial proposal, was provided by Elgin P.D.). Model v0.1 proved to be highly predictive of involvement in crime; however, we realized that it was too generic, in that it made no effort to differentiate criminal involvement in terms of extent or severity of the crimes. Suppose, for example, that Person A will be involved in a simple assault next year, while Person B will be involved in one shooting, two aggravated battery incidents, and an armed robbery.
In this case, v0.1 would classify both persons identically, because both were involved in at least one “serious” crime. But, this would fail to capture the obvious conclusion that Person B is likely to experience a much greater extent and severity of crimes than Person A.

**Prediction model Version 0.2:** To address this shortcoming, we developed model v0.2, in which we shifted to predicting an overall crime-risk index (CRI) for each individual, based on both the extent and severity of crime involvement. To capture the notion of severity, we have adopted a numeric scale, in which each type of crime is assigned a crime-severity value (CSV).

An individual’s overall crime-risk index is then computed as a count of crime incidents, with each incident weighted by its crime-severity value. For example, suppose that homicides are assigned a crime-severity value of 100, while the less severe crime of robbery is represented by a crime-severity value of 25. In this case, an individual involved in one homicide and two robberies would have an overall crime-risk index of $1 \times 100 + 2 \times 25 = 150$.

This approach makes legal sense, because the legal system is based on this very concept, classifying crimes into subcategories (e.g., various classes of felonies and misdemeanors), and assigning different penalties to crime types according to their perceived severity. For example, assault carries a maximum penalty of 30 days’ imprisonment, while murder can result in life imprisonment. Crimes of similar type can also be differentiated based on the circumstances of the crime. For example, an assault is elevated to a charge of aggravated assault if, for example, a deadly weapon is used, or if the victim is a child. Correspondingly, aggravated assault is associated with harsher penalties, and will be associated in our approach with higher CSVs. As a proof of concept for our new modeling approach, we assigned a coarse, preliminary system of CSVs for various categories of crime, as shown in Table 1. We trained a model to predict an individual’s CRI for the upcoming year, based on crime involvement in the past two years. Historically, we found that individuals can have CRI as high as 580.

![Figure 1. Variables used to assess risk. Bars represent relative importance of the variables.](image)

We identified 25 variables that might be predictive of crime risk, and trained and tested numerous potential models based on these variables. The best-performing model proved to be one known as the LASSO, which not only produced the best NDCG, but also required only 8 of the variables to be effective, thus yielding a readily interpretable model. Figure 1 shows the selected variables and their relative importance. Figure 2 shows two findings on one graph for v0.2:

1) The blue histogram shows the numbers of individuals in various ranges of predicted CRI (see scale on the left); 2) the red curve shows the probability that an individual will be involved in a serious crime in the following year. This shows that a small number of individuals (around 50) have a nearly 45% chance of being involved in a serious crime, with another 600 individuals having risk in

<table>
<thead>
<tr>
<th>Crime severity value</th>
<th>Crime category</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Homicide</td>
</tr>
<tr>
<td>70</td>
<td>Higher-severity violent</td>
</tr>
<tr>
<td>50</td>
<td>Lower-severity violent</td>
</tr>
<tr>
<td>25</td>
<td>Robbery / Agg. Stalking</td>
</tr>
</tbody>
</table>

*Table 1. Preliminary crime severity values*
the range from 15-30%. Our goal is to rank individuals according to their crime risk, so as to prioritize delivery of social services. Thus, we have adopted as our performance metric the normalized discounted cumulative gain (NDCG), which is a way to test the accuracy of a ranking scheme. This metric is on a scale of 0% to 100%, where 100% represents perfect ranking. We used this metric as a guide to optimize performance of model v0.2, and to test the models. We found that v0.2 achieved NDCG of 62%. To progress from v0.2 to the first deliverable version (v1.0), we will need to examine the predictions more carefully, align the CSV scale more precisely with that implied by the legal system, and think further about the optimization metric, to ensure that the results are as solid, dependable, and interpretable as possible. We also want to see whether we can better differentiate individuals at the highest end of the risk scale to ensure that they are prioritized as well as possible for social services. Once we and Elgin P.D. are satisfied with the model, it will be designated as v1.0. We will then put the model into practice, turning our attention to user training, and deployment of the pilot interventions in Elgin, while continuing in the lab toward v2.0 of the model.