ILLINOIS INSTITUTE OF TECHNOLOGY

impact

Nayar Prize II Phase I Quarterly Progress Report

January 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

The goal of this project is to design, implement, and deploy a flexible, new model for crime prevention that can be translated to a wide array of communities in the United States and beyond, thereby achieving far-reaching societal impact. In the first quarter, we met our goals of assessing the community needs, gathering data for the algorithm, interacting with experts, and analyzing Fourth Amendment legal constraints affecting the use of predictive algorithms. We shared our knowledge about the project and about predictive policing in general with media, Illinois Institute of Technology alumni, Illinois Tech classes, and police. We also participated in an academic/community group that is a clearinghouse and an incubator for research on, and services for, preventing gun violence.

Understanding Elgin's Needs and Resources

The Elgin Police Department's commitment to the project arises from the recognition that 10% of offenders account for 55% of crimes. During Fall 2016, we met with the Elgin Police Department (Elgin PD) to devise an overall strategy for the program. We determined that, once we develop the algorithm to predict who is at highest risk of committing future crimes, Elgin will deploy a social services unit that would act as case managers for the high-risk people who voluntarily participate in the program. Members of the social services unit were identified within Elgin PD during the Fall 2016. The social services unit will partner with various outside organizations to provide services to these high-risk individuals. Partners will include alcohol and drug addiction counseling and rehabilitation, mental health, job placement, and education services. Through our interaction with the Elgin Police Department in Fall 2016, we are convinced that they have the commitment and resources to undertake the implementation phase of the project.

Steps Toward Development of the Algorithm

We have obtained from the Elgin Police Department (Elgin PD) a crime database using Microsoft SQL Server. This database contains a total of 6,828 crime incidents during the period from January 1, 2012 to July 25, 2016. It involves a total of 22,049 individuals. For privacy protection, the sensitive information concerning these individuals such as the name, sex, race, address, etc. has been removed from the database, so that it will not be used in our model development.

CODE	DESCRIPTION
	Homicide
0110	1st-degree murder
0115	Homicide, unborn child
0120	Voluntary manslaughter, unborn child
0130	2nd-degree murder
0141	Involuntary manslaughter
0142	Reckless homicide
0150	Justifiable homicide
0160	Concealment of a homicide death
0165	Involuntary manslaughter & reckless homicide of an unborn child
0170	Drug induced homicide
0190	Solicitation for murder or murder for hire
	Criminal sexual assault
0260	Criminal sexual assault
0261	Aggravated criminal sexual assault
0262	Forcible sodomy
0280	Predatory criminal sexual assault of a child
0281	Criminal sexual assault with an object
	Robbery
0310	Armed robbery
0320	Robbery
0325	Vehicular hijacking
0326	Aggravated vehicular hijacking
0330	Aggravated robbery
	Battery
0410	Aggravated battery
0460	Battery
0470	Reckless conduct
0475	Battery of an unborn child
0480	Heinous battery
0485	Aggravated battery of a child
0486	Domestic battery
0487	Aggravated battery of an unborn child
0488	Aggravated domestic battery
0491	Aggravated stalking
0495	Aggravated battery of a senior citizen
	Assault
0510	Aggravated assault
0560	Assault

Table 1. Crimes identified for risk model development.

In the database, the crime incidents are stored in various data tables, including general crimes, narcotic arrests, warrants, gang affiliations, bookings, etc. Through discussions with Elgin PD, we have identified a set of “serious” crimes for developing our risk model of an individual becoming an offender.

From the crime tables, we are beginning to construct a social network to describe the offender-offender and victim-offender relationships. In this network, an individual (offender or victim) is identified by a unique number called JacketID. In the database, for each crime incident, all parties involved are identified, including offenders, victims, witnesses, and even parents if the crime is committed by a minor. This will allow us easily to identify both the offender(s) and the victim(s) of a specific crime (along with the type of crime) and to establish a link between them in the social network. Furthermore, the database also includes a link table created by Elgin PD, which provides information about how offenders, victims, family, and friends are socially connected. For example, an entry in this table could be “JacketID_1 is connected to JacketID_2 and the relationship is parent..” We are incorporating this table to introduce more connections in the social network. In our risk model, we will assign a risk score only to those who have been arrested at least once in the past. The development of our risk model involves training and testing phases.

To establish the platform for this process, we will use the crime data from April 1, 2012 to July 7, 2016 (4 years and 3 months). We will create two datasets, each of which consists of two years of data plus one year ahead for creating class labels. The two datasets overlap for a period of nine months. Specifically, the two datasets consist of the following: 1) Dataset 1: 11,165 individuals among which 1,101 have positive labels; and 2) Dataset 2: 10,738 individuals among which 1,130 have positive labels. Using these datasets, we have identified 23 variables that we will use in the predictive algorithm.

The Legal Analysis

We analyzed 100 court cases that dealt with law enforcement prediction of criminality and recorded the variables used in the predictions and analyzed how courts determined whether it was legally permissible to use each variable. We found a total of 123 variables in four different settings that law enforcement officers said they used to predict that a particular individual was likely to commit a crime or to have committed a crime.

The cases ranged in how many variables were involved in the particular prediction from the use of one variable to the use of 11 variables. Our findings provide a basis for a legal assessment of the algorithm we are developing. The court cases, which ranged from cases at the trial court level to cases decided by the U.S. Supreme Court, emphasized the importance of context, suggesting that the use of multiple variables in tandem is the most appropriate approach. For example, while a certain single variable used alone (such as location) is deemed inappropriate by courts, it can be permissible when used with other predictive variables.

We found that courts were more lenient toward law enforcement officials with respect to the factors they used to predict potential dangerousness at airports than elsewhere in public. We determined that the precedents for the public, non-airport settings were more appropriate for our use. Those cases emphasize that it is legally improper to use race as a predictive variable, and we are designing our algorithm in a way that does not use race itself as a variable, nor does it use factors that are linked closely to race. Even beyond its usefulness for the development of the algorithm, the extensive analysis of court cases about police prediction that a particular individual was likely to commit a crime or to have committed a crime is sufficiently important and unique that we will expand upon it and publish it as an article to guide judges in their work and to guide police departments in their training of officers.