HunchLab

Data Guide

Azavea | 340 N 12th St, Suite 402, Philadelphia, PA
info@azavea.com | T 215.925.2600 | F 215.925.2663
www.hunchlab.com

Introduction

HunchLab is a predictive policing solution that helps police departments to use their resources more effectively by leveraging advanced forecasts of crime. HunchLab's forecasting methodology fuses many crime theories and data sets into one picture of risk. The system automatically determines how to incorporate concepts such as recent crime events, temporal cycles such as day of week and season, the weather, and geographic locations such as bars and schools to produce a single forecast. The system uses these crime patterns when appropriate without requiring a police department to have a statistician on staff. This approach not only generates robust forecasts of crime but also provides insights into the dynamics of crime patterns.

The forecasting engine uses ensemble machine learning approaches that can incorporate the following crime patterns into a single prediction of criminal risk:

- Baseline crime levels
  - Similar to traditional hotspot maps
- Near repeat patterns
  - Event recency (contagion)
- Risk Terrain Modeling
  - Proximity and density of geographic features (points, lines, and polygons)
- Routine activity theory
  - Offender: proximity and concentration of known offenders
  - Guardianship: police presence (historic AVL / GPS data)
  - Targets: measures of exposure such as population, parcels, or automobiles
- Collective Efficacy
  - Socioeconomic indicators, neighborhood heterogeneity, etc.
- Temporal cycles
  - Seasonality, time of month, day of week, time of day, etc.
- Recurring temporal events
  - Holidays, sporting events, etc.
- Weather
  - Temperature, precipitation, etc.

Types of Information

Event Data

To forecast a space-time event such as a crime, HunchLab requires several years of historic data for the event to build both the outcome variable to be forecasted and several input covariates. At a minimum, HunchLab requires 5 years of crime (event) data for any event being modeled. This quantity of data is necessary to (1) "warm-up" variables that reach into the past up to 1 year, (2) include 3 years of examples to properly model seasonal patterns, and (3) hold back recent data to test accuracy.

This is the only required data set. Reasonably accurate models of crime can be generated with simply this data, but such models do not reveal insights into crime dynamics beyond crime events leading to more crime events.

Event data should be provided for at least the entire area for which forecasts will be used. If data can be provided for a buffer around this region, this can also be included. A buffer of up to 1000m can be useful within the modeling process. A reason to include additional data from nearby areas is that it may increase the overall data volume increasing predictive power. For instance, a small jurisdiction may not incur many violent crimes, but by including violent crimes from nearby jurisdictions more information is presented to the modeling process. Keep in mind that other data sets used in modeling must also be available for the buffered area.

Geographic Data

Geographic layers provide environmental context to the locations at which crimes occur. These datasets change slowly over long periods of time. While HunchLab's analysis is based on a raster format, geographic layers can be provided as points, lines, or polygon layers. HunchLab then builds variables based upon the distance to and concentration of these features. A given geographic layer may be split into multiple layers for the purposes of building covariates. For instance, given a street network for a city, each street segment may be of a different type – highways, highway onramps, residential streets, footpaths, etc. The distance to any street network feature may be a useful feature overall, but building variables for the distance to the nearest highway onramp or footpath may be useful as distinct variables. HunchLab can automatically split geographic layers on 'type' attributes to support this concept.

Implementation staff can easily take static extracts of geographic layers in ShapeFile format for inclusion in HunchLab. This approach requires no integration and therefore incurs no integration fees. Alternatively, a client may desire HunchLab to directly ingest GIS layers from a source such as an ArcGIS Server instance or other web API in GeoJSON format. In such cases, each system from which HunchLab pulls is considered one data connection and the relevant integration fee applies. Ingesting multiple GIS layers from a system does not incur additional fees.

While most geographic data provided by clients is in vector format, HunchLab can also leverage raster layers as variables. For instance, a city may have land cover data in raster format. Such data sets are transformed into a set of covariates and are resampled at the resolution of the HunchLab analysis.

Temporal Data

Temporal data sets provide information about the state of the entire jurisdiction and are considered "global" across the jurisdiction. For instance, a temporal data set may represent when the public school system is in session or the current air temperature. These data sets are provided in CSV format with the relevant time period, variable name, and a numeric value. School being in session may be represented as binary values

with a value of 1 when in session and 0 when not in session. The air temperature may be represented as a numeric value in degrees Fahrenheit. Alternatively, the severity of activity between two feuding gangs may be represented as integers: 0 for no activity, 1 for mild activity, 2 for severe activity.

It is important to realize that any temporal data used in forecasts must be available both historically for several years and for at least 48 hours into the future. The need for future values of the variable necessitates the use of variables that can be manually uploaded far in advance (such as the school schedule) or automation of updates (such as for weather). Temporal data sets that are uploaded into HunchLab manually do not incur integration fees. If instead, HunchLab was configured to automatically pull temporal data from a custom source, then a data connection fee would apply.

Other Variables

HunchLab also leverages variables that are not based upon specific data sets but are, instead, calculated. For instance, the day of the week and day of the month are simply calculated from the date. The moon phase, sunrise and sunset time, and season are other examples of variables calculated in a similar manner.

HunchLab Provided Data

HunchLab has processes in place to automatically manage the inclusion of common data sources if desired by the client. It should be noted that the use of these data sets is not required. For instance, a client may not desire any socioeconomic variables to be used in the forecasts even if academic research suggests it is useful.

Natural Terrain

Elevation data can be automatically loaded into HunchLab. This data set is transformed into several variables that describe the nature of the physical terrain such as the slope and aspect. This data is useful in identifying natural geographic structures that impact settlement patterns.

US Census

The US Census Bureau's American Community Survey provides up-to-date information about the US population based upon a sampled survey of residents. The data is available at the Census blockgroup level. HunchLab can automate the transformation of this data into relevant variables. For instance, this data set can provide measures of the collective efficacy and social cohesion of a neighborhood based upon socioeconomic indicators such as income and the prevalence of renters. The data set also includes information about potential targets of crimes such population density, automobile ownership, and home values.

Weather

Weather data provides a rich source of information about the conditions in a jurisdiction. For instance, seasonal patterns are often found in violent crimes, but these patterns may be more due to the conditions outside (warm temperatures) than the time of year itself. HunchLab can maintain historic weather data and upcoming forecasts automatically for inclusion in models. Variables include such items as the air temperature, humidity, perceived temperature, and precipitation.

Open Street Map

Open Street Map (OSM) is an online, collaborative project to create an editable map of the world. The OSM database includes detailed information about street networks and major points of interest such as schools, libraries, and transportation hubs. HunchLab can use this data if such layers are not readily available from the client.

Thinking About Data Sets

In addition to the above data sets, clients can provide geographic and temporal data for inclusion in HunchLab's models. While more information is often better in building predictive models, a few well-chosen data sets can go a long way to building an accurate and insightful predictive model of crime. We encourage clients to think about this process in an iterative manner as additional data sets can be added over time.

When evaluating a potential data set for inclusion in HunchLab, there are a few key questions to ask:

- Is the data already available? If not, what will be the cost to generate and maintain the data?
    - For instance, a geographic layer that changes infrequently may cost little to maintain while one that changes more often may be burdensome.
- How strongly connected to crime is the data?
    - For instance, if the crime within a jurisdiction drastically changes based upon changes in the student population at a local university, then data sets related to the university are likely quite important.
- Are there synergies between this and other data sets? In other words, does 1 + 1 = 3?
    - The locations of public schools may be useful by itself. The school schedule may also be useful by itself. By providing both school locations and the school schedule, the system can fully identify when and where school may be having an impact. Such related data sets may warrant evaluation as a group.
- Does one set of data represent many ideas?
    - For instance, a city's parcel database may include zoning or land use information that provides information about residential developments, hotels, fast food locations and more.

Ideas for Data Sets

Here are some ideas of data sets that may be useful to include within HunchLab:

- Where people congregate
  - Restaurants, fast food, bars, liquor licenses, nightclubs, places of worship, tourist attractions, movie theaters, exotic clubs
- Where people live
  - University dorms, fraternities, public housing, apartment complexes
- How people move around
  - Bus stops, bus stations, train stations, recreational paths, highway onramps
- Venues for particular types of crimes
  - Pawn shops, retail stores, malls, convenience stores, motels/hotels, ATMs, banks, parking lots, bike parking
- Government buildings
  - Police and fire stations, libraries, post offices
- Problem places
  - Abandoned buildings, vacant lots, foreclosed houses

Additional ideas may be gleaned from the literature reviews of relevant factors for each crime type available for download from the Rutgers University website at http://rutgerscps.weebly.com/publications.html