Predictive Analytics in Enforcement: Population Uncertainty and Lack of Information

Posted by China Layne

This is the fourth installment of our new blog series: Predictive Analytics in Enforcement. See our previous posts: What is Predictive Analytics?, Searching for Regulatory Violations, and Technical Challenges with Non-Random Investigative Data

Last week, we discussed the technical challenges with using non-random investigative data for enforcement. In this post, we will explore the challenges of population uncertainty and information gaps.

Challenge: Uncertainty about the population under regulatory enforcement

Many agencies only periodically identify and collect information on all organizations under their regulatory purview; or, they only do this for the specific organizations which they investigate. For instance, the agency may know that all businesses with employees fall under their regulatory enforcement, but may not know exactly how many businesses this includes, or the specific identity of these businesses. For these agencies, a predictive model that focuses only on estimating an organization's probability of violation will be of limited utility, because the model's results cannot be used to target for investigation organizations that are not identified in the agency's data.

Resolution: Leverage the explanatory power of the model's predictors

Summit has found that expanding the priorities of the model—beyond simply predicting the probability of violation—improves its usefulness for agencies that are uncertain about their regulatory population. The model would also emphasize accurate estimates of the correlation of specific organization characteristics with the probability of violation. With this information, the agency could target investigations on organizations, both currently identified and unidentified, which possess those characteristics. In a current project, Summit's model will allow an agency to target for investigation all organizations within a specific county with characteristics (such as a high unemployment rate) that are shown to be predictive of an organization's increased probability of violation.

Challenge: Lack of organization-specific information

Many agencies' administrative data collection is designed primarily for case management or performance measurement purposes. As such, it often includes a limited amount of information about the investigated organization. Many kinds of organization-specific information, such as indicators of financial health (which would be useful in accurately predicting an organization's probability of violation), may not be available from the agency's administrative data.

Resolution: Incorporate external data sources

Summit has found that linking agencies' administrative data to data from external sources can greatly improve the accuracy of the resultant predictive model. In recent work predicting the probability of an organization to be delinquent in paying fines, we incorporated financial data from Experian. This allowed us to control for the effect of an organization's ability to pay on their probability of delinquent payment. In a current project, we are linking an agency's administrative data with information from the American Community Survey (ACS) about the economic characteristics of the counties in which organizations operate. This allows us to control for the effect of regional economic climate on organizations' probability of violation.

Join us next week to explore the use of random sampling in enforcement.

Topics: Summit Blog, Predictive Analytics