Predictive Analytics in Enforcement: Technical Challenges with Non-Random Investigative Data

Posted by China Layne

This is the third installment of our new blog series: Predictive Analytics in Enforcement. See our previous posts: What is Predictive Analytics? and Searching for Regulatory Violations

Government agencies often use predictive analytics to improve the effectiveness and efficiency of regulatory enforcement. Their goal: minimize the time and cost required to find violators and change their behavior.

Administrative data on a particular agency's investigation and enforcement activities is often the primary–and best–data source for predictive modeling. However, using an agency's own administrative data can present unique challenges to developing an accurate and useful predictive model. Through our previous and current work in predictive analytics, Summit has developed several effective strategies for resolving these challenges.

In this post, we'll address the first of these challenges: the non-randomness of investigation data. Next week, we'll talk about population uncertainty and information gaps.

Challenge: Non-random investigation data

At most agencies, investigations are not conducted randomly. Agencies generally target their investigations based on national or regional level policy priorities, direct complaints, and referrals from other agencies. Because of this, there is usually some amount of selection bias in an agency's investigation data. The organizations that are investigated are likely to be very different from the organizations that are not investigated; these differences may be associated with an organization's probability of violating regulations. A predictive model based only on these non-randomly investigated organizations will produce inaccurate (biased) estimates of an organization's probability of violating regulations.

Solution: Augment the investigation data

There are several ways to augment an agency's investigation data and reduce the amount of selection bias in a predictive model based on that data.

- **Random selection**: The agency could conduct supplemental investigations on a small sample of randomly selected organizations. The data on this set of randomly-investigated organizations can be used to produce unbiased population estimates of violation rates for all organizations and for subsets of organizations. These unbiased estimates can then be used to quantify and correct for bias in the predictive model's estimate of an organization's probability of violating regulations. Summit used this method of augmenting non-random investigation data as part of refining a model to predict a pension plan's risk of violating Employee Retirement Income Security Act (ERISA) standards.
- **Reject Inference Method**: There are occasions when conducting an additional set of randomly selected investigations would be cost or time prohibitive for an agency. In these situations, the Reject Inference Method is a useful way to augment the agency's investigation data and reduce selection bias in the predictive model. In Reject Inference, organizations that have not been investigated are imputed an investigation outcome based on the specification of the predictive model developed using the investigation data. These non-investigated organizations with imputed outcomes are added to the investigation data and the predictive model is refined and re-estimated in an iterative process based on this new, expanded set of data. In this way, the estimates produced by the predictive model will better reflect an organization's probability of violating regulations in the full population of organizations, not just those that were investigated.

Join us next week to discuss how agencies can address population uncertainty and information gaps in their enforcement efforts.

Topics: Summit Blog, Predictive Analytics