ISSUES WITH FRAUD DETECTION

Fraud Is Rare

Fraud is usually a rare event and often exceedingly so. Identifying fraud is very difficult because of its rarity and because its very nature is stealthy. This stealthy action is directed against an external individual or organization (public or private) for the purposes of some sort of gain. The vast majority of the records (i.e., 99.9%) may be legitimate. Only 0.1% of the records may be fraudulent. It may be relatively easy to build a fraud model on these records that is 99% accurate (overall). For other modeling problems in business, this accuracy would be exceeding high. But this model would miss 9 out of 10 fraudsters! Much more time must be spent to identify many more of the 9 fraudsters that would be missed. Often, the extra accuracy is associated with higher cost, but the cost of not doing so may be much higher.

Fundamentally, fraud is a form of human response that can be modeled in ways very similar to customer response in business. But because of its rare and stealthy nature, the fraud signal is very diffuse and must be detected with much more rigorous methods than the more conventional responses of attrition and cross-sell/up-sell discussed in Chapter 16 on customer response modeling.

Fraud Is Evolving!

Fraudsters may adapt quickly to many fraud detection methods, by devising novel and increasingly subtle ways to get away with it. Also, fraud detection schemes must evolve also to try to keep up with (and get ahead of) fraudsters. This process is very much like the way flu viruses evolve. Flu vaccine designers try to craft new vaccines not just to confer immunity to strains of flu viruses they know, but to get ahead of the next epidemic. Fraud detection is a lot like that.

Large Data Sets Are Needed

Large credit card issuers like Capital One may process billions of transactions per year. Even a very small percentage of fraud among these billions of transactions can result in proportionately large losses. AT&T processed almost 300 million telephone calls each day in 1998 (Cortes and Pregebon, 1998). Phone fraud was one of the major incentives that prompted AT&T Bell Labs to develop Hancock, a large database computer system capable of analyzing huge volumes of call detail records. In addition to the fast computer systems, you must use fast and efficient algorithms to process all these data in time to make actionable any information related to fraud.

The Fact of Fraud Is Not Always Known During Modeling

Sometimes you can identify fraudsters, and sometimes you can’t. When you can “tag” a certain group of records as fraudulent, the analyses to model them are called supervised. The training of the model is supervised by the known identity of the fraudster. If you can’t