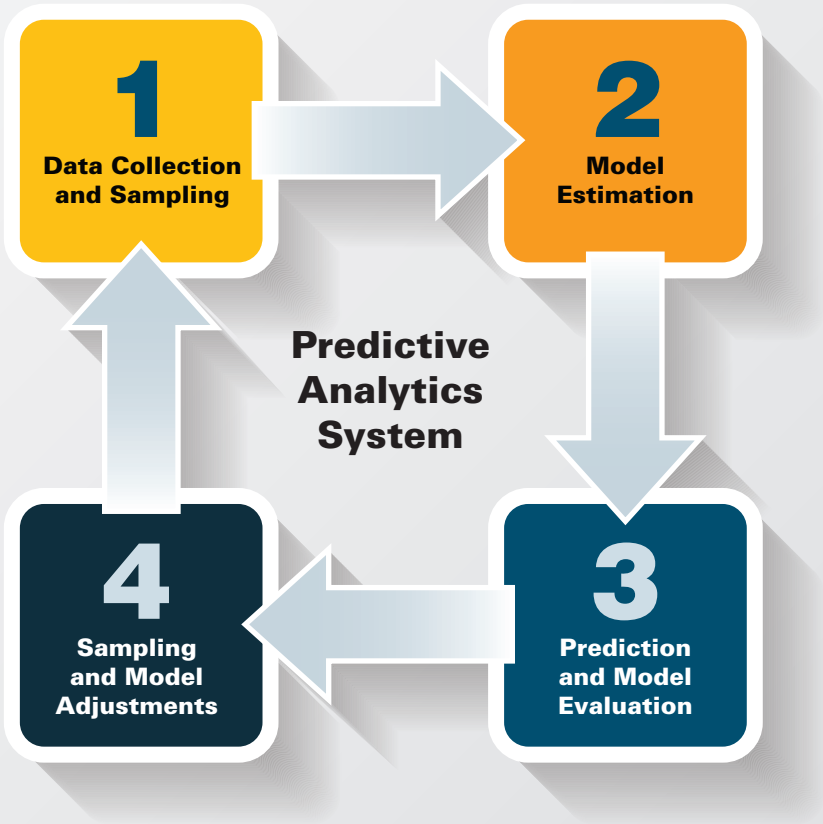
PREDICTIVE ANALYTICS:

the New Tool to Combat Fraud, Waste and Abuse

By: Albert J. Lee, Ph.D.

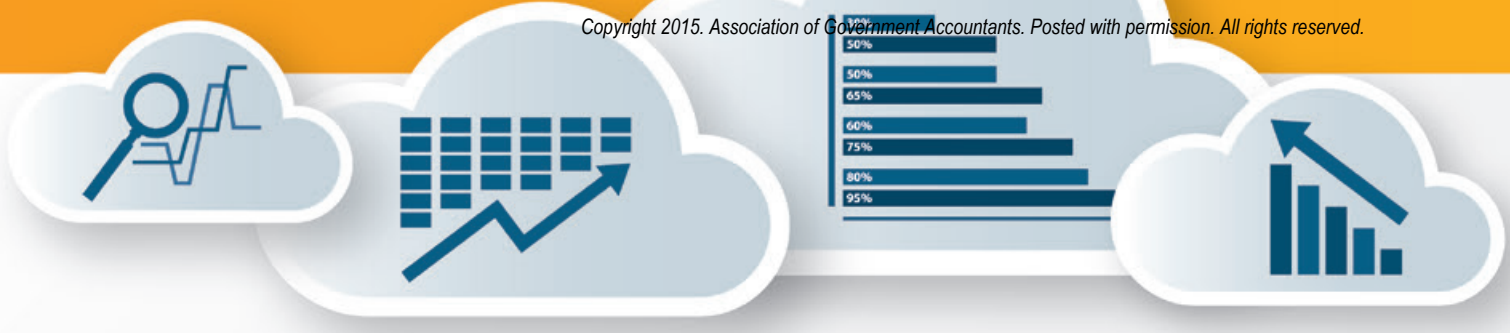
Government entities tasked with both regulatory enforcement and data analysis have an increasing number of data sources at their disposal. However, the data now at their fingertips can be increasingly complex, unstructured and unmanageable. To effectively manage raw data that may be coming from different data streams, a systematic approach needs to be taken for projects that require data modeling. Predictive analytics is a process that encompasses a series of methodologies that can successfully manage large-scale, data-driven problems that many government entities face. It is an iterative process that meshes the statistical methods of sampling, model estimation, model prediction and evaluation to form a cohesive system for targeting fraud, waste, abuse and other outcomes of interest to government agencies. Predictive analytics is a powerful tool that can assist agencies with decision- and policy-making in areas ranging from audit selection to regulatory enforcement.



Firms are implementing varied predictive analytics systems in both the federal and private sectors to a great degree of success. In the federal sector, structured modeling techniques have been used as predictive analytics solutions. Examples of this include a probabilistic simulation tool that determines the impact of various economic scenarios on mortgage insurance fund performance, a multistate model that predicts the number of defaulted loans in a federal credit agency's large loan portfolio, and a risk-ranking model that assists with enforcement of regulatory compliance by ranking enforcement subjects in terms of probability of compliance violation. Private-sector examples of predictive analytics include more unstructured work in text mining analytics to assist a document management corporation in categorizing its collection of written and scanned documents. Methodologies in sampling, model estimation and evaluation differ greatly in each of the aforementioned examples. However, it is clear that data and the effective use of data are leading to an increase in efficiency and transparency in both the public and private sectors.

SAMPLING AND DATA COLLECTION

Data collection is a key first step of predictive analytics. Many of the data sources from administrative records are triggered by specific events, which could be unrepresentative of the underlying population at large. Crime rate statistics are a good example. The identification of a crime presupposes an investigation. To the extent that investigations are uneven across



crimes, crime statistics are biased toward crimes that favor investigations. This would be different if investigations were statistically random. The resulting statistics would be unbiased, and present a true reflection of population patterns.

Consequently, a key component of any good predictive analytics process is a data collection and sampling design grounded in statistical theory and applicable to the problem at hand. The initial step of data collection must be well thought out, with all necessary variables and outcomes of interest documented in a data dictionary and stored in a secure database. A comprehensive database containing all potential data items that could be used in a predictive analytics tool is preferable. One can easily remove unneeded variables from a study; however, re-gathering information of interest can be expensive or impossible.

The way in which data are gathered (or sampled) must be understood if one is to gain insight into the population of interest. Sampling is defined as the selection of a subset of data within a larger target population.[1] The target population (e.g., all households in the United States) must be defined prior to data collection. Various sampling methods can be used to target subpopulations of interest, or to ensure observations from underrepresented subpopulations are present in the sample.

For example, administrative data on an agency's enforcement activities are typically non-random, and understanding how the data are collected is especially important when building a predictive analytics tool. The enforcement data gathering process is usually triggered by a precedent such as a referral from another government agency or a customer complaint. Additionally, enforcement subjects usually comprise a very small fraction of an agency's total enforcement population. When the non-random nature of these data is left unaccounted for, predictive methods will only detect what the existing investigations already look for, rather than true malfeasance in the enforcement population. Here, sampling design and associated sampling weights can improve the predictive analytics process.

There are three general sampling methods employed in data collection: simple random sampling, stratified sampling and multi-stage sampling.[2] A simple random sample occurs when data collection is such that every observation in the target population has an equal probability of being selected into the dataset for model estimation. Stratified sampling is most commonly put into practice for agency-related sampling designs. This is due to the ability to take a specific number of observations for a subpopulation of the population of interest. Multi-stage sampling is often used in geographical surveys.

PRACTICAL APPLICATION

Sampling in Practice

Large federal credit agencies and private banks often face the issue of identifying improper payments during financial audits, such as a loan that is disbursed to an ineligible entity. Sampling experts assist these agencies by first creating a random stratified sampling design, then stratifying payment amount, which allows for the calculation of loan origination and disbursement errors. Finally, the errors are projected out using sampling weights to estimate the actual dollar amount of improper payments for the audit population. Sampling will allow the agency to focus on the source of the eligibility problem. This can range from false documentation (fraud) to poor oversight or inefficient business processes (waste).

Many agencies, especially those tasked with regulatory enforcement, integrate sampling into their predictive analytics processes. For example, an agency that has over a million enforcement subjects cannot realistically investigate each subject for potential malfeasance. With limited staff hours and limited resources that can be devoted to enforcement investigations, the agency must select an investigation subpopulation, usually less than 1 percent of the total population. In these instances, it is very important to implement a good sampling design. Investigations are resource-intensive, and a poor sampling design could lead to even greater inefficiency.

MODEL ESTIMATION

Model estimation is considered the "bread and butter" of a predictive analytics tool, yet it is also the step in the predictive analytics process that takes the least amount of time, relative to data preparation and model evaluation. The majority of model estimation involves determining which type of model would best serve the problem at hand, identifying factors that would influence an outcome, and incorporating the sampling design into the algorithm. Model outputs can be continuous numbers, discrete categories or ordered categories.

After model selection, feature selection and/or feature engineering can be used to determine an optimal set of features relating to the outcome of interest. Next, an experimental design must be implemented to achieve consistent output from the model. To ensure that model estimates generalize well to any new incoming data, one can simulate a "current" dataset and a "future" dataset (also known as training and testing[3] datasets, respectively) by splitting the data into two groups, provided that the future data have a similar distribution to the current data. A standard split of current and future datasets involves designating

two-thirds of the data as the training set and the remaining one-third as the test set,[4] with two-thirds of the data used for model estimation training and remaining one-third for testing how well the model predicts.

Following data setup, the actual model building process occurs. Frequently used models, such as ordinary least squares regression models, are computationally quick and highly interpretable. More non-linear models must iterate over a set of tunable parameters to get an output. For these tunable models, a validation set must be used to determine the best model. To do this, the training set is further split into a training set and a validation set. This is done because a conflict of interest would occur if evaluation of the training set on the test set was used to determine the optimal model.

PRACTICAL APPLICATION

Model Building in Practice

Large grant-making agencies often face the problem of grant fraud. A risk ranking model is a commonly used tool for identifying and evaluating potentially fraudulent grants. This model tries to predict the risk of fraud from individual grants in the grant population so agencies can efficiently allocate their audit resources. For example, an agency first determines which model type outputs a continuous score or probabilistic response, so that individual grants in the population can be sorted in order of "riskiness." Agency stakeholders then identify a cutoff point for the risk score corresponding to the agency's available audit and/or enforcement resources. This process was particularly important during sequestration when federal agencies were required to reduce spending in every account by 10 percent, which invariably had a negative impact on oversight.

MODEL PREDICTION AND EVALUATION

Model prediction and evaluation includes the outcome predicted by the model and other resulting output. During this step, performance metrics are produced and diagnostic tests are performed.[5] Understanding a model's output and why this output was produced is an essential part of the iterative process of predictive analytics. This process allows the model to be improved over time, either due to the changing distribution of incoming data or the inadequacies in the selected model.

The obvious output of the model is the predicted outcome, which can be categorical, continuous, a set of group assignments, or any other outcome(s) of interest. For continuous outcomes, general model accuracy can be determined by examining how far off predicted values are from the actual values. For categorical responses, one can measure accuracy by seeing how many correct predictions occur relative to the total number of predictions.

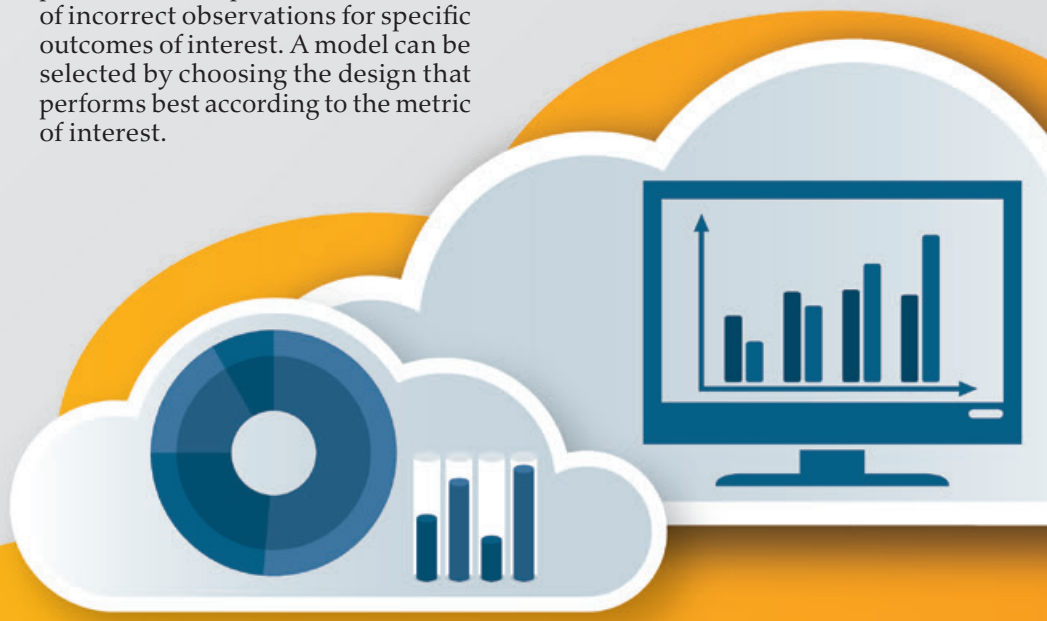
Depending on what one is looking to maximize or predict, different metrics can be looked at to evaluate model performance. For models with continuous responses, a common set of metrics relates to the model's complexity and the accuracy of its predictions. This set of metrics works more effectively with models that are simple yet still explain the data well. For categorical models, most metrics involve the number of correct predictions compared to the number of incorrect observations for specific outcomes of interest. A model can be selected by choosing the design that performs best according to the metric of interest.

PRACTICAL APPLICATION

Model Evaluation in Practice

Federal agencies holding large portfolios of loans often need independent validation and verification (IV&V) of their portfolios' financial health. This allows for a second review of the performance of their loans and provides the opportunity to ensure the correct valuation of outstanding debt. Independent auditors test the IV&V model to ensure that it correctly uses the input data and provides accurate calculations that perform the intended operations. With this appropriate level of independent financial review, the recent financial scandals on Wall Street, which in turn led to numerous federal financial bailouts, could have been avoided.

For all models, the main diagnostics revolve around whether the model suffers from high bias or high variance. High bias occurs when the model cannot approximate the underlying pattern in the data. This means that the model itself is inadequate, resulting in a high error rate. High variance means that the model overfits to the training data and does not generalize to the entire target population well.



ADVANCED ANALYTICS

Advanced analytics ties together the previously mentioned tools into an iterative system of data-driven solutions. These processes feed into each other to fine-tune the analytics model over time. By employing these advanced analytics techniques, large amounts of data can be processed more efficiently and with higher predictive power.

After data are collected, the sampling weights of the observations influence the model; observations with higher weights are treated as more important. The predictions that a model makes are based off of the type of model chosen and estimated. Relevant metrics are determined by what type of output the model gives — numeric, categorical or otherwise. Lastly, after examination of the predictions and residual output, one can determine if the model or data collection method used is insufficient for a given problem. If the model is insufficient, other models can be estimated and the process of model estimation and model prediction/evaluation repeated until an adequate model is produced. If the model is sufficient, the data must be recollected according to a newly designed plan, and the sampling, model estimation and model prediction process is repeated.

Even if the model performs well initially, this system needs to be continually updated to account for changing behaviors. More recent data could shift the distribution of the data, making the old model inadequate as time goes on. Therefore, new data must be added to the model estimation process and old data taken out to further refine the model. It is generally good practice to add new data into the model estimation process as it becomes available, since model performance usually improves with additional data.

Recent advances in model learning could be additional avenues for improving model performance. Model boosting allows the models to learn from their prediction errors by over-weighting false positives and negatives.[6,7] This learning process is initiated by incorporating past errors in the evaluation of incoming or new observations so previous data directs the prediction of new data and changes the nature of the model. This algorithm has been shown to improve model prediction performance in the long run.

To complete the modeling cycle, it is beneficial to generate a statistical sample in which selection probabilities are proportional to model predictions. This would steer the sample towards subjects with a greater propensity for an outcome, such as non-compliance, while maintaining its statistical randomness. The results from the evaluation of this sample usher in another round of modeling estimation and refinements.

It is important for advanced analytics systems to be well documented so that anyone can interpret and use the system through a user interface or business intelligence tool. These tools allow a complex advanced analytics system to be easily understood and tracked. Bugs and complaints should be recorded to improve the user experience and make necessary corrections.

PRACTICAL APPLICATION

Predictive Analytics in Practice

For large enforcement agencies, business intelligence tools can provide staff with a practical user interface for accessing model results and related data. These tools, when combined with the internal knowledge of enforcement staff, enable agencies to leverage data-driven relationships uncovered by predictive models to detect regulatory or criminal violations. In fact, predictive policing is being employed across the country to fight crime by using large data sets combining identical data elements from successful cases.

Endnotes

1. Thompson, S. K. (2012). *Sampling*. John Wiley & Sons.
2. Ibid.
3. Webb, A. (2002). *Statistical Pattern Recognition*. John Wiley & Sons.
4. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI (Vol. 14, No. 2, pp. 1137-1145).
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (Vol. 2, No. 1). New York: Springer.
6. See Endnote 3.
7. See Endnote 4.

***Albert J. Lee, Ph.D.** is the founding principal and expert economist of Summit. He has extensive experience managing consulting engagements that involve diverse teams of academic experts, economists and other professionals. Lee is an expert in econometric modeling and statistical sampling.*