

Jim Godfrey (Cont')

Partner/Family:

I am currently single, and my only close family member is my mom.

Pets:

I have three dogs: Chico, Honey and Harley. I also have a Senegal parrot named Kenya and an African grey parrot named Zulu.

Sports/Hobbies:

I belong to a medieval reenactment society where I study heavy armored fighting. I also scuba dive and play paintball.

What your ideal weekend would be:

My idea weekend would be a private jet to Cozumel, Mexico, for two days of diving on the incredible reefs!

Favourite foods:

I like ANY Indian or Thai food...oh, and sushi!

If I could be anything at all (besides a SAS consultant), I would be...

...a dolphin trainer! My first degree was in marine biology, and I always wanted to work with dolphins.

When I'm not involved on SAS consulting projects, I like to ...

Continue my SAS education through the courses offered by SAS Education and attend data mining conferences like the M-series of conferences sponsored by SAS and held every year in Las Vegas.

Something I've learned out in the field doing consulting work for SAS that I feel would benefit other SAS users is ...

...the power of applying oversampling to binary predictive modeling problems, particularly where the overall number of modeling records is very high and the number of binary cases of interest are rare. I will illustrate the concept of oversampling with an example below:

Consider a predictive modeling problem where you are trying to predict, say, the likelihood that a noncredit card customer at a bank will pick up a credit card product. Furthermore, let's say that there are 2 million potential credit card customers in your modeling population and only 2,000 of these customers became credit card customers within the defined target window (usually one month).

Now let's say that in this dataset of 2 million customers, 2,000 of which have the event of interest, there are 1,500 potential input variables that include customer characteristics such as demographic information and past behaviour in other banking products, such as account balances and transaction history.

Now it is possible to use this dataset in its raw form for predictive modeling and you will, eventually, get pretty good modeling results. However, this is a very large dataset and using it in its raw form for modeling would require very large amounts of processing power, space and time. In addition, the target event is very rare with only 0.1 percent of customers experiencing the event of interest! Many predictive modeling techniques use measures of prediction accuracy as the default modeling objective. Therefore, a model would be 99.9 percent accurate by just predicting that no customers will take the credit card product. This is a very accurate and totally useless result! Wouldn't it be better if the target event weren't so rare and the data size were not so large? Well, the solution is oversampling!

Oversampling involves creating a sample from the modeling population where the density of event classes is enriched. This is done by selecting all of your event records and a random subset of your nonevent records, thus enriching your population with event cases. For the example above, we would first select all 2,000 of the records where the event of interest is present (i.e., the customer became a credit card customer) and a random subset of the nonevent records (i.e., the customer did not become a credit card customer). Therefore, if we wanted to create an oversampled population where 50 percent of the records are event records, we would select all of the 2,000 event records and 2,000 of the 1,998,000 remaining non-event records, thus creating a modeling set that has 50 percent events and 50 percent nonevents. We could also create a modeling dataset where one-third of the records are event records and two-thirds are nonevent records by selecting all 2,000 event records and 4,000 nonevent records. By doing this we have 1) reduced the data size from a daunting 2 million observations to a mere 4,000 observations that can easily be processed on a standalone PC, and 2) enriched the target event population which makes differences between event and nonevent cases easier to detect.

Simulation studies have shown that models built using oversampled populations perform, on average, just as well as models that are built on the entire dataset. Therefore, you can get about the same results using much less computing resources. However, the models that you will build using the oversampled dataset will give results applicable to the oversampled population and NOT to the original 0.1 percent event rate population. Luckily, this model "bias" is easily corrected using the following simple formula:

$$p_1 = \frac{\tilde{p}_1(\pi_1/\rho_1)}{\tilde{p}_0(\pi_0/\rho_0) + \tilde{p}_1(\pi_1/\rho_1)}$$

Where \tilde{p}_1 is the predicted probability of event from the model built on the oversampled population, \tilde{p}_0 is the predicted probability of a nonevent from the model built in the oversampled population, π_1 is the actual population probability of event (in this case $2,000/2,000,000 = 0.001$), π_0 is the actual population probability of nonevent (in this case $1 - 0.001 = 0.999$), ρ_1 is the probability of event in the OVERSAMPLED population (this would be 0.33333 in the case of a 1/3 oversample, or 0.5 in an 50/50 oversample), and finally ρ_0 is the probability of nonevent in the OVERSAMPLED population ($1 - \rho_1$).