Modeling Challenges:

Rare Event Modeling – Credit Default Portfolios – Methods from a Practitioner

Clayton Botkin

When modeling rare events such as credit default, there are very few 'events' or defaults. This is especially true considering changes in credit 'tightening' that we've seen following the financial crisis. We may be dealing with portfolios with just a few or less than 100 defaults. When modeling in this situation, it's often difficult to properly estimate models that are meaningful or even produce models that 'fit' at all. Statistical procedures such as logistic regression can significantly underestimate the probability of rare events. This is exasperated when dealing with certain asset-classes such as Residential Mortgage where a panel dataset may have millions of observations – and only a 1 or 2 percent default rate.

Since 1994, I've been using stratified sampling and weighting methods in SAS and recently in R to handle rare event modeling. My applied econometric background focuses on the simplicity of the technique.

This briefing provides our perspective on modeling methods and determining alternative methods and techniques that are useful and defensible to address challenges when modeling rare events.

As a leader in Model Risk Management, Montana Analytics has been active in developing models and utilizing rigorous analytical methods for examining models since 2002.

Rare Event Modeling – Credit Default Portfolios

Often, when modeling rare events such as credit default, there are very few 'events' or defaults. This is especially true considering changes in credit tightening that we've seen following the financial crisis. We may be dealing with portfolios with just a few or less than 100 defaults. When modeling in this situation, it's often difficult to properly estimate models that are meaningful or even produce models that 'fit' at all. Statistical procedures such as logistic regression can significantly underestimate the probability of rare events. This is exasperated when dealing with certain asset-classes such as Residential Mortgage where a panel dataset may have millions of observations – and only a 1 or 2 percent default rate.

Rare events can often be described as low frequency, high severity events and there are many examples: default in credit risk for various loan types, insurance fraud, stock market crashes, and disease outbreaks. With loan defaults, even when the event results in a low severity outcome, the loss often far outweighs the marginal revenue of the loan.

Since 1994, I've been using stratified sampling and weighting in SAS and recently in R to handle rare event modeling. A solid academic analysis of this method is provided by King (2001) in "Logistic Regression in Rare Events Data". This method is also useful in OLS modeling.

Weighting is a procedure that weights the data to compensate the differences in sample and population. Modelers use the term "1's" or "bads" broadly in modeling events of interest while "0's" or "goods" are non-events. So, we then sample all the 1's (rare events) and a fraction of 0's (non-events). In such cases we have to weigh the observations accordingly. Calculating weights is simple. If we know the sampling fraction for each case, the weight is the inverse of the sampling fraction.

Weight = 1 / (sampling fraction)

Example: In a population of 1,000,000 loans there are X defaults. In this case you would:

1. Sample all known 'event' observations (i.e. 100% of the bads = x)
2. Create a weight variable 'wt'. Assign a '1' on each bad observation.
   Design Weight = 1 / (sampling fraction) = 1 / 100% = 1

3. Sample 10% of the 'non-event' observations (10% of 1,000,000 = 10,000)
4. Using the same weight variable 'wt'. Assign a '10' on each good observation.
   Design Weight = 1 / (sampling fraction) = 1 / 10% = 10

The weighting is related to the sample proportions and the modeling is called the Weighted Maximum Likelihood method. You do not need to create a 'balanced' data set with an equal number of 1's and 0's. All you need is a sufficient number of 1's for the maximum likelihood to converge.

A basic experiment can demonstrate identical outcomes:

Balanced Sample: 10 % of the 1's and 10% of the 0's – use a weight of 10 for both
Weighted Sample: 100% of the 1's and 10% of the 0's – use a weight of 1 for the 1's and 10 for the 0's

In both cases, you will get identical estimates. Again, the idea of weighting is related to sampling. If you are using the complete (natural) data set you should not weight it.

The sampling fraction could also be the over-sampling amount for a given group or subpopulation. We've built *stratified sampling* methods that included weighting observations across many dimensions to ensure all product types, states or other dimensions were appropriately included (weighted) in the sample. The outcome ensured the model results were robust in each of these dimensions. Multiple dimensions can get complicated – so we will keep this discussion as is – over-sampling of the rare events (bads) – as we've defined above.

In SAS, here is the sample code for the logistic application:

```
proc logistic data=training;
        model y(event="1")=x; weight wt; output out=out p=probs;
        run;
```

In R, we use glm. Here is the sample code:

```
glm(y ~ x1 + x2, weights=wt, data=training, family=binomial("logit"))
```


