A Monte Carlo study on methods for handling class imbalance

Mark H. White II | markhwhiteii@gmail.com

Method

Data Generating Process

Two class data were simulated by adapting the `caret::twoClassSim` R function (Kuhn, 2008):

- Two multivariate normal predictors (A and B) are generated. A and B are correlated at $r = .65$. These two variables contributed to the log-odds by $4A + 4B + 2AB$.
- Another variable, $J \sim U(-1, 1)$, was generated. This variable further added to the log-odds by $J^3 + 2 \times \exp(-6 \times (J - 0.3)^2)$.
- Two more variables, $K \sim U(0, 1)$ and $L \sim U(0, 1)$, were generated and contributed to the log-odds by $2 \times \sin(K \times L)$.
- For each data set, a number X was selected, where $X \sim N(50, 7)$. Another number, Y, was selected, where $Y \sim N(.15, .033)$. $Z = X - (X \times Y)$ variables were generated from a $N(0, 1)$ distribution. Each of these Z variables further added to the log-odds in a simple additive fashion, where coefficients were (a) of alternating signs and (b) evenly spaced from 2.50 to 0.25.
- $\frac{Y}{2}$ variables were generated from a $N(0, 1)$ distribution and did not contribute to the log-odds.
- The log-odds for each case were converted to probabilities. For each data set, a positive (i.e., minority) class proportion, M, was sampled from $N(.03, .007)$. Probabilities were sorted from lowest to highest. The difference between the probability for the $1 - M$th highest probability and M was calculated, and this constant was added to the probability for each case.
- Lastly, the number of cases in each data set were randomly drawn from a distribution $N(40000, 5000)$. 500 data sets were generated, and sixteen combinations of sampling techniques and algorithms were fit to each of these data sets.

Sampling Techniques

Algorithms

Metrics

Results