Using weights to improve the reliability of classification metrics with survey data

Adway S. Wadekar and Jerome P. Reiter
Department of Statistical Science, Duke University, Durham, NC

Abstract
Surveys are commonly used to facilitate empirical social science research. Due to several constraints, they are often not simple random samples. Therefore, respondents are usually assigned weights indicative of their relative worth in a statistical procedure. It has been proven that using weights produces unbiased estimates of population totals and accurate explanatory models of outcomes. However, predictive modeling, which has become popular in the social sciences, does not traditionally incorporate representative weighting in model development or assessment. This research investigates whether weighted performance measures on survey testing data, used with well-established model development approaches, produce reliable estimates of population performance. We test this using simulated stratified sampling, both under known relationships between predictors and outcomes and with real-world data. We show that unweighted metrics on sample testing data for models subject to default train/test cycles do not represent population performance, but weighted metrics do. We also show that the same holds for models trained using methods directly orthogonal to population representation, such as upsampling for mitigating class imbalance. Our results suggest that regardless of development procedure, weighted metrics should be used when evaluating performance on sample test data.