Incorporating survey weights within tree-based and other classification methods

Summer Research Proposal

Student: Adway Wadekar
Mentor: Prof. Jerome Reiter

May 5, 2022

1 Background

When sampling from a population of interest, it is difficult to sample in such a way that the observations survey are representative of said population due to logistical, cost and other constraints (Pfeffermann, 1996). As a result, survey weights have long been used to generate samples that are representative of the population from which observations are drawn. Usually, they take the form of numerical values assigned to each observation, which is an indicator of its worth or importance in a statistical procedure (Johnson, 2008). They were originally used to develop weighted sample statistic estimates (such as proportions and means) that would be analogous to those calculated if the sample is a simple random sample from the population of interest. The inclusion of survey weights has also been extended to regression techniques such as multiple linear regression (Dumouchel & Duncan, 1983; Reiter, Zanutto, & Hunter, 2005; Faiella, 2010).

However, researchers have found that a substantial proportion of studies that use data from complex survey designs dont use survey weights in their statistical analyses (Bell et al., 2012), and other researchers highlight a debate as to whether or not survey weights are really needed in such analyses (Bollen, Biemer, Karr, Tueller, & Berzofsky, 2016). Moreover, a recent trend in epidemiological studies–specifically within the field of psychiatric epidemiology–has been to use techniques beyond generalized linear models to analyze data (Jiang, Gradus, & Rosellini, 2020). Researchers now often use classification and regression trees (CART) for their analyses, and the use of black-box machine learning models such as neural networks is also on the rise. With this has come a shift from simply conducting inferential modeling to the use of models for classification tasks. However, traditional metrics used to evaluate the strength of predictive models, such as sensitivity, specificity, positive and negative predictive value, etc. do not consider the respective value (defined through survey weights) of each observations classification decision. In addition, techniques used to train classifiers to perform predictive tasks, such as synthetically generating data to handle class imbalance, go directly against the grain of weighting observations to represent the population of interest.

2 Project Design

Therefore, the goal of my summer research project is to understand the role of survey weights in these various contexts. I will begin by learning, in a systematic way, about how survey weights are derived. Then, I will analyze data to begin to answer two research questions:

R1: Is it possible to create performance metrics that consider the respective importance of each observation? When these new metrics are applied to previous studies which do not consider survey weights, does prediction accuracy change?

R2: Is the consideration of survey weights necessary to train tree-based and other classifiers. Tree-based classifiers currently are able to consider survey weights through the case.weight parameter, but are traditional methods for balancing classes (up-sampling, down-sampling and SMOTE) better choices for weighted accuracy?

R3: From a simulated population, when we take random samples from the population according to a sampling design and fit an unweighted tree to it, what are the predictive accuracy measures?

R4: How do we incorporate survey weighting into algorithmic fairness measures?

To test these of these questions, I would like to conduct an experiment similar to one which addressed measurement error in random forests using quantitative bias analysis (Jiang, Gradus, Lash, & Fox, 2021). This study used the National Comorbidity Survey Replication (NCS-R) to test whether accounting for the misclassification of predictors yields a change in predictive performance and variable importance. Using the same data set and a similar predictive task, I would like to test whether the inclusion of survey weights changes the true predictive performance of the model.

References

Bell, B. A., Onwuegbuzie, A. J., Ferron, J. M., Jiao, Q. G., Hibbard, S. T., & Kromrey, J. D. (2012, July). Use of Design Effects and Sample Weights in Complex Health Survey Data: A Review of Published Articles Using Data From 3 Commonly Used Adolescent Health Surveys. *American Journal of Public Health*, *102*(7), 1399–1405. doi: 10.2105/AJPH.2011.300398

Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., & Berzofsky, M. E. (2016). Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. *Annual Review of Statistics and Its Application*, *3*(1), 375–392. doi: 10.1146/annurev-statistics-011516-012958

Dumouchel, W. H., & Duncan, G. J. (1983, September). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, *78*(383), 535–543. doi: 10.1080/01621459.1983.10478006

Faiella, I. (2010, January). *The Use of Survey Weights in Regression Analysis* (SSRN Scholarly Paper No. 1601936). Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.1601936

Jiang, T., Gradus, J. L., Lash, T. L., & Fox, M. P. (2021, September). Addressing Measurement Error in Random Forests Using Quantitative Bias Analysis. *American Journal of Epidemiology*, *190*(9), 1830–1840. doi: 10.1093/aje/kwab010

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020, September). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. doi: 10.1016/j.beth.2020.05.002

Johnson, D. R. (2008, November). *Using Weights in the Analysis of Survey Data.*

Pfeffermann, D. (1996, September). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, *5*(3), 239–261. doi: 10.1177/096228029600500303

Reiter, J. P., Zanutto, E. L., & Hunter, L. W. (2005, October). Analytical Modeling in Complex Surveys of Work Practices. *ILR Review*, *59*(1), 82–100. doi: 10.1177/001979390505900105