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

Adway S. Wadekar and Jerome P. Reiter, Ph.D.
Department of Statistical Science, Duke University

Introduction & Motivation

- **Empirical research in the social sciences, particularly in disciplines such as public health rely on data collected through surveys.**
  - When sampling from a population of interest, it is difficult to sample in such a way that the observations selected are representative of said population due to logistical, cost and other constraints (Pfefferman, 1996).
  - Non-representative samples are collected, and weights are assigned to each observation in the survey and are a measure of its relative worth in inferential statistical procedures (Johnson, 2008).
  - Weighting has been extended to techniques like regression to decipher relationships between outcomes and predictors (Dumochel, 1983).
- **There has been a shift in focus in the social sciences from explanatory and inferential modeling to prediction (with explanation done along the way).**
  - The goal is now to develop models based on sample data to identify those in the population that might be at risk for a disease, etc.
  - Predictive modeling does not traditionally incorporate weighting in model development or assessment, and some techniques for development go directly against the grain of weighting for representation.
- **How should weights be incorporated into the new paradigm of predictive analysis?**

Background, Definitions, Research Questions

Derivation of Weights

- **Commonly-used method for deriving weights:** inverse probability sampling.
  - In a stratified design, each observation has a probability of selection into the survey, based on relative proportions of the strata. Their weights are simply the inverse of this probability of selection (Lohr, 2009).
- $p_i = \frac{n_k}{N_k}; w_i = \frac{1}{p_i}$, are the probabilities and weights for individual i, where k is individual i's stratum. n_k is the number of observations in stratum k in the survey and N_k is the number of observations in stratum k for the population.

Horvitz-Thompson Estimator of Totals

- Horvitz and Thompson (1952) developed an estimator for population totals that use survey weights, which, on average equals the population total.
- The estimate for totals multiplies the survey weights by the count for each observation and adds them up.

Traditional and Proposed Performance Metrics

Confusion Matrix

		Observed (Ground Truth)	
		Positive	Negative
Predicted	Positive	True Positive $(TP_i = 1)$	False Positive $(FP_i = 1)$
	Negative	False Negative $(FN_i = 1)$	True Negative $(TN_i = 1)$

$$Sensitivity = \frac{\sum_{i=1}^{n_t}(w_i)TP_i}{\sum_{i=1}^{n_t}(w_i)TP_i + (w_i)FN_i}$$

$$Specificity = \frac{\sum_{i=1}^{n_t}(w_i)TN_i}{\sum_{i=1}^{n_t}(w_i)TN_i + (w_i)FP_i}$$

- **Sensitivity (True Positive Rate):** Ability of the classifier to correctly identify those who are positive.
- **Specificity (True Negative Rate):** Ability of the classifier to correctly identify those who are negative.
- **Area under ROC curve (AUROC):** Analyze the tradeoff between sensitivity and specificity at various decision thresholds.
- Traditionally, in these metrics, each observation is treated equally $(w_i = 1)$.

Research Questions

- **R1:** Can we create more reliable Horvitz-Thompson-like estimates $(w_i = \frac{1}{p_i})$ of population accuracy of models?
- **R2:** Do the estimators of accuracy remain reliable when the analyst uses complex classification training methods (up/downsampling, SMOTE, etc.)?

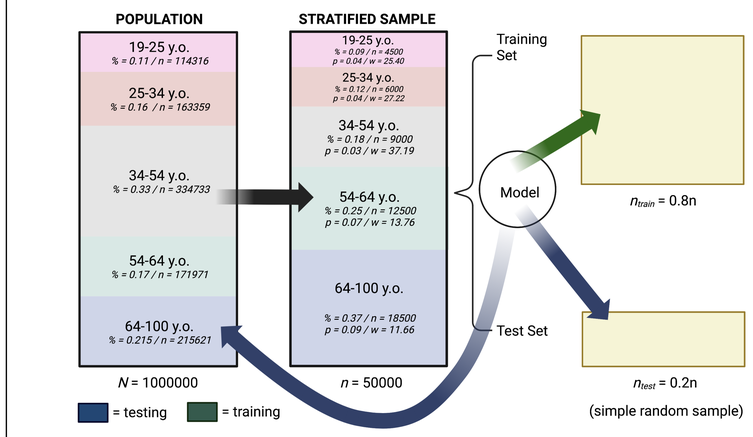
Methods

Theoretical Results

- **Assuming a constant model**, we show that the expectations of the (up)weighted Horvitz-Thompson test set counts of true positives, false positives, true negatives and false negatives are **equal to the population counts**, if the model were applied to the entire population.
- Sensitivity and specificity are **ratios of these estimators**. The bias of these **ratios is negligible** with large n.
- **In reality, the models are never held constant** and are dependent upon the train/test cycle.

Initial Simulation Design

- **Simulate a population** of 1,000,000 records in which the connection between the outcome and the correlated three explanatory variables is determined by a logistic regression.
- Simulated (skewed towards older ages) stratified sampling by age category.
- Used logistic regression to "predict" a fairly balanced outcome, across **50 train/test cycles from a constant stratified sample.**
  - Generated estimates and errors for the average "true" (population-level) performance across all possible models and the same for the average "extrapolated" (unweighted and weighted) performance in the non-representative test set.
- Using the same model ensures that there is an engineered relationship to be found. However, **it also introduces the train/test cycle variability allowing for the model to change** during each run.



Schematic of sampling design and train-test framework.

- The average weighted test sensitivity and specificity were 0.92 and 0.90 respectively, which matched the average population performance of 0.92 sensitivity and 0.90 specificity.
- The average unweighted sensitivity was 0.94 and specificity was 0.87.

Real Data Application

- Use real data from the National Survey on Drug Use and Health (NSDUH) and **treat it as population data.**
  - "Population" was data of respondents 18+ from 2016, 2017 and 2018 surveys. Outcome was any illicit drug use in past year and there were 17 predictors, all of which are categorical.
- There was **no predetermined, simulated relationship**, introducing potential new variability in model fit.
- Outcome **was also imbalanced,** which provided a test bed to see if weighted predictive performance is **robust to highly altered model training.**
- Repeated **a similar experimental design** with both logistic regression and random forests 50 times. Kept the population and sample fixed throughout .
  - In random forests (where balanced training is often applied) we used up-sampling to **balance the training set** to better predict the minority class. We left the non-representative test set imbalanced. We also repeated the experiments without balancing the training set, to see if the reliability of weighted metrics was affected.

Results – Data Application

Metrics	Population	Weighted	Unweighted
LOGISTIC REGRESSION			
Avg. Sensitivity	0.50 (± 0.0001)	0.49 (± 0.005)	0.44 (± 0.005)
Avg. Specificity	0.91 (± < 0.0001)	0.92 (± 0.002)	0.94 (± 0.001)
Avg. AUROC	0.81 (± 0.0001)	0.82 (± 0.002)	0.83 (± 0.002)
RANDOM FOREST			
Avg. Sensitivity	0.48 (± 0.0008)	0.47 (± 0.006)	0.42 (± 0.006)
Avg. Specificity	0.92 (± 0.0004)	0.92 (± 0.002)	0.94 (± 0.002)
Avg. AUROC	0.81 (± 0.0002)	0.80 (± 0.003)	0.81 (± 0.003)
BALANCED RANDOM FOREST			
Avg. Sensitivity	0.67 (± 0.001)	0.65 (± 0.005)	0.60 (± 0.005)
Avg. Specificity	0.80 (± 0.001)	0.80 (± 0.003)	0.85 (± 0.002)
Avg. AUROC	0.81 (± 0.0002)	0.80 (± 0.003)	0.81 (± 0.003)

Point estimates and errors of average true and predicted population metrics. n = 50 trials. 95% confidence.

- **For sensitivity and specificity** at a 50% decision threshold, we find that the average weighted (Horvitz-Thompson) non-representative test set performance falls much closer to the average population performance than the unweighted test set estimators.
  - The ranges in which we are 95% confident the average unweighted test set performance metrics lie across models do not tend to intersect with the same ranges for weighted test set performance.
  - The ranges in which we are 95% confident the average population performance metrics across all models lie are much closer to the same range for average weighted test set performance across all models (sometimes intersecting) compared to the range for average unweighted test set performance.
- **For AUROC,** the weighted test set average estimates do not substantially differ from the population average estimates, nor from the unweighted test set estimates.
- **This suggests that, on average, the inclusion of weights for metrics at commonly-used decision thresholds will provide a better indication of how a model will perform in the population than not including weights.**

Conclusions, Reflections, Future Work

- **The key takeaway is that the inclusion of weights matters when assessing model performance.**
  - We suggest that survey analysts report weighted performance metrics.
  - This includes when using methodologies to alter the training data for more accurate results. Although the variance of the weighted estimates increases, it is still substantially closer to the expected range of those models' population performance.
- **Future directions include** understanding the role of these weighted estimators across non-strata domains, replicating previous predictive survey studies, and developing a pipeline for weighted prediction into an R package.

Acknowledgments

- I am very grateful for the support and guidance of my mentor, Prof. Reiter, which allowed this project idea to come to life.
- This work was supported by the Huang Fellows Program and the Duke Initiative for Science and Society.

Primary References

1. Lohr, S. L. (2009). *Sampling: Design and Analysis* (2nd edition). Cengage Learning.
2. Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663–685.