Predicting Infant Mortality: Minimizing False Negatives

In 2015, India failed to meet one of its Millennium Development Goals to bring Infant Mortality Rate (IMR) down to 26 in 1000 live births. The national average has hovered around 36 for the last five years. Despite the fact that there are over 20 schemes currently in operation that aim to reduce IMR, India's decline in IMR has been consistently slowing down.

For my dissertation, I want to build a predictive model that uses a set of easily obtained variables to classify pregnant mothers into high-risk and low-risk categories, in order to narrowly target the benefits of public health efforts. Lemon et al. (2003) outline two approaches traditionally employed to segment out part of a population that is at high-risk for a particular health condition. The first is to simply compute the likelihood of observing the health issue conditional upon belonging to a particular pre-defined subgroup of the population. While this is useful for descriptive purposes, it does not allow for a simultaneous consideration of several independent factors. The second is regression analysis, in this case, usually logistic regression because the outcome variable is dichotomous (Hosmer & Lemeshow, 2000). Regression analyses compute the average effect of an explanatory variable on our outcome of interest and hence, when policy is developed from these results, they are targeted at the average member of the population, without accounting for the fact that certain subgroups are disproportionately vulnerable to some health risks (Forthofer & Bryant, 2000). Even though we can explore the impact of interaction terms, interpretation becomes progressively more difficult as more variables are interacted together.

With growing evidence that the actual relationships between health outcomes and their explanatory variables are complex and nonlinear (Song et al. 2004), recent studies in epidemiology have begun to use decision trees and other modern prediction methods for identifying high-risk groups vulnerable to bacterial infections among infants (Bachur & Harper, 2001), colon cancer (Camp & Slattery, 2002), coronary heart disease (Carmelli, et al. 2007), etc.

Tesfaye, et al. (2017) created a model to predict under-5 mortality using the Ethiopian demographic and health survey data. Breast-feeding, maternal education, family planning, preceding birth interval, occurrence of diarrhea, father's education, birth weight and mother's age were found to be predictors of child mortality. They found that a pruned decision trees method has greater accuracy of prediction than a logistic regression approach with an accuracy of 90.38% and area under ROC (Receiver Operating Characteristic curve) of 94.8%. This model was written into a web-based algorithm for use in areas without well-trained health professionals, where users can enter certain key measureable pieces of information and then the model classifies the child as being high-risk or low-risk.

Chen, et al. (2011) took a novel approach to building a predictive model for preterm births, one

of the biggest causes of new-born deaths, using a combination of a neural network and a decision tree. They collected data on thousands of variables covering medical history, lifestyle factors, socio-economic variables for both parents and first used a neural network to identify the 15 most important factors that affect the likelihood of a preterm birth. Following this, Chen et al. used a decision tree to arrive at a set of rules for classification into high-risk and low-risk categories based on these fifteen variables. They find that multiple births, paternal drinking, and smoking, previous preterm births and low body weight for the mother are some of the best predictors of preterm births. They arrived at a set of ten different rules for classification, which are easy to interpret algorithmically, with precision ranging from 80% to 100%.

My proposed study is distinct from this literature in that my focus is specifically on minimizing false negatives during prediction. Given that the objective of this model is to help target public health resources to save more infant lives, we are most concerned with minimizing false negatives, i.e., we want to reduce the number of at-risk babies being misclassified as healthy. This is challenging because the training data available is severely imbalanced. Among the children in the dataset I propose to use in my study, less than 10% actually died before the age of 1. So, even if a model 'trained' on this data blindly predicted 'Survive' for every child, it would still achieve 'accuracy' of around 90%. I argue that the more appropriate metric for model evaluation in this context is false negative rate.

Data

The National Family Health Survey (NFHS) is a nationally representative survey that is conducted in India every ten years. It is carried out by the Ministry of Health and Family Welfare, Government of India, along with the International Institute for Population Sciences, Mumbai. The last two rounds were carried out in 2015-16 and 2005-06, respectively. Funding for the NH-3 was obtained from the United States Agency for International Development (USAID), the Department for International Development (DFID), the Bill and Melinda Gates Foundation, UNICEF, the United Nations Population Fund, and the Government of India. For this study, I use data from NHFS-3 (2005-06).

Typically, Demographic Health Surveys (DHS), publish data that can be accessed via an application to the USAID, in four different datasets - household data, individual women's data, children's data, and household listing data. Most of the variables of interest to this study come from the individual women's dataset, which covers a range of questions on deliveries, infant-care, maternal health-care, nutrition, a few questions on the woman's agency and safety in the household and as of the NFHS-3, even a host of anthropometric measures such as height, weight, hemoglobin levels for women and children. Of key relevance, is the section on birth history, where each respondent can describe the details of the birth of up to 20 children they have had. All respondents are women of reproductive age, between 15 and 49.

The NFHS-3 consists of 124,385 recorded surveys in the individual women's dataset. Out of these, all observations about women who had never given birth to a child are dropped, bringing the sample down to 84,609. For each woman interviewed, information is recorded on up to 20 past births. This data when extracted, provides information on 256,782 recorded births.The DHS surveys only go into further detail regarding the specific circumstances of a delivery for each respondent's last five births. This means that if we want to figure out whether variables such as the baby's birth weight, the place of delivery, the number of prenatal doctor visits and breastfeeding are good predictors of infant mortality, we are restricted to a much smaller subset of the full dataset. We thus drop all observations that do not record this data, bringing the number of observations down to 28,999. Further, using household IDs from the individual women's dataset, household-level variables such as access to drinking water, sanitation and wealth index, have been merged into the dataset.

Methods

The nature of this dataset poses what is called the Class Imbalance Problem in Machine Learning, meaning that the training data 'teaches' the model disproportionately about the majority class, to the point where the model does not 'know' how to recognize a member of the minority class. Liu, et al. (2009) suggest the use of a Random Forest along with some techniques to 'correct' the imbalance to address this issue. Batista (2004) explores a host of different ways to do this, including random undersampling, random oversampling, Tomek links, synthetic minority oversampling (SMOTE). My preliminary results evaluate two balance correction techniques. The variables I use for all three new Random Forest models are drawn from the literature outlined above and include individual characteristics of the birth such as birth order, mother's age, mother's anemia and household level socio-demographic variables such as caste, religion, state of residence, wealth index, etc.

Table 1 : Model Evaluation Metrics

	Area under ROC	False Positive Rate	False Negative Rate
Model A: Logit	0.67	0.61	0.93
Model B: RF	0.66	0.07	0.85
Model C: RF, Upsampled	0.67	0.07	0.48
Model D: RF, Downsampled	0.66	0.08	0.55

Model A is the original, unchanged dataset fitted to a logistic regression. Model B is the same dataset fitted to a classification-based random forest model with a depth of 50 trees. Model C uses an 'upsampled' modification of the dataset, where the data on deceased children is randomly sampled with repetition until we have as many observations as we do about children that are alive. Model D uses a 'downsampled' modification of the dataset, where a subset of the data on children that are alive is chosen, that only has as many observations as are available on deceased children. Table 1 displays basic evaluation metrics for all three models. For a standard of comparison, Tesfay (2017) achieve an accuracy of prediction of over 90% using pruned decision trees, but does not discuss the issue of an imbalanced sample - the 'useless predictor' that serves as baseline is not clear.

We find that in all the Random Forest models, False Negative Rates have indeed fallen, especially modelsC and D, and False Positive Rates are highly minimized. Area under the ROC remains roughly the same. These preliminary results should be treated cautiously. Clearly, downsampling methods require intense cross-validation given how dependent they are on the specific subset of the majority class data that is sampled out for that iteration. For the purposes of a quick back-of- the-envelope calculation, I ran the random forest model with downsampling (Model D) using a different seed, to find a False Negative Rate of 0.73. Upsampling, on the other hand, inordinately weighs each observation in the minority class because of the vast number of duplicates created. This has two consequences - first, that 'more' data does not necessarily translate to more information in this case, and second, that outliers could get undue consideration in training the model.

Moving forward, I want to use classification methods that allow me to use a custom loss function that penalizes misclassification resulting in false negatives more harshly than misclassification resulting in false positives. I intend to compare and contrast results obtained from both random forest models and support vector machines with non-linear kernels that lend themselves to using asymmetric loss functions.

REFERENCES

1. Bachur, Richard G., and Marvin B. Harper. ”Predictive model for serious bacterial infection among infants younger than 3 months of age.” Pediatrics 108, no. 2 (2001): 311-316.
2. Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. ”A study of the behavior of several methods for balancing machine learning training data.” ACM Sigkdd Explorations Newsletter 6, no. 1 (2004): 20-29
3. Camp, Nicola J., and Martha L. Slattery. ”Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States).” Cancer Causes and Control 13, no. 9 (2002): 813-823.
4. Carmelli, Dorit, Heping Zhang, and Gary E. Swan. ”Obesity and 33-year follow-up for coronary heart disease and cancer mortality.” Epidemiology (2007): 378-383.
5. Chen, Hsiang-Yang, Chao-Hua Chuang, Yao-Jung Yang, and Tung-Pi Wu. ”Ex- ploring the risk factors of preterm birth using data mining.” Expert Systems with Applications 38, no. 5 (2011): 5384-5387.
6. Forthofer, Melinda S., and Carol A. Bryant. ”Using audience-segmentation tech- niques to tailor health behavior change strategies.” American Journal of Health Behavior 24, no. 1 (2000): 36-43.
7. Hosmer, David W., and Stanley Lemeshow. ”Special topics.” Applied Logistic Regression, Second Edition (2000): 260-351.
8. Liu, David, Yudie Yuan, and Shufang Liao. ”Artificial neural networks for opti- mization of gold-bearing slime smelting.” Expert Systems with Applications 36, no. 9 (2009): 11671-11674.
9. Song, Xiaowei, Arnold Mitnitski, Jafna Cox, and Kenneth Rockwood. ”Comparison of machine learning techniques with classical statistical models in predicting health outcomes.” Medinfo 11, no. Pt 1 (2004):736-40.
10. Tesfaye, Brook, Suleman Atique, Noah Elias, Legesse Dibaba, Syed-Abdul Shabbir, and Mihiretu Kebede. ”Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach.” Computer Methods and Programs in Biomedicine 140 (2017): 45-51.