Predicitng Infant Mortality in India

Sushmita V Gopalan*

June 4, 2017

Perspectives in Computational Research

Abstract

In this paper, I attempt to build a simple model to predict the risk of a child dying with a year of its birth, given a host of variables about their mother, socioeconomic and demographic data. The variables found to be significant by a logistic regression approach are consistent with the literature on the determinants of infant and child mortality in developing countries. However, the model has very high False Negative Rates, the metric we seek to minimize, because of being highly imbalanced in terms of available data towards children who are alive. I test two methods- random oversampling and random undersampling, to address the class imbalance problem, using random forests, and achieve moderately better results in terms of False Negative Rates.

keywords: mortality, public health, machine learning

JEL classification: D91, E21, H30

*University of Chicago, M.A. Computational Social Science, sushmitavgopalan@uchicago.edu.

1 Introduction

In 2015, India failed to meet one its Millenium 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 5 years. Even though there existed, as of 2013, over 20 schemes in India that targeted a reduction in IMR (Press Information Bureau, 2013), decline in IMR has been consistently slowing down, suggesting that there is a need to go beyond disease-, program- and sector-specific approaches (Claeson et al. 2000).

These schemes aim to achieve a variety of goals, having a range of different target populations. The Janani Suraksha Yojana, for instance, is aimed at promoting institutional deliveries. The Navjaat Shishu Suraksha Karyakram (NSSK) is a programme that trains healthcare providers in essential newborn care and resuscitation. One scheme offers pregnant mothers a Mother and Child Protection Card to help monitor healthcare service delivery through the Ministry of Women and Child Development. These schemes are not uniformly implemented across states either. There appears to be no overarching logic to the collection of schemes nor have there been rigorous independent evaluations of the impact of each scheme or any explorations into their interplay with each other. This essentially means that it is not clear that tax dollars are being utilized to effectively serve the most vulnerable populations.

It is as integral to good policy-making to scientifically determine who needs intervention most desperately, as it is to rigorously analyse whether these interventions serve their stated aims or not. In this paper, I attempt to come up with a predictive model based on a set of variables that are easily obtained, in order to narrowly target the benefits of public health efforts. Tesfaye et al.(2017), built a simple model to predict in Ethiopia that was over 90% accurate in its predictions and wrote that into a rudimentary web-based algorithm that could be used as a black box tool by public health workers in remote locations, to classify children into risk-categories and allocate expenditure and resources accordingly. In this paper, I use data from the National Family Health Survey in India from 2005-06, part of the family of Demographic Health Surveys maintained by the United States Agency for International Development.

2 Literature Review

2.1 Determinants of Infant Mortality in India

What is now a canonical framework to analyze child mortality was put forth by Chen and Mosely in 1984, who called for variables that are specific to the child (such as anthropomorphic data, medical information) to be studied along with socio-economic, demographic and environmental-level factors, in the exploration of the determinants of infant mortality. I focus my analysis to the socioeconomic determinants of mortality - variables basic enough that any ASHA (Accredited Social Health Activist) worker could collect and plug into a model that yields a risk-score.

Income is generally considered to be the variable most correlated with infant mortality (Barbus, 2011 and Hobcraft, 1985) . Maesham et al. (1999) trace out the role of income in changing infant mortality rates between 1975 and 1990. While income did have statistically significant influence on IMR, other factors such as technical progress and education levels had more substantial impact. Less than 25% of Indias reduction in IMR in this period can be explained by income growth. When compared to other developing countries, they find that Indias reduction in IMR over this time period is lower than would be predicted by the corresponding increase in income. In this paper, which uses DHS data that does not include household or individual income, I use the DHS's Wealth Index measure, which divides the sample into five quantiles as a proxy.

A World Bank report from 1999 finds that while the poorest Indian states have the worst infant mortality rates, the richest states do not have the best (Maesham et al. 1999). The states with the best indicators - Kerala and Tamil Nadu rank seventh and eleventh, respectively, in terms of per capita income. On the other hand, Delhi and Goa, which are the two states with highest per capita income do not even feature in the list of ten states with lowest infant mortality rates. Non-income factors such as maternal and child health interventions are found to play more significant roles in reducing infant mortality (Claeson et al. 2000). I thus include two dummy variables in my analyses, indicating whether or not the child in question lives in one of the five states with highest HDI or the five states will lowest HDI (as per the Public Affairs Index[1])

Tamil Nadu forms an interesting case study to explore the impact of state-driven initiatives to improve maternal and child health services. Relative to the rest of India, infant mortality decreased rapidly in Tamil Nadu, from 80 in 1995 to 21 in 2007 (Padmanabhan et al. 2009). Concerted efforts to improve infrastructure such as standardizing a maternal death registration and audit, setting up certified obstetric and newborn-care centres, changes in incentive structure to attract medical officers to rural areas were found to be the primary driver of this reduction in IMR (Padmanabhan et al. 2009).

Consistent with this, Cleaeson et al. (2000) find that there is a significant positive relationship between lowered infant mortality rates and certain child health interventions like oral rehydration therapy, care seeking for acute respiratory infections, and immunization rates. Other important factors have been found to be nutrition status, age of the mother, employment status of the mother, whether or not the delivery was institutional, access to healthcare during pregnancy and time since previous birth (Saabneh, 2017)

A WHO report from 2005[2] provides a set of guidelines on factors associated with infant and child mortality, from the mother's education levels, anemia levels, to the sex of the baby, time elapsed since the mother's previous pregnancy, etc.

To briefly summarize the literature on the determinants and causes of infant mortality in developing countries, the variables that have been found in the past, to be associated with infant, and in general, child mortality, can be divided into

[1] http://pacindia.org/2016/07/28/measuring-the-quality-of-governance-of-indian-states/

[2] WHO (2005): World Health Report 2005: Make every mother and child count. Geneva - http://www.who.int/whr/2005/en/

three categories- 1) Personal and biological characteristics of the mother and the child - weight with respect to median, malnutrition, age of the mother, time since previous birth, etc. 2) Parentss health status and behaviour - smoking, drinking, dietary habits, tetanus, anemia, attitude towards health seeking, awareness levels 3) Community - sanitation, public health facilities, access to insurance, communicable diseases, cultural attitudes toward health care, womens empowerment, etc.

2.2 Overview of Commonly Used Methods

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 provide 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 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), as described above, 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 diarrhoea, fathers education, birth weight and mothers age were found to be predictors of child mortality. They find that a pruned decision trees method has greater accuracy of prediction than a logistic regression approach or a decision tree without pruning, with an accuracy of 90.38% and area under ROC 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) take 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 arrive at a set of ten different rules for classification, which are easy to interpret algorithmically, with precision ranging from 80% to 100%.

The gap in the literature that my research seeks to plug, apart from its policy-related goals, is is building a predictive model to classify at-risk infants and target policy efforts accordingly, in the Indian context.

3 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 womans data, childrens data, and household listing data. Most of the variables of interest to this study come from the individual womans dataset, which covers a range of questions on deliveries, infant-care, maternal health-care, nutrition, a few questions on the womans 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.

At the outset, there were 124,385 recorded surveys in the individual womans dataset. Out of these, all observations about women who had never given birth to a child were dropped.. This brought the sample down to 84,609. Further, for each woman interviewed, information was recorded on up to 20 past births. This data was extracted, and converted from wide to long, to result in a dataset that recorded 256,782 births. Further, using household IDs from the individual womans dataset, household-level variables such as access to drinking water, sanitation and wealth index, were merged into the dataset. The creation of the master dataset was done using Python 3.7. I also merged Human Development Index (HDI) rankings for each Indian state into the dataset. Table 3 shows the prevalence and associated Infant mortality rate (number of children who die before the age of 1 in every 1000 live births). From this, two separate datasets were created.

Dataset A: Since my variable of interest is dichotomous, i.e. whether or not a child survives up to the age of 1, and Im using a logistic model for my analysis, the dataset has to be censored. Children who are alive and under 1 year of age at the time of the interview are dropped (bringing the number of observations down to 197,884) because they have not as yet been fully exposed to mortality risk - that is, they could still have died after the interview. Children who died over the age of 1 are dropped as well, resulting in 195,660 observations. Further, missing data on caste (down to 188,810), access to water and sanitation (down to 185,361), anemia (down to 169,445), and finally, parents education levels, brings the number of observations down to 167,714.

Dataset B: The DHS surveys only go into further detail regarding the specific circumstances of a delivery for each respondents last five births. This means that if we want to figure out whether variables such as the babys 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, when it comes to the impact of breastfeeding on mortality, Palloni and Millman (1986) caution against the danger of a reverse causation bias because babies who die young would have a very short period of breastfeeding recorded, even if they were breastfed for the entire duration of their short lives. To complicate things more, the smallest unit of measurement for 'duration breastfed' in the DHS is 1 month, which means that all children (approximately 3000 in my sample) who died within a months of their birth, would have 'duration breastfed' recorded as 0. Taking into account the dropping of observations for children who are alive and below the age of 1, this resulted in information of breastfeeding being available for less than a 1000 children. Hence, despite the fact that extensive literature on infant mortality places primacy on the positive impact of breastfeeding on infant survival, I do not include it as part of my analysis.

In order to make for more straightforward interpretation of regression coefficients, some variables were recoded. My response variable is set to take 0 if the child died on or before their first birthday and 1, otherwise. Recall that observations on children who died beyond the age of 1 have been dropped from the analysis. and The variables on access to drinking water and sanitation facilities are recoded into Improved and Non-Improved facilities, based on guidelines given by the UNICEF(2006). Eight different questions on tobacco usage are combined into a single dummy variable that takes the value 1 if the respondent smokes and 0, otherwise. Following the approach used by Kaldewei (2010), I recode the respondents age at the time of the childs birth into two dummy variables for under 20 years of age and over 35 years of age. Several studies have found a U-shaped relationship between mothers age and child mortality, dropping past the age of 20 and rising again in the mid-thirties (George & Ahmad (1992), McCormick et al. (1984). A categorical variable was created for the childs year of birth, representing each decade before the survey. Finally, note that the variable containing information about anemia is recorded as being from 1 to 4 with decreasing severity, and not, as is intuitive, from 1 to 4 with increasing severity. I did not recode this in order to be able to compare easily with other studies usind DHS data.

4 Model

Since my response variable is binary and I'm interested in being able to predict whether a child is at high risk of infant mortality or not, my logical choice of model is a logistic regression, as specified below.

$$P(Y = 1|X = x) = \frac{e^{x\beta}}{1 + e^{x\beta}} \tag{1}$$

The probability that event Y = 1 will occur, conditional on a covariate vector x is determined by a logistic function of the vector x and the vector of coefficients .

Similar to Kaldewi(2010), I run two separate logistic regressions first. Using a set of 23 different variables drawn from the literature outlined above, I run a logit regressions on training subsets of Dataset A (N= 138,024), the Dataset B (N=18,287). Of these 23 variables, only 18 are available for Dataset A. Then, from the the variable importance table for Model 2, I pick the top ten variables and run

Model 3, using those, again, on a training subset of Dataset B.

Recall, at this point, that the response variable is set to take 0 if the child survives and 1 if they do not. So, variables that are statistically significant and whose coefficients have positive signs, are estimated to decrease the log-odds of survival.

5 Results

We first examine the results of the regressions from Models 1 and 2. As the coefficient plot in Figure 1, and the output of the regression models in Table 4 show, the variables identified as significant by both models are not dissimilar. In line with the literature and our expectations, the household's Wealth Index, Mother's Education Level, living in a state with a high HDI index (as a proxy for infrastructure) and a longer preceding birth interval,are shown to positively impact the log-odds of survival. Multiple births is both significant with a p-value ¡ 0.01 and has a high, negative magnitude. Being born to a young mother, further down the birth order, in a rural region, in a low-HDI state are all shown to have significant, negative impacts on the log-odds of a child's survival.

Curiously, deliveries in public institutions, private institutions and at home are also shown to be significant in Model 1. A closer look at Table 3 however, shows that these are less than 1% (rounded to 0) of the sample, with a very high IMR. We find later, though, that only 'delivery in public institutions' is among the ten most important variables, but it becomes negative and no longer statistically significant in Model 3. Further exploration of other interacting factors is required to work out why this is the case.

Potentially owing to the far larger sample available to Model 1, it is clear from Figure 1, that all the coefficients lie within tighter confidence intervals. In Model 3, we find that nearly all of the variables that were deemed most important by Model 2 are statistically significant, with the exception of 'delivery in public institution',

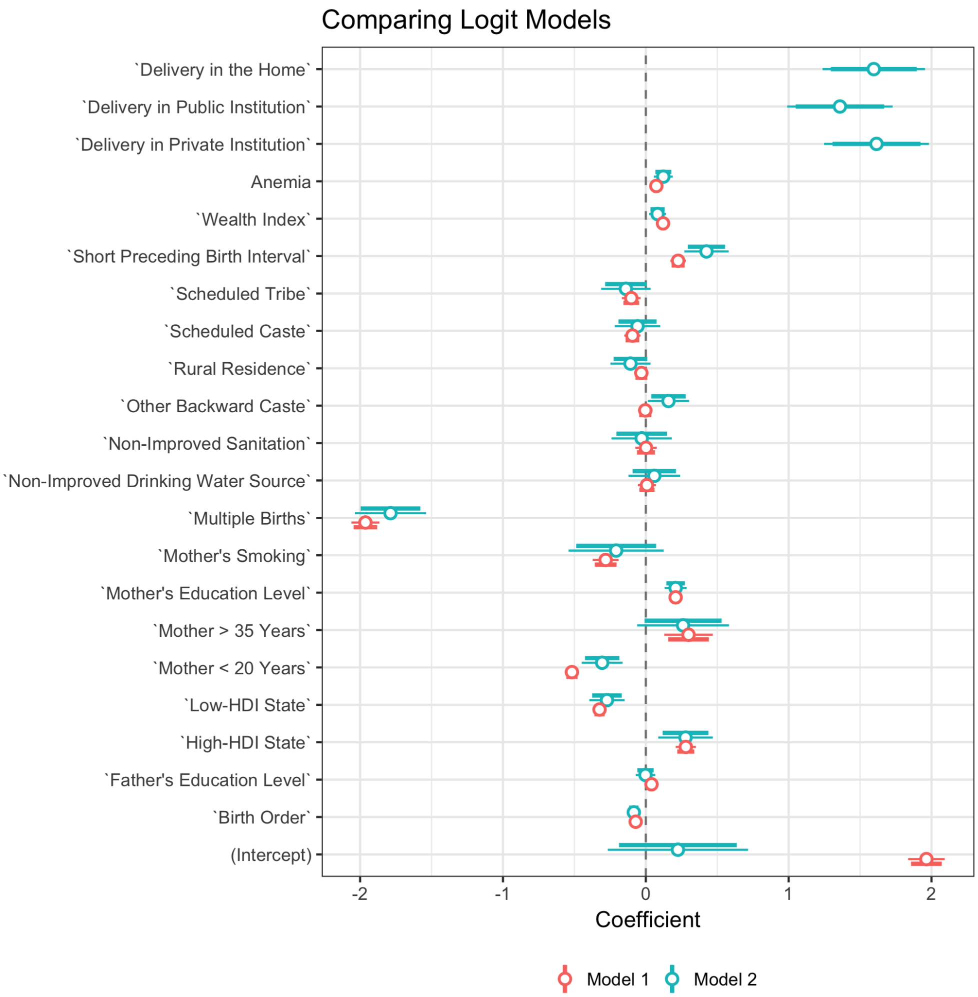

Figure 1: Regression Coefficient Plot: Models 1 and 2

as I already mentioned.

Now, I want to explore the predictive abilities of the three models. Recall that I had only run the models on approximately 70% training subsets and held the remaining 30% aside for the purposes of testing. I first evaluated the accuracy of prediction for each model and found that each predicted whether or not a child would survive up to the age of 1 with an accuracy of over 90%. However, closer inspection of other evaluation metrics quickly revealed that the my sample is so imbalanced (only 8% of Dataset A and 9% of the Dataset B) is about children who died before the age of 1) that even if my model was to blindly predict 'Survive' all the time, it would still be right around 90% of the time. Just as a stopped clock is

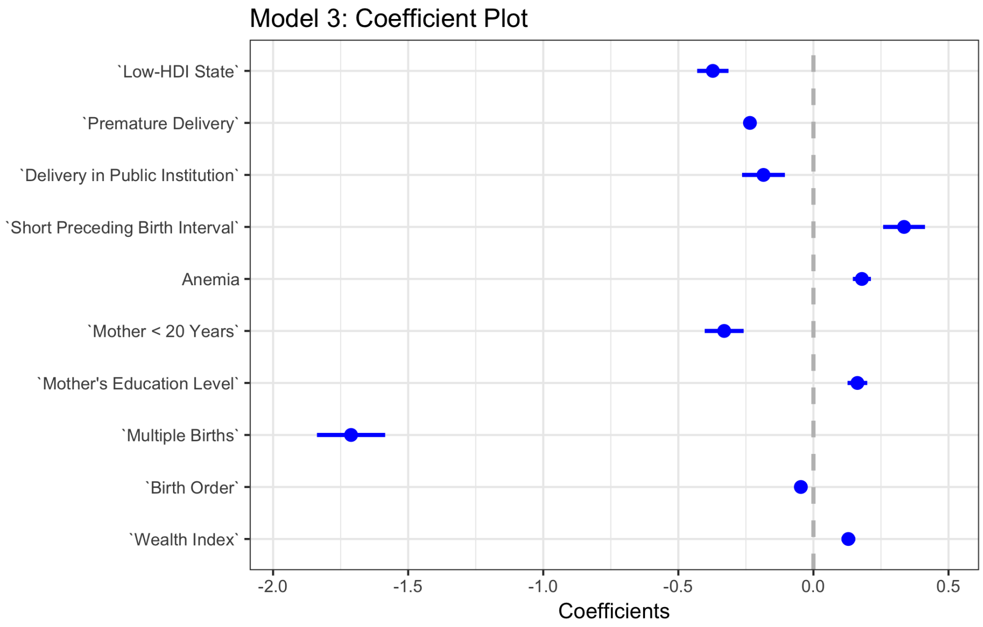

Figure 2: Regression Coefficient Plot: Model 3

right twice a day, a prediction model trained on a highly imbalanced sample is also very often correct, merely by chance.

What we really need to look at, keeping in mind, particularly, my motivation for this study - to come with a good predictive model that can identify at-risk babies and focus public health efforts on them, is the False Negative Rate. A high false negative rate would mean that babies that were actually at high risk, were not being identified as such, and hence, fell out of the purview of preventive and curative measures, resulting in illness, or worse. I find, predictably, that the false negative rate for Models 1, 2 and 3 is higher than 92% and hence, clearly unacceptable for our purposes.

While a high false positive rate is also unsuitable for our purposes (public health resources are scarce and highly competed for in developing countries), the trade-off is certainly in favour of minimizing the false negative rate.

This 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

Table 1: Prediction Evaluation Metrics

	AUROC	Accuracy	False Positive Rate	False Negative Rate
Model 1	0.66	0.92	0.61	0.93
Model 2	0.67	0.92	0.48	0.92
Model 3	0.67	0.91	0.55	0.92

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). While this is an entire body of literature in its own regard, I take a first pass as trying two balance correction techniques, and compare it to a Random Forest with the original dataset used in Model 3. The variables I use for all three new Random Forest models (Models 4,5 and 6) are the same as those used in Model 2 and the new, semi-synthetic datasets are created from Dataset B.

Model 4 is the original, unchanged Dataset B, fitted to a classification-based random forest model with a depth of 50 trees. Model 5 uses an 'upsampled' modification of Dataset B, 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 6 uses a 'downsampled' modification of Dataset B, where I randomly choose a subset of the data on children that are alive, that only has as many observations as I have on deceased children. Table 2 displays basic evaluation metrics. For a standard of comparison, Tesfay(2017) achieved 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 three cases, False Negative Rates have indeed fallen, especially models 5 and 6, and False Positve Rates are highly minimized. Area under the ROC remains roughly the same. These preliminary results should be treated very cautiously, because upsampling and downsampling are not very rigorous methods to correct for an imbalance problem. Clearly, downsampling methods require intense

Table 2: Prediction Evaluation Metrics (Random Forests)

	AUROC	False Positive Rate	False Negative Rate
Model 4	0.66	0.07	0.85
Model 5	0.67	0.07	0.48
Model 6	0.66	0.08	0.55

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 5) 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.

6 Conclusion

In this paper, I used data from the National Family Healthy Survey in India, and attempted to build a predictive model for infant mortality that could take in variables easily measured and recorded by public health workers and return a risk prediction, with a low rate of false negatives. To this end, I first fit three logistic models to determine which variables had most impact, and were most predictive of infant mortality. I found the variables flagged as significant by my analysis to be consistent with the literature - wealth index, mother's education, living in a state with good health infrastructure, being born early in the birth order, etc. However, these models do not serve my predictive goals, because they have an extremely high False Negative Rate and their predictive ability beyond a 'useless classifier' is barely discernible - one reason for this could be that all my training data is highly imbalanced in favour of children that are alive (92%).

I then try two simple, common techniques to address the problem of imbalance

- upsampling and downsampling, along with random forests that are 50 trees deep. While false negative rates did indeed fall, serious cross-validation is required for these estimates to be credible. My next step would be to use classification methods that can be modified to weigh false negatives as more 'costly' than false positives, such as cost-sensitive neural networks (Zhou, 2006) and support vector machines (Akbani et al.)

7 References

Akbani, Rehan, Stephen Kwek, and Nathalie Japkowicz. "Applying support vector machines to imbalanced datasets." In European conference on machine learning, pp. 39-50. Springer Berlin Heidelberg, 2004.

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.

Barbus, Alexandra . 2011. "DETERMINANTS OF INFANT MORTALITY." Europolis 5, no.

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

2: 154-178. Political Science Complete, EBSCOhost (accessed April 24, 2017).

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.

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.

Chen, Hsiang-Yang, Chao-Hua Chuang, Yao-Jung Yang, and Tung-Pi Wu. "Exploring the risk factors of preterm birth using data mining." Expert Systems with Applications 38, no. 5 (2011): 5384-5387.

Claeson, Mariam, Eduard R. Bos, Tazim Mawji, and Indra Pathmanathan. "Reducing child mortality in India in the new millennium." Bulletin of the World Health Organization 78, no. 10 (2000): 1192-1199.

Forthofer, Melinda S., and Carol A. Bryant. "Using audience-segmentation techniques to tailor health behavior change strategies." American Journal of Health Behavior 24, no. 1 (2000): 36-43.

Hobcraft, John N., John W. McDonald, and Shea O. Rutstein. "Demographic determinants of infant and early child mortality: a comparative analysis." Population studies 39, no. 3 (1985): 363-385

Hosmer, David W., and Stanley Lemeshow. "Special topics." Applied Logistic Regression, Second Edition (2000): 260-351.

Kaldewei, Cornelia. "Determinants of Infant and Under-five mortalitythe Case of Jordan." Development Policy and Analysis Division of the United Nations Department of Economic and Social Affairs Technical Note (2010): 1-31.

Kayode, Gbenga A., Victor T. Adekanmbi, and Olalekan A. Uthman. "Risk factors and a predictive model for under-five mortality in Nigeria: evidence from Nigeria demographic and health survey." BMC pregnancy and childbirth 12, no. 1 (2012): 10.

Liu, David, Yudie Yuan, and Shufang Liao. "Artificial neural networks for optimization of gold-bearing slime smelting." Expert Systems with Applications 36, no. 9 (2009): 11671-11674.

Measham, Anthony R., Krishna D. Rao, Dean T. Jamison, Jia Wang, and Alaka Singh. "Reducing Infant Mortality and Fertility, 1975-1990: Performance at All-India and State Levels." Economic and Political Weekly 34, no. 22 (1999): 1359-367.

Mosley, W. Henry, and Lincoln C. Chen. "An analytical framework for the study of child survival in developing countries." Population and development review 10 (1984): 25-45.

Padmanaban, P., Parvathy Sankara Raman, and Dileep V. Mavalankar. "Innovations and challenges in reducing maternal mortality in Tamil Nadu, India." Journal of Health, Population and Nutrition (2009): 202-219.

Press Information Bureau. Government of India. Ministry of Health and Family Welfare.

Rajan, K., Vennila Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan. "Automatic classification of Tamil documents using vector space model and artificial neural network." Expert Systems with Applications 36, no. 8 (2009): 10914-10918.

Saabneh, Ameed. "The association between maternal employment and child survival in India, 199899 and 200506." Asian Population Studies 13, no. 1 (2017): 67-85.

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.

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.

United Nations Childrens Fund 2015 , Fact Sheet: Infant and Child Mortality in India: Levels, Trends and Determinants

Zhou, Zhi-Hua, and Xu-Ying Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem." IEEE Transactions on Knowledge and Data Engineering 18, no. 1 (2006): 63-77.

Table 3: Summary Statistics (N= 250,990)

Variable	Category	Prevalance	IMR
Sex	Male	48.00	76.00
	Female	52.00	83.00
Wealth Index	First Quintile	16.00	126.00
	Second Quintile	18.00	105.00
	Third Quintile	21.00	81.00
	Fourth Quintile	23.00	65.00
	Fifth Quintile	23.00	42.00
Mother Smoking	No	97.00	78.00
	Yes	3.00	124.00
Religion	Hindu	71.00	85.00
	Muslim	16.00	74.00
	Christian	8.00	55.00
Caste Category	Scheduled Caste	18.00	95.00
	Scheduled Tribe	14.00	86.00
	Other Backward Caste	34.00	86.00
Region	Urban	40.00	62.00
	Rural	60.00	91.00
Type of Birth	Single	99.00	76.00
	Multiple	1.00	341.00
Access to Drinking Water	Improved Sources	67.00	82.00
	Non-Improved Sources	33.00	74.00
Access to Sanitation	Improved Sources	79.00	79.00
	Non-Improved Sources	21.00	80.00
Place of Delivery[3]	Home	11.00	61.00
	Private Facility	5.00	38.00
	Public Facility	4.00	41.00
	Other	0	53.00
Anemia	No Anemia	42.00	74.00
	Mild Anemia	35.00	83.00
	Moderate Anemia	13.00	95.00
	Severe Anemia	2.00	109.00
Preceding Birth Interval	Short	67.0	77.00
	Adequate	33.00	84.00
State	Low HDI State	29.00	111.00
	High HDI State	15.00	44.00

11.5

Table 4: Logistic Regression Results

	Model 1	Model 2	Model 3
‘Wealth Index‘	0.112***	0.085***	0.105***
	(0.010)	(0.031)	(0.025)
‘Mother’s Smoking‘	-0.301***	0.030	
	(0.046)	(0.183)	
‘Father’s Education Level‘	0.057***	0.015	
	(0.012)	(0.035)	
‘Birth Order‘	-0.070***	-0.058***	-0.045**
	(0.008)	(0.020)	(0.019)
‘Multiple Births‘	-1.989***	-1.952***	-1.803***
	(0.050)	(0.125)	(0.127)
Anemia	0.090***	0.178***	0.209***
	(0.013)	(0.033)	(0.033)
‘Rural Residence‘	-0.042	-0.124*	
	(0.025)	(0.072)	
‘Mother’s Education Level‘	0.214***	0.157***	0.182***
	(0.015)	(0.040)	(0.037)
‘Scheduled Caste‘	-0.091***	-0.154*	
	(0.029)	(0.081)	
‘Scheduled Tribe‘	-0.105***	-0.105	
	(0.034)	(0.090)	
‘Other Backward Caste‘	-0.009	0.152**	
	(0.026)	(0.074)	
‘Non-Improved Drinking Water Source‘	0.026	0.023	
	(0.033)	(0.093)	
‘Non-Improved Sanitation‘	-0.025	-0.043	
	(0.039)	(0.108)	
‘Mother <20 Years‘	-0.499***	-0.304***	-0.353***
	(0.024)	(0.073)	(0.072)
‘Mother >35 Years‘	0.341***	0.207	
	(0.087)	(0.165)	
‘Short Preceding Birth Interval‘	0.235***	0.360***	0.305***
	(0.028)	(0.079)	(0.078)
‘Delivery in Private Institution‘		1.739***	
		(0.178)	
‘Delivery in Public Institution‘		1.464***	-0.079
		(0.179)	(0.081)
‘Delivery in the Home‘		1.539***	
		(0.172)	
‘Low-HDI State‘	-0.320***	-0.302***	-0.383***
	(0.022)	(0.063)	(0.058)
‘High-HDI State‘	0.290***	0.216**	
	(0.036)	(0.098)	
‘Premature Delivery‘			-0.253***
			(0.020)

*p<0.1; **p<0.05; ***p<0.01