Predicting Infant Mortality: Minimizing False Negatives

Sushmita V. Gopalan*

April 2018

Abstract

One of India's key public health challenges has been its plateauing infant mortality rate. Early experiments from other countries suggest that targeting public health resources on high-risk pregnant mothers can help reduce the risk of infant death in cost-effective ways. In this study, we apply machine learning to train a classifier to identify pregnant mothers who are at high risk for infant death, without requiring the input of a medical professional. In our dataset, around 94% of the children have survived past the age of 1, making it highly imbalanced. The sparseness of observations about children who have died makes it difficult for a classifier to learn to identify them. Since we do not want to misclassify at-risk children as being healthy, we focus on minimizing false negatives. We test five different resampling methods to address the class imbalance problem and find that removing Tomek links before using a random forest classifier helps reduce the false negative rate from 74% to 7%, while keeping the false positive rate at 6%.

keywords Machine learning, class imbalance, prediction, infant mortality, public health

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

1 Introduction

In 2015, India failed to meet one 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 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 and have a range of different target populations. The Janani Suraksha Yojana is aimed at promoting institutional deliveries. The Navjaat Shishu Suraksha Karyakram (NSSK) is a program 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 however, not uniformly implemented across states. 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 significant explorations into their interplay with each other. It is thus not clear that tax dollars are being utilized effectively to serve the most vulnerable populations.

According to the National Health Profile, 2017, India has just about a million doctors catering to its 1.2 billion large population. Of these, only 10% work in the public health sector. This translates to roughly 1 government doctor for every 10,000 citizens. Further, a report by the World Health Organization in 2001 estimated that

three in five allopathic doctors do not possess legitimate medical degrees, with 31.4% only having completed secondary education. Under these circumstances, the role of informal healthcare workers gains tremendous importance. In nearly every Indian village, the National Rural Health Mission appoints an Accredited Social Health Activist (ASHA) worker, a literate female resident of the village, between 25 to 45 years old, to serve as a first point of contact, recommend and facilitate access to professional medical care.

Das et al. (2016) conducted a randomized controlled trial in India and found that even nine months of informal training can improve the quality of care provided by healthcare providers without medical degrees. Leveraging machine learning and artificial intelligence to aid in diagnosis and risk prediction could potentially be highly valuable to such non-medical healthcare workers. Misdiagnoses were reduced by 14% after the training period.

In this paper, we train an algorithm to identify pregnant mothers who are at high risk for infant death. While trying to tune classification models to achieve highest accuracy, we need to keep in mind the fact that a high overall accuracy rate is not enough for our purposes. Since far more children survive than die, we have an imbalanced training dataset. The sparseness of observations about children who die, relative to children who survive, hinders the classifier's ability to learn to recognize children who are likely to die. Since it is not catastrophic for a low-risk child to receive extra medical attention, but could potentially be fatal for a high-risk baby to be classified as healthy, we want to focus on minimizing false negatives. At the same time however, false positives cannot rise unduly since that would defeat our goal of

narrowly targeting the recipients of public health expenditure. We test five different methods to address the class imbalance problem using five different classification algorithms. Using a method proposed by Tomek (1976), which uses the k-nearest neighbours algorithm to create more distinct class boundaries, we are able to reduce the false negative rate from 74% to 7% using random forests, while keeping the false positive rate at 6%.

1.1 Literature Review

Previous studies have found that the primary determinants of infant and child mortality rates can be divided into three categories. The first category of explanatory variables is 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. The second is the parents's health status and behaviour - smoking, drinking, dietary habits, tetanus, anemia, attitude towards health seeking, awareness levels. Finally, there are community-specific variables - sanitation, public health facilities, access to insurance, communicable diseases, cultural attitudes toward health care, women's empowerment, etc.

A World Health Organization (WHO) report from 2005 [1] 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.

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

Income is generally considered to be the variable most correlated with infant mortality Barbus (2011) and Hobcraft et al. (1985) . Measham 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 India's reduction in IMR in this period can be explained by income growth. When compared to other developing countries, they find that India's 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, we 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 Measham 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).

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 (Padmanaban 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 (Padmanaban et al., 2009). Consistent with this, Claeson 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).

What is now a canonical framework to analyze child mortality was put forth by Mosley and Chen (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. We focus this analysis to the socioeconomic determinants of mortality - variables basic enough that any ASHA worker could collect and plug into a model that yields a risk-score.

The literature outlined above serves as a guide to select the features (variables) that will be used as inputs to classification algorithms. The methods used in these papers are regression-based econometric methods used to determine and estimate the impact of different variables on infant mortality. They seek to indentify and quantify the specific ways in which these variables affect the probability of infant death. The focus of this paper, however, is prediction. We want to find ways to extrapolate

patterns found in historical data to predict outcomes for observations whose outcomes are yet to be realized. Machine learning algorithms are the ideal tool for this. They are designed to seek generalizable structure in underlying data and are evaluated on their ability to find it (Mullainathan and Spiess, 2017). For a clearer understanding of the benefits machine learning has to offer, Section 1.2 describes some of the ways machine learning algorithms have been applied in the context of disease and mortality prediction.

1.2 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 and 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 and 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 and Harper, 2001), colon cancer (Camp and Slattery, 2002), coronary heart disease (Carmelli et al., 1997), 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 diarrhoea, father's education, birth weight and mother's 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 the receiving operating characterisitc curve (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. (2011) 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 fill, apart from its policy-related goals, is building a predictive model to classify at-risk infants and target policy efforts accordingly, in the Indian context.

2 Data and Methods

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, we use data from NHFS-4 (2015-16).

Typically, Demographic Health Surveys (DHS), publish data that can be accessed via an application to the USAID, in four different datasets - household data, individual

Table 1: Summary Statistics

Variable	Category	Prevalance	IMR
Wealth Index	First Quartile	23.53	106.56
	Second Quartile	23.46	82.69
	Third Quartile	20.58	66.17
	Fourth Quartile	17.54	54.22
	Fifth Quartile	14.86	39.84
Smoking	Mother Does Not Smoke	83.43	71.30
	Mother Smokes	16.56	84.83
Multiple Births	Single Birth	98.66	69.85
	Multiple Births	1.33	346.35
Sex	Male Child	50.65	85.99
	Female Child	49.34	60.77
Anemia	No Anemia	47.48	68.03
	Mild Anemia	38.05	75.87
	Moderate Anemia	11.75	87.31
	Severe Anemia	1.04	106.14
Type of Residence	Urban	25.88	55.52
	Rural	74.11	79.84
Water Sources	Non-Improved	83.68	74.14
	Improved	16.31	70.49
Sanitation	Improved	51.76	55.88
	Non-Improved	48.23	92.49
Preceding Birth Interval	Adequate	63.54	70.36
	Short	36.45	79.09
Caste	Scheduled Caste	80.77	75.15
	Scheduled Tribe	15.01	69.23
	Other	3.86	57.14
Mother's Education	No Education	50.65	85.99
	Primary Education	16.07	73.93
	Secondary Efucation	29.58	56.23
	Higher Education	3.68	39.81
Birth Order	First, Second or Third Child	82.18	69.96
	Fourth or Higher	17.81	90.06
Religion	Hindu	74.07	78.13
	Muslim	14.39	71.27
	Christian	7.25	48.01
	Sikh	1.85	46.98
	Other	2.42	43.58

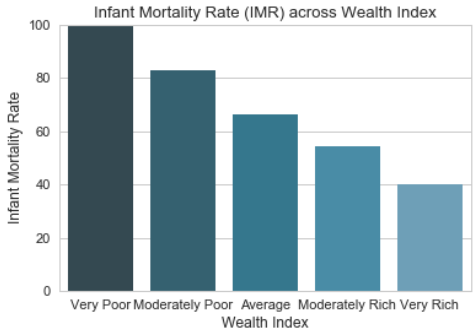
woman's data, children's data, and household listing data. Most of the variables of interest to this study come from the individual woman's dataset, which covers a range of questions on deliveries, infant-care, maternal health-care, nutrition, a few questions on the woman's agency and safety in the household 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.

Using household IDs from the individual woman's 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. 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).

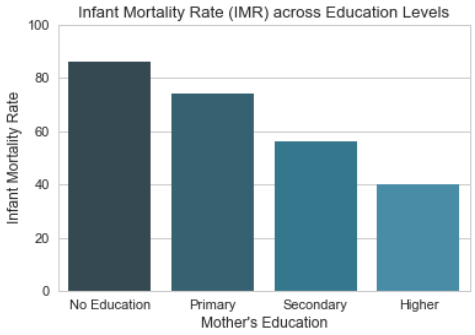
Further, when it comes to the impact of breastfeeding on mortality, Palloni and Millman (1986) caution against the danger of a reverse causation bias because babies who die young would have a very short period of breastfeeding recorded, even if they were breastfed for the entire duration of their short 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 month 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.

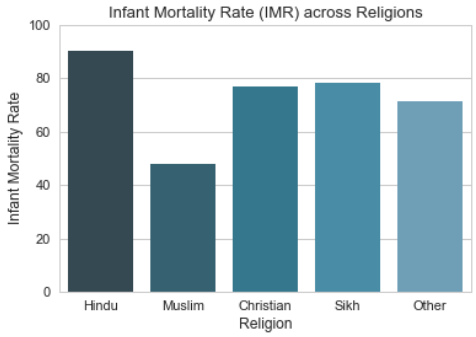
In order to make for more straightforward interpretation of regression coefficients, some variables were recoded. My response variable is set to take 0 if the child died on or before their first birthday and 1, otherwise. The variables on access to drinking water and sanitation facilities are recoded 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. Several studies have found a U-shaped relationship between mother's age and child mortality, dropping past the age of 20 and rising again in the mid-thirties (Fischl et al., 1992). A categorical variable was created for the child's 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 using DHS data. Table 1 shows the infant mortality rates across various explanatory variables, along with how prevalant each category is in the data. Figure 1 visualizes bivariate relationships between infant mortality rate and important explanatory variables.



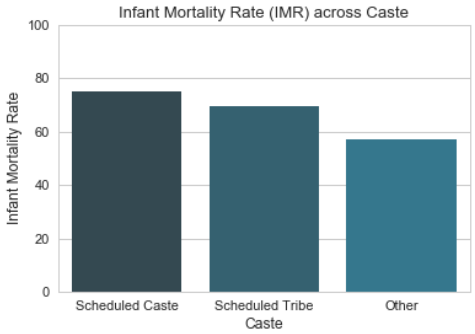
(a) Wealth Index



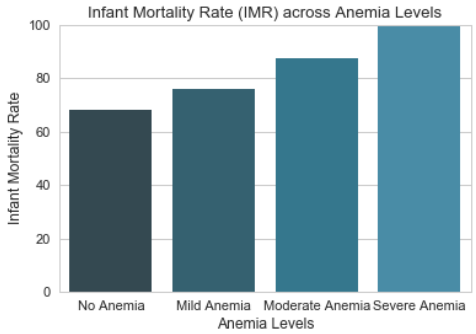
(b) Education



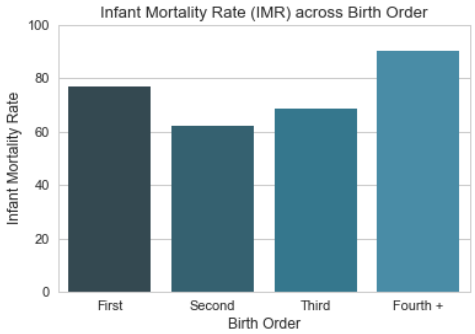
(c) Religion



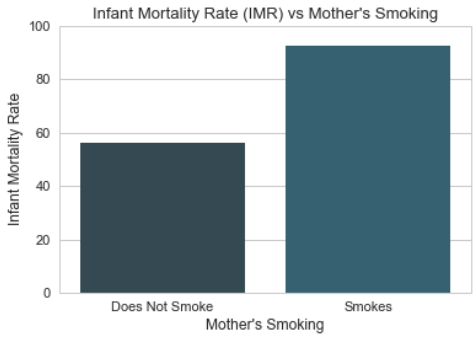
(d) Caste



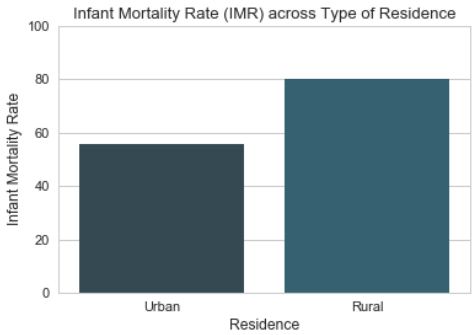
(e) Anemia



(f) Birth Order



(g) Mother's Smoking



(h) Region

Figure 1: Bivariate Relationships between Infant Mortality and Predictors

2.1 Classification Methods

First, since our response variable is binary, we run a logistic regression to ensure that the features we have chosen are meaningful and statistically significant predictors of infant mortality. The probability that the event $Y = 1$ will occur, conditional on a covariate vector x is determined by the logistic function of the vector x and the vector of coefficients β. The model is specified as below.

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

Tree-based classification algorithms have been found to achieve high accuracy in disease-prediction contexts. A decision tree is essentially a way to break up a complex decision involving numerous variables into an amalgamation of smaller decisions involving one variable at a time. There are two significant reasons for their success in this context. The first is that they can achieve 100% training accuracy on data that is not linearly-separable and the second is that they implicitly account for interactions between features. In regression-based classification methods like logistic regressions, we would have to give up many degrees of freedom to fully account for all possible interactions between variables. Having categorical explanatory variables complicates this further.

However, decision trees do not always generalize well to out-of-sample data because they can easily overfit to training data. We also use two different tree-based classification models that account for this problem - random forest models and adap-

tive boosting models. A random forest model is an ensemble classification method which uses a number of single decision trees, each fit on a subset of features, and outputs the mode of the classes predicted by each tree. This is done in order to introduce an element of randomness into the training data and improve generalizability. We expect training accuracy on random forest models to be lower than that for a single decision tree, but testing accuracy to be higher from a random forest.

Another classification method we use is an AdaBoost tree, where AdaBoost stands for adaptive boosting. AdaBoost is also an ensemble classifier where weak learners are iteratively trained at each step. After a weak classifier is learned, the algorithm boosts or increases the weight of previously misclassified examples so that the next weak classifier is optimized to correct these errors. It can be proven that so long as each weak classifier is better than a random guess, the final weighted model converges to a strong classifier. XGBoost is very similar to AdaBoost but provides more flexibility in terms of hyperparameter-tuning.

2.2 Methods to Address Class Imbalance

7.3% of the children in our dataset died within the age of 1, while the remaining 93.6% survived. Clearly, children who died form the minority class and children who survived form the majority class. The sparseness of observations in the minority class can complicate a classifier's ability to learn to identify it. In the machine learning literature, this problem is called the class imbalance problem. The class imbalance problem is encountered in a diverse range of applications detection of oil

spills from satellite radar images (Kubat et al., 1998), detection of fraudulent phone calls (Fawcett and Provost, 1997) and diagnoses of rare medical conditions like thyroid disorders (Murphy and Aha, 1994).

Typically, the performance of a classifier is assessed by its accuracy, i.e. the proportion of test cases that are correctly classified by the algorithm. However, in the case of an imbalanced dataset, accuracy is a misleading measure. For instance, in this dataset, if a classifier failed to identify even a single at-risk child and predicts that they will all survive, it would still be correct 93 out of 100 times and hence, have an accuracy of 0.93. Other performance measures that are commonly used include precision and recall. Precision is the ratio of the number of observations that are correctly predicted to be at-risk to the total number of observations that are predicted to be as at-risk. Recall is the ratio of the number of observations that are correctly predicted to be at-risk to the total number of observations that are actually at-risk. Often, the geometric mean of precision and recall, called the F-score, is used as a measure that combines these two concerns.

Kubat et al. (1997) suggest using the geometric mean of the accuracies measured separately on each class. The receiver operating characteristic (ROC) curve is a good way to visualize the fact that sometimes, classifiers are biased towards one particular class and that the degree of this bias can be controlled. The ROC curve plots accuracy on the positive class against error rate on the negative class. This helps us understand what trade-offs are being made as we attempt to increase accuracy on one of the classes.

The choice of performance metric for model evaluation is obviously context-specific.

For the purposes of this study, we are most interested in minimizing false negative rates. Classifying a baby highly likely to be healthy as at-risk (false positive) is not a catastrophic error, while classifying a sick baby as not-at-risk (false negative) could potentially have fatal consequences. However, we do need to control the trade-off because classifying too many healthy babies as at-risk defeats our intentions of identifying at-risk babies to focus public health resources on. Our goal is hence, to minimize false negatives as much as we can without allowing false positive rates to rise unduly. In this paper, we evaluate five different methods to address the class imbalance problem.

Random oversampling This is a non-heuristic resampling method, where we randomly sample from the minority class with replacement, until we have as many examples as there are in the majority class. An obvious drawback of this approach is that by simply replicating observations we are not actually training the classifier with more information. It is equivalent to up-weighting each observation from the minority class and down-weighting each observation from the majority class. This also places undue emphasis on outliers because the classifier repeatedly encounters examples that are outliers and learns that they are typical instances of the minority class, leading to poor prediction accuracy on held-out data. This approach makes the classifier too specific, it learns to recognize the examples it has encountered very well, but fails to generalize to out-of-sample examples.

Random undersampling : This is also a non-heuristic resampling method where

we train with a random subset of examples from the majority class and all of the data from the minority class, such that the two are equal in number. This method is also not robust to cross-validation because of the fact that not all small subsets of a large dataset follow the same probability distribution. Averaging across enough folds of the training data to achieve stable metrics is computationally intensive.

Asymmetric penalties : Most machine learning classifiers use an iterative process of minimizing loss at each step. Loss is essentially some function of the difference between actual and predicted values of a target variable. What this approach does is levy a higher penalty on misclassifications on the minority class than misclassifications on the majority class. So, in order to minimize this new loss function, a different set of parameters are chosen by the classifier, that takes into account the asymmetric penalties.

Generating Synthetic Samples : Chawla et al. (2002) came up with the Synthetic Minority Oversampling Technique (SMOTE) where we don't merely replicate examples from the minority class, but generate new, synthetic examples. This is done by interpolating between several minority class examples that lie together. Intuitively, this allows the classifier to build larger decision regions that contain nearby instances from the minority class (García et al., 2012) and helps avoid overfitting to the very few examples available from the minority class.

Removing Tomek Links : In order to help the classifier better distinguish between the two classes, Tomek (1976) suggests that borderline examples and those

suffering from class-label noise be removed. Consider two observations x and y, each having a different class label. Let $\delta(x, y)$ be the Euclidian distance between x and y. Then, a pair (x, y) is considered a Tomek link if and only if no other example z exists such that $\delta(x, z) < \delta(x, y)$ or $\delta(y, z) < \delta(y, x)$. In other words, two observations form a Tomek link if they have different class labels and are each other's nearest neighbour. Observations that belong to Tomek links are thus considered either borderline or noisy. The idea here is that the presence of borderline or noisy examples occludes the classifier's ability to learn what an example from the minority class looks like. By removing the majority class example from each Tomek link, we create clearer decision boundaries.

3 Results

3.1 Determinants of Infant Mortality

First, we examine the logistic regression estimates. Figure 2 depicts the results of the logistic regression. The predictors are arranged in order of magnitude of the coefficient. The variables at the top of the chart have a highly positive impact on the log-odds of infant death while the variables at the bottom of the chart have a highly negative impact on the log-odds of infant death. So, all the variables to the right of the $x = 0$ line increase the odds of infant death, while those to the left lower the odds.

We see immediately that the risk of infant death is very high in cases of multiple

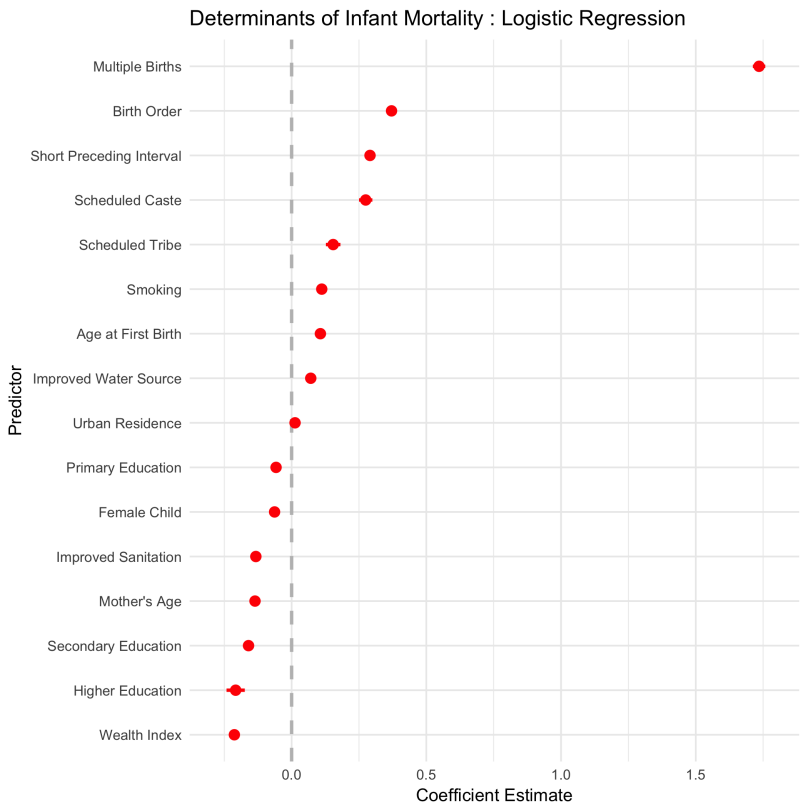


Figure 2: Regression Coefficient Plot

births, i.e., twins, triplets, etc. After the third child, the odds of infant death increase sharply. A gap of less than 6 months since the mother's previous pregnancy also increases risk because the mother's body has not had enough time to heal and regain strength from her previous pregnancy. As we would expect, a mother who smokes some form of tobacco leaves her child more prone to sickness.

Belonging either to either a scheduled caste or a scheduled tribe significantly increases vulnerability to infant death. Examining the relationship between caste and infant mortality while controlling for wealth reveals interesting patterns. We see that across the all five wealth quantiles, infant mortality is highest for scheduled castes, followed by scheduled tribes. In fact, the infant mortality for scheduled castes in each wealth quantile is the same the same as the infant mortality for *upper caste* families from the next-lowest wealth index. For example, scheduled castes in the *Poor* wealth index have the same infant mortality rate as *upper caste* families in the *Very Poor* wealth index. This highlights a system of cultural inequities that mediate access to healthcare, even while controlling for wealth.

We see that the mother's age at the time of their first pregnancy positively impacts infant mortality but the mother's age at the time of their current pregnancy negatively impacts it. As intuition and the past literature would suggest, the mother's having increasingly higher levels of education reduces infant mortality further. Improved sources of sanitation decrease the risk of infant death, but contrary to expectations, improved water sources actually increase risk. The categories used in the DHS surveys are meant to be standardized across the world. Perhaps the way *improved* and *non-*

Table 2: Determinants of Infant Mortality

	Coefficient	Standard Error
Wealth Index	-0.213^{***}	(0.005)
Smoking	0.112^{***}	(0.012)
Birth Order	0.371^{***}	(0.006)
Multiple Births	1.735^{***}	(0.023)
Female Child	-0.063^{***}	(0.009)
Moderate Anemia	-0.185^{***}	(0.039)
Mild Anemia	-0.339^{***}	(0.037)
No Anemia	-0.418^{***}	(0.037)
Urban Residence	0.013	(0.013)
Mother's Age at First Birth	0.107^{***}	(0.003)
Improved Water Source	0.071^{***}	(0.012)
Improved Sanitation	-0.133^{***}	(0.012)
Mother's Age	-0.136^{***}	(0.003)
Short Preceding Interval	0.291^{***}	(0.012)
Scheduled Caste	0.275^{***}	(0.025)
Scheduled Tribe	0.154^{***}	(0.027)
Primary Education	-0.058^{***}	(0.013)
Secondary Education	-0.160^{***}	(0.012)
Higher Education	-0.208^{***}	(0.034)
Intercept	-1.738^{***}	(0.056)
Observations	695,704	
Note:	*p<0.1; **p<0.05; ***p<0.01	

improved are categorized map more accurately onto actually cleaner sources in some countries over others. The mother's having severe anemia significantly increases risk of infant death as well. Table 2 also shows that all of the features used in the model are statistically significant at the 0.01 level, except for whether or not the family resides in an urban area.

3.2 Prediction

Since the review of past literature and the logistic regression results show that the features chosen are statistically significant and meaningful, we now want to find the combination of resampling method and classification algorithm with best predictive power. In this context, that means a low false negative rate and a low false positive rate. We bear in mind that a decrease in false negative rate could be achieved at the cost of an increase in false positive rate and we want to control this trade-off to the extent possible. Each algorithm has been evaluated using 10-fold cross-validation and averaging model metrics across results.

In this analysis, we evaluate combinations of five different classification algorithms and resampling methods outlined in Section 2 on our data. In order to set out a baseline we will endeavor to improve upon, we first run all five algorithms on the original, unaltered data. For each, we record False Negative Rate (FNR), False Positive Rate (FPR) and the area under the ROC curve (AUROC) for the training subset and the testing subset.

We see that the tree-based methods all achieve fairly low false negative and false

Table 3: Comparison of Classifiers: Original Data

	Training			**Testing**		
	FNR	FPR	AUROC	FNR	FPR	AUROC
Random Forest	0.05	0.06	0.63	0.74	0.07	0.52
AdaBoost	0.05	0.06	0.63	0.74	0.07	0.51
XGBoost	0.40	0.07	0.51	0.40	0.07	0.51
Decision Tree	0.03	0.06	0.63	0.82	0.07	0.52
Logistic Regression	0.42	0.07	0.51	0.41	0.07	0.51

positive rates on the training set but have very high false negative rates on the testing data. This clearly points to the fact that the models are overfit on the minority class in the training data. The XGBoost algorithm and the logistic regression both outperform the other classifiers on this data. Note that both these methods use a regularization term to restrict the size of the parameter vector. All data have been centered and standardized in order to avoid biases while using regularization penalties. The decision tree has the lowest training error rate and the highest testing error rate because even for highly non-linear and non-separable data, a decision tree can, given enough iterations and a small enough minimum node size, fit the training set perfectly. As a result, it does not perform well on out-of-sample examples. The Random Forest, which increases variances across the different trees it is composed of, does slightly better on the testing data, as does the AdaBoost, which up-weights misclassified examples at each iteration.

Next, we evaluate the effects of randomly oversampling from the minority class. Notice that the false negative rates on the testing data have decreased on average, because the classifier has encountered more examples on the minority class than with

Table 4: Comparison of Classifiers: Random Oversampling

	Training			**Testing**		
	FNR	FPR	AUROC	FNR	FPR	AUROC
Random Forest	0.23	0.16	0.80	0.25	0.17	0.78
AdaBoost	0.05	0.06	0.63	0.74	0.07	0.51
XGBoost	0.38	0.39	0.61	0.39	0.39	0.61
Decision Tree	0.23	0.16	0.80	0.26	0.18	0.78
Logistic Regression	0.39	0.39	0.61	0.39	0.39	0.61

Table 5: Comparison of Classifiers: Random Undersampling

	Training			**Testing**		
	FNR	FPR	AUROC	FNR	FPR	AUROC
Random Forest	0.18	0.17	0.83	0.43	0.42	0.57
AdaBoost	0.39	0.39	0.61	0.39	0.39	0.61
XGBoost	0.38	0.38	0.62	0.39	0.39	0.61
Decision Tree	0.14	0.2	0.83	0.44	0.44	0.56
Logistic Regression	0.39	0.39	0.61	0.40	0.39	0.60

the original data. However, this also comes at the cost of far higher false positive rates. Specifically, with the random forest that achieves lowest false negatives on testing, the false positive rate has risen to 17% - given the context of this analysis, having such a high false positive rate would defeat the intentions of identifying at-risk mothers in order to focus public health efforts on them. Public health resources are too scarce to accommodate a 17% misclassification rate on babies that are likely to be perfectly healthy.

Random undersampling from the majority has fairly different effects on model accuracy than random oversampling. Both false negative rates and false positive rates are much higher when we use this method because in effect, we're only using a small

Table 6: Comparison of Classifiers: SMOTE

	Training			Testing		
	FNR	FPR	AUROC	FNR	FPR	AUROC
Random Forest	0.13	0.11	0.88	0.88	0.07	0.55
AdaBoost	0.21	0.29	0.75	0.87	0.06	0.58
XGBoost	0.15	0.23	0.81	0.86	0.06	0.57
Decision Tree	0.13	0.12	0.88	0.88	0.07	0.55
Logistic Regression	0.39	0.39	0.61	0.89	0.05	0.61

fraction of available data on the majority class. Close to 90% of the data on children that survive is not used at all while training the models and hence, it stands to reason that false positive rates are very high. The notable exception is AdaBoost which up-weights misclassified samples and hence achieves lower training error rates. This too however, has an unacceptably high false negative rate. 10-fold cross-validation on classifiers using randomly undersampled data also has high variance because model metrics are influenced strongly by which subset of the majority class data happened to be chosen for each round of cross-validation.

Using the synthetic minority oversampling technique (SMOTE) gives us very low false positive rates, but very high false negative rates, even exacerbating the imbalance from the original data. Since SMOTE creates synthetic samples by interpolating within the feature space and not extrapolating, it performs better in contexts where out-of-sample examples are very similar to training samples, but fails in contexts where extrapolation to unseen examples in required.

Finally, we evaluate the effects of removing Tomek Links. In this implementation of the algorithm, we identify all pairs of observations that form Tomek Links, i.e., all

Table 7: Comparison of Classifiers: Tomek Links

	Training			Testing		
	FNR	FPR	AUROC	FNR	FPR	AUROC
Random Forest	0.06	0.06	0.62	0.07	0.06	0.62
AdaBoost	0.40	0.07	0.51	0.45	0.07	0.50
XGBoost	0.40	0.07	0.51	0.43	0.07	0.51
Decision Tree	0.03	0.06	0.62	0.07	0.06	0.62
Logistic Regression	0.41	0.07	0.51	0.43	0.07	0.651

pairs of observations that are each other's nearest neighbour and have different class labels. We then proceed to drop only the majority class sample from each pair. We see that false negatives using random forests have dropped to 6%, down from 74% using random forests on the original data and 41% using logistic regression on the original data. This decrease is substantial and useful because the false positive rate stays at 6% as well. We obtain the same results from the decision tree algorithm as well. Typically, we would expect ensemble classifiers to outperform their pure counterparts, but in cases where the data is large and stable, this need not be the case.

Of all the resampling methods we tested, we achieve best results from removing Tomek Links. Using the original data, the best false negative rate we achieved was 41% using a logistic regression. Using random forests after removing Tomek Links reduces this by 85 % and brings it down to 6%. The reason that the removal of Tomek Links outperforms other resampling methods are severalfold. First, there is very little loss of information. Unlike in random undersampling, we do not simply discard large subsets of the training data from the majority class. Out of 741,296 examples of surviving children in the training dataset, only 1433 are identified as

belonging to a pair of Tomek Links. Removing these only discards 0.2% of available data on surviving children. Second, it is robust to cross-validation unlike random undersampling where which subset of the majority class data is selected for training sways results non-trivially. Third, the intuition behind this method is the idea that when two observations form a Tomek Link, either they are both borderline or one of them is noise. This makes sense in this context because it is highly plausible that among two sick babies, one will survive and one will die for a variety of reasons. Removing these confusing examples helps overcome the fact that we do not have data that can distinguish between sick babies that survived and healthy babies. Finally, this method outperforms generating synthetic samples from the minority class because the SMOTE technique creates new samples within a convex feature space and does not generalize well unless out-of-sample observations are nearly identical to training observations.

4 Conclusion

This paper uses machine learning to train a classifier to identify pregnant mothers at risk for infant death in India. It is difficult to train a classifier when far more data is available with one class label than another. Since 94% of the children in our dataset survived, we have relatively little data on children who died, resulting in a class imbalance problem. Given the context of the study, where it is potentially fatal to misclassify a baby who is at risk as healthy, we focus on minimizing false negatives. We describe and test five different methods to address this issue and find

that removing Tomek links provides best improvement in predictive ability. Using a random forest classifier after removing Tomek links, we find that the false negative rate drops by almost 90%, from 74% to 7%. The false positive rate stays the same at 6%. The baseline used for comparison is by replicating the methods used by Tesfaye et al. (2017) on this datasets.

It is worth recalling at this point that the Tomek link removal method is premised on the idea that if two observations form a Tomek link, either both are borderline or one of them is noise. In this implementation of the method, we remove only the majority class example from each Tomek pair. The assumption that the majority class sample is noise is a fairly strong assumption and if this is not true, we could be creating artificial class boundaries. In this context however, if two children fall sick due to similar reasons, any number of unobserved factors could lead to different survival outcomes for each child, ranging from the mother's resourcefulness to one family happening to have a friendly, knowledgable adult in the neighbourhood. What happens when we remove a sick, surviving baby which looks very similar all aspects to a sick baby that died is that we prevent the classifier from learning that this is an example of a healthy, low-risk child.

Further improvements could potentially be made by adding more features that have been found to be significant in determining infant mortality in past literature such as breast-feeding and place of delivery. District and state level variables could also be important. Unlike econometric methods where interaction terms are added explicitly to account for the interplay of effects between variables, tree-based algorithms implicitly account for interaction because each split made on the dataset is

based on one feature at a time.

In a country that is experiencing a severe shortfall of medical professionals, simple machine-learning based tools with high predictive ability could aid informally trained healthcare professionals, reduce mistakes and potentially save lives.

References

Bachur, Richard G and Marvin B Harper, "Predictive model for serious bacterial infections among infants younger than 3 months of age," *Pediatrics*, 2001, *108* (2), 311–316.

Barbus, Alexandra, "Determinants of Infant Mortality.," *Europolis*, 2011, *5* (2).

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 & Control*, 2002, *13* (9), 813–823.

Carmelli, Dorit, Heping Zhang, and Gary E Swan, "Obesity and 33-year follow-up for coronary heart disease and cancer mortality," *Epidemiology*, 1997, pp. 378–383.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, 2002, *16*, 321–357.

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*, 2011, *38* (5), 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*, 2000, *78*, 1192–1199.

Das, Jishnu, Abhijit Chowdhury, Reshmaan Hussam, and Abhijit V Banerjee, "The impact of training informal health care providers in India: A randomized controlled trial," *Science*, 2016, *354* (6308), aaf7384.

Fawcett, Tom and Foster Provost, "Adaptive fraud detection," *Data mining and knowledge discovery*, 1997, *1* (3), 291–316.

Fischl, Margaret A, Raj B Uttamchandani, George L Daikos, Rita B Poblete, Jose N Moreno, Ricardo R Reyes, Ahmad M Boota, Lisa M Thompson, Timothy J Cleary, and Shenghan Lai, "An outbreak of tuberculosis caused by multiple-drug-resistant tubercle bacilli among patients with HIV infection," *Annals of internal Medicine*, 1992, *117* (3), 177–183.

Forthofer, Melinda S and Carol A Bryant, "Using audience-segmentation techniques to tailor health behavior change strategies," *American Journal of Health Behavior*, 2000, *24* (1), 36–43.

García, Vicente, Javier Salvador Sánchez, and Ramón Alberto Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, 2012, *25* (1), 13–21.

Hobcraft, John N, John W McDonald, and Shea O Rutstein, "Demographic determinants of infant and early child mortality: a comparative analysis," *Population studies*, 1985, *39* (3), 363–385.

Hosmer, David W and Stanley Lemeshow, "Special topics," *Applied Logistic Regression, Second Edition*, 2000, pp. 260–351.

Kubat, Miroslav, Robert C Holte, and Stan Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning*, 1998, *30* (2-3), 195–215.

_ , Robert Holte, and Stan Matwin, "Learning when negative examples abound," in "European Conference on Machine Learning" Springer 1997, pp. 146–153.

Lemon, Stephenie C, Jason Roy, Melissa A Clark, Peter D Friedmann, and William Rakowski, "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression," *Annals of behavioral medicine*, 2003, *26* (3), 172–181.

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*, 1999, pp. 1359–1367.

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

Mullainathan, Sendhil and Jann Spiess, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Murphy, PM and DW Aha, "UCI repository of machine learning databases, University of California, Department of Information and Computer Science, Irvine, CA," 1994.

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, *27* (2), 202.

Palloni, Alberto and Sara Millman, "Effects of inter-birth intervals and breastfeeding on infant and early childhood mortality," *Population Studies*, 1986, *40* (2), 215–236.

Saabneh, Ameed, "The association between maternal employment and child survival in India, 1998–99 and 2005–06," *Asian Population Studies*, 2017, *13* (1), 67–85.

Song, Xiaowei, Arnold Mitnitski, Jafna Cox, and Kenneth Rockwood, "Comparison of machine learning techniques with classical statistical models in predicting health outcomes.," in "Medinfo" 2004, pp. 736–740.

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*, 2017, *140*, 45–51.

Tomek, Ivan, "Two modifications of CNN," *IEEE Trans. Systems, Man and Cybernetics*, 1976, *6*, 769–772.