ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 05, pp.51094-51097, May, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

PREDICTION OF MORBIDITY PATTERN FOR VERY LOW BIRTH WEIGHT BABIES BASED ON DATA MINING TECHNIQUES

[1,*]Dr. Leo Alexander, T. and [2]Ciril Jenuvariyus, T. C.

[1]Associate Professor, Department of Statistics, Loyola College, Chennai – 600 034, India
[2]Research Scholar, Department of Statistics, Loyola College, Chennai – 600 034, India

ARTICLE INFO

Article History:
Received 25th February, 2017
Received in revised form
21st March, 2017
Accepted 14th April, 2017
Published online 31st May, 2017

Key words:
Binary Logistic Regression,
CART- Classification and Regression Tree, Gestational age,
Infants, Random forest,
Very low birth weight (VLBW),
WHO- World Health Organization.

ABSTRACT

According to the WHO, a baby which weighs less than 1500 gm at birth is termed very low birth weight baby. Birth weight and gestational age is an important parameter that predicts the outcome of the baby. Very low birth weight babies (VLBW) are at increased risk of a number of complications both immediate and late. Worldwide it has been observed that these babies contribute to a significant extent to neonatal mortality and morbidity. The common causes of mortality in these babies include sepsis, RDS and extreme prematurity. The risk of developing complications is inversely related to the gestational age and birth weight. The following paper is done to review our data related to the VLBW babies to identify the morbidity pattern and factors that influence mortality. This may help us to plan strategies to reduce mortality and prevent minimize morbidity in these babies.

Citation: Dr. Leo Alexander, T. and Ciril Jenuvariyus, T.C. 2017. "Prediction of morbidity pattern for very low birth weight babies based on data mining techniques", *International Journal of Current Research*, 9, (05), 51094-51097.

1. INTRODUCTION

Very Low birth weight has long been used as an important public health indicator. Low birth weight is not a proxy for any one dimension of either maternal or perinatal health outcomes. Globally, the indicator is a good measure of a multifaceted public health problem that includes long-term maternal malnutrition, ill health, hard work and poor pregnancy health care. On an individual basis, low birth weight is an important predictor of health; efforts must therefore go into measuring it, as accurately as possible at birth, organizing and planning infant care. Accordingly, the smaller the baby, the more important it is to monitor his or her growth in the weeks after birth. This is particularly important for infants at high risk of poor feeding and inadequate growth. Countries should therefore be encouraged to ensure accurate and reliable weighing of infants as close to birth as possible.This study is a combined referral work from various journals and books; "Socio-Economic and Nutritional Determinants of Low Birth Weight in India, N Am J Med Sci. 2014 Jul; 6(7): 302–308", "Risk Factors for Low Birth Weight (LBW) Babies and its MedicoLegal Significance, J Indian Acad Forensic Med, 32(3) ISSN 0971-0973", "Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis, 2001 by Frank E. Harrell Jr", "Classification and Regression Trees, by Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen; January 1, 1984 by Chapman and Hall/CRC", "Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression; Stephenie C. Lemon, Jason Roy, Melissa A. Clark, Peter D. Friedmann, William Rakowski (2003)", "Random Forests and Decision Trees Jehad Ali , Rehanullah Khan , Nasir Ahmad , Imran Maqsood; IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012". In Sections 2 and 3, we have discussed the methodology and the results of the techniques used for the study on very low birth weight data. Finally a comparison and conclusion as to which model is best has been discussed in Section 4.

OBJECTIVES OF THE STUDY

- To assess the various demographic factors and comparisons among the variables
- Identify the morbidity pattern and factors that influence mortality,

Corresponding author: **Dr. Leo Alexander, T.**
Associate Professor, Department of Statistics, Loyola College, Chennai – 600 034, India.

- Strategies to reduce mortality and morbidity in these babies,
- To study various risk factors that influence the conversion of mortality using the statistical techniques,
- To identify variables associated with mortality among very low birth weight infants admitted to a Neonatal Intensive Care Unit in Chennai, India.

2. STATISTICAL TECHNIQUES

2.1 LOGISTIC REGRESSION

In statistics, logistic regression, a regression model where the dependent variable is categorical. This article covers the case of a binary dependent variable that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. An explanation of logistic regression can begin with an explanation of the standard **logistic function**. The logistic function is useful because it can take any **real** input t, ($t \in R$), whereas the output always takes values between zero and one] and hence is interpretable as a probability. The logistic function σ(t) is defined as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Let us assume that t is a linear function of a single explanatory variable x, we can express t as follows:

$$t = \beta_0 + \beta_1 x \quad ,$$
$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The above F(x) is the logistic function.

The odds ratio

$$\mathrm{OR} = \frac{\mathrm{odds}(x+1)}{\mathrm{odds}(x)} = \frac{\left(\frac{F(x+1)}{1-F(x+1)}\right)}{\left(\frac{F(x)}{1-F(x)}\right)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

The above exponential relationship provides (OR), an interpretation for β_1: The odds multiple by exp(β_1) for every 1-unit increase in x.

2.2 CLASSIFICATION AND REGRESSION TREE (CART)

Classification and Regression Trees is a classification method which uses historical data to construct so-called decision trees. Decision trees are then used to classify new data. In order to use CART we need to know number of classes a priori.. For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations. Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts. CART algorithm will search for all possible variables and all possible values in order to find the best split – the question that splits the data into two parts with maximum homogeneity. The process is then repeated for each of the resulting data fragments.

2.3 RANDOM FOREST

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = (x_1 \ldots x_n)$ with responses $Y = (y_1 \ldots y_n)$ bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1 \ldots B$:

- Sample, with replacement, B training examples from X, Y; call these X_b, Y_b.
- Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples X' can be made by averaging the predictions from all the individual regression trees on X' as follows,

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

by taking the majority vote in the case of decision trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.An optimal number of trees B can be found using cross-validation, or by observing the *out-of-bag error*: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample. The training and test error tend to level off after some number of trees have been fit.

This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. An analysis of how bagging and random subspace projection contribute to accuracy gains under different conditions is given by Ho.

3. REVIEWING THE DATA USING STATISTICAL TECHNIQUES

The data is randomly divided into 70% and 30%. The former is known as the data for ‘development’ and the latter is known as the data for ‘validation’. We call these as the ‘train’ and ‘test’ respectively. Usually the techniques are tried and developed on the train data (development data) and the goodness is checked on the test data (validation data). here my dependent variable (outcome) recoded as 0=”death”, 1=”alive”.

3.1 BINARY LOGISTIC REGRESSION

The logistic regression model was first built on the train data and the resulting model provided the following output. Some of the important independent variables which turned up as significant on the basis of p-value are listed here in Table 3.1.

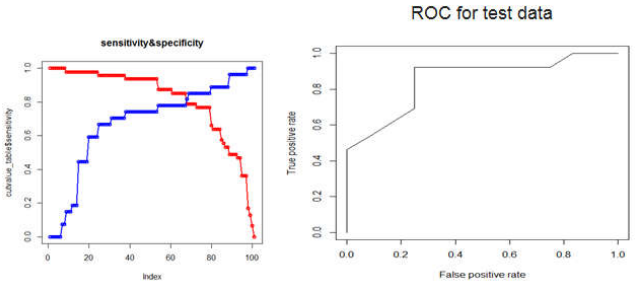
Table 3.1. List of significance independent variables from binary logistic regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.270119	4.055347	2.039	0.04142 *
GA	-0.081140	0.151057	-0.537	0.59116
Birth.weight	- 0.005766	0.002208	-2.612	0.00901 **
shock	3.122606	0.737855	4.232	2.32e-05 ***

It can be noted that he significant independent variable based on their p-value another way of evaluating the fit of the logistic regression model is through a classification Table 3.2.

Table 3.2. Classification table (logistic regression)

	0	1
0	22	5
1	10	37

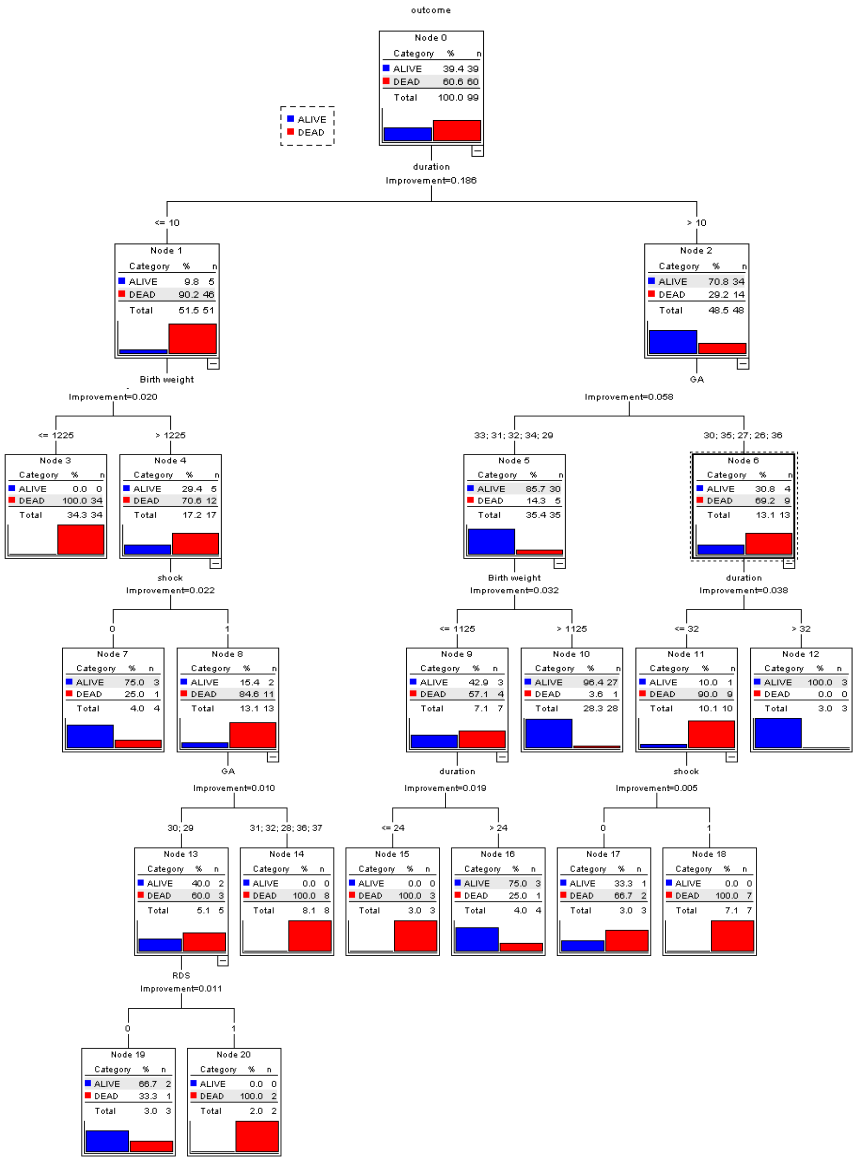


From the above classification Table 3.2, we see that 22 births out of 27 are correctly classified as births5 births are misclassified as deaths 37 out of 47 deaths are correctly classified as deaths and 10 deaths are misclassified as births. We obtained the required sensitivity and specificity curve.

Now we determine the sensitivity and specificity so as to obtain the most efficient cut value. Sensitivity is the proportion of 1's correctly classified as 1's and Specificity is the proportion of 0"s correctly classified as 0's.Now the minimum difference between the sensitivity and specificity is our most efficient cut value. Here our minimum difference is .02 which corresponds to cut value 0.67. Using this cut value we predict our dependent variable outcome (alive=0, death=1). Using ROC e can find that the Model accuracy is 85%.To obtain the important risk factors, that affect very low birth weight babies can be seen in the following CART method.

3.2 CLASSIFICATION AND REGRESSION TREE (CART)

CART model in R is a decision tree model which makes development data as input with 24 variable. It performs a univariate split with respect to independent variables. Which gives maximum information gain from root node to child node. The class method deals with the case when response variable is



a categorical variable. From the above Classification and Regression Tree, we can arrive at a conclusion that if the duration of the babies at the Neo-natal care is greater than 10 days, then the chance of surviving of those babies is the minimum.

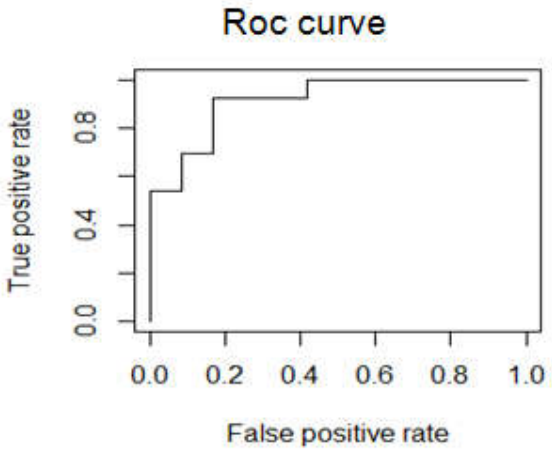
The same structure if the Maternal Gestation period varies between (33, 31, 32, 34, 29) weeks, then the chance of survival of those babies is greater. This is followed by the next condition that if the weight of the babies are greater than 1125 grams then there is a higher survival chance. On the other hand if the gestation period takes the other values (in weeks), then the chance of death is higher and if the duration of these babies in the neo-natal unit is less than 32 days then the death chances of those babies are high. Since the AUC value obtained for the CART model is 95%, we try to improve the prediction accuracy by applying Random Forest which is discussed in section 3.3.

3.3 RANDOM FOREST

The relative importance of the independent variables based on the mean decrease Gini index have been obtained by applying the random forest model to the very low birth weight data. We have the classification table given below.

Table 3.3. Classification table for test data

	Y	N	class.error
Y	43	4	0.08510638
N	7	20	0.25925926



From Table 3.3, we see that 43 births out of 47 are correctly classified as births 4 births are misclassified as deaths. We make use of classification Table 3.3 to visualize the performance of the CART model by the ROC curve summarize its performance in a single number by computing the AUC. Since the AUC value obtained for the Using random forest we get 91.66 (92%) accuracy. It means that our independent variables explains 92% our dependent variable.

4 Conclusion

Among all the three models, classification and regression tree (CART) has given the highest accuracy (95%). So we can conclude that CART method is best among all the models for this low birth weight data. We can arrive at a conclusion that if the duration of the babies at the Neo-natal care is greater than 10 days, then the chance of surviving of those babies is the minimum. This is followed by the next condition that if the weight of the babies are greater than 1125 grams then there is a higher survival chance. On the other hand if the gestation period takes the other values (in weeks), then the chance of death is higher and if the duration of these babies in the neo-natal unit is less than 32 days then the death chances of those babies are high.

REFERENCES

Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression; Stephenie C. Lemon, Jason Roy, Melissa A. Clark, Peter D. Friedmann, William Rakowski (2003).

Classification and Regression Trees, by Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen; January 1, 1984 by Chapman and Hall/CRC.

Random Forests and Decision Trees Jehad Ali , Rehanullah Khan , Nasir Ahmad , Imran Maqsood; IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis, 2001 by Frank E. Harrell Jr.

Risk Factors for Low Birth Weight (LBW) Babies and its MedicoLegal Significance, J Indian Acad Forensic Med, 32(3) ISSN 0971-0973.

Socio-Economic and Nutritional Determinants of Low Birth Weight in India, N Am J Med Sci. 2014 Jul; 6(7): 302–308.

****